GEOMETRIC-STEERING MUON FOR ONLINE LARGE LANGUAGE MODEL FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning is a fundamental approach for adapting pretrained large language models (LLMs) to downstream tasks and evolving application requirements. Although Muon has recently attracted increasing attention as a matrix-structured optimizer for LLM training and fine-tuning, existing studies have largely focused on conventional offline settings, leaving its behavior under online adaptation comparatively underexplored. In online LLM fine-tuning, evolving objectives can render accumulated momentum increasingly inconsistent with the current optimization signal, yet Muon's Newton–Schulz (NS) transformation primarily reshapes the spectral structure inherited from momentum without explicitly adapting its update orientation to the current objective. We therefore propose Geometric-Steering Muon (GS-Muon), which introduces Geometric-Steering Newton–Schulz (GS-NS) to explicitly couple Muon's spectral transformation with the current optimization signal. GS-NS decouples orientation adaptation from spectral orthogonalization: a short NS prefix establishes a near-orthogonal geometry, a tangent-inspired steering step redirects the momentum-inherited orientation toward the current optimization signal, and a final NS correction restores the orthogonalization structure. Theoretically, we establish geometric stability properties of GS-NS and derive variation-dependent nonconvex regret guarantees for GS-Muon under both first-order and zeroth-order signal access, explicitly characterizing the effects of momentum tracking and zeroth-order signal estimation. Finally, we establish a unified streaming evaluation protocol across two model scales and eleven NLP datasets, covering single-dataset streams, sequential within-family task shifts, and heterogeneous cross-task streams.

1 INTRODUCTION

Fine-tuning is a standard paradigm for adapting pretrained large language models (LLMs) to downstream tasks and domain-specific applications (Raffel et al., 2020; Wei et al., 2021; Ouyang et al., 2022). Conventional full-parameter fine-tuning typically relies on backpropagation-based first-order optimization, whose memory and computational costs become increasingly restrictive as model scale grows (Dettmers et al., 2023). To reduce the number of trainable parameters, parameter-efficient fine-tuning (PEFT) methods adapt pretrained models through compact parameterizations while keeping most model weights frozen (Houlsby et al., 2019; Hu et al., 2021; Lester et al., 2021). Along a complementary axis, zeroth-order (ZO) fine-tuning estimates optimization signals from function evaluations without backpropagation, providing a memory-efficient alternative for adapting LLMs (Malladi et al., 2023). Recent ZO methods further exploit low-dimensional or low-rank structures to alleviate the dimensionality and variance challenges of gradient estimation in large models (Chen et al., 2025; Yu et al., 2025). Beyond trainable parameterization and signal access, optimizer design provides another important dimension for efficient model adaptation, motivating growing interest in matrix-structured optimizers (Liu et al., 2025; Bernstein & Newhouse, 2024; Dong & Sawin, 2026).

Among recent matrix-structured optimizers, Momentum Orthogonalized by Newton–Schulz (Muon) has emerged as a promising alternative to conventional coordinate-wise optimizers, applying Newton–Schulz (NS) iterations to approximately orthogonalize momentum updates for matrix-valued parameters (Jordan et al., 2024). By exploiting matrix structure rather than treating parameters coordinate-wise, Muon has demonstrated strong empirical efficiency in LLM fine-tuning, achieving substantially improved compute efficiency over AdamW and scaling successfully to billion-parameter models (Liu et al., 2025). Recent studies have further improved Muon's practicality through low-precision

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.

optimizer states, memory-efficient variants, and refined matrix-aware update designs, highlighting growing interest in Muon as a scalable optimizer for LLMs (Huang et al., 2025; Gupta et al., 2025; Du & Su, 2026). Despite this progress, existing Muon studies have predominantly considered conventional pretraining or fixed-dataset fine-tuning settings, where optimization proceeds repeatedly over pre-collected data under a comparatively stable training protocol (Page et al., 2025; Dong & Sawin, 2026). Accordingly, Muon has received considerably less attention in online settings, where models must continually adapt to sequentially arriving data and evolving distributions, which is becoming increasingly important for LLM deployment in dynamic environments.

Online fine-tuning differs fundamentally from conventional offline optimization because the objective can evolve over time, making historical optimization information progressively less representative of the current objective (Fang et al., 2026). This issue is particularly relevant to Muon, which first accumulates gradient information through momentum and then applies a finite number of Newton–Schulz iterations to approximately orthogonalize the resulting momentum matrix (Kim & Oh, 2026). Crucially, finite Newton–Schulz primarily reshapes the singular-value spectrum while preserving the corresponding left and right singular-vector subspaces (Li & Tsuchiya, 2026). Consequently, when the accumulated momentum becomes inconsistent with the current optimization signal, standard NS can reshape its spectrum but cannot independently steer the update orientation toward the current objective (Nguyen et al., 2026). This limitation motivates an online Muon update that preserves the benefits of spectral orthogonalization while explicitly incorporating the current optimization signal into the orientation of the transformed update.

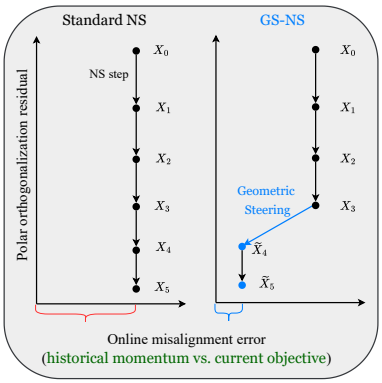


Figure 1: GS-NS vs. Standard NS

To address this limitation, we propose *Geometric-Steering Muon* (GS-Muon), an online Muon framework that explicitly corrects the momentum-inherited update orientation using the current optimization signal within the Newton–Schulz transformation. Unlike standard NS, which primarily reshapes the spectrum of accumulated momentum while preserving its inherited orientation, GS-Muon enables the transformed update to respond directly to evolving objectives. At its core, *Geometric-Steering Newton–Schulz* (GS-NS) decouples orientation adaptation from spectral orthogonalization: a short NS prefix first establishes a near-orthogonal reference geometry, a tangent-inspired steering step then redirects the update toward the current optimization signal, and a final NS correction restores the near-orthogonal structure perturbed by steering (Figure 1). This design combines current-objective responsiveness with Muon's matrix-structured spectral geometry, while reusing the available optimization signal.

Our main contributions are summarized as follows:

- We introduce GS-Muon for online LLM fine-tuning and develop GS-NS, which explicitly incorporates the current optimization signal into Muon's Newton–Schulz trajectory. Unlike standard NS, which reshapes the singular-value spectrum while preserving the orientation inherited from accumulated momentum, GS-NS directly adapts this orientation to the current objective and subsequently restores the near-orthogonal update geometry.
- We establish geometric guaranties for GS-NS, including nonnegative alignment with the available optimization signal, control of the finite-prefix tangent defect, and local contraction of the final orthogonality residual. Building on these properties, we derive variation-dependent nonconvex regret guaranties for GS-Muon under both FO and ZO signal access.
- We establish a unified streaming evaluation protocol for online LLM fine-tuning across two LLM backbones of different scales and eleven NLP datasets. The protocol covers single-dataset streams, sequential within-family task shifts, and heterogeneous cross-task streams, enabling systematic evaluation under progressively more challenging forms of online adaptation.

2 RELATED WORK

LLM Fine-tuning Fine-tuning adapts pretrained LLMs to downstream tasks by optimizing either the full parameter space or a restricted set of trainable parameters. Conventional FO fine-tuning relies on backpropagation to obtain gradient information and can achieve strong adaptation performance, but its activation and optimizer-state memory costs become increasingly restrictive as model scale grows (Zhang et al., 2022; Bai et al., 2023). PEFT alleviates this burden by freezing most pretrained weights and optimizing compact adaptation modules or low-rank parameterizations, with LoRA and its variants serving as representative approaches (Hu et al., 2021; Dettmers et al., 2023; Liu et al., 2024). Although PEFT substantially reduces the number of trainable parameters, most PEFT methods still rely on backpropagation and therefore retain part of the memory overhead associated with gradient computation. ZO optimization provides a complementary route by estimating update directions solely from function evaluations, avoiding backpropagation and reducing the memory footprint of LLM adaptation (Malladi et al., 2023). MeZO adapts classical ZO optimization to language-model fine-tuning and approaches inference-level memory usage, while subsequent methods improve ZO adaptation through variance reduction, curvature information, sparse updates, and momentum-guided perturbations (Gautam et al., 2024; Zhao et al., 2025). LOZO constructs low-rank ZO estimators, SubZero performs optimization in randomized low-rank subspaces, and related extensions further improve scalability or exploit low-rank geometry (Chen et al., 2025; Yu et al., 2025).

Gradient-Orthogonalized Methods Traditional adaptive optimizers such as Adam process parameters largely through coordinate-wise statistics, which can underutilize the matrix structure of neural-network weight matrices (Kingma & Ba, 2014). Matrix-aware optimizers instead exploit structured geometry or spectral information; representative methods such as Shampoo and SOAP employ matrix preconditioning to rescale gradients along structured directions (Gupta et al., 2018; Vyas et al., 2025). More recently, Muon has attracted substantial interest by applying a short NS transformation to matrix-valued momentum, producing approximately orthogonalized updates that exploit matrix geometry rather than treating the update as a vector (Jordan et al., 2024; Liu et al., 2025). Norm-based and preconditioning interpretations further connect Muon-style updates to steepest-descent and spectral optimization perspectives (Bernstein & Newhouse, 2024; Lau et al., 2025). Finite-depth NS is a deliberate optimization primitive rather than an exact polar-decomposition surrogate, and recent analyzes characterize the behavior of truncated polynomial iterations (Kim & Oh, 2026; Li & Tsuchiya, 2026). AMO further allocates NS computation according to observed operator-wise geometry (Zhuang et al., 2026). Other variants modify scaling or update magnitude, including AdaMuon and OptMuon (Si et al., 2025; Yuan, 2026). Gradient-orthogonalized optimization has also been extended to ZO settings: JAGUAR-Muon, ZO-Muon, and ZO-MOPI combine function-value feedback with spectral or partial orthogonalization, often exploiting low-rank structure to control the noise and cost of ZO estimates (Petrov et al., 2025; Lang et al., 2026; Chen & Ma, 2026).

Online Learning Online learning considers sequential decision problems in which the learner repeatedly receives new data or feedback and updates its model over time, in contrast to conventional batch optimization over a fixed dataset. In online convex optimization, performance is commonly measured through static or dynamic regret, with path length and loss variation quantifying the difficulty of tracking a changing comparator or objective (Hazan, 2016; Yang et al., 2016). Adaptive online methods further adjust optimization geometry according to observed gradients, providing an important foundation for time-varying optimization (Duchi et al., 2011). Extending online optimization beyond convex objectives requires different notions of performance. Local regret and related stationarity measures have been developed for online nonconvex optimization, while dynamic variants explicitly account for concept drift and temporally changing objectives (Hazan et al., 2017; Hallak et al., 2021b; Gao et al., 2018). Online methods have also been studied under restricted feedback. Gradient-based and bandit variants of normalized online optimization address settings with either FO information or only function values, and ZO online methods further extend these ideas to limited-feedback and distributed environments (Guan et al., 2023; Roy et al., 2019). Communication-efficient and residual-feedback ZO algorithms have additionally been developed for distributed online nonconvex optimization (Kaya et al., 2023; Fang et al., 2026). A key challenge in nonstationary online optimization is to balance the stability provided by historical information with responsiveness to the current objective.

3 PRELIMINARIES

3.1 LARGE LANGUAGE MODELS FINE-TUNING

We consider fine-tuning the matrix-valued parameters $\mathbf{W} = (W^{(\ell)})_{\ell=1}^{L}$ of a pretrained LLM, where $W^{(\ell)} \in \mathbb{R}^{m_\ell \times n_\ell}$ denotes the ℓ-th trainable matrix block. The fine-tuning objective is formulated as

$$\min_{\mathbf{W}} F(\mathbf{W}) := \mathbb{E}_{\xi \sim \mathcal{D}} \left[f(\mathbf{W}; \xi) \right], \tag{1}$$

where ξ denotes a training sample drawn from distribution $\mathcal{D}$ and $f(\mathbf{W}; \xi)$ is the corresponding loss.

FO fine-tuning obtains the gradient $\nabla F(\mathbf{W})$ through backpropagation. While effective, backpropagation requires storing or recomputing intermediate activations, resulting in substantial memory overhead as LLMs scale. In contrast, ZO fine-tuning estimates $\nabla F(\mathbf{W})$ solely from function evaluations, without requiring explicit gradient computation or backpropagation (Ghadimi & Lan, 2013; Malladi et al., 2023). Given a smoothing parameter $\mu > 0$ and a random perturbation matrix Z, a standard symmetric two-point estimator is

$$\widehat{\nabla} F(\mathbf{W}) = \frac{F(\mathbf{W} + \mu Z) - F(\mathbf{W} - \mu Z)}{2\mu} Z. \tag{2}$$

By relying only on objective evaluations, ZO optimization reduces the memory cost associated with backpropagation and remains applicable in black-box settings where gradient information or access to internal backpropagation is unavailable. Recent ZO fine-tuning methods further restrict random perturbations to low-dimensional or low-rank subspaces, reducing the effective search dimension and improving gradient-estimation efficiency for LLMs (Chen et al., 2025; Yu et al., 2025).

3.2 MUON AND FINITE NEWTON–SCHULZ ORTHOGONALIZATION

Muon is a matrix-structured optimizer that exploits the geometry of matrix-valued parameters rather than treating their entries independently (Jordan et al., 2024; Liu et al., 2025). Given an gradient $\nabla F_t(\mathbf{W}_t)$ at round t, Muon first accumulates it through momentum,

$$\mathbf{M}_1 = \nabla F_1(\mathbf{W}_1), \qquad \mathbf{M}_t = \beta \mathbf{M}_{t-1} + \nabla F_t(\mathbf{W}_t), \quad t \geq 2, \tag{3}$$

where $\beta \in [0, 1)$ is the momentum coefficient. Muon then approximately orthogonalizes the momentum matrix, producing a more balanced spectral update than conventional coordinate-wise optimization (Bernstein & Newhouse, 2024; Lau et al., 2025).

In practice, the orthogonalization is implemented using a small number of NS iterations. For a nonzero momentum matrix $\mathbf{M}_t$, we initialize $\mathbf{X}_{t,0} = \mathbf{M}_t / \|\mathbf{M}_t\|_F$, and apply

$$\mathbf{X}_{t,j+1} = \frac{1}{2} \mathbf{X}_{t,j} \left(3\mathbf{I} - \mathbf{X}_{t,j}^\top \mathbf{X}_{t,j} \right), \qquad j = 0, \ldots, K_{\mathrm{NS}} - 1. \tag{4}$$

The resulting $\mathbf{X}_{t,K_{\mathrm{NS}}}$ is then used as the update direction.

3.3 ONLINE LEARNING AND REGRET

We consider LLM fine-tuning under the standard online learning protocol, where training data and supervision arrive sequentially over time (Hazan, 2016). At each round $t = 1, \ldots, T$, the learner maintains model parameters $\mathbf{W}_t$ and makes a prediction before observing the current feedback. Once the feedback is revealed, it induces a round-wise objective $F_t(\mathbf{W}_t)$, from which the gradient $\nabla F_t(\mathbf{W}_t)$ becomes available. The model is then updated according to

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \nabla F_t(\mathbf{W}_t), \tag{5}$$

where $\eta > 0$ denotes the learning rate. After the update, the current examples are discarded and not revisited in subsequent rounds. Unlike convex online optimization, where performance is commonly characterized relative to a global comparator, fine-tuning LLMs generally gives rise to nonconvex objectives for which local stationarity provides a more natural measure of optimization progress. Motivated by gradient-based stationarity criteria in online nonconvex optimization, we consider the cumulative squared gradient norm (Hazan et al., 2017; Aydore et al., 2019; Hallak et al., 2021a):

$$\mathcal{R}_T := \sum_{t=1}^{T} \|\nabla F_t(\mathbf{W}_t)\|_F^2 . \tag{6}$$

This quantity measures the cumulative FO stationarity of the online trajectory, with a smaller $\mathcal{R}_T$ indicating that the iterates remain closer to stationary points of the sequentially revealed objectives.

4 METHODS

4.1 ORIENTATION GAP IN ONLINE MUON

Online fine-tuning requires continual adaptation to objectives that may evolve as new data arrive over time (Yang et al., 2016). In such nonstationary settings, historical gradient information can improve update stability, but it may become less representative of the current objective. This issue is particularly relevant to Muon because its momentum explicitly aggregates optimization signals across rounds. Following Eq. (3), the momentum at round t can be written as

$$\mathbf{M}_t = \nabla F_t(\mathbf{W}_t) + \beta^{t-1}\nabla F_1(\mathbf{W}_1) + \sum_{s=2}^{t-1} \beta^{t-s}\nabla F_s(\mathbf{W}_s), \tag{7}$$

The remaining terms aggregate historical gradients from earlier rounds. As the objective evolves over time, these past gradients may become less representative of the current optimization landscape and induce an orientation that is no longer well aligned with the current gradient $\nabla F_t(\mathbf{W}_t)$.

This temporal lag poses a particular challenge to Muon's finite Newton–Schulz transformation. Let $\nabla F_t(\mathbf{W}_t) = \mathbf{P}_t\mathbf{\Gamma}_t\mathbf{Q}_t^\top$, $\mathbf{M}_t = \mathbf{U}_t\mathbf{\Sigma}_t\mathbf{V}_t^\top$ be the singular value decompositions of the current gradient and accumulated momentum, respectively. We characterize their orientation discrepancy by

$$\Delta_t := \left\|\mathbf{U}_t\mathbf{V}_t^\top - \mathbf{P}_t\mathbf{Q}_t^\top\right\|_F. \tag{8}$$

As the online objective evolves, gradient information accumulated from previous rounds may become stale relative to the current objective, leading to a non-negligible orientation gap Δ_t between the accumulated momentum and the current gradient.

Finite NS acts only on the singular values while preserving the singular-vector subspaces of its input (Kim & Oh, 2026; Li & Tsuchiya, 2026). For the classical NS iteration, each singular value follows the scalar map $\phi(\sigma) = \frac{1}{2}\sigma(3-\sigma^2)$. Thus, if $\mathbf{X}_{t,0} = \mathbf{U}_t\mathbf{\Sigma}_{t,0}\mathbf{V}_t^\top$, then after K_{NS} iterations,

$$\mathbf{X}_{t,K_{\mathrm{NS}}} = \mathbf{U}_t \cdot \phi^{\circ K}\left(\mathbf{\Sigma}_{t,0}\right) \cdot \mathbf{V}_t^\top, \tag{9}$$

where $\phi^{\circ K}$ denotes K repeated applications of ϕ to each diagonal entry of $\mathbf{\Sigma}_{t,0}$. The orientation discrepancy after finite NS therefore satisfies $\Delta_t^{\mathrm{NS}} = \Delta_t$. Thus, finite NS can rebalance the singular-value spectrum of the accumulated momentum, but it cannot independently correct the orientation mismatch caused by historical information that may lag behind the current objective.

4.2 GEOMETRIC-STEERING NEWTON–SCHULZ

The orientation gap identified above motivates introducing the current gradient directly into the NS trajectory, rather than allowing it to influence the transformation only through the accumulated momentum. We first normalize the current gradient as $\overline{\mathbf{G}}_t = \nabla F_t(\mathbf{W}_t)/\|\nabla F_t(\mathbf{W}_t)\|_F$. Starting from $\mathbf{X}_{t,0} = \mathbf{M}_t/\|\mathbf{M}_t\|_F$, GS-NS first applies three standard NS iterations,

$$\mathbf{X}_{t,j+1} = 0.5 \cdot \mathbf{X}_{t,j}\left(3\mathbf{I} - \mathbf{X}_{t,j}^\top\mathbf{X}_{t,j}\right), \qquad j = 0, 1, 2, \tag{10}$$

producing a near-orthogonal iterate $\mathbf{X}_{t,3}$ that still inherits the orientation of $\mathbf{M}_t$.

We then steer $\mathbf{X}_{t,3}$ using the current gradient through

$$\mathbf{T}_{t,3} = \overline{\mathbf{G}}_t - \mathbf{X}_{t,3} \cdot \mathrm{sym}\left(\mathbf{X}_{t,3}^\top\overline{\mathbf{G}}_t\right), \qquad \text{where } \mathrm{sym}(\mathbf{A}) = 0.5(\mathbf{A} + \mathbf{A}^\top), \tag{11}$$

which coincides with the tangent-space projection of the current gradient when $\mathbf{X}_{t,3}^\top\mathbf{X}_{t,3} = \mathbf{I}$ (Edelman et al., 1998). With finite NS iterations, however, $\mathbf{X}_{t,3}$ is only approximately orthogonal, so $\mathbf{T}_{t,3}$ is interpreted as a tangent-inspired steering direction. Rather than introducing an absolute steering scale, we adapt its magnitude to the local progress of the NS trajectory by matching it to the most recent NS correction:

$$\gamma_t = \alpha\frac{\|\mathbf{X}_{t,3} - \mathbf{X}_{t,2}\|_F}{\|\mathbf{T}_{t,3}\|_F}, \qquad \widetilde{\mathbf{X}}_{t,4} = \mathbf{X}_{t,3} + \gamma_t\mathbf{T}_{t,3}, \tag{12}$$

where $0 < \alpha \leq 1$ controls the steering strength.

Finally, one NS correction is applied to restore the orthogonalization structure,

$$\mathbf{X}_{t,5} = 0.5 \cdot \widetilde{\mathbf{X}}_{t,4} \left(3\mathbf{I} - \widetilde{\mathbf{X}}_{t,4}^{\top} \widetilde{\mathbf{X}}_{t,4}\right). \tag{13}$$

GS-NS therefore decouples orientation adaptation from spectral orthogonalization: the current gradient directly steers the intermediate orientation, while the final NS step restores the near-orthogonal geometry. The resulting three-step NS prefix, steering stage, and final NS correction maintain a comparable major matrix-multiplication budget to the standard five-step NS transformation.

4.3 Geometric-Steering Muon

GS-Muon provides a unified framework for both FO and ZO online fine-tuning. The two regimes differ only in how the current optimization signal is obtained while sharing the same momentum accumulation and GS-NS transformation. Algorithm 1 and Figure 2 summary the complete procedure.

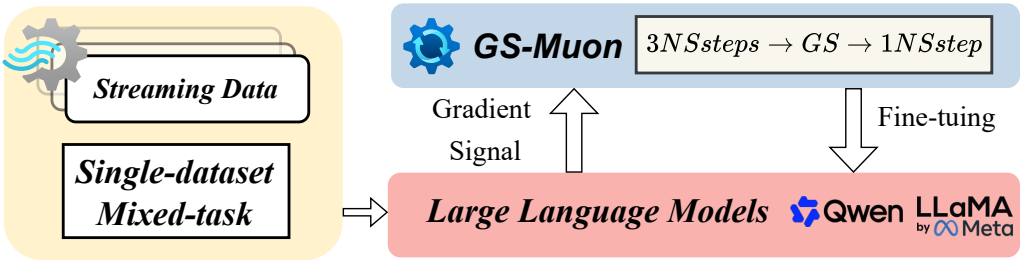


Figure 2: The frameworks of GS-Muon.

Algorithm 1 GS-Muon

Require: Initial parameters $\mathbf{W}_1$, momentum coefficient β, steering strength α,learning rates η

Initialize $\mathbf{M}_0 \leftarrow \mathbf{0}$
for $t = 1, \ldots, T$ **do**
  Obtain the current optimization signal: $\mathbf{G}_t \leftarrow \begin{cases} \nabla F_t(\mathbf{W}_t), & \text{FO}, \\ \widehat{\nabla} F_t(\mathbf{W}_t), & \text{ZO} \end{cases}$
  Update Muon momentum: $\mathbf{M}_t \leftarrow \beta \mathbf{M}_{t-1} + \mathbf{G}_t$
  Normalize: $\mathbf{X}_{t,0} \leftarrow \mathbf{M}_t / \|\mathbf{M}_t\|_F$
  Normalize the current signal: $\overline{\mathbf{G}}_t \leftarrow \mathbf{G}_t / \|\mathbf{G}_t\|_F$
  for $j = 0, 1, 2$ **do**
    $\mathbf{X}_{t,j+1} \leftarrow 0.5 \cdot \mathbf{X}_{t,j} \left(3\mathbf{I} - \mathbf{X}_{t,j}^{\top} \mathbf{X}_{t,j}\right)$
  end for
  Compute the tangent-inspired steering direction:

$$\mathbf{T}_{t,3} \leftarrow \overline{\mathbf{G}}_t - \mathbf{X}_{t,3} \cdot \mathrm{sym}\left(\mathbf{X}_{t,3}^{\top} \overline{\mathbf{G}}_t\right)$$

  Compute the steering scale:

$$\gamma_t \leftarrow \alpha \frac{\|\mathbf{X}_{t,3} - \mathbf{X}_{t,2}\|_F}{\|\mathbf{T}_{t,3}\|_F}$$

  Steer the intermediate iterate: $\widetilde{\mathbf{X}}_{t,4} \leftarrow \mathbf{X}_{t,3} + \gamma_t \mathbf{T}_{t,3}$
  Apply the final NS correction:

$$\mathbf{X}_{t,5} \leftarrow 0.5 \cdot \widetilde{\mathbf{X}}_{t,4} \left(3\mathbf{I} - \widetilde{\mathbf{X}}_{t,4}^{\top} \widetilde{\mathbf{X}}_{t,4}\right)$$

  Update the model: $\mathbf{W}_{t+1} \leftarrow \mathbf{W}_t - \eta \mathbf{X}_{t,5}$
end for
return $\mathbf{W}_{T+1}$

5 REGRET ANALYSIS

We analyze the cumulative squared-gradient measure $\mathcal{R}_T$ in Eq. (6), which serves as a stationarity criterion for evolving nonconvex objectives (Roy et al., 2022). We condition on a fixed data stream and initialization, so that all expectations below are taken only over the randomness of the ZO perturbations. Our analysis relies on three standard assumptions that ensure smooth descent, bounded gradient magnitude, and controlled temporal variation (Ghadimi & Lan, 2013; Hazan et al., 2017; Roy et al., 2022). We state these conditions in global form so that the function-variation budget can also be used to control inter-round gradient drift:

Assumption 5.1 (Smoothness and bounded initial gap)**.** Each $F_t : \mathbb{R}^d \to \mathbb{R}$ has an L-Lipschitz gradient and satisfies $F_t(\mathbf{W}) \geq F_{\inf}$ for all $\mathbf{W}$.

Assumption 5.2 (Bounded gradients)**.** $\|\nabla F_t(\mathbf{W}_t)\|_F \leq G$ at every t, almost surely under ZO access.

Assumption 5.3 (Bounded temporal variation)**.** The fixed loss sequence has a finite deterministic variation budget V_T^F such that $\sum_{t=2}^{T} \sup_{\mathbf{W}\in\mathbb{R}^d} |F_t(\mathbf{W}) - F_{t-1}(\mathbf{W})| \leq V_T^F$.

These assumptions control the local smoothness of each objective, the magnitude of the optimization signal, and the temporal drift of the online loss sequence. Combined with the alignment properties of GS-NS, they allow us to quantify how objective variation and momentum-induced tracking error affect online optimization. We first establish the following guaranty under exact FO access:

Theorem 5.4 (FO nonconvex regret)**.** *Under Assumptions 5.1–5.3, let $d = \sum_\ell m_\ell n_\ell$ and $k = \sum_\ell \min(m_\ell, n_\ell)$. We denote $F_1(\mathbf{W}_1) - F_{\inf} \leq \Delta_F < \infty$, and let $\eta > 0$, $0 \leq \beta < 1$, and $0 \leq \alpha < 1/4$. With exact FO signals, GS-Muon satisfies*

$$\mathcal{R}_T \leq \frac{G}{1-4\alpha}\left[\frac{\Delta_F + V_T^F}{\eta} + 2L\eta kT + \frac{2\sqrt{k}\,\beta}{1-\beta}\left(G + 2L\eta\sqrt{k}(T-1) + \sqrt{8L(T-1)V_T^F}\right)\right]. \quad (14)$$

Remark 5.5. The bound separates the effects of the initial optimality gap, temporal objective variation, model dimensions, and optimizer parameters, with L and G entering as problem-dependent constants. The factor $\beta/(1-\beta)$ captures the amplification of momentum-tracking error as historical information is retained more strongly, whereas $1 - 4\alpha$ provides a sufficient lower bound on the alignment preserved by the complete GS-NS transformation. GS-Muon itself allows $\alpha \in [0, 1]$. For fixed $\beta < 1$ and $\alpha < 1/4$, bounded V_T^F, and $\eta = \Theta(T^{-1/2})$, Eq. (14) yields $\mathcal{R}_T = \mathcal{O}(\sqrt{T})$, $\frac{\mathcal{R}_T}{T} = \mathcal{O}(T^{-1/2})$, showing vanishing average FO stationarity along the online trajectory.

Under ZO access, use the symmetric estimator in Eq. (2) with one fresh $\mathcal{N}(0, I_d)$ direction and two evaluations of the same current loss. Lemma B.1 in Appendix B.1 establishes its smoothing bias and conditional variance. Analyzing the accumulated noise in momentum, following the proof strategy of Cutkosky & Mehta (2020), gives the following explicit bound.

Theorem 5.6 (ZO nonconvex regret)**.** *Under Assumptions 5.1–5.3, let $\eta, \mu > 0$, $0 \leq \beta < 1$, and $0 \leq \alpha < 1/4$. With the single-direction Gaussian estimator above, GS-Muon satisfies*

$$\begin{aligned}\mathbb{E}[\mathcal{R}_T] \leq \frac{G}{1-4\alpha}\Bigg[&\frac{\Delta_F + V_T^F}{\eta} + 2L\eta kT + \frac{2\sqrt{k}\,\beta}{1-\beta}\left(G + 2L\eta\sqrt{k}(T-1) + \sqrt{8L(T-1)V_T^F}\right) \\ &+ 2\sqrt{k}\,T\left(L\mu\sqrt{d} + \sqrt{\frac{1-\beta}{1+\beta}\left[2(d+1)G^2 + \frac{L^2\mu^2 d(d+2)(d+4)}{2}\right]}\right)\Bigg].\end{aligned} \quad (15)$$

Remark 5.7. Compared with the FO bound, the additional ZO terms explicitly capture the smoothing bias, through $L\mu\sqrt{d}$, and the stochastic estimation noise, whose magnitude is attenuated by $\sqrt{(1-\beta)/(1+\beta)}$. This reveals a tracking–averaging trade-off in the momentum coefficient: increasing β strengthens temporal averaging and suppresses ZO noise, but simultaneously amplifies the momentum-tracking error through $\beta/(1-\beta)$, making the update less responsive to an evolving objective. For a stationary objective, i.e., $V_T^F = 0$, and fixed problem dimensions, choosing $1 - \beta = T^{-1/2}$, $\eta = \Theta(T^{-3/4})$, and $\mu = \mathcal{O}(T^{-1/4})$, with fixed $\alpha < 1/4$, yields $\mathbb{E}[\mathcal{R}_T] = \mathcal{O}(T^{3/4})$, $\frac{\mathbb{E}[\mathcal{R}_T]}{T} = \mathcal{O}(T^{-1/4})$. For bounded temporal variation $V_T^F = \mathcal{O}(1)$, the choices $1 - \beta = T^{-1/3}$, $\eta = \Theta(T^{-2/3})$, and $\mu = \mathcal{O}(T^{-1/6})$ give $\mathbb{E}[\mathcal{R}_T] = \mathcal{O}(T^{5/6})$, $\frac{\mathbb{E}[\mathcal{R}_T]}{T} = \mathcal{O}(T^{-1/6})$.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Datasets and Models. We evaluate on eleven NLP datasets from GLUE, SuperGLUE, and established question-answering benchmarks: SST-2, RTE, CB, BoolQ, WSC, WiC, MultiRC, COPA, ReCoRD, SQuAD, and DROP (Wang et al., 2018; 2019; Rajpurkar et al., 2016; Dua et al., 2019). Together, these datasets cover a broad range of language understanding and reasoning tasks, including sentiment classification, natural language inference, coreference resolution, commonsense reasoning, multiple-choice prediction, and reading comprehension. We evaluate two LLM backbones at different scales, Qwen2.5-1.5B (Qwen Team, 2024) and Llama-3-8B (Grattafiori et al., 2024). To emulate evolving online environments, we organize examples into sequential streams comprising single-dataset streams, sequential within-family task shifts, and heterogeneous cross-task streams. At each round, the model receives the current input, makes a prediction before observing its feedback, and then performs a single update using the newly revealed example.

Baselines. We organize the compared methods into training-free, FO, and ZO baselines. Zero-shot inference and in-context learning (ICL) are included as training-free references, where model parameters remain unchanged throughout the stream. For FO fine-tuning, we compare GS-Muon (FO) with standard Muon fine-tuning (FT-Muon) (Jordan et al., 2024). For ZO fine-tuning, we compare GS-Muon (ZO) with MeZO (Malladi et al., 2023), HiZOO (Zhao et al., 2025), SubZero (Yu et al., 2025), and LOZO (Chen et al., 2025). Whenever applicable, compared optimization methods use the same trainable matrices, data streams, task objectives, and hyperparameter tuning budgets. More experimental details appear in Appendix A.1.

6.2 MAIN EXPERIMENTS

Single-Dataset Online Streams. Tables 1 and 2 report the online fine-tuning performance on eleven individual datasets across the two model scales. Under FO access, GS-Muon consistently improves over the matched FT-Muon baseline and achieves the strongest overall performance, demonstrating that explicitly steering the momentum-inherited orientation toward the current optimization signal is beneficial even within a single-task stream. Under ZO access, GS-Muon likewise achieves the best average performance across both model scales and outperforms representative ZO baselines, including MeZO, HiZOO, SubZero, and LOZO, on the majority of datasets.

Table 1: Online fine-tuning performance on Qwen2.5-1.5B across eleven NLP datasets.

Method	**Classification**							**Choice**		**Generation**		**Average**
	SST-2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	
Zero-shot	58.8	81.0	16.4	69.0	50.1	50.5	58.0	77.3	55.2	74.5	30.6	56.5
ICL	89.3	82.0	32.5	79.1	50.6	54.5	68.0	78.8	54.1	53.8	12.8	59.6
FT-Muon	86.7	83.9	80.0	78.9	50.4	51.4	76.3	80.0	56.6	85.1	48.6	70.7
GS-Muon (FO)	88.5	84.6	82.4	79.1	52.9	61.5	78.7	81.5	57.0	85.9	49.8	**72.9**
MeZO	61.7	68.1	46.0	66.8	52.5	55.2	57.2	76.8	53.0	70.0	28.8	58.3
HiZOO	76.0	69.3	65.2	64.4	54.3	50.4	57.7	75.3	52.5	68.4	29.1	60.2
SubZero	61.6	77.8	36.4	66.0	50.0	56.9	62.9	77.0	51.6	65.0	27.9	57.6
LOZO	68.5	72.7	29.2	69.3	50.7	54.5	61.1	76.5	52.1	66.2	32.3	57.6
GS-Muon (ZO)	76.3	78.3	66.4	69.8	54.9	55.1	63.0	77.3	54.0	69.5	31.1	**63.2**

Table 2: Online fine-tuning performance on Llama-3-8B across eleven NLP datasets.

Method	**Classification**							**Choice**		**Generation**		**Average**
	SST-2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	
Zero-shot	57.9	52.7	49.6	68.7	53.8	52.7	56.8	73.5	62.6	53.6	31.9	55.8
ICL	69.3	66.3	48.8	69.9	56.9	59.9	56.2	78.9	62.9	55.1	32.7	59.7
MeZO	69.6	52.6	65.6	65.5	53.3	51.7	57.1	82.8	69.4	73.5	38.0	61.7
HiZOO	56.5	49.8	63.2	58.0	49.3	51.3	54.2	81.5	54.5	69.4	32.0	56.3
SubZero	73.5	52.2	57.2	63.0	54.5	52.6	57.6	84.3	67.9	74.2	35.6	61.1
LOZO	68.5	50.6	56.0	66.5	53.8	52.2	56.4	83.3	68.1	73.3	34.0	60.2
GS-Muon (ZO)	78.4	53.6	62.0	69.9	55.1	58.7	58.7	84.8	71.1	74.7	35.4	**63.9**

Cross-Dataset and Cross-Task Online Streams. Table 3 reports the results on sequential task-shift streams using Qwen2.5-1.5B, covering within-family classification and generation shifts as well as a heterogeneous cross-task shift (Each source dataset contains 500 examples.). Compared with the corresponding Muon and ZO baselines, GS-Muon demonstrates strong performance in these

nonstationary settings, where accumulated optimization information can become less representative after task transitions. By explicitly steering the momentum-inherited orientation toward the current optimization signal, GS-Muon improves responsiveness to newly arriving tasks while retaining Muon's matrix-structured spectral transformation.

Table 3: Online performance on task-shift streams. Each source dataset contributes 500 examples.

Model	Method	Classification Shift	Generation Shift	Cross-Task Shift	Average
		SST-2 → WiC	SQuAD → DROP	WSC → ReCoRD → SQuAD	
Qwen2.5-1.5B	Zero-shot	56.8	35.5	50.2	47.4
	ICL	71.2	33.3	52.4	52.3
	FT-Muon	68.4	62.3	59.5	63.4
	GS-Muon (FO)	72.1	66.4	62.6	**67.0**
	MeZO	55.8	43.2	51.4	50.1
	HiZOO	53.2	46.4	54.1	51.2
	SubZero	54.3	40.3	50.4	48.4
	LOZO	55.2	48.8	54.6	52.9
	GS-Muon (ZO)	58.5	49.8	56.6	**55.0**

6.3 Ablation Studies

GS-NS Components. We conduct ablation studies on three representative datasets, CB, COPA, and WSC, using Qwen2.5-1.5B to isolate the contribution of the proposed GS-NS transformation (Table 4). We consider two variants of GS-Muon. First, **w/o Steering** removes the geometric-steering stage while retaining the three-step NS prefix and the final NS correction, resulting in four consecutive standard NS steps. This variant isolates the effect of explicitly incorporating the current optimization signal into the NS trajectory. Second, **w/o GS-NS** removes the GS-NS transformation entirely and directly uses the accumulated momentum as the update direction, allowing us to assess the contribution of the matrix-structured geometric transformation as a whole.

Table 4: Ablation study of GS-NS on Qwen2.5-1.5B under FO and ZO access.

Access	Variant	CB	COPA	WSC	Avg.
FO	w/o GS-NS	51.2	76.8	51.1	59.7
	w/o Steering	72.4	77.5	48.4	66.1
	GS-Muon	**82.4**	**81.5**	**52.9**	**72.3**
ZO	w/o GS-NS	28.4	54.5	51.1	44.7
	w/o Steering	14.8	77.0	54.0	48.6
	GS-Muon	**66.4**	**77.3**	**54.9**	**66.2**

Hyperparameter Sensitivity. We further investigate the sensitivity of GS-Muon to three key hyperparameters under ZO access: the learning rate η, the momentum coefficient β, and the steering strength α. Specifically, we vary $\eta \in \{1\times10^{-5}, 3\times10^{-5}, 1\times10^{-4}, 3\times10^{-4}\}$ on SST-2, $\beta \in \{0.8, 0.9, 0.95, 0.99\}$ on MultiRC, and $\alpha \in \{0.05, 0.1, 0.15, 0.2\}$ on RTE. As shown in Figure 3 in Appendix A.2, the learning-rate study on SST-2 shows a clear advantage for the larger setting, with $\eta = 3\times10^{-4}$ consistently achieving the strongest cumulative online performance after the initial adaptation stage. In contrast, the four momentum settings on MultiRC exhibit highly similar trajectories, indicating that GS-Muon is relatively insensitive to β within the examined range. For the steering strength, $\alpha \in \{0.05, 0.1, 0.15\}$ yields comparable performance on RTE, whereas the more aggressive setting $\alpha = 0.2$ consistently underperforms, suggesting that moderate steering is robust while excessive steering can degrade online adaptation.

7 Conclusion

We proposed GS-Muon for online LLM fine-tuning. GS-NS inserts the current optimization signal into a finite Newton–Schulz trajectory: a short NS prefix establishes the reference geometry, a tangent-inspired step redirects the momentum-derived orientation, and a final NS step locally reduces the resulting orthogonality residual when the steered iterate lies in the NS contraction region. The same transformation accepts an exact FO gradient or a ZO estimate, making signal access an application setting rather than a separate optimizer design. Our analysis identifies what the finite transformation guarantees. The steering stage has nonnegative linear alignment with the available signal, its deviation from a tangent direction is controlled by the prefix residual, and the final NS stage contracts the residual locally. The proposed streaming protocol across three model scales and eleven datasets will test whether these geometric properties improve online adaptation in practice.

AI USE STATEMENT

In this work, generative AI tools were used to assist with language polishing, manuscript revision, and aspects of theoretical development, including exploring proof strategies and checking intermediate derivations. All mathematical arguments, proofs, equations, notation, references, and technical claims were subsequently examined and verified by the authors. Generative AI was not used to conduct experiments, produce experimental measurements, or alter empirical results; human-subject data collection and participant interaction are not applicable to this work. The tools served only as research and writing assistants, while the formulation of the methods, validation of the theoretical results, experimental design, analysis, interpretation, and scientific conclusions remained the responsibility of the authors. The authors take full responsibility for all content in the final manuscript, including material developed with the assistance of generative AI.

REFERENCES

Sergul Aydore, Tianhao Zhu, and Dean P Foster. Dynamic local regret for non-convex online forecasting. *Advances in neural information processing systems*, 32, 2019.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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

Jiahe Chen and Ziye Ma. Accelerating zeroth-order spectral optimization with partial orthogonalization from power iteration. *arXiv preprint arXiv:2605.09034*, 2026.

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*, volume 2025, pp. 62581–62607, 2025.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2260–2268. PMLR, 2020. URL https://proceedings.mlr.press/v119/cutkosky20b.html.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

Yihe Dong and Will Sawin. Muon ^{p}: Muon with fractional spectral powers. *arXiv preprint arXiv:2606.13867*, 2026.

Zhehang Du and Weijie Su. The newton-muon optimizer. *arXiv preprint arXiv:2604.01472*, 2026.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2368–2378. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246/.

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

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Jinjie Fang, Runwen You, Wanli Shi, Wenkang Wang, Ganyu Wang, Haozhen Zhang, Yi Chang, and Bin Gu. Online black-box prompt optimization with regret guarantees under noisy feedback. In *International Conference on Learning Representations*, volume 2026, pp. 121524–121555, 2026.

Xiand Gao, Xiaobo Li, and Shuzhong Zhang. Online learning with non-convex losses and non-stationary regret. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 235–243. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/gao18a.html.

Tanmay Gautam, Youngsuk Park, Hao Zhou, Parameswaran Raman, and Wooseok Ha. Variance-reduced zeroth-order methods for fine-tuning language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15180–15208. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/gautam24a.html.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ziwei Guan, Yi Zhou, and Yingbin Liang. Online nonconvex optimization with limited instantaneous oracle feedback. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 3328–3355. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/guan23a.html.

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.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Nadav Hallak, Panayotis Mertikopoulos, and Volkan Cevher. Regret minimization in stochastic non-convex learning via a proximal-gradient approach. In *International Conference on Machine Learning*, pp. 4008–4017. PMLR, 2021a.

Nadav Hallak, Panayotis Mertikopoulos, and Volkan Cevher. Regret minimization in stochastic non-convex learning via a proximal-gradient approach. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4008–4017. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/hallak21a.html.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2 (3-4):157–325, 2016.

Elad Hazan, Karan Singh, and Cyril Zhang. Efficient regret minimization in non-convex games. In *International Conference on Machine Learning*, pp. 1433–1441. PMLR, 2017.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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

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

Ege C Kaya, Mehmet Berk Sahin, and Abolfazl Hashemi. Communication-efficient zeroth-order distributed online optimization: Algorithm, theory, and applications. *IEEE Access*, 11:61173–61191, 2023.

Gyu Yeol Kim and Min-hwan Oh. Convergence of muon with newton-schulz. In *International Conference on Learning Representations*, volume 2026, pp. 45672–45718, 2026.

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

Yicheng Lang, Changsheng Wang, Yihua Zhang, Mingyi Hong, Zheng Zhang, Wotao Yin, and Sijia Liu. Powering up zeroth-order training via subspace gradient orthogonalization. *arXiv preprint arXiv:2602.17155*, 2026.

Tim Tsz-Kit Lau, Qi Long, and Weijie Su. Polargrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

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.

Mingyi Li and Taira Tsuchiya. Muon with finite newton-schulz: The smoothing benefit in nonsmooth nonconvex optimization, 2026. URL https://arxiv.org/abs/2608.26288.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-Decomposed Low-Rank Adaptation. In *International Conference on Machine Learning*, volume 235, pp. 32100–32121, 2024. URL https://mlanthology.org/icml/2024/liu2024icml-dora/.

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

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. doi: 10.1007/s10208-015-9296-2. URL https://doi.org/10.1007/s10208-015-9296-2.

Tien-Phat Nguyen, Truong Nguyen, Minh-Phuc Truong, Tuc Nguyen, James Bailey, and Trung Le. Spectral flattening is all muon needs: How orthogonalization controls learning rate and convergence, 2026. URL https://arxiv.org/abs/2605.13079.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Saurabh Page, Advait Joshi, and SS Sonawane. Muonall: Muon variant for efficient finetuning of large language models. *arXiv preprint arXiv:2511.06086*, 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.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Abhishek Roy, Krishnakumar Balasubramanian, Saeed Ghadimi, and Prasant Mohapatra. Multi-point bandit algorithms for nonstationary online nonconvex optimization. *arXiv preprint arXiv:1907.13616*, 2019.

Abhishek Roy, Krishnakumar Balasubramanian, Saeed Ghadimi, and Prasant Mohapatra. Stochastic zeroth-order optimization under nonstationarity and nonconvexity. *Journal of Machine Learning Research*, 23(64):1–47, 2022. URL https://jmlr.org/papers/v23/19-750.html.

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

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 93423–93444, 2025.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355. Association for Computational Linguistics, 2018. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning*, pp. 449–457. PMLR, 2016.

Ziming Yu, Pan Zhou, Sike Wang, Jia Li, Mi Tian, and Hua Huang. Zeroth-order fine-tuning of llms in random subspaces. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4475–4485. IEEE, 2025.

Ganzhao Yuan. Optmuon: Closed-loop orthogonalized momentum methods for stochastic optimization with zero-noise optimality. *arXiv preprint arXiv:2606.08783*, 2026.

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.

Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor Tsang. Second-order fine-tuning without pain for llms: A hessian informed zeroth-order optimizer. In *International Conference on Learning Representations*, volume 2025, pp. 43496–43520, 2025.

Xinlin Zhuang, Panyi Ouyang, Yichen Li, Jiangming Shi, Yizhang Chen, Shuman Liu, Ying Qian, Weiyang Liu, Haibo Zhang, and Imran Razzak. Amo: Adaptive muon orthogonalization. *arXiv preprint arXiv:2605.17806*, 2026.

A Supplementary Materials for the Experiment

A.1 Supplementary Experimental setup

Datasets and task interface. We evaluate on eleven NLP datasets from GLUE, SuperGLUE: SST-2, RTE, CB, BoolQ, WSC, WiC, MultiRC, COPA, ReCoRD, SQuAD , and DROP. These tasks cover sentiment classification, natural language inference, coreference, word-sense reasoning, multiple-choice reasoning, cloze completion, and extractive or generative question answering. Dataset-specific official metrics are used, including accuracy, macro-F1, exact match, and normalized F1 where appropriate. All optimizers share the same text interface, task instructions, response markers, label verbalizers, and response parsers. Classification and choice tasks are scored using response-token-normalized log-probabilities, while generative tasks use deterministic greedy decoding.

Online stream construction. To emulate the sequential nature of online LLM fine-tuning, we construct streaming training sequences in which examples arrive over time and are processed once in their observed order, without replaying past examples. Under this streaming protocol, we consider three levels of online adaptation difficulty. Single-dataset streams contain examples from a single task and evaluate continual adaptation within a relatively stable task environment. Within-family streams sequentially combine related tasks from classification, choice or cloze, and question-answering families without resetting the optimizer at task boundaries. Heterogeneous streams further combine tasks from different families under the same response-token training objective, creating more pronounced shifts in both task structure and data distribution.

Online evaluation protocol. We follow a sequential online evaluation protocol. At round t, the model receives the current input and makes a prediction using $\mathbf{W}_t$ before observing its label. The feedback is then revealed, the prediction is scored, and the current objective F_t is constructed. FO methods compute $\nabla F_t(\mathbf{W}_t)$, whereas ZO methods estimate the current optimization signal through symmetric function evaluations. The optimizer subsequently performs a single update, after which the current example is discarded and cannot be revisited. To characterize performance throughout the stream, we report cumulative average task metrics over all examples observed up to the current round.

Computational cost. Each single-direction ZO update requires two evaluations of the current training objective, and GS-NS reuses the resulting estimator without additional training-loss queries. For a tall matrix, one ordinary NS step requires approximately two major matrix multiplications. Thus, a five-step NS transformation requires approximately ten such products. GS-NS uses six products in its three-step NS prefix, approximately two in the geometric-steering stage, and two in the final NS correction, giving the same coarse total of approximately ten major matrix products.

Baseline details. Detailed descriptions of the compared baselines are provided below:

- **Zero-shot** directly evaluates the pretrained model on each downstream task without providing task-specific demonstrations or updating model parameters. It serves as a training-free reference for the model's inherent task-solving capability before online adaptation.
- **In-Context Learning (ICL)** (Dong et al., 2024) provides several task demonstrations in the input context to guide prediction while keeping all model parameters fixed. In contrast to fine-tuning, adaptation is achieved entirely through contextual conditioning rather than parameter optimization.
- **FT-Muon** (Jordan et al., 2024) applies standard Muon to first-order fine-tuning. Muon first accumulates gradient information through momentum and then transforms the momentum matrix using a finite Newton–Schulz iteration to obtain an approximately orthogonalized update direction. We use FT-Muon as the primary first-order reference for evaluating the effect of introducing geometric steering into the Muon transformation.
- **MeZO** (Malladi et al., 2023) performs zeroth-order fine-tuning without backpropagation. At each optimization step, it evaluates the objective under opposite parameter perturbations and uses the resulting loss difference to construct a gradient-like update. Since only forward evaluations are required, MeZO avoids storing the intermediate activations needed for standard backpropagation.

- **HiZOO** (Zhao et al., 2025) incorporates curvature information into zeroth-order fine-tuning. It estimates diagonal second-order information and uses it to adapt the scaling of perturbations and parameter updates across coordinates, allowing the optimizer to account for heterogeneous local curvature while retaining a backpropagation-free optimization procedure.
- **SubZero** (Yu et al., 2025) performs zeroth-order optimization in a randomly constructed low-dimensional subspace instead of perturbing the full parameter space. Restricting optimization to a smaller search space reduces the effective perturbation dimension and helps control the estimation difficulty associated with high-dimensional ZO fine-tuning.
- **LOZO** (Chen et al., 2025) exploits the matrix structure of LLM parameters by constructing zeroth-order perturbations with a low-rank form. The resulting structured search directions reduce the effective optimization dimension and provide an alternative to unrestricted full-space perturbations for memory-efficient LLM fine-tuning.

Hyperparameter selection. To ensure a fair comparison, all optimization methods are tuned under the same hyperparameter search budget whenever applicable. For GS-Muon, we search the learning rate over $\eta \in \{1 \times 10^{-5}, 3 \times 10^{-5}, 1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}\}$, the momentum coefficient over $\beta \in \{0.8, 0.9, 0.95\}$, and the steering strength over $\alpha \in \{0, 0.1, 0.25\}$. Under ZO access, we additionally search the perturbation scale over $\mu \in \{1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}\}$. For the remaining baselines, method-specific hyperparameters are tuned over comparable search budgets following their recommended configurations.Hyperparameters are selected separately for each model and signal-access regime,while the trainable parameter set, online data stream, and task objective are kept fixed across compared methods.

A.2 Supplementary Experimental results

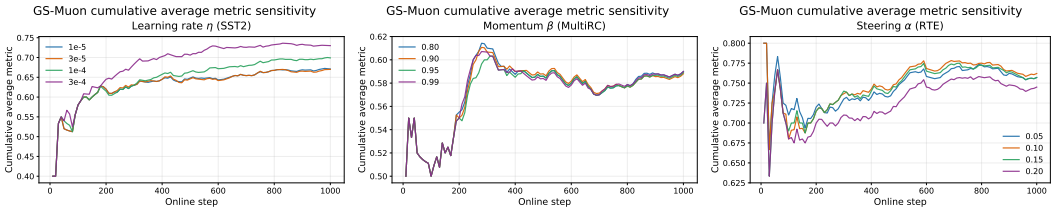


Figure 3: Sensitivity analysis for η, β and α on ZO case

B Theoretical analysis

B.1 Gaussian smoothing and estimation error

We use the standard Gaussian-smoothing framework of (Nesterov & Spokoiny, 2017; Ghadimi & Lan, 2013). The former gives the symmetric two-point expectation identity; the latter summarizes smoothing estimates for a forward-difference oracle. The theorem below specializes these facts to the single-direction estimator in Eq. (2). Its explicit Euclidean bias and centered-variance constants are derived below, rather than quoted as identical constants from those references. Fix a round, condition on $\mathcal{H}_t$, and suppress its index. Let w be the vectorization of the current matrices and $f = F_t$. Define

$$f_\mu(w) = \mathbb{E}_z f(w + \mu z).$$

The one-pair estimator is

$$h(z) = \frac{f(w + \mu z) - f(w - \mu z)}{2\mu} z, \qquad z \sim \mathcal{N}(0, I_d).$$

All expectations in this subsection are conditional on the same current objective and iterate.

Lemma B.1 (Gaussian two-point estimation; adapted from Nesterov & Spokoiny (2017); Ghadimi & Lan (2013))**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ have an L-Lipschitz gradient and $\|\nabla f(w)\|_2 \le G$. Conditionally on*

the current f and w, draw one $z \sim \mathcal{N}(0, I_d)$ and evaluate the same f at $w + \mu z$ and $w - \mu z$, with $\mu > 0$. Then the estimator satisfies $\mathbb{E}_t h(z) = \nabla f_\mu(w)$ and

$$\|\nabla f_\mu(w) - \nabla f(w)\|_2 \le L\mu\sqrt{d}, \qquad \mathbb{E}_t\|h(z) - \nabla f_\mu(w)\|_2^2 \le 2(d+1)G^2 + \frac{L^2\mu^2 d(d+2)(d+4)}{2}. \tag{16}$$

The bound on the current gradient is supplied by Assumption 5.2.

Proof. Global Lipschitz continuity of ∇f gives the growth bounds

$$\|\nabla f(w+u)\|_2 \le \|\nabla f(w)\|_2 + L\|u\|_2, \qquad |f(w+u)| \le |f(w)| + \|\nabla f(w)\|_2\|u\|_2 + \tfrac{L}{2}\|u\|_2^2.$$

Thus the relevant functions and their derivatives have at most polynomial growth and are integrable against a Gaussian density. Differentiation under the expectation is justified by these bounds and yields $\nabla f_\mu(w) = \mathbb{E}_z[\nabla f(w + \mu z)]$. For completeness, let $p(z)$ be the standard Gaussian density. Coordinatewise integration by parts, using $\partial_j p(z) = -z_j p(z)$ and the vanishing Gaussian boundary term, gives

$$\mathbb{E}[f(w+\mu z) z_j] = -\int_{\mathbb{R}^d} f(w+\mu z)\, \partial_j p(z)\, dz = \int_{\mathbb{R}^d} \mu\, \partial_j f(w+\mu z)\, p(z)\, dz = \mu\, \partial_j f_\mu(w).$$

The substitution $z \mapsto -z$ also gives $\mathbb{E}[f(w-\mu z)z] = -\mathbb{E}[f(w+\mu z)z]$. Combining the two identities proves

$$\mathbb{E}[h(z)] = \frac{1}{\mu}\mathbb{E}[f(w+\mu z)z] = \nabla f_\mu(w).$$

The preceding gradient representation, Jensen's inequality, and L-smoothness imply

$$\begin{aligned}
\|\nabla f_\mu(w) - \nabla f(w)\|_2 &= \|\mathbb{E}[\nabla f(w+\mu z) - \nabla f(w)]\|_2 \\
&\le \mathbb{E}\|\nabla f(w+\mu z) - \nabla f(w)\|_2 \\
&\le L\mu\mathbb{E}\|z\|_2 \\
&\le L\mu\sqrt{\mathbb{E}\|z\|_2^2} \\
&= L\mu\sqrt{d}.
\end{aligned}$$

The two Taylor remainders obey

$$f(w \pm \mu z) = f(w) \pm \mu\langle \nabla f(w), z\rangle + \rho_\pm(z), \qquad |\rho_\pm(z)| \le \tfrac{L}{2}\mu^2\|z\|_2^2.$$

Therefore, with the ordinary definitions

$$a_\mu(z) = \frac{\rho_+(z) - \rho_-(z)}{2\mu}, \qquad A(z) = \langle \nabla f(w), z\rangle z, \qquad V(z) = a_\mu(z) z,$$

we have

$$h(z) = A(z) + V(z), \qquad |a_\mu(z)| \le \tfrac{L\mu}{2}\|z\|_2^2, \qquad \|V(z)\|_2^2 \le \tfrac{L^2\mu^2}{4}\|z\|_2^6.$$

First, $\mathbb{E}[A(z)] = \mathbb{E}[zz^\top]\nabla f(w) = \nabla f(w)$. If $\nabla f(w) \neq 0$, rotation invariance allows $\nabla f(w) = \|\nabla f(w)\|_2 e_1$ for computing its second moment. The scalar moments $\mathbb{E}[z_i^2] = 1$, $\mathbb{E}[z_i^4] = 3$, and independence give

$$\begin{aligned}
\mathbb{E}\|A(z)\|_2^2 &= \|\nabla f(w)\|_2^2 \mathbb{E}\left[z_1^2 \sum_{j=1}^d z_j^2\right] \\
&= \|\nabla f(w)\|_2^2 \left(\mathbb{E}[z_1^4] + \sum_{j=2}^d \mathbb{E}[z_1^2]\mathbb{E}[z_j^2]\right) \\
&= (d+2)\|\nabla f(w)\|_2^2.
\end{aligned}$$

The identity also holds at $\nabla f(w) = 0$. Subtracting the squared mean yields

$$\mathbb{E}\|A - \nabla f(w)\|_2^2 = \mathbb{E}\|A\|_2^2 - \|\nabla f(w)\|_2^2 = (d+1)\|\nabla f(w)\|_2^2.$$

Similarly, using $\mathbb{E}z_i^6 = 15$ and expanding the cube gives

$$\begin{aligned}\mathbb{E}\|z\|_2^6 &= \sum_i \mathbb{E}z_i^6 + 3\sum_{i\neq j}\mathbb{E}[z_i^4 z_j^2] + 6\sum_{i<j<r}\mathbb{E}[z_i^2 z_j^2 z_r^2] \\ &= 15d + 9d(d-1) + d(d-1)(d-2) = d(d+2)(d+4).\end{aligned}$$

Let $v = \mathbb{E}[V(z)]$, so that $\nabla f_\mu(w) = \nabla f(w) + v$. The two centered components need not be independent. The deterministic inequality $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$ is enough:

$$\begin{aligned}\mathbb{E}\|h - \nabla f_\mu(w)\|_2^2 &= \mathbb{E}\|(A - \nabla f(w)) + (V - v)\|_2^2 \\ &\le 2\mathbb{E}\|A - \nabla f(w)\|_2^2 + 2\mathbb{E}\|V - v\|_2^2 \\ &\le 2(d+1)\|\nabla f(w)\|_2^2 + 2\mathbb{E}\|V\|_2^2 \\ &\le 2(d+1)G^2 + \tfrac{L^2\mu^2}{2}d(d+2)(d+4).\end{aligned}$$

□

B.2 LEMMAS

Lemma B.2 (Three-step prefix)**.** *Let $K \in \mathbb{R}^{m\times n}$ be nonzero, with positive singular values $s_1, \dots, s_r$, where $r = \operatorname{rank}(\mathbf{K})$. Let Y be the three-step classical NS output initialized at $\mathbf{K}/\|\mathbf{K}\|_F$. With*

$$\phi(a) = \frac{a(3-a^2)}{2}, \qquad y_i = \phi^{\circ 3}(s_i/\|K\|_F),$$

the output has the same left and right singular-vector pairs as $\mathbf{K}$ and satisfies

$$\frac{s_i}{\|\mathbf{K}\|_F} \le y_i \le 1 \quad (s_i > 0), \qquad y_i = 0 \quad (s_i = 0). \tag{17}$$

Furthermore, $\|\mathbf{Y}\|_{\mathrm{op}} \le 1$, $\|\mathbf{Y}\|_F \le \sqrt{r}$, and $\langle \mathbf{K}, \mathbf{Y}\rangle_F \ge \|\mathbf{K}\|_F$.

Proof. Use the compact singular value decomposition

$$\mathbf{K} = U_r \Sigma V_r^\top, \qquad \Sigma = \operatorname{diag}(s_1, \dots, s_r), \qquad U_r^\top U_r = V_r^\top V_r = I_r.$$

Write $\mathbf{X}_j = U_r A_j V_r^\top$, with diagonal A_j. Initially, $A_0 = \Sigma/\|\mathbf{K}\|_F$. If this form holds at step j, then

$$\mathbf{X}_j^\top \mathbf{X}_j = V_r A_j^2 V_r^\top, \qquad V_r^\top(3I_n - V_r A_j^2 V_r^\top) = (3I_r - A_j^2)V_r^\top.$$

Substituting these two identities into the NS update gives

$$\mathbf{X}_{j+1} = \tfrac{1}{2}U_r A_j V_r^\top(3I_n - V_r A_j^2 V_r^\top) = U_r\left[\tfrac{1}{2}A_j(3I_r - A_j^2)\right]V_r^\top. \tag{18}$$

Thus every positive singular value follows $a_{j+1} = \phi(a_j)$, the singular vectors remain unchanged, and zero singular values stay zero because $\phi(0) = 0$. Induction for $j = 0, 1, 2$ gives the stated three-step expression for $\mathbf{Y} = \mathbf{X}_3$.

For $0 \le a \le 1$, we can get:

$$\phi(a) - a = \tfrac{1}{2}a(1-a^2) \ge 0, \qquad 1 - \phi(a) = \tfrac{1}{2}(1-a)^2(a+2) \ge 0.$$

Since $0 < s_i/\|\mathbf{K}\|_F \le 1$, each positive scalar trajectory is nondecreasing and remains in $(0, 1]$. This proves Eq. (17) and $\|\mathbf{Y}\|_{\mathrm{op}} \le 1$.

Since $\mathbf{Y}$ has r positive singular values in $(0, 1]$, its squared Frobenius norm satisfies

$$\|\mathbf{Y}\|_F^2 = \sum_{i=1}^r y_i^2 \le r.$$

Orthogonality of the singular vectors also imply

$$\langle \mathbf{K}, \mathbf{Y}\rangle_F = \operatorname{tr}\big(V_r \operatorname{diag}(s_i y_i) V_r^\top\big) = \sum_{i=1}^r s_i y_i \ge \frac{\sum_{i=1}^r s_i^2}{\|\mathbf{K}\|_F} = \|\mathbf{K}\|_F.$$

□

Lemma B.3 (Finite-prefix steering geometry)**.** *For the following proof define*

$$\mathbf{X} = \mathbf{X}_{t,3}, \qquad \mathbf{H} = \overline{\mathbf{G}}_t, \qquad \mathbf{A} = \mathbf{X}^\top \mathbf{H}, \qquad \mathbf{S} = \tfrac{1}{2}(\mathbf{A} + \mathbf{A}^\top), \qquad \mathbf{C} = \tfrac{1}{2}(\mathbf{A} - \mathbf{A}^\top),$$
$$\mathbf{R} = \mathbf{I}_n - \mathbf{X}^\top \mathbf{X}, \qquad \mathbf{T} = \mathbf{H} - \mathbf{X}\mathbf{S}, \qquad \mathbf{E} = \gamma_t \mathbf{T}, \qquad \mathbf{Z} = \mathbf{X} + \mathbf{E}.$$

Here $\mathbf{S}$ *and* $\mathbf{C}$ *are respectively the symmetric and skew-symmetric parts of* $\mathbf{A}$*. The normalized signal satisfies* $\mathbf{H} = \mathbf{G}_t / \|\mathbf{G}_t\|_F$*, whether* $\mathbf{G}_t$ *is obtained by FO or ZO access. The quantities above satisfy*

$$\mathbf{X}^\top \mathbf{T} + \mathbf{T}^\top \mathbf{X} = \mathbf{R}\mathbf{S} + \mathbf{S}\mathbf{R}, \qquad \langle \mathbf{H}, \mathbf{T} \rangle_F = \|\mathbf{H}\|_F^2 - \|\mathbf{S}\|_F^2 \geq 0. \tag{19}$$

Consequently, $\left\|\mathbf{X}^\top \mathbf{T} + \mathbf{T}^\top \mathbf{X}\right\|_F \leq 2 \|\mathbf{R}\|_{\mathrm{op}} \|\mathbf{S}\|_F$*. The steering scale in Eq.* (12) *obeys* $\|\mathbf{E}\|_F \leq \alpha \|\mathbf{X}_{t,3} - \mathbf{X}_{t,2}\|_F$*.*

Proof. Since $\mathbf{S}^\top = \mathbf{S}$ and $\mathbf{C}^\top = -\mathbf{C}$, transposition of a scalar trace gives

$$\operatorname{tr}(\mathbf{S}\mathbf{C}) = \operatorname{tr}((\mathbf{S}\mathbf{C})^\top) = \operatorname{tr}(\mathbf{C}^\top \mathbf{S}^\top) = -\operatorname{tr}(\mathbf{C}\mathbf{S}) = -\operatorname{tr}(\mathbf{S}\mathbf{C}).$$

Thus $\langle \mathbf{S}, \mathbf{C} \rangle_F = \operatorname{tr}(\mathbf{S}\mathbf{C}) = 0$, and $\mathbf{A} = \mathbf{S} + \mathbf{C}$ implies $\|\mathbf{A}\|_F^2 = \|\mathbf{S}\|_F^2 + \|\mathbf{C}\|_F^2$. The cross term in the steering alignment is

$$\langle \mathbf{H}, \mathbf{X}\mathbf{S} \rangle_F = \operatorname{tr}(\mathbf{H}^\top \mathbf{X}\mathbf{S}) = \operatorname{tr}(\mathbf{A}^\top \mathbf{S}) = \langle \mathbf{S} + \mathbf{C}, \mathbf{S} \rangle_F = \|\mathbf{S}\|_F^2 .$$

Subtracting this term from $\langle \mathbf{H}, \mathbf{H} \rangle_F$ gives the equality in Eq. (19). To verify its sign, the prefix bound $\|\mathbf{X}\|_{\mathrm{op}} \leq 1$ implies

$$\langle \mathbf{H}, \mathbf{T} \rangle_F = \left(\|\mathbf{H}\|_F^2 - \left\|\mathbf{X}^\top \mathbf{H}\right\|_F^2\right) + \|\mathbf{C}\|_F^2 \geq (1 - \|\mathbf{X}\|_{\mathrm{op}}^2) \|\mathbf{H}\|_F^2 + \|\mathbf{C}\|_F^2 \geq 0. \tag{20}$$

This sign argument does not require full column rank or $m \geq n$.

The definition of $\mathbf{T}$, the symmetry of $\mathbf{S}$, and $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_n - \mathbf{R}$ give

$$\begin{aligned} \mathbf{X}^\top \mathbf{T} + \mathbf{T}^\top \mathbf{X} &= \mathbf{X}^\top \mathbf{H} + \mathbf{H}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{X}\mathbf{S} - \mathbf{S}\mathbf{X}^\top \mathbf{X} \\ &= 2\mathbf{S} - (\mathbf{I}_n - \mathbf{R})\mathbf{S} - \mathbf{S}(\mathbf{I}_n - \mathbf{R}) \\ &= \mathbf{R}\mathbf{S} + \mathbf{S}\mathbf{R}. \end{aligned} \tag{21}$$

The product norm inequalities then give

$$\|\mathbf{R}\mathbf{S} + \mathbf{S}\mathbf{R}\|_F \leq \|\mathbf{R}\mathbf{S}\|_F + \|\mathbf{S}\mathbf{R}\|_F \leq 2 \|\mathbf{R}\|_{\mathrm{op}} \|\mathbf{S}\|_F .$$

If $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_n$, this defect vanishes. More explicitly,

$$\mathbf{T} = (\mathbf{I}_m - \mathbf{X}\mathbf{X}^\top)\mathbf{H} + \mathbf{X}\mathbf{C}.$$

The two terms are orthogonal because $\mathbf{X}^\top(\mathbf{I}_m - \mathbf{X}\mathbf{X}^\top) = 0$. Moreover, any tangent matrix $\mathbf{V}$ satisfies $\mathbf{X}^\top \mathbf{V} + \mathbf{V}^\top \mathbf{X} = 0$, so $\mathbf{X}^\top \mathbf{V}$ is skew-symmetric and $\langle \mathbf{X}\mathbf{S}, \mathbf{V} \rangle_F = \langle \mathbf{S}, \mathbf{X}^\top \mathbf{V} \rangle_F = 0$. Thus $\mathbf{H} = \mathbf{T} + \mathbf{X}\mathbf{S}$ is the decomposition into tangent and normal components for the Euclidean embedding of the Stiefel manifold (Edelman et al., 1998).

For $\mathbf{T} \neq 0$, the algorithm's definition of γ_t gives

$$\|\mathbf{E}\|_F = \alpha \|\mathbf{X}_{t,3} - \mathbf{X}_{t,2}\|_F .$$

When $\mathbf{T} = 0$, the convention $\gamma_t = 0$ gives $\mathbf{E} = 0$, so the same displacement bound also holds. □

Lemma B.4 (Residual contraction in the local NS region)**.** *Define the steered residual and the final output by*

$$\widetilde{\mathbf{R}} = \mathbf{I}_n - \mathbf{Z}^\top \mathbf{Z}, \qquad \mathbf{X}_5 = \tfrac{1}{2}\mathbf{Z}(3\mathbf{I}_n - \mathbf{Z}^\top \mathbf{Z}), \qquad \rho = \left\|\widetilde{\mathbf{R}}\right\|_{\mathrm{op}} .$$

The final NS correction satisfies the exact identities and bound

$$\widetilde{\mathbf{R}} = \mathbf{R} - \gamma_t(\mathbf{R}\mathbf{S} + \mathbf{S}\mathbf{R}) - \gamma_t^2 \mathbf{T}^\top \mathbf{T}, \tag{22}$$

$$\mathbf{I}_n - \mathbf{X}_5^\top \mathbf{X}_5 = \tfrac{3}{4}\widetilde{\mathbf{R}}^2 + \tfrac{1}{4}\widetilde{\mathbf{R}}^3, \tag{23}$$

$$\|\mathbf{X}_5 - \mathbf{X}\|_F \leq \alpha \|\mathbf{X}_{t,3} - \mathbf{X}_{t,2}\|_F + \tfrac{1}{2} \|\mathbf{Z}\|_{\mathrm{op}} \left\|\widetilde{\mathbf{R}}\right\|_F . \tag{24}$$

If $\rho < 1$*, then*

$$\left\|\mathbf{I}_n - \mathbf{X}_5^\top \mathbf{X}_5\right\|_F \leq \frac{\rho(3+\rho)}{4} \left\|\widetilde{\mathbf{R}}\right\|_F < \left\|\widetilde{\mathbf{R}}\right\|_F \quad \textit{when } \widetilde{\mathbf{R}} \neq 0. \tag{25}$$

The identities are valid for every shape.

Proof. Multiplying the two factors in $\mathbf{Z}^\top\mathbf{Z}$ separately gives:

$$\mathbf{Z}^\top\mathbf{Z} = (\mathbf{X}+\gamma_t\mathbf{T})^\top(\mathbf{X}+\gamma_t\mathbf{T}) = \mathbf{X}^\top\mathbf{X} + \gamma_t(\mathbf{X}^\top\mathbf{T}+\mathbf{T}^\top\mathbf{X}) + \gamma_t^2\mathbf{T}^\top\mathbf{T}.$$

Subtract from $\mathbf{I}_n$ and substitute the tangent-defect identity in Lemma B.3 to obtain Eq. (22).

Substituting $\mathbf{Z}^\top\mathbf{Z} = \mathbf{I}_n - \widetilde{\mathbf{R}}$ into the final step gives

$$\mathbf{X}_5 = \mathbf{Z}(\mathbf{I}_n + \widetilde{\mathbf{R}}/2).$$

Since $\widetilde{\mathbf{R}}$ is symmetric and each factor below is a polynomial in $\widetilde{\mathbf{R}}$, the factors commute. Consequently,

$$\begin{aligned}\mathbf{X}_5^\top\mathbf{X}_5 &= (\mathbf{I}_n + \widetilde{\mathbf{R}}/2)(\mathbf{I}_n - \widetilde{\mathbf{R}})(\mathbf{I}_n + \widetilde{\mathbf{R}}/2)\\ &= (\mathbf{I}_n - \widetilde{\mathbf{R}})(\mathbf{I}_n + \widetilde{\mathbf{R}} + \widetilde{\mathbf{R}}^2/4)\\ &= \mathbf{I}_n + \widetilde{\mathbf{R}} + \widetilde{\mathbf{R}}^2/4 - \widetilde{\mathbf{R}} - \widetilde{\mathbf{R}}^2 - \widetilde{\mathbf{R}}^3/4\\ &= \mathbf{I}_n - 3\widetilde{\mathbf{R}}^2/4 - \widetilde{\mathbf{R}}^3/4. \end{aligned} \tag{26}$$

Subtracting this expression from $\mathbf{I}_n$ proves Eq. (23) without any local assumption.

The norm product inequalities give

$$\left\|\widetilde{\mathbf{R}}^2\right\|_F \le \rho\left\|\widetilde{\mathbf{R}}\right\|_F, \qquad \left\|\widetilde{\mathbf{R}}^3\right\|_F \le \rho^2\left\|\widetilde{\mathbf{R}}\right\|_F.$$

Applying the triangle inequality to the residual polynomial gives the first inequality in Eq. (25). For $0 \le \rho < 1$, its factor is strictly below one because

$$1 - \frac{\rho(3+\rho)}{4} = \frac{(1-\rho)(4+\rho)}{4} > 0.$$

If $\widetilde{\mathbf{R}} = 0$, the final step fixes $\mathbf{Z}$ exactly; otherwise the contraction is strict. The eigenvalues of $\widetilde{\mathbf{R}}$ are $1-\sigma_i(\mathbf{Z})^2$, together with a unit eigenvalue for each missing column direction. Thus $\rho < 1$ is equivalent to $0 < \sigma_i(\mathbf{Z}) < \sqrt{2}$ for all n column directions.

The final update has the exact decomposition

$$\mathbf{X}_5 - \mathbf{X} = (\mathbf{Z}-\mathbf{X}) + (\mathbf{X}_5 - \mathbf{Z}) = \gamma_t\mathbf{T} + \tfrac{1}{2}\mathbf{Z}\widetilde{\mathbf{R}}.$$

Therefore

$$\|\mathbf{X}_5 - \mathbf{X}\|_F \le \|\gamma_t\mathbf{T}\|_F + \tfrac{1}{2}\left\|\mathbf{Z}\widetilde{\mathbf{R}}\right\|_F \le \alpha\|\mathbf{X}_{t,3} - \mathbf{X}_{t,2}\|_F + \tfrac{1}{2}\|\mathbf{Z}\|_{\mathrm{op}}\left\|\widetilde{\mathbf{R}}\right\|_F.$$

□

The local condition is an explicit condition on the realized steered matrix, not an extra step of the algorithm. A sufficient check follows directly from Eq. (22):

$$\left\|\widetilde{\mathbf{R}}\right\|_{\mathrm{op}} \le \|\mathbf{R}\|_{\mathrm{op}} + 2\gamma_t\|\mathbf{R}\|_{\mathrm{op}}\|\mathbf{S}\|_{\mathrm{op}} + \gamma_t^2\|\mathbf{T}\|_{\mathrm{op}}^2 < 1.$$

The specified relative steering scale does not by itself imply this inequality. In particular, the algebraic results must not be turned into an unconditional orthogonality claim for arbitrary input spectra.

Finally, taking the inner product of $\mathbf{X}_5 - \mathbf{X} = \gamma_t\mathbf{T} + \mathbf{Z}\widetilde{\mathbf{R}}/2$ with the current signal yields

$$\begin{aligned}\langle\mathbf{G}_t, \mathbf{X}_5 - \mathbf{X}\rangle_F &= \gamma_t\|\mathbf{G}_t\|_F(\|\mathbf{H}\|_F^2 - \|\mathbf{S}\|_F^2) + \tfrac{1}{2}\left\langle\mathbf{G}_t, \mathbf{Z}\widetilde{\mathbf{R}}\right\rangle_F\\ &\ge \gamma_t\|\mathbf{G}_t\|_F(\|\mathbf{H}\|_F^2 - \|\mathbf{S}\|_F^2) - \tfrac{1}{2}\|\mathbf{G}_t\|_F\|\mathbf{Z}\|_{\mathrm{op}}\left\|\widetilde{\mathbf{R}}\right\|_F. \end{aligned} \tag{27}$$

Here the last step is Cauchy–Schwarz followed by $\left\|\mathbf{Z}\widetilde{\mathbf{R}}\right\|_F \le \|\mathbf{Z}\|_{\mathrm{op}}\left\|\widetilde{\mathbf{R}}\right\|_F$. The corresponding one-block alignment loss obeys

$$[\langle\mathbf{G}_t, \mathbf{Z} - \mathbf{X}_5\rangle_F]_+ \le \tfrac{1}{2}\|\mathbf{G}_t\|_F\|\mathbf{Z}\|_{\mathrm{op}}\left\|\widetilde{\mathbf{R}}\right\|_F.$$

These bounds connect the geometric identities to the subsequent cumulative squared-gradient analysis.

Lemma B.5 (Uniform norm and alignment bounds)**.** *For matrix collections,* $\langle \mathbf{A}, \mathbf{B} \rangle_F = \sum_\ell \langle \mathbf{A}^{(\ell)}, \mathbf{B}^{(\ell)} \rangle_F$ *and* $\|\mathbf{A}\|_F^2 = \sum_\ell \|\mathbf{A}^{(\ell)}\|_F^2$*. Let* $k = \sum_\ell \min(m_\ell, n_\ell)$ *for the trainable matrix blocks, and* $0 \le \alpha < 1/4$*. For every realized optimization signal, the complete GS-NS transformation in Algorithm 1 satisfies*

$$\|\mathbf{X}_{t,5}\|_F \le 2\sqrt{k}. \tag{28}$$

Moreover, for the actual momentum $\mathbf{M}_t$ *and any* $0 \le \beta < 1$*,*

$$\langle \nabla F_t(\mathbf{W}_t), \mathbf{X}_{t,5} \rangle_F \ge (1 - 4\alpha)\|\nabla F_t(\mathbf{W}_t)\|_F - 2\sqrt{k}\,\|\nabla F_t(\mathbf{W}_t) - (1-\beta)\mathbf{M}_t\|_F. \tag{29}$$

Proof. Set $\mathbf{K} = (1-\beta)\mathbf{M}$. Since $1 - \beta > 0$, normalizing $\mathbf{K}$ and $\mathbf{M}$ gives the same NS input. For $\mathbf{K} \neq 0$, this input has Frobenius norm one. The scalar map $\phi(s) = s(3 - s^2)/2$ preserves $[0, 1]$ and satisfies $0 \le \phi(s) \le 3s/2$ there. The singular-value representation in Lemma B.2 therefore yields

$$\|\mathbf{X}_2\|_F \le (3/2)^2 \|\mathbf{X}_0\|_F = \frac{9}{4}, \qquad \|\mathbf{X}_2\|_{\mathrm{op}} \le 1, \qquad \|\mathbf{X}_3\|_{\mathrm{op}} \le 1.$$

Every eigenvalue of $\mathbf{X}_2^\top \mathbf{X}_2$ lies in $[0, 1]$, consequently:

$$\|\mathbf{X}_3 - \mathbf{X}_2\|_F = \frac{1}{2}\|\mathbf{X}_2(\mathbf{I} - \mathbf{X}_2^\top \mathbf{X}_2)\|_F \le \frac{1}{2}\|\mathbf{X}_2\|_F \|\mathbf{I} - \mathbf{X}_2^\top \mathbf{X}_2\|_{\mathrm{op}} \le \frac{9}{8}.$$

Write $\mathbf{Y} = \mathbf{X}_3$ and $\mathbf{E} = \gamma \mathbf{T}_3$ for the steering displacement. For $\mathbf{T}_3 \neq 0$, the Eq. (12) gives

$$\|\mathbf{E}\|_F = \alpha \|\mathbf{X}_3 - \mathbf{X}_2\|_F \le \frac{9\alpha}{8},$$
$$\|\mathbf{E}\|_{\mathrm{op}} \le \|\mathbf{E}\|_F \le \frac{9\alpha}{8} \le \frac{9}{32}.$$

When $\mathbf{T}_3 = 0$, the zero-steering convention yields $\mathbf{E} = 0$, so these bounds still hold. If $\mathbf{K} = 0$, the prefix and the relative steering scale are zero, so $\mathbf{Y} = \mathbf{E} = 0$ and the same inequalities hold.

For matrices, write $\phi(\mathbf{A}) = 3\mathbf{A}/2 - \mathbf{A}\mathbf{A}^\top \mathbf{A}/2$ and define the comparison matrix $\mathbf{Q} = \phi(\mathbf{Y})$. The actual output is $\mathbf{X}_5 = \phi(\mathbf{Y} + \mathbf{E})$. Expanding every term of the cubic polynomial gives

$$\mathbf{X}_5 - \mathbf{Q} = \frac{3}{2}\mathbf{E} - \frac{1}{2}\big(\mathbf{Y}\mathbf{Y}^\top \mathbf{E} + \mathbf{Y}\mathbf{E}^\top \mathbf{Y} + \mathbf{E}\mathbf{Y}^\top \mathbf{Y} + \mathbf{Y}\mathbf{E}^\top \mathbf{E} + \mathbf{E}\mathbf{Y}^\top \mathbf{E} + \mathbf{E}\mathbf{E}^\top \mathbf{Y} + \mathbf{E}\mathbf{E}^\top \mathbf{E}\big). \tag{30}$$

Set $e = \|\mathbf{E}\|_{\mathrm{op}}$ within this step. Because $\|\mathbf{Y}\|_{\mathrm{op}} \le 1$, each of the first three mixed terms has Frobenius norm at most $\|\mathbf{E}\|_F$; each of the next three has norm at most $e\|\mathbf{E}\|_F$; and the final term has norm at most $e^2\|\mathbf{E}\|_F$. These estimates follow from $\|\mathbf{A}\mathbf{B}\mathbf{C}\|_F \le \|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{B}\|_F\|\mathbf{C}\|_{\mathrm{op}}$, with the Frobenius factor placed at an occurrence of $\mathbf{E}$ as appropriate. Hence

$$\|\mathbf{X}_5 - \mathbf{Q}\|_F \le \left(3 + \frac{3}{2}e + \frac{1}{2}e^2\right)\|\mathbf{E}\|_F \le \left(3 + \frac{27}{64} + \frac{81}{2048}\right)\frac{9\alpha}{8}\frac{7089}{2048}\frac{9\alpha}{8} = \frac{63801}{16384}\alpha \le 4\alpha, \tag{31}$$

where the numerical constant $63801/16384$ is strictly smaller than 4. This estimate does not bound the final correction in isolation; it compares the complete perturbed transformation with an ordinary NS output whose alignment can be controlled directly.

Restore the block index and let $r_\ell = \min(m_\ell, n_\ell)$. The comparison matrix $\mathbf{Q}^{(\ell)}$ has at most r_ℓ nonzero singular values, all in $[0, 1]$. Since $r_\ell \ge 1$,

$$\|\mathbf{X}_5^{(\ell)}\|_F \le \|\mathbf{Q}^{(\ell)}\|_F + \|\mathbf{X}_5^{(\ell)} - \mathbf{Q}^{(\ell)}\|_F \le \sqrt{r_\ell} + 4\alpha \le (1 + 4\alpha)\sqrt{r_\ell}.$$

Squaring and summing over blocks proves

$$\|\mathbf{X}_{t,5}\|_F^2 \le (1 + 4\alpha)^2 \sum_\ell r_\ell \le 4k,$$

For this step only, let $\mathbf{H}^{(\ell)} = \nabla_{\mathbf{W}^{(\ell)}} F_t(\mathbf{W}_t)$ and $\mathbf{K}^{(\ell)} = (1-\beta)\mathbf{M}_t^{(\ell)}$. For a nonzero $\mathbf{K}^{(\ell)}$, the fourth ordinary NS step preserves its singular-vector pairs and has singular values at least those of $\mathbf{K}^{(\ell)}/\|\mathbf{K}^{(\ell)}\|_F$. Thus

$$\langle \mathbf{K}^{(\ell)}, \mathbf{Q}^{(\ell)} \rangle_F \ge \|\mathbf{K}^{(\ell)}\|_F, \qquad \|\mathbf{Q}^{(\ell)}\|_F \le \sqrt{r_\ell}.$$

The first inequality is also valid when $\mathbf{K}^{(\ell)} = 0$. Using Eq. (31), Cauchy–Schwarz, and then the reverse triangle inequality gives

$$\begin{aligned}\langle \mathbf{H}^{(\ell)}, \mathbf{X}_5^{(\ell)} \rangle_F &= \langle \mathbf{K}^{(\ell)}, \mathbf{Q}^{(\ell)} \rangle_F + \langle \mathbf{H}^{(\ell)} - \mathbf{K}^{(\ell)}, \mathbf{Q}^{(\ell)} \rangle_F + \langle \mathbf{H}^{(\ell)}, \mathbf{X}_5^{(\ell)} - \mathbf{Q}^{(\ell)} \rangle_F \\ &\geq \|\mathbf{K}^{(\ell)}\|_F - \sqrt{r_\ell}\|\mathbf{H}^{(\ell)} - \mathbf{K}^{(\ell)}\|_F - 4\alpha\|\mathbf{H}^{(\ell)}\|_F \\ &\geq (1-4\alpha)\|\mathbf{H}^{(\ell)}\|_F - (1+\sqrt{r_\ell})\|\mathbf{H}^{(\ell)} - \mathbf{K}^{(\ell)}\|_F.\end{aligned}$$

Let b denote the number of blocks within this proof. Since $r_\ell \geq 1$, we have $b \leq k$. A second application of Cauchy–Schwarz gives

$$\sum_\ell (1+\sqrt{r_\ell})\|\mathbf{H}^{(\ell)} - \mathbf{K}^{(\ell)}\|_F \leq \left(\sum_\ell (1+\sqrt{r_\ell})^2\right)^{1/2} \|\mathbf{H} - \mathbf{K}\|_F,$$

$$\sum_\ell (1+\sqrt{r_\ell})^2 = b + k + 2\sum_\ell \sqrt{r_\ell} \leq b + k + 2\sqrt{bk} = (\sqrt{b}+\sqrt{k})^2 \leq 4k.$$

Finally, $1 - 4\alpha > 0$ and $\sum_\ell \|\mathbf{H}^{(\ell)}\|_F \geq \|\mathbf{H}\|_F$. Summing the blockwise alignment inequality and substituting the original gradient and momentum therefore proves Eq. (29). □

Lemma B.6 (Gradient drift)**.** *Under Assumptions 5.1 and 5.3, the actual GS-Muon trajectory with* $0 \leq \alpha < 1/4$ *satisfies*

$$\sum_{t=2}^{T} \|\nabla F_t(\mathbf{W}_t) - \nabla F_{t-1}(\mathbf{W}_{t-1})\|_F \leq 2L\eta\sqrt{k}(T-1) + \sqrt{8L(T-1)V_T^F}. \tag{32}$$

Proof. Let $h_t(\mathbf{W}) = F_t(\mathbf{W}) - F_{t-1}(\mathbf{W})$ and $v_t = \sup_{\mathbf{W}\in\mathbb{R}^d} |h_t(\mathbf{W})|$. The gradient of h_t is $2L$-Lipschitz, and the triangle inequality gives

$$\begin{aligned}\|\nabla h_t(\mathbf{W}) - \nabla h_t(\mathbf{V})\|_F &\leq \|\nabla F_t(\mathbf{W}) - \nabla F_t(\mathbf{V})\|_F + \|\nabla F_{t-1}(\mathbf{W}) - \nabla F_{t-1}(\mathbf{V})\|_F \\ &\leq 2L\|\mathbf{W} - \mathbf{V}\|_F.\end{aligned}$$

Apply its descent inequality at the auxiliary point $\mathbf{W} - \nabla h_t(\mathbf{W})/(2L)$, which belongs to the global domain of both objectives. We obtain

$$h_t\left(\mathbf{W} - \frac{\nabla h_t(\mathbf{W})}{2L}\right) \leq h_t(\mathbf{W}) - \frac{\|\nabla h_t(\mathbf{W})\|_F^2}{2L} + L\left\|\frac{\nabla h_t(\mathbf{W})}{2L}\right\|_F^2 = h_t(\mathbf{W}) - \frac{\|\nabla h_t(\mathbf{W})\|_F^2}{4L}.$$

Since $-v_t \leq h_t \leq v_t$ everywhere, rearrangement gives

$$\|\nabla F_t(\mathbf{W}) - \nabla F_{t-1}(\mathbf{W})\|_F^2 = \|\nabla h_t(\mathbf{W})\|_F^2 \leq 8Lv_t. \tag{33}$$

Add and subtract $\nabla F_t(\mathbf{W}_{t-1})$, use smoothness and Eq. (33), and then substitute the actual update and Lemma B.5:

$$\begin{aligned}\|\nabla F_t(\mathbf{W}_t) - \nabla F_{t-1}(\mathbf{W}_{t-1})\|_F &\leq L\|\mathbf{W}_t - \mathbf{W}_{t-1}\|_F + \sqrt{8Lv_t} \\ &= L\eta\|\mathbf{X}_{t-1,5}\|_F + \sqrt{8Lv_t} \\ &\leq 2L\eta\sqrt{k} + \sqrt{8Lv_t}.\end{aligned}$$

Summing for $t = 2, \ldots, T$ and applying Cauchy–Schwarz gives

$$\sum_{t=2}^{T} \sqrt{8Lv_t} \leq \sqrt{8L(T-1)\sum_{t=2}^{T} v_t} \leq \sqrt{8L(T-1)V_T^F}.$$

□

Lemma B.7 (Momentum tracking)**.** *Under the assumptions of Theorem 5.4, with fixed* $\beta \in [0, 1)$ *and* $\mathbf{M}_0 = 0$*, the FO iterates satisfy*

$$\sum_{t=1}^{T} \|\nabla F_t(\mathbf{W}_t) - (1-\beta)\mathbf{M}_t\|_F \leq \frac{\beta}{1-\beta}\left[G + 2L\eta\sqrt{k}(T-1) + \sqrt{8L(T-1)V_T^F}\right]. \tag{34}$$

For the single-direction Gaussian oracle in Theorem 5.6, the corresponding bound is

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}\|\nabla F_t(\mathbf{W}_t)-(1-\beta)\mathbf{M}_t\|_F\\ &\le\frac{\beta}{1-\beta}\left[G+2L\eta\sqrt{k}(T-1)+\sqrt{8L(T-1)V_T^F}\right]\\ &\quad+TL\mu\sqrt{d}+T\sqrt{\frac{1-\beta}{1+\beta}\left[2(d+1)G^2+\frac{L^2\mu^2d(d+2)(d+4)}{2}\right]}.\end{aligned}\tag{35}$$

Here V_T^F is the deterministic function-variation budget in Assumption 5.3.

Proof. Let $\mathcal{H}_t$ contain the fixed loss sequence, the fixed initial iterate, and all Gaussian directions drawn before round t. Thus $\mathbf{W}_t$ and $\mathbf{M}_{t-1}$ are $\mathcal{H}_t$-measurable, whereas the current direction is fresh and conditionally Gaussian. Write $\mathbb{E}_t[\cdot]=\mathbb{E}[\cdot\mid\mathcal{H}_t]$. For this proof only, set

$$\mathbf{K}_t=(1-\beta)\mathbf{M}_t,\qquad \mathbf{b}_t=\mathbb{E}_t\mathbf{G}_t-\nabla F_t(\mathbf{W}_t),\qquad \boldsymbol{\xi}_t=\mathbf{G}_t-\mathbb{E}_t\mathbf{G}_t.$$

These definitions leave the algorithm unchanged: multiplying every momentum block by the same positive constant $1-\beta$ leaves its normalized NS initialization unchanged. The oracle result in Lemma B.1 gives, almost surely,

$$\|\mathbf{b}_t\|_F\le L\mu\sqrt{d},\qquad \mathbb{E}_t\boldsymbol{\xi}_t=0,\qquad \mathbb{E}_t\|\boldsymbol{\xi}_t\|_F^2\le 2(d+1)G^2+\frac{L^2\mu^2d(d+2)(d+4)}{2}.$$

Under FO access, $\mathbf{b}_t=\boldsymbol{\xi}_t=0$ at every round.

Multiplying the actual momentum recurrence by $1-\beta$ and unrolling from $\mathbf{K}_0=0$ yields

$$\mathbf{K}_t=(1-\beta)\sum_{s=1}^{t}\beta^{t-s}\left[\nabla F_s(\mathbf{W}_s)+\mathbf{b}_s+\boldsymbol{\xi}_s\right].$$

Subtracting the current gradient and collecting consecutive gradient differences gives the exact identity

$$\begin{aligned}\mathbf{K}_t-\nabla F_t(\mathbf{W}_t)=&-\beta^t\nabla F_1(\mathbf{W}_1)+\sum_{s=2}^{t}\beta^{t-s+1}\left[\nabla F_{s-1}(\mathbf{W}_{s-1})-\nabla F_s(\mathbf{W}_s)\right]\\ &+(1-\beta)\sum_{s=1}^{t}\beta^{t-s}\mathbf{b}_s+(1-\beta)\sum_{s=1}^{t}\beta^{t-s}\boldsymbol{\xi}_s.\end{aligned}\tag{36}$$

For $t\ge 2$, the coefficient of $\nabla F_1(\mathbf{W}_1)$ in the first two terms is $-\beta^t+\beta^{t-1}=(1-\beta)\beta^{t-1}$; each interior gradient has coefficient $\beta^{t-s}-\beta^{t-s+1}=(1-\beta)\beta^{t-s}$, and the current gradient has coefficient $-\beta$.

Since $\|\nabla F_1(\mathbf{W}_1)\|_F\le G$, geometric summation gives

$$\sum_{t=1}^{T}\beta^t\|\nabla F_1(\mathbf{W}_1)\|_F\le\frac{\beta G}{1-\beta}.$$

For the second term of Eq. (36), the triangle inequality and interchanging two finite sums give

$$\begin{aligned}&\sum_{t=1}^{T}\left\|\sum_{s=2}^{t}\beta^{t-s+1}[\nabla F_{s-1}(\mathbf{W}_{s-1})-\nabla F_s(\mathbf{W}_s)]\right\|_F\\ &\le\sum_{s=2}^{T}\|\nabla F_{s-1}(\mathbf{W}_{s-1})-\nabla F_s(\mathbf{W}_s)\|_F\sum_{t=s}^{T}\beta^{t-s+1}\\ &\le\frac{\beta}{1-\beta}\sum_{s=2}^{T}\|\nabla F_{s-1}(\mathbf{W}_{s-1})-\nabla F_s(\mathbf{W}_s)\|_F\\ &\le\frac{\beta}{1-\beta}\left[2L\eta\sqrt{k}(T-1)+\sqrt{8L(T-1)V_T^F}\right],\end{aligned}$$

where the final inequality is Lemma B.6. It holds for every realized trajectory, so taking expectations does not change this deterministic upper bound. When $\beta = 0$, both terms are zero. In the FO case there are no bias or noise terms, and these two estimates already prove Eq. (34).

For each round, the nonnegative weights have total mass $(1-\beta)\sum_{s=1}^{t}\beta^{t-s} = 1-\beta^t \le 1$. Consequently,

$$\left\|(1-\beta)\sum_{s=1}^{t}\beta^{t-s}\mathbf{b}_s\right\|_F \le (1-\beta)\sum_{s=1}^{t}\beta^{t-s}L\mu\sqrt{d} \le L\mu\sqrt{d}.$$

Summing this bound contributes at most $TL\mu\sqrt{d}$.

For $r < s$, the matrix $\boldsymbol{\xi}_r$ is $\mathcal{H}_s$-measurable. The tower property therefore implies

$$\mathbb{E}\langle \boldsymbol{\xi}_r, \boldsymbol{\xi}_s\rangle_F = \mathbb{E}\langle \boldsymbol{\xi}_r, \mathbb{E}_s\boldsymbol{\xi}_s\rangle_F = 0.$$

The conditional second-moment bound makes these inner products integrable. Expanding the squared norm, cancelling the cross terms, and summing the squared geometric weights yields

$$\begin{aligned}
&\mathbb{E}\left\|(1-\beta)\sum_{s=1}^{t}\beta^{t-s}\boldsymbol{\xi}_s\right\|_F^2 \\
&= (1-\beta)^2\sum_{s=1}^{t}\beta^{2(t-s)}\mathbb{E}\|\boldsymbol{\xi}_s\|_F^2 \\
&\le \frac{(1-\beta)^2(1-\beta^{2t})}{1-\beta^2}\left[2(d+1)G^2 + \frac{L^2\mu^2 d(d+2)(d+4)}{2}\right] \\
&\le \frac{1-\beta}{1+\beta}\left[2(d+1)G^2 + \frac{L^2\mu^2 d(d+2)(d+4)}{2}\right].
\end{aligned} \tag{37}$$

Cauchy–Schwarz bounds the expected norm by the square root of Eq. (37). Applying the triangle inequality to Eq. (36), taking expectations, and summing the four bounds proves Eq. (35). □

B.3 Proof of the regret theorem

Theorem 5.4 (FO nonconvex regret) Under Assumptions 5.1–5.3, we denote $F_1(\mathbf{W}_1) - F_{\inf} \le \Delta_F < \infty$, and let $\eta > 0$, $0 \le \beta < 1$, and $0 \le \alpha < 1/4$. With exact FO signals, GS-Muon satisfies

$$\mathcal{R}_T \le \frac{G}{1-4\alpha}\left[\frac{\Delta_F + V_T^F}{\eta} + 2L\eta kT + \frac{2\sqrt{k}\,\beta}{1-\beta}\left(G + 2L\eta\sqrt{k}(T-1) + \sqrt{8L(T-1)V_T^F}\right)\right]. \tag{38}$$

Proof of Theorem 5.4. Let the temporary displacement be $D = -\eta\mathbf{X}_{t,5}$. The fundamental theorem of calculus and L-smoothness imply

$$\begin{aligned}
F_t(\mathbf{W}_t + D) - F_t(\mathbf{W}_t) &= \int_0^1 \langle \nabla F_t(\mathbf{W}_t + sD), D\rangle_F\, ds \\
&\le \langle \nabla F_t(\mathbf{W}_t), D\rangle_F + \int_0^1 Ls\|D\|_F^2\, ds \\
&= -\eta\langle \nabla F_t(\mathbf{W}_t), \mathbf{X}_{t,5}\rangle_F + \frac{L\eta^2}{2}\|\mathbf{X}_{t,5}\|_F^2.
\end{aligned}$$

Apply both conclusions of Lemma B.5 and rearrange:

$$(1-4\alpha)\|\nabla F_t(\mathbf{W}_t)\|_F \le \frac{F_t(\mathbf{W}_t) - F_t(\mathbf{W}_{t+1})}{\eta} + 2L\eta k + 2\sqrt{k}\|\nabla F_t(\mathbf{W}_t) - (1-\beta)\mathbf{M}_t\|_F. \tag{39}$$

Adding and subtracting the previous objective at each current iterate gives the exact identity

$$\begin{aligned}&\sum_{t=1}^{T}[F_t(\mathbf{W}_t) - F_t(\mathbf{W}_{t+1})]\\ &= F_1(\mathbf{W}_1) - F_T(\mathbf{W}_{T+1}) + \sum_{t=2}^{T}[F_t(\mathbf{W}_t) - F_{t-1}(\mathbf{W}_t)]\\ &\le F_1(\mathbf{W}_1) - F_{\inf} + V_T^F \le \Delta_F + V_T^F.\end{aligned}$$

Assumption 5.1 controls the terminal loss and bounds the initial loss gap by Δ_F; Assumption 5.3 controls the remaining sum uniformly over all iterates.

Sum Eq. (39) and substitute Eq. (34) to obtain

$$(1-4\alpha)\sum_{t=1}^{T}\|\nabla F_t(\mathbf{W}_t)\|_F \le \frac{\Delta_F + V_T^F}{\eta} + 2L\eta kT + \frac{2\sqrt{k}\,\beta}{1-\beta}\left[G + 2L\eta\sqrt{k}(T-1) + \sqrt{8L(T-1)V_T^F}\right].$$

Finally, Assumption 5.2 implies

$$\mathcal{R}_T = \sum_{t=1}^{T}\|\nabla F_t(\mathbf{W}_t)\|_F^2 \le G\sum_{t=1}^{T}\|\nabla F_t(\mathbf{W}_t)\|_F.$$

Multiplying the preceding deterministic upper bound by $G/(1-4\alpha)$ proves Eq. (14). This argument requires no convexity or unobserved initial gradient. □

Remark B.8. Under FO access, suppose $V_T^F = \mathcal{O}(1)$ and $\beta < 1$ is fixed. Choosing

$$\eta = \Theta(T^{-1/2}) \tag{40}$$

makes both the initial-gap and smoothness contributions at most $\mathcal{O}(\sqrt{T})$. The momentum contribution has the dominant terms

$$\mathcal{O}\left(\eta T + \sqrt{TV_T^F}\right) = \mathcal{O}(\sqrt{T}), \tag{41}$$

and hence

$$\mathcal{R}_T = \mathcal{O}(\sqrt{T}), \qquad \frac{\mathcal{R}_T}{T} = \mathcal{O}(T^{-1/2}). \tag{42}$$

Theorem 5.6 (ZO nonconvex regret) Under Assumptions 5.1–5.3, let $\eta, \mu > 0$, $0 \le \beta < 1$, and $0 \le \alpha < 1/4$. With the single-direction Gaussian estimator above, GS-Muon satisfies

$$\begin{aligned}\mathbb{E}[\mathcal{R}_T] \le \frac{G}{1-4\alpha}\Bigg[&\frac{\Delta_F + V_T^F}{\eta} + 2L\eta kT + \frac{2\sqrt{k}\,\beta}{1-\beta}\left(G + 2L\eta\sqrt{k}(T-1) + \sqrt{8L(T-1)V_T^F}\right)\\ &+ 2\sqrt{k}\,T\left(L\mu\sqrt{d} + \sqrt{\frac{1-\beta}{1+\beta}\left[2(d+1)G^2 + \frac{L^2\mu^2 d(d+2)(d+4)}{2}\right]}\right)\Bigg].\end{aligned} \tag{43}$$

Proof of Theorem 5.6. Algorithm 1 uses $\mathbf{W}_{t+1} = \mathbf{W}_t - \eta\mathbf{X}_{t,5}$. The descent inequality and Lemma B.5 therefore give, for every realized Gaussian sequence,

$$\begin{aligned}F_t(\mathbf{W}_{t+1}) &\le F_t(\mathbf{W}_t) - \eta\langle\nabla F_t(\mathbf{W}_t), \mathbf{X}_{t,5}\rangle_F + \frac{L\eta^2}{2}\|\mathbf{X}_{t,5}\|_F^2\\ &\le F_t(\mathbf{W}_t) - \eta(1-4\alpha)\|\nabla F_t(\mathbf{W}_t)\|_F + 2L\eta^2 k + 2\eta\sqrt{k}\|\nabla F_t(\mathbf{W}_t) - (1-\beta)\mathbf{M}_t\|_F.\end{aligned}$$

The geometric inequality is pathwise for the available signal; it does not require the normalized ZO direction to be unbiased. Rearranging and summing produces

$$\begin{aligned}&(1-4\alpha)\sum_{t=1}^{T}\|\nabla F_t(\mathbf{W}_t)\|_F\\ &\le \frac{1}{\eta}\sum_{t=1}^{T}[F_t(\mathbf{W}_t) - F_t(\mathbf{W}_{t+1})] + 2L\eta kT + 2\sqrt{k}\sum_{t=1}^{T}\|\nabla F_t(\mathbf{W}_t) - (1-\beta)\mathbf{M}_t\|_F.\end{aligned} \tag{44}$$

The loss differences satisfy the exact identity

$$\begin{aligned}&\sum_{t=1}^{T}[F_t(\mathbf{W}_t) - F_t(\mathbf{W}_{t+1})]\\ &= F_1(\mathbf{W}_1) - F_T(\mathbf{W}_{T+1}) + \sum_{t=2}^{T}[F_t(\mathbf{W}_t) - F_{t-1}(\mathbf{W}_t)]\\ &\le F_1(\mathbf{W}_1) - F_{\inf} + V_T^F \le \Delta_F + V_T^F.\end{aligned}$$

The last inequality uses the initial-gap bound in Assumption 5.1.

Take expectations in Eq. (44), apply the preceding telescoping estimate, and substitute Eq. (35) gives

$$\begin{aligned}&(1-4\alpha)\sum_{t=1}^{T}\mathbb{E}\|\nabla F_t(\mathbf{W}_t)\|_F\\ &\le \frac{\Delta_F + V_T^F}{\eta} + 2L\eta kT + \frac{2\sqrt{k}\beta}{1-\beta}\left[G + 2L\eta\sqrt{k}(T-1) + \sqrt{8L(T-1)V_T^F}\right] \qquad (45)\\ &\quad + 2\sqrt{k}\,T\left\{L\mu\sqrt{d} + \sqrt{\frac{1-\beta}{1+\beta}\left[2(d+1)G^2 + \frac{L^2\mu^2 d(d+2)(d+4)}{2}\right]}\right\}.\end{aligned}$$

Finally, the original regret in Eq. (6) and the bounded-gradient assumption imply

$$\mathbb{E}[\mathcal{R}_T] = \sum_{t=1}^{T}\mathbb{E}\|\nabla F_t(\mathbf{W}_t)\|_F^2 \le G\sum_{t=1}^{T}\mathbb{E}\|\nabla F_t(\mathbf{W}_t)\|_F.$$

Multiplying the right-hand side of Eq. (45) by $G/(1-4\alpha)$ proves Theorem 5.6, since $0 \le \alpha < 1/4$. Every term in the final bound is specified by the initial-gap bound Δ_F, variation budget, smoothness and gradient bounds, matrix dimensions, horizon, and algorithmic parameters.

The parameter schedules below are horizon-dependent choices specified before each run and kept fixed throughout the run; they do not modify Algorithm 1. All asymptotic rates treat L, G, k, d, Δ_F, and $\alpha < 1/4$ as constants. □

Remark B.9. For stationary ZO objectives, $V_T^F = 0$. Choosing

$$1-\beta = T^{-1/2}, \qquad \eta = \Theta(T^{-3/4}), \qquad \mu = \mathcal{O}(T^{-1/4}), \qquad (46)$$

gives an $\mathcal{O}(T^{3/4})$ initial-gap term and an $O(T^{1/4})$ smoothness term. The momentum contribution satisfies

$$\mathcal{O}\left(\frac{1+\eta T}{1-\beta}\right) = \mathcal{O}(T^{1/2} + T^{3/4}). \qquad (47)$$

Moreover, the ZO bias contributes $\mathcal{O}(T\mu) = \mathcal{O}(T^{3/4})$. For fixed d, the centered-variance coefficient in Eq. (15) remains bounded, yielding

$$\mathcal{O}\left(T\sqrt{1-\beta}\right) = \mathcal{O}(T^{3/4}) \qquad (48)$$

for the centered-noise contribution. Therefore, with two function evaluations per round,

$$\mathbb{E}[\mathcal{R}_T] = \mathcal{O}(T^{3/4}), \qquad \frac{\mathbb{E}[\mathcal{R}_T]}{T} = \mathcal{O}(T^{-1/4}). \qquad (49)$$

For nonstationary ZO objectives with bounded $V_T^F = \mathcal{O}(1)$, we instead choose

$$1-\beta = T^{-1/3}, \qquad \eta = \Theta(T^{-2/3}), \qquad \mu = \mathcal{O}(T^{-1/6}). \qquad (50)$$

The initial-gap and smoothness terms become $\mathcal{O}(T^{2/3})$ and $\mathcal{O}(T^{1/3})$, respectively, while the momentum term is

$$\mathcal{O}\left(\frac{1+\eta T+\sqrt{T}}{1-\beta}\right) = \mathcal{O}(T^{1/3} + T^{2/3} + T^{5/6}). \qquad (51)$$

Both the ZO bias and centered-noise contributions are $\mathcal{O}(T^{5/6})$. Consequently,

$$\mathbb{E}[\mathcal{R}_T] = \mathcal{O}(T^{5/6}), \qquad \frac{\mathbb{E}[\mathcal{R}_T]}{T} = \mathcal{O}(T^{-1/6}). \qquad (52)$$