DISPLACEMENT-AWARE SUBSPACE ADAPTATION FOR MUON-STYLE ZEROTH-ORDER FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

First-order fine-tuning of large language models (LLMs) is often constrained by the substantial memory overhead of backpropagation, whereas zeroth-order (ZO) fine-tuning offers a memory-efficient alternative based solely on forward evaluations. However, high-dimensional ZO estimates can be noisy, making it difficult to reliably exploit matrix structure through Muon-style spectral updates. Recent ZO-Muon methods alleviate this issue by restricting ZO estimation to low-rank search subspaces, but repeatedly refreshing these subspaces at random may discard directions that have consistently driven recent parameter updates. Consequently, subsequent Muon steps may operate in less informative search spaces, as polar normalization can only exploit directions already captured by the selected bases. To address this issue, we propose Displacement-Aware Subspace ZO-Muon (DAZM), which uses accumulated parameter displacements produced by recent Muon updates to adapt the low-rank search subspace. Specifically, we accumulate the realized Muon updates over each interval, allowing directions that are repeatedly reinforced along the optimization trajectory to become prominent in the resulting displacement. We then retain the leading left and right singular directions as trajectory-informed exploitation bases, while filling the remaining dimensions with random directions orthogonal to the retained subspaces to preserve exploration. Theoretically, we show that under the temporal persistence of the local optimization geometry, displacement-aware subspace adaptation retain more informative gradient structure for subsequent Muon updates than purely random refreshes. Extensive experiments on multiple LLMs and downstream tasks demonstrate that our method achieves consistently superior performance over existing ZO-Muon and non-Muon ZO baselines.

1 INTRODUCTION

Fine-tuning large language models (LLMs) is a widely adopted approach for adapting pretrained models to downstream tasks (Brown et al., 2020; Li & Liang, 2021). Despite its effectiveness, conventional first-order (FO) fine-tuning is highly memory-intensive, requiring intermediate activations for backpropagation, as well as gradients and optimizer states during training (Chen et al., 2016; Rajbhandari et al., 2020). This memory bottleneck becomes particularly severe for modern LLMs with billions of parameters, often making full-parameter fine-tuning impractical on resource-constrained hardware (Hu et al., 2022; Dettmers et al., 2023). Although parameter-efficient fine-tuning methods substantially reduce the number of trainable parameters, they typically still require backpropagation through the model and therefore retain non-negligible activation and gradient-related memory overhead (Houlsby et al., 2019; Lialin et al., 2023). This memory bottleneck is further exacerbated in long-context settings, where activation storage grows substantially with sequence length, further increasing the memory gap between training and inference (Luo et al., 2024; Li et al., 2023).

This memory challenge has motivated zeroth-order (ZO) optimization as a promising approach to memory-efficient fine-tuning, as it estimates update directions from function evaluations alone and avoids explicit backpropagation (Malladi et al., 2023; Zhang et al., 2024). However, standard ZO optimization can suffer from slow convergence due to the high variance of finite-difference gradient estimates in high-dimensional parameter spaces (Nesterov & Spokoiny, 2017; Chen et al., 2019). For matrix-valued parameters, conventional ZO updates directly follow the estimated gradient, so their update geometry inherits the potentially anisotropic singular-value spectrum of the

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.

estimate (Lau et al., 2025). Muon-style optimization provides a natural alternative by replacing the raw matrix estimate with its polar-normalized direction, thereby balancing its singular directions and exploiting the spectral geometry of matrix-valued updates (Petrov et al., 2025; Liu et al., 2025). This normalization, however, becomes problematic when applied directly to noisy full-space ZO estimates (Nesterov & Spokoiny, 2017; Chen et al., 2019). When noise contaminates the spectral structure of the estimated gradient, polar normalization can assign comparable weight to weak or noise-dominated singular directions, reducing the reliability of the resulting update (Lang et al., 2026; Chen & Ma, 2026).

To mitigate this issue, ZO-Muon combines Muon-style optimization with ZO estimation in randomly generated low-dimensional subspaces, rather than directly applying spectral normalization to noisy full-space ZO gradient estimates (Lang et al., 2026). The key idea is that restricting perturbations to a low-dimensional subspace reduces the effective dimension of gradient estimation and consequently alleviates the dimension-dependent variance of ZO estimators (Chen et al., 2025; Yu et al., 2025). Within the reduced search space, Muon applies spectral normalization to the projected ZO estimate, allowing its matrix update to exploit lower-dimensional spectral structure rather than directly orthogonalizing a noisy full-space estimate (Lang et al., 2026; Chen & Ma, 2026). This coupling makes Muon-style spectral optimization substantially more effective in the ZO setting, yielding improved accuracy and runtime efficiency over conventional ZO baselines.

Despite recent progress in combining Muon with low-dimensional ZO estimation, random subspace refresh remains a common component of existing ZO-Muon methods (Lang et al., 2026; Chen & Ma, 2026). Although random refreshes reduce the estimation dimension and promote exploration, they do not explicitly preserve information about which directions have consistently contributed to optimization process. This issue is especially consequential for Muon-style updates: polar normalization can only act on spectral directions exposed by the selected subspace and equalizes the active singular values of the projected estimate, so weak or noise-dominated modes that remain in the subspace can receive comparable normalized weight (Liu et al., 2025; Lau et al., 2025; He et al., 2025). These observations motivate a fundamental question: how can we construct more informative low-rank search subspaces than those obtained by purely random sampling.

To address this limitation, we propose *Displacement-Aware Subspace ZO-Muon* (DAZM), a Muon-style ZO fine-tuning method that adapts the low-rank search subspace using parameter displacements accumulated from recent Muon updates. Our key insight is that accumulated displacement records the update modes actually realized within the current search space, allowing directions that persist across optimization process to be coherently reinforced (Figure 1). Based on this principle, DAZM periodically adapts the search subspace by retaining the leading left and right singular directions of the accumulated displacement and replenishing the remaining dimensions with random directions orthogonal to the retained bases. Unlike purely random subspace adaptation, DAZM reuses information already generated during optimization to retain recurrent update modes, providing subsequent Muon steps with a more informative search-space prior. This displacement-aware adaptation reuses optimization history without additional function evaluations or full-matrix displacement storage, yielding an efficient trajectory-informed alternative to purely random subspace refreshes. Our main contributions are summarized as follows:

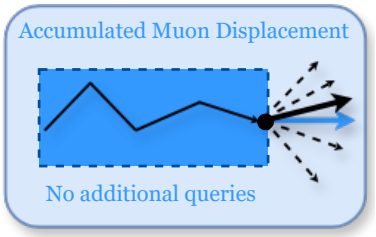


Figure 1: Displacement-Aware

- We propose DAZM, which replaces purely random subspace adaptation with a trajectory-informed strategy based on accumulated Muon displacements. DAZM retains leading displacement modes from the current search space, while operating entirely in reduced coordinates without additional function evaluations.
- We develop a theoretical characterization of displacement-aware subspace adaptation. We establish conditions under which temporal persistence of the local optimization geometry enables displacement-guided subspaces to capture more gradient signal for subsequent Muon updates than purely random subspaces.
- We conduct extensive experiments across multiple LLMs and downstream tasks. The results demonstrate that DAZM improves the accuracy-efficiency trade-off over existing ZO-Muon and non-Muon ZO baselines.

2 Related Work

Memory-efficient Fine-tuning. As LLMs continue to scale, adapting them to downstream tasks with conventional FO optimization becomes increasingly constrained by memory consumption (Zhang et al., 2022; Touvron et al., 2023; Bai et al., 2023). To address this issue, memory-efficient fine-tuning has attracted growing attention, aiming to reduce training overhead while preserving the adaptability of large pretrained models. Parameter-efficient fine-tuning methods, such as LoRA (Hu et al., 2022) and its variants (Dettmers et al., 2023; Liu et al., 2024a; Qin et al., 2024), reduce the number of trainable parameters by introducing lightweight adaptation modules or low-rank updates. Although effective, these methods still generally rely on backpropagation and therefore cannot fully avoid the memory cost associated with gradient computation. ZO optimization offers an alternative route to memory-efficient fine-tuning by estimating update directions only through function evaluations (Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017). MeZO (Malladi et al., 2023) adapts the classical ZO-SGD framework to LLM fine-tuning and achieves a memory footprint close to inference by eliminating backpropagation. Sparse-MeZO (Liu et al., 2024b) reduces the effective optimization dimension through sparse parameter updates, while MeZO-SVRG (Gautam et al., 2024) introduces variance reduction to improve the quality of ZO gradient estimates. HiZOO (Zhao et al., 2024) incorporates diagonal Hessian information to stabilize optimization under non-uniform curvature, and LMAO (Zhang et al., 2026) combines LoRA-style adaptation with ZO updates to improve fine-tuning flexibility.

Low-rank ZO Optimization Recent studies have shown that the parameter updates or gradients involved in LLM fine-tuning often exhibit strong low-rank characteristics, suggesting that the essential optimization information may lie in a much lower-dimensional subspace (Hu et al., 2022; Dettmers et al., 2023; Wang et al., 2024). This observation has motivated the use of low-rank structures to improve the efficiency of both FO and ZO fine-tuning. In standard ZO methods such as MeZO, however, gradient estimates are typically constructed from full-space random perturbations, which may be inefficient for capturing the intrinsic low-rank structure of LLM adaptation (Malladi et al., 2023; Gautam et al., 2024; Peng et al., 2025). This mismatch between full-space ZO perturbations and low-rank fine-tuning dynamics can increase estimation variance and contribute to the performance gap between ZO and FO methods (Chen et al., 2025). To alleviate this issue, recent low-rank ZO methods restrict gradient estimation to low-dimensional subspaces, aiming to reduce the effective search dimension while better matching the low-rank structure of LLM updates. LOZO (Chen et al., 2025) constructs low-rank ZO estimators to improve the alignment between ZO updates and the intrinsic low-rank gradient structure. SubZero (Yu et al., 2025) formulates ZO fine-tuning in randomized low-rank subspaces and applies low-rank perturbations to reduce estimation variance and improve scalability. RoZO (Song & Li, 2026) further introduces a Riemannian ZO optimization framework by constraining updates on the LoRA manifold.

Gradient-orthogonalized Methods. Traditional adaptive optimizers, such as Adam (Kingma, 2014), typically treat neural network parameters as vectors and update them in an element-wise manner, which may underutilize the matrix structure of weight parameters. To better exploit matrix geometry, recent studies have explored spectral or structure-aware optimization methods that operate on matrix-valued gradients. Representative examples include Shampoo (Gupta et al., 2018) and SOAP (Vyas et al., 2025), which use matrix preconditioning to rescale gradients along structured directions. More recently, Muon-style optimizers (Jordan et al., 2024; Liu et al., 2025) have attracted increasing attention by applying gradient orthogonalization or polar updates to matrix-valued momentum, producing spectrally normalized update directions with improved step efficiency. Such updates can be interpreted as selecting descent directions under spectral-norm constraints, enabling larger effective descent per step and faster convergence in FO training (Lau et al., 2025; Kovalev, 2025). Despite their success in FO optimization, extending gradient-orthogonalized methods to ZO fine-tuning is nontrivial. Directly applying Muon-style polar normalization to full-space ZO estimates can be ineffective because ZO gradients are high-variance and their spectral structures may be unreliable. For example, JAGUAR-Muon (Petrov et al., 2025) attempts to incorporate Muon-style updates into ZO optimization, but direct orthogonalization of noisy full-space estimates may provide limited gains. Recent ZO-Muon methods address this issue by coupling gradient orthogonalization with low-rank subspace ZO estimation, so that Muon operates on lower-dimensional matrix estimates rather than highly noisy full-space gradients (Lang et al., 2026; Chen & Ma, 2026).

3 PRELIMINARIES

3.1 ZEROTH-ORDER FINE-TUNING

We consider fine-tuning a matrix-valued parameter $\mathbf{W} \in \mathbb{R}^{m\times n}$:

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

where ξ denotes a training sample drawn from the data distribution $\mathcal{D}$ and $\mathcal{F}(\mathbf{W};\xi)$ denotes the corresponding sample loss. FO methods compute explicit gradients $\nabla\mathcal{F}(\mathbf{W})$ through backpropagation, which typically requires storing or recomputing intermediate activations.

In contrast, ZO methods construct gradient estimates using only function evaluations, without explicit backpropagation (Ghadimi & Lan, 2013; Malladi et al., 2023). A standard one-sided randomized gradient estimator (Nesterov & Spokoiny, 2017) draws i.i.d. perturbation matrices $\mathbf{Z}_i \in \mathbb{R}^{m\times n}$:

$$\widehat{\nabla}\mathcal{F}_\mu(\mathbf{W}) = \frac{1}{N}\cdot\sum_{i=1}^{N}\frac{\mathcal{F}(\mathbf{W}+\mu\mathbf{Z}_i)-\mathcal{F}(\mathbf{W})}{\mu}\cdot\mathbf{Z}_i, \tag{2}$$

where $\mu > 0$ is the smoothing radius and N denotes the number of independent perturbation directions. Since the estimation error of randomized finite-difference methods generally worsens with the ambient dimension mn, full-space ZO estimation can become highly noisy for the large parameter matrices encountered in LLMs, motivating optimization over lower-dimensional search spaces (Wang et al., 2020).

3.2 LOW-RANK SUBSPACE ZO OPTIMIZATION

Low-rank ZO optimization reduces the effective dimension of gradient estimation by restricting perturbations to a low-dimensional structured matrix subspace (Chen et al., 2025; Yu et al., 2025). Specifically, let $\mathbf{U} \in \mathbb{R}^{m\times r}$ and $\mathbf{V} \in \mathbb{R}^{r\times n}$ be two-sided factors satisfying $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_r$ and $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_r$, where the columns of $\mathbf{U}$ and $\mathbf{V}^\top$ span the left and right search subspaces, respectively. These bases define the two-sided matrix search space: $\mathcal{S}(\mathbf{U},\mathbf{V}) := \{\mathbf{U}\mathbf{D}\mathbf{V} : \mathbf{D} \in \mathbb{R}^{r\times r}\}$. Instead of perturbing $\mathbf{W}$ with a full-space random matrix, we draw a reduced-coordinate perturbation $\mathbf{\Sigma}_i \in \mathbb{R}^{r\times r}$ and lift it to the ambient parameter space as $\mathbf{Z}_i = \mathbf{U}\mathbf{\Sigma}_i\mathbf{V} \in \mathcal{S}(\mathbf{U},\mathbf{V})$. The corresponding subspace ZO estimator is:

$$\widehat{\nabla}\mathcal{F}_\mu(\mathbf{W};\mathbf{U},\mathbf{V}) := \mathbf{U}\widehat{\mathbf{G}}\mathbf{V}, \quad \widehat{\mathbf{G}} = \frac{1}{N}\cdot\sum_{i=1}^{N}\frac{\mathcal{F}\left(\mathbf{W}+\mu\mathbf{U}\mathbf{\Sigma}_i\mathbf{V}\right)-\mathcal{F}(\mathbf{W})}{\mu}\cdot\mathbf{\Sigma}_i, \tag{3}$$

where $\widehat{\mathbf{G}} \in \mathbb{R}^{r\times r}$ denotes the reduced-coordinate ZO estimate. Since $\mathcal{S}(\mathbf{U},\mathbf{V})$ forms an r^2-dimensional linear subspace of $\mathbb{R}^{m\times n}$, the effective perturbation dimension is reduced from mn to r^2. Consequently, ZO estimation is carried out entirely in the reduced $r\times r$ coordinate space, while the corresponding lifted estimator in the ambient space has rank at most r.

3.3 RANDOM SUBSPACE GRADIENT ORTHOGONALIZATION

Muon under low-rank subspaces. Muon-style optimization exploits matrix structure by replacing matrix-valued gradients or momenta with spectrally normalized directions obtained through polar decomposition (Jordan et al., 2024; Liu et al., 2025). For analysis, let $\mathbf{H}_t := \nabla\mathcal{F}(\mathbf{W}_t)$ denote the full FO gradient and define its exact projection onto the current two-sided search space as:

$$\mathbf{G}_t := \mathbf{U}_t^\top\mathbf{H}_t\mathbf{V}_t^\top \in \mathbb{R}^{r\times r}. \tag{4}$$

To expose the underlying geometry, we consider the projected Muon update with a step size $\eta > 0$:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta\cdot\mathbf{U}_t\,\mathrm{Polar}(\mathbf{G}_t)\mathbf{V}_t. \tag{5}$$

In ZO optimization, $\mathbf{G}_t$ is not directly accessible and is instead approximated by the reduced-coordinate estimator $\widehat{\mathbf{G}}_t$ introduced in Section 3.2. Accordingly, practical ZO-Muon applies the Muon transformation to $\widehat{\mathbf{G}}_t$ or its corresponding momentum (Lang et al., 2026).

Crucially, the steepest-descent property of the polar update is conditional on the selected search subspace. Within the current search space $\mathcal{S}(\mathbf{U}_t, \mathbf{V}_t)$, consider the operator-norm constrained update set $\mathcal{S}_{\mathrm{op}}(\mathbf{U}_t, \mathbf{V}_t) := \{\mathbf{U}_t\mathbf{D}\mathbf{V}_t : \|\mathbf{D}\|_{\mathrm{op}} \le 1\}$, the maximum FO descent signal attainable is:

$$\max_{\|\mathbf{D}\|_{\mathrm{op}}\le 1} \langle \mathbf{H}_t, \mathbf{U}_t\mathbf{D}\mathbf{V}_t\rangle_F = \max_{\|\mathbf{D}\|_{\mathrm{op}}\le 1} \left\langle \mathbf{U}_t^\top \mathbf{H}_t \mathbf{V}_t^\top, \mathbf{D}\right\rangle_F = \left\|\mathbf{U}_t^\top \mathbf{H}_t \mathbf{V}_t^\top\right\|_*. \tag{6}$$

By the duality between the operator and nuclear norms, a maximizer is $\mathbf{D}_t^\star = \mathrm{Polar}(\mathbf{G}_t)$. This motivates the subspace capture ratio:

$$\rho_t := \frac{\|\mathbf{U}_t^\top \mathbf{H}_t \mathbf{V}_t^\top\|_*}{\|\mathbf{H}_t\|_*} \in [0, 1]. \tag{7}$$

The captured signal directly determines the FO descent available to the projected Muon update. For clarity, consider the ideal projected Muon direction $\mathbf{D}_t^{\mathrm{m}} = \mathbf{U}_t\,\mathrm{Polar}(\mathbf{G}_t)\mathbf{V}_t$. Assume that $\mathcal{F}$ satisfies the smoothness condition with respect to the operator norm:

$$\begin{aligned}&\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t)\\ &\le -\eta\,\langle \mathbf{H}_t, \mathbf{D}_t^{\mathrm{m}}\rangle_F + \frac{L\eta^2}{2}\,\|\mathbf{D}_t^{\mathrm{m}}\|_{\mathrm{op}}^2 \le -\eta\left\|\mathbf{U}_t^\top \mathbf{H}_t \mathbf{V}_t^\top\right\|_* + \frac{L\eta^2}{2} = -\eta\cdot\rho_t\cdot\|\mathbf{H}_t\|_* + \frac{L\eta^2}{2}.\end{aligned} \tag{8}$$

Equation 8 shows that the amount of gradient signal captured by the selected subspace directly controls the linear descent term available to Muon: although the polar factor is optimal within the current search space, a small capture ratio ρ_t proportionally reduces this term.

Limitation of random subspaces. Random subspaces reduce the effective estimation dimension, but when sampled independently of the current gradient they do not preferentially capture gradient energy. For independent orthonormal bases $\mathbf{U}_t$ and $\mathbf{V}_t$, we have

$$\mathbb{E}\left[\left\|\mathbf{U}_t^\top \mathbf{H}_t \mathbf{V}_t^\top\right\|_F^2\right] = \mathbb{E}\left[\mathrm{tr}\big(\mathbf{U}_t^\top \mathbf{H}_t \mathbf{V}_t^\top \mathbf{V}_t \mathbf{H}_t^\top \mathbf{U}_t\big)\right] = \frac{r}{n}\,\mathrm{tr}\big(\mathbf{H}_t\mathbf{H}_t^\top \mathbb{E}[\mathbf{U}_t\mathbf{U}_t^\top]\big) = \frac{r^2}{mn}\|\mathbf{H}_t\|_F^2. \tag{9}$$

Beyond the projected Frobenius energy in Equation 9, we are particularly interested in the nuclear-norm signal accessible to Muon. Let $\mathbf{H}_t = \sum_j \sigma_j \mathbf{a}_j \mathbf{b}_j^\top$. Then

$$\mathbb{E}[\rho_t] = \frac{\mathbb{E}[\|\mathbf{U}_t^\top \mathbf{H}_t \mathbf{V}_t^\top\|_*]}{\|\mathbf{H}_t\|_*} \le \frac{1}{\|\mathbf{H}_t\|_*}\sum_j \sigma_j \mathbb{E}\big[\|\mathbf{U}_t^\top \mathbf{a}_j\|_2 \|\mathbf{V}_t \mathbf{b}_j\|_2\big] \le \frac{r}{\sqrt{mn}}. \tag{10}$$

The detailed derivation is provided in Lemma 11 of Appendix C. Hence, when $r \ll \min\{m, n\}$, a purely random low-rank subspace is expected to capture only a limited fraction of the gradient signal available to the subsequent Muon update. Although Muon produces the optimal polar direction within the selected subspace, it cannot exploit gradient components that lie outside the randomly sampled left and right spans. Therefore, in the low-dimensional regime, the quality of subspace selection becomes a central factor: *the selected bases determine which gradient modes are exposed to the subsequent Muon transformation*.

4 METHODS

4.1 FROM RANDOM SUBSPACES TO ADAPTIVE SUBSPACES

The analysis in Section 3.3 shows that the FO descent signal accessible to subspace Muon is determined by the amount of projected gradient signal captured by the selected search subspace. Specifically, for $\mathbf{H}_t = \nabla\mathcal{F}(\mathbf{W}_t)$, the maximum linearized descent signal accessible through a two-sided rank-r search space is $\left\|\mathbf{U}^\top \mathbf{H}_t \mathbf{V}^\top\right\|_*$, as established in Equation 6. This motivates the following oracle subspace selection problem:

$$(\mathbf{U}_t^\star, \mathbf{V}_t^\star) \in \arg\max_{\mathbf{U},\mathbf{V}} \left\|\mathbf{U}^\top \mathbf{H}_t \mathbf{V}^\top\right\|_*. \tag{11}$$

By the variational characterization of singular values, an optimizer is given by the leading r left and right singular subspaces of $\mathbf{H}_t$. Thus, the usefulness of a search subspace depends on which gradient modes it exposes to the Muon update. However, solving Equation 11 directly requires access

to the full gradient $\mathbf{H}_t$, which is unavailable in the ZO setting. Computing such a gradient through backpropagation would defeat the memory-saving purpose of ZO fine-tuning, while accurately reconstructing it through function evaluations would require prohibitive query cost. We therefore seek an alternative signal, available from the optimization process itself, that can inform how the search subspace should be adapted.

Consider a subspace-refresh interval $\mathcal{I}_k = \{\tau_k, \dots, \tau_{k+1} - 1\}$ during which the bases $(\mathbf{U}_k, \mathbf{V}_k)$ remain fixed. Let $\mathbf{O}_t \in \mathbb{R}^{r\times r}$ denote the reduced-space Muon direction actually applied at step t. We define the accumulated descent displacement over this interval as:

$$\boldsymbol{\Delta}_k := \mathbf{W}_{\tau_k} - \mathbf{W}_{\tau_{k+1}} = \sum_{t\in\mathcal{I}_k} \eta_t \mathbf{U}_k \mathbf{O}_t \mathbf{V}_k = \mathbf{U}_k \boldsymbol{\Psi}_k \mathbf{V}_k, \text{ where } \boldsymbol{\Psi}_k := \sum_{t\in\mathcal{I}_k} \eta_t \mathbf{O}_t \in \mathbb{R}^{r\times r}. \quad (12)$$

Because the bases are fixed over $\mathcal{I}_k$, the full displacement is represented exactly by the reduced core $\boldsymbol{\Psi}_k \in \mathbb{R}^{r\times r}$. Unlike a single ZO estimate, $\boldsymbol{\Delta}_k$ aggregates the realized Muon update directions that have actually contributed to parameter movement throughout the refresh interval. We therefore use the leading singular modes of the accumulated displacement to identify directions within the current search space that have contributed most coherently to recent parameter movement. When the useful local optimization geometry evolves gradually, these persistent recent update modes provide a natural prior for constructing the search space used in subsequent steps. Importantly, this information is available without additional function evaluations or backpropagation: it suffices to accumulate the small core $\boldsymbol{\Psi}_k$ from updates already generated during optimization.

4.2 Displacement-Aware Low-Rank Subspace Adaptation

We next describe how to extract the leading singular modes of the accumulated displacement and use them to adapt the search subspace, without explicitly materializing the full matrix $\boldsymbol{\Delta}_k$. Recall from Equation 12 that $\boldsymbol{\Delta}_k = \mathbf{U}_k \boldsymbol{\Psi}_k \mathbf{V}_k$, where $\boldsymbol{\Psi}_k \in \mathbb{R}^{r\times r}$ is the accumulated displacement core. Consider the SVD operation $\boldsymbol{\Psi}_k = \mathbf{A}_k \boldsymbol{\Lambda}_k \mathbf{B}_k^\top$, where $\mathbf{A}_k, \mathbf{B}_k \in \mathbb{R}^{r\times r}$ are orthogonal and the singular values in $\boldsymbol{\Lambda}_k$ are arranged in nonincreasing order. Since $\mathbf{U}_k$ has orthonormal columns and $\mathbf{V}_k$ has orthonormal rows, $\boldsymbol{\Delta}_k = (\mathbf{U}_k \mathbf{A}_k) \boldsymbol{\Lambda}_k (\mathbf{B}_k^\top \mathbf{V}_k)$ is a thin SVD of $\boldsymbol{\Delta}_k$. Therefore, the leading p left and right singular modes of the accumulated displacement can be obtained:

$$\mathbf{U}_{k+1}^{\text{exploit}} = \mathbf{U}_k \mathbf{A}_{k,1:p}, \quad \mathbf{V}_{k+1}^{\text{exploit}} = \mathbf{B}_{k,1:p}^\top \mathbf{V}_k. \quad (13)$$

By the Eckart–Young–Mirsky theorem (Eckart & Young, 1936), these subspaces yield the optimal rank-p approximation of $\boldsymbol{\Delta}_k$ in Frobenius norm. Thus, extracting the retained modes requires only an SVD of the $r \times r$ core $\boldsymbol{\Psi}_k$, without decomposing the full $m \times n$ displacement matrix.

Because $\boldsymbol{\Delta}_k$ lies entirely within the current two-sided search spans, constructing the next subspace solely from its singular modes would prevent the bases from expanding beyond those spans. We therefore reserve the remaining $r - p$ basis dimensions for randomized exploration beyond the retained displacement modes. For the left basis, we draw $\boldsymbol{\Omega}_U \in \mathbb{R}^{m\times(r-p)}$ with i.i.d. standard Gaussian entries and project it onto the orthogonal complement of the retained left subspace:

$$\widetilde{\mathbf{U}}_{k+1}^{\text{explore}} = \left(\mathbf{I} - \mathbf{U}_{k+1}^{\text{exploit}} (\mathbf{U}_{k+1}^{\text{exploit}})^\top\right) \boldsymbol{\Omega}_U. \quad (14)$$

Let $\mathrm{QR}(\cdot)$ be the orthonormal factor of QR decomposition, then we obtain the exploratory basis:

$$\mathbf{U}_{k+1}^{\text{explore}} = \mathrm{QR}\left(\widetilde{\mathbf{U}}_{k+1}^{\text{explore}}\right). \quad (15)$$

For the right basis, it is convenient to work with the column-orthonormal representation $\mathbf{R}_{k+1}^{\text{exploit}} := (\mathbf{V}_{k+1}^{\text{exploit}})^\top \in \mathbb{R}^{n\times p}$. Drawing $\boldsymbol{\Omega}_V \in \mathbb{R}^{n\times(r-p)}$ with i.i.d. standard Gaussian entries, we construct

$$\widetilde{\mathbf{R}}_{k+1}^{\text{explore}} = \left(\mathbf{I} - \mathbf{R}_{k+1}^{\text{exploit}} (\mathbf{R}_{k+1}^{\text{exploit}})^\top\right) \boldsymbol{\Omega}_V, \quad \mathbf{R}_{k+1}^{\text{explore}} = \mathrm{QR}\left(\widetilde{\mathbf{R}}_{k+1}^{\text{explore}}\right), \quad (16)$$

and set $\mathbf{V}_{k+1}^{\text{explore}} = (\mathbf{R}_{k+1}^{\text{explore}})^\top$. The updated left and right search bases are then

$$\mathbf{U}_{k+1} = \left[\mathbf{U}_{k+1}^{\text{exploit}}, \mathbf{U}_{k+1}^{\text{explore}}\right], \quad \mathbf{V}_{k+1} = \begin{bmatrix} \mathbf{V}_{k+1}^{\text{exploit}} \\ \mathbf{V}_{k+1}^{\text{explore}} \end{bmatrix}, \quad (17)$$

which define the next search space $\mathcal{S}(\mathbf{U}_{k+1}, \mathbf{V}_{k+1})$.

4.3 DISPLACEMENT-AWARE ZO-MUON OPTIMIZATION

We integrate displacement-aware subspace adaptation with low-rank ZO estimation and Muon-style polar updates. For each layer ℓ, the search bases $(\mathbf{U}_{\ell,k}, \mathbf{V}_{\ell,k})$ are kept fixed within each interval of ν iterations and adapted at the interval boundary using the accumulated reduced displacement $\mathbf{\Psi}_{\ell,k}$. At each iteration, ZO estimation is performed in the $r \times r$ reduced space, and the resulting estimate is incorporated into the reduced-space momentum before polar normalization. The normalized direction is then lifted back to the original parameter space through the current bases, while its realized update is accumulated into $\mathbf{\Psi}_{\ell,k}$ for the next subspace adaptation. When the bases are refreshed, the reduced momentum is reprojected onto the new coordinates. The detailed low-rank subspace adaptation procedure is provided in Algorithm 2 in Appendix A, while the complete DAZM framework is summarized in Algorithm 1 and Figue 2.

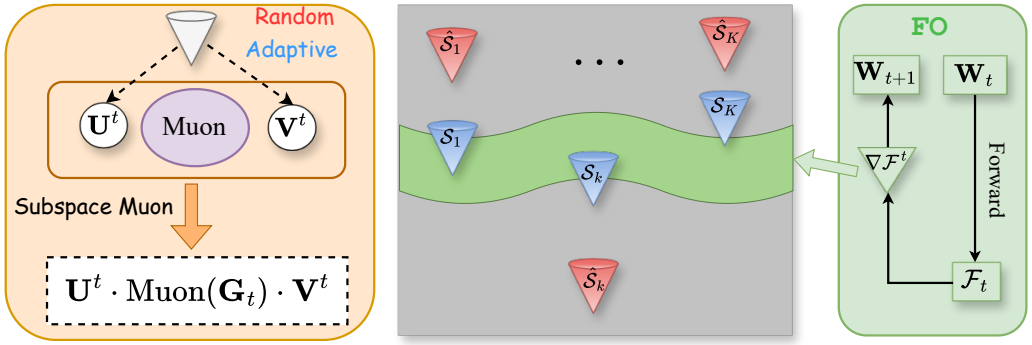


Figure 2: Comparison of the random and adaptive subspaces.

Algorithm 1 DAZM

Require: Initial weights $\mathbf{W}_0 = \{\mathbf{W}_{\ell,0}\}_{\ell=1}^{\mathcal{L}}$, loss function $\mathcal{F}$, learning rate η, perturbation scale μ, momentum coefficient β, step budget T, update interval ν, rank r, exploitation rank p, and number of perturbations N.

Initialize $\mathbf{U}_{\ell,0} \in \mathbb{R}^{m_\ell \times r}$ and $\mathbf{V}_{\ell,0} \in \mathbb{R}^{r \times n_\ell}$ such $\mathbf{U}_{\ell,0}^\top \mathbf{U}_{\ell,0} = \mathbf{I}_r$ and $\mathbf{V}_{\ell,0}\mathbf{V}_{\ell,0}^\top = \mathbf{I}_r$ for all ℓ
Initialize $\mathbf{\Psi}_{\ell,0} \leftarrow \mathbf{0}$, $\mathbf{M}_{\ell,-1} \leftarrow \mathbf{0}$ for all ℓ, and set $k \leftarrow 0$
for $t = 0$ to $T-1$ **do**
  Sample ξ_t and evaluate $\mathcal{F}_t \leftarrow \mathcal{F}(\mathbf{W}_t; \xi_t)$
  for $i = 1$ to N **do**
    for $\ell = 1$ to $\mathcal{L}$ **do**
      Sample $\mathbf{\Sigma}_{\ell,t,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{r\times r})$, $\quad \widetilde{\mathbf{W}}_{\ell,t,i} \leftarrow \mathbf{W}_{\ell,t} + \mu \mathbf{U}_{\ell,k}\mathbf{\Sigma}_{\ell,t,i}\mathbf{V}_{\ell,k}$
    end for
    $\mathcal{F}_{t,i}^+ \leftarrow \mathcal{F}(\widetilde{\mathbf{W}}_{t,i}; \xi_t)$, $\quad \Phi_{t,i} \leftarrow (\mathcal{F}_{t,i}^+ - \mathcal{F}_t)/\mu$
  end for
  for $\ell = 1$ to $\mathcal{L}$ **do**
    $\widehat{\mathbf{G}}_{\ell,t} \leftarrow \frac{1}{N}\sum_{i=1}^N \Phi_{t,i}\mathbf{\Sigma}_{\ell,t,i}$, $\quad \mathbf{M}_{\ell,t} \leftarrow \beta \mathbf{M}_{\ell,t-1} + \widehat{\mathbf{G}}_{\ell,t}$, $\quad \mathbf{O}_{\ell,t} \leftarrow \mathrm{Polar}(\mathbf{M}_{\ell,t})$
    $\mathbf{W}_{\ell,t+1} \leftarrow \mathbf{W}_{\ell,t} - \eta\, \mathbf{U}_{\ell,k}\mathbf{O}_{\ell,t}\mathbf{V}_{\ell,k}$, $\quad \mathbf{\Psi}_{\ell,k} \leftarrow \mathbf{\Psi}_{\ell,k} + \eta \mathbf{O}_{\ell,t}$
  end for
  if $(t+1) \bmod \nu = 0$ **then**
    for $\ell = 1$ to $\mathcal{L}$ **do**
      $(\mathbf{U}_{\ell,k+1}, \mathbf{V}_{\ell,k+1}) \leftarrow \textsc{LowRankSubspaceAdaptation}(p, r, \mathbf{U}_{\ell,k}, \mathbf{V}_{\ell,k}, \mathbf{\Psi}_{\ell,k})$
      $\mathbf{M}_{\ell,t} \leftarrow \mathbf{U}_{\ell,k+1}^\top \mathbf{U}_{\ell,k}\mathbf{M}_{\ell,t}\mathbf{V}_{\ell,k}\mathbf{V}_{\ell,k+1}^\top$, $\quad \mathbf{\Psi}_{\ell,k+1} \leftarrow \mathbf{0}$
    end for
    $k \leftarrow k+1$
  end if
end for
  return $\mathbf{W}_T$

4.4 Convergence Analysis

Assumption 1 (Sample-wise spectral-norm smoothness) *For almost every sample ξ, the stochastic loss $\mathcal{F}(\cdot;\xi)$ is continuously differentiable and satisfies, for all $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{m\times n}$,*

$$\|\nabla\mathcal{F}(\mathbf{W}_1;\xi) - \nabla\mathcal{F}(\mathbf{W}_2;\xi)\|_* \le L\,\|\mathbf{W}_1 - \mathbf{W}_2\|_{\mathrm{op}}. \tag{18}$$

Assumption 2 (Unbiased stochastic gradients with bounded variance) *For every $\mathbf{W}$,*

$$\mathbb{E}_\xi\left[\nabla\mathcal{F}(\mathbf{W};\xi)\right] = \nabla\mathcal{F}(\mathbf{W}), \quad \mathbb{E}_\xi\left[\|\nabla\mathcal{F}(\mathbf{W};\xi) - \nabla\mathcal{F}(\mathbf{W})\|_F^2\right] \le \sigma^2. \tag{19}$$

Assumption 3 (Bounded initial optimality gap) *Let $\mathbf{W}^\star$ be a global minimizer of $\mathcal{F}$. There exists $H < \infty$ such that*

$$\mathcal{F}(\mathbf{W}_0) - \mathcal{F}(\mathbf{W}^\star) \le H. \tag{20}$$

Theorem 1 (Convergence of DAZM) *Let $\mathbf{N}_t = (1-\beta)\mathbf{M}_t$ denote the transported reduced momentum and let $\varepsilon_q := \sup_t \varepsilon_{t,q} < 1$ be the Newton–Schulz approximation error. Under the above assumptions, choosing $\eta = \sqrt{H/LT(1+\varepsilon_q)^2}$, $\beta = 1 - \min\{T^{-1/4}, 1\}$, gives*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{G}_t\|_* = \mathcal{O}\left(\frac{1}{1-\varepsilon_q}\begin{bmatrix}(1+\varepsilon_q)\sqrt{LH}\,T^{-1/2} + \sqrt{rLH}\,T^{-3/4} \\ +\sqrt{rLH}\,T^{-1/4} + \sqrt{r\Theta} + \sqrt{r}\,T^{1/4}\Gamma_{1,T}\end{bmatrix}\right), \tag{21}$$

where $\Gamma_{1,T}$ measures the average momentum-transport mismatch caused by subspace refreshes.

Remark 1 (Comparison with ZO-SGD with momentum) *When the search bases are fixed, $\Gamma_{1,T} = 0$. For comparison, ZO-SGD with momentum satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{G}_t\|_* = \mathcal{O}\Big(\sqrt{r}\left[\sqrt{LH}\,T^{-1/4} + (LH\Theta)^{1/4}T^{-1/4} + \sqrt{LH}\,T^{-3/8} + \sqrt{\Theta}\right]\Big) \tag{22}$$

While both methods have a worst-case $\mathcal{O}(T^{-1/4})$ optimization term, ZO-Muon admits a larger-order certified learning rate. Specifically, Theorem 1 uses $\eta_{\mathrm{M}} = \mathcal{O}(T^{-1/2})$, whereas for ZO-SGD with momentum in Theorem 3 in Appendix D.3, $\eta_{\mathrm{S}} = \mathcal{O}(T^{-3/4})$.

Theorem 2 (Lower bound of displacement-aware gradient capture) *Consider the refresh interval $\mathcal{I}_k = \{\tau_k, \dots, \tau_{k+1} - 1\}$. Let $\mathbf{H}_{\tau_{k+1}} := \nabla F(\mathbf{W}_{\tau_{k+1}})$ denote the gradient at the beginning of the next refresh interval, with leading p left and right singular bases $\mathbf{A}_{p,k+1}$ and $\mathbf{B}_{p,k+1}$. Let $\varepsilon_{\mathrm{exploit},k} := \max\{\|\sin\Theta(\mathbf{U}_{k+1}^{\mathrm{exploit}}, \mathbf{A}_{p,k+1})\|_{\mathrm{op}}, \|\sin\Theta(\mathbf{R}_{k+1}^{\mathrm{exploit}}, \mathbf{B}_{p,k+1})\|_{\mathrm{op}}\} < 1$ measure the alignment error of the retained displacement modes, and define $\gamma_{p,k+1} := \sum_{j=1}^{p}\sigma_j(\mathbf{H}_{\tau_{k+1}})/\|\mathbf{H}_{\tau_{k+1}}\|_*$ as the fraction of nuclear-norm mass contained in the leading p singular modes. Then the capture ratio of the refreshed DAZM subspace satisfies*

$$\rho_{\tau_{k+1}} := \frac{\left\|\mathbf{U}_{k+1}^\top\mathbf{H}_{\tau_{k+1}}\mathbf{V}_{k+1}^\top\right\|_*}{\|\mathbf{H}_{\tau_{k+1}}\|_*} \ge \gamma_{p,k+1} - \varepsilon_{\mathrm{exploit},k}^2. \tag{23}$$

Remark 2 (Comparison with random subspace refresh) *Theorem 2 shows that the benefit of displacement-aware adaptation is governed jointly by the concentration of gradient mass in the leading p singular modes and the alignment between the retained displacement modes and the gradient structure at the next refresh. In comparison, Equation* (10) *shows that an independent random rank-r two-sided subspace satisfies $\mathbb{E}[\rho_{\mathrm{rnd}}] \le r/\sqrt{mn}$. Therefore, whenever $\gamma_{p,k+1} - \varepsilon_{\mathrm{exploit},k}^2 > r/\sqrt{mn}$, the certified capture lower bound of DAZM exceeds the expected capture benchmark of an independent random rank-r refresh. Figure 4 in Appendix B.7 empirically tracks this lower bound throughout optimization and compares it with the corresponding random-subspace benchmark.*

5 Experiments

5.1 Experimental Setup

Datasets and Models. To evaluate optimization performance across diverse language understanding and reasoning tasks, we consider GLUE (Wang et al., 2018) and SuperGLUE (Sarlin et al.,

2020), including SST-2, RTE, CB, BoolQ, WSC, WiC, MultiRC, COPA, ReCoRD, SQuAD, and DROP. We conduct experiments on Qwen2.5-7B (Bai et al., 2023) and OPT-13B (Zhang et al., 2022). Code is available at `https://anonymous.4open.science/r/DAZM-D24F`.

Baselines. To comprehensively evaluate DAZM, we compare it against a diverse set of memory-efficient ZO, non-training, and FO baselines. The ZO baselines include MeZO (Malladi et al., 2023), MeZO-LoRA (Malladi et al., 2023), SubZero (Yu et al., 2025), LOZO (Chen et al., 2025), LOZO+ (Fang et al., 2026), and ZO-Muon (Lang et al., 2026). We additionally report two non-training baselines, Zero-shot inference (Zero-shot) and in-context learning (ICL) (Dong et al., 2024), as well as FO baselines based on Adam fine-tuning (FT) (Kingma, 2014) and LoRA fine-tuning (FT-LoRA) (Hu et al., 2022). Detailed configurations are provided in Appendix B.1.

5.2 Main results

Tables 1 and 2 report the results on Qwen2.5-7B and OPT-13B, respectively. Overall, DAZM achieves the best average performance among the zeroth-order methods on both models, reaching 83.37 on Qwen2.5-7B and 70.41 on OPT-13B. Compared with ZO-Muon, it improves the average score by 2.33 and 1.81 points, respectively, while showing consistent gains across all tasks.

We provide additional results to further evaluate the efficiency, robustness, and effectiveness of DAZM. Appendix B.3 reports the GPU memory consumption in Tables 4 and 5, respectively, while Appendix B.4 compares the wall-clock training time in Tables 6 and 7. Appendix B.5 investigates the sensitivity of DAZM to the η, r, and ν in Figure 3. Appendix B.6 further examines the contribution of the adaptive subspace mechanism through the ablation results in Table 8. Finally, Figure 4 in Appendix B.7 empirically verifies the displacement-aware capture analysis in Theorem 2.

Table 1: Performance comparison on Qwen2.5-7B across different tasks.

Task Type	Classification							Multiple Choice		Generation		Average
	SST-2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	
Zero-shot	63.65	84.48	62.50	81.90	68.27	60.97	77.80	86.00	75.20	85.08	60.78	73.33
ICL	95.30	83.39	91.07	85.40	66.35	64.42	86.70	86.00	76.30	74.82	61.72	79.22
FT-LoRA	79.60	89.50	78.60	80.40	73.10	65.80	87.70	86.00	77.00	89.80	61.90	79.04
FT	94.95	89.89	96.43	87.60	75.96	75.55	86.60	90.00	80.00	91.54	69.48	85.27
MeZO	88.88	88.45	89.29	83.30	65.38	68.18	87.90	90.00	76.00	86.73	58.58	80.24
MeZO-LoRA	67.66	85.20	85.71	82.20	72.12	64.89	85.40	85.00	75.30	81.76	61.59	76.98
SubZero	87.50	88.09	92.86	84.00	69.23	66.46	88.20	89.00	74.50	85.91	61.61	80.67
LOZO	88.07	87.36	91.07	83.70	71.15	66.93	87.40	87.00	74.30	87.12	61.82	80.54
LOZO+	90.37	89.17	89.29	82.80	65.38	67.08	87.30	90.00	77.10	88.34	62.14	80.82
ZO-Muon	93.23	89.53	87.50	84.70	65.38	69.75	88.10	88.00	79.00	89.72	62.28	81.56
DAZM	**93.92**	**89.53**	**92.86**	**84.80**	**75.00**	**70.69**	**88.30**	**90.00**	**79.60**	**89.89**	**62.51**	**83.37**

Table 2: Performance comparison on OPT-13B across different tasks.

Task Type	Classification							Multiple Choice		Generation		Average
	SST-2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	
Zero-shot	58.80	59.60	46.40	59.00	38.50	55.00	46.90	80.00	81.00	46.20	14.60	53.27
ICL	87.00	62.10	57.10	66.90	39.40	50.50	53.10	87.00	82.30	75.90	29.50	62.80
FT-LoRA	92.43	62.09	60.71	66.90	40.38	58.15	58.00	84.00	81.10	81.32	29.38	64.95
FT	92.00	70.80	83.90	77.10	63.50	70.10	71.10	79.00	74.10	84.90	31.30	72.53
MeZO	90.10	60.30	67.90	66.10	62.50	54.70	57.70	87.00	81.10	79.60	30.40	67.04
MeZO-LoRA	86.58	60.29	67.86	63.70	52.88	55.02	50.10	85.00	81.00	78.05	31.18	64.70
SubZero	91.30	61.90	67.90	66.10	63.50	55.90	57.30	86.00	81.90	80.70	30.50	67.55
LOZO	90.30	62.50	67.90	65.60	63.50	55.30	56.90	86.00	81.80	80.50	30.10	67.31
LOZO+	92.43	61.37	69.64	69.00	60.58	57.68	58.40	87.00	80.80	82.81	30.38	68.19
ZO-Muon	91.28	67.15	69.64	66.50	55.77	59.09	56.70	88.00	81.60	83.07	30.08	68.08
DAZM	**92.55**	**70.76**	**71.43**	**70.50**	**65.38**	**59.40**	**58.60**	**88.00**	**82.10**	**84.10**	**31.73**	**70.41**

6 Conclusion

We presented DAZM, a memory-efficient Muon-style ZO fine-tuning method that adapts low-rank search subspaces using parameter displacements accumulated from recent updates. Rather than refreshing the search space randomly, DAZM retains the leading singular modes of the realized displacement, thereby incorporating recent optimization information into subsequent subspace construction. Our theoretical analysis characterizes the gradient signal accessible to subspace Muon and establishes conditions under which temporal persistence enables displacement-guided subspaces to capture more useful gradient signals than purely random subspaces. Extensive experiments on multiple LLMs and downstream tasks demonstrate that our method achieves consistently superior performance over existing ZO-Muon and non-Muon ZO baselines.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, manuscript refinement, and improving the clarity and presentation of technical arguments. We did not use generative AI tools to conduct experiments, generate experimental results, or fabricate or modify empirical data, and human-subject data collection and participant interaction are not applicable to this work. We carefully reviewed all AI-assisted content. References, equations, notation, and technical claims were also manually reviewed for correctness and consistency. Generative AI was used only as an assistive tool, and all final scientific decisions, analyses, interpretations, and conclusions were made by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

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

Yiming Chen, Yuan Zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order finetuning for language models with low-rank structures. In *International Conference on Learning Representations*, volume 2025, pp. 62581–62607, 2025.

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.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Jinjie Fang, Chengxun Jin, Tianxing Man, Yi Chang, and Bin Gu. Lozo+: Provably efficient zeroth-order fine-tuning via greedy low-rank subspace selection. In *Forty-third International Conference on Machine Learning*, 2026.

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

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

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

Chuan He, Zhanwang Deng, and Zhaosong Lu. Low-rank orthogonalization for large-scale matrix optimization with applications to foundation model training. *arXiv preprint arXiv:2509.11983*, 2025.

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, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

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

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

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

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.

Shenggui Li, Fuzhao Xue, Chaitanya Baranwal, Yongbin Li, and Yang You. Sequence parallelism: Long sequence training from system perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2391–2404, 2023.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

Vladislav Lialin, Vijeta Deshpande, Xiaowei Yao, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*, 2023.

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 *Forty-first International Conference on Machine Learning*, 2024a.

Yong Liu, Zirui Zhu, Chaoyu Gong, Minhao Cheng, Cho-Jui Hsieh, and Yang You. Sparse mezo: Less parameters for better performance in zeroth-order llm fine-tuning. *arXiv preprint arXiv:2402.15751*, 2024b.

Cheng Luo, Jiawei Zhao, Zhuoming Chen, Beidi Chen, and Anima Anandkumar. Mini-sequence transformers: Optimizing intermediate memory for long sequences training. *Advances in Neural Information Processing Systems*, 37:97299–97327, 2024.

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.

Yuezhang Peng, Yuxin Liu, Yao Li, Sheng Wang, Fei Wen, and Xie Chen. Zo-asr: Zeroth-order fine-tuning of speech foundation models without back-propagation. *arXiv preprint arXiv:2512.01267*, 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.

Haotong Qin, Xudong Ma, Xingyu Zheng, Xiaoyang Li, Yang Zhang, Shouda Liu, Jie Luo, Xianglong Liu, and Michele Magno. Accurate lora-finetuning quantization of llms via information retention. *arXiv preprint arXiv:2402.05445*, 2024.

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

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4938–4947, 2020.

Zichen Song and Weijia Li. Rozo: Geometry-aware zeroth-order fine-tuning on low-rank adapters for black-box large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1773–1783, 2026.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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 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, 2018.

Shaowen Wang, Linxi Yu, and Jian Li. Lora-ga: Low-rank adaptation with gradient approximation. *Advances in Neural Information Processing Systems*, 37:54905–54931, 2024.

Zhongruo Wang, Krishnakumar Balasubramanian, Shiqian Ma, and Meisam Razaviyayn. Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *arXiv preprint arXiv:2001.07819*, 2020.

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

Jia Zhang, Yu Bai, Hualin Zhang, Tianshuo Chen, Zhaogeng Liu, Yi Chang, Bin Gu, et al. Three forward, one backward: Memory-efficient full-rank fine-tuning of large models via extra forward passes. In *The Fourteenth International Conference on Learning Representations*, 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.

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

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

APPENDIX

A ALGORITHM

Algorithm 2 Low-Rank Subspace Adaptation

Require: Exploitation rank p, subspace rank r, bases $\mathbf{U}_k, \mathbf{V}_k$, accumulated displacement $\mathbf{\Psi}_k$
Ensure: Updated bases $\mathbf{U}_{k+1}, \mathbf{V}_{k+1}$

$\mathbf{\Psi}_k = \mathbf{A}_k \mathbf{\Lambda}_k \mathbf{B}_k^\top \leftarrow \mathrm{SVD}(\mathbf{\Psi}_k)$

$\mathbf{U}_{k+1}^{\text{exploit}} \leftarrow \mathbf{U}_k \mathbf{A}_{k,1:p}, \quad \mathbf{V}_{k+1}^{\text{exploit}} \leftarrow \mathbf{B}_{k,1:p}^\top \mathbf{V}_k$

Sample $\mathbf{\Omega}_U \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ with $\mathbf{\Omega}_U \in \mathbb{R}^{m \times (r-p)}$

$\widetilde{\mathbf{U}}_{k+1}^{\text{explore}} \leftarrow \left(\mathbf{I} - \mathbf{U}_{k+1}^{\text{exploit}} \mathbf{U}_{k+1}^{\text{exploit}\top}\right) \mathbf{\Omega}_U, \quad \mathbf{U}_{k+1}^{\text{explore}} \leftarrow \mathrm{QR}\left(\widetilde{\mathbf{U}}_{k+1}^{\text{explore}}\right)$

Sample $\mathbf{\Omega}_V \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ with $\mathbf{\Omega}_V \in \mathbb{R}^{n \times (r-p)}$

$\mathbf{R}_{k+1}^{\text{exploit}} \leftarrow \mathbf{V}_{k+1}^{\text{exploit}\top}, \quad \widetilde{\mathbf{R}}_{k+1}^{\text{explore}} \leftarrow \left(\mathbf{I} - \mathbf{R}_{k+1}^{\text{exploit}} \mathbf{R}_{k+1}^{\text{exploit}\top}\right) \mathbf{\Omega}_V$

$\mathbf{R}_{k+1}^{\text{explore}} \leftarrow \mathrm{QR}\left(\widetilde{\mathbf{R}}_{k+1}^{\text{explore}}\right), \quad \mathbf{V}_{k+1}^{\text{explore}} \leftarrow \mathbf{R}_{k+1}^{\text{explore}\top}$

$\mathbf{U}_{k+1} \leftarrow \left[\mathbf{U}_{k+1}^{\text{exploit}}, \mathbf{U}_{k+1}^{\text{explore}}\right], \quad \mathbf{V}_{k+1} \leftarrow \begin{bmatrix} \mathbf{V}_{k+1}^{\text{exploit}} \\ \mathbf{V}_{k+1}^{\text{explore}} \end{bmatrix}$

return $\mathbf{U}_{k+1}, \mathbf{V}_{k+1}$

B EXPERIMENTAL SUPPLEMENTARY

B.1 BASELINE DESCRIPTIONS

Detailed descriptions of all baselines are provided below:

- **Zero-shot** directly applies the pretrained model to downstream tasks without introducing task-specific examples or performing any parameter updates.
- **In-Context Learning (ICL)** (Dong et al., 2024) conditions the model on a few task demonstrations included in the input context and performs prediction without modifying the pretrained parameters.
- **Fine-tuning with Adam** (Kingma, 2014) serves as the full-parameter FO baseline, where all trainable model weights are optimized with Adam using the same training-data budget.
- **FT-LoRA** (Hu et al., 2022) performs FO parameter-efficient fine-tuning by freezing the pretrained backbone and introducing trainable low-rank factors into selected Transformer layers. Since optimization is restricted to these low-dimensional adapters, LoRA substantially reduces the number of trainable parameters and the associated memory overhead compared with full fine-tuning.
- **MeZO** (Malladi et al., 2023) enables backpropagation-free fine-tuning through zeroth-order gradient estimation. It evaluates the objective under symmetric parameter perturbations and constructs an update using only forward evaluations. By avoiding activation storage and backward-graph construction, MeZO maintains a memory footprint close to that required for inference.
- **MeZO-LoRA** (Malladi et al., 2023) combines MeZO with LoRA by applying zeroth-order optimization only to the low-rank adapter parameters while leaving the pretrained backbone unchanged. This substantially reduces the dimension of the search space and alleviates the difficulty of zeroth-order estimation in large models.
- **SubZero** (Yu et al., 2025) performs zeroth-order optimization within a randomly sampled low-dimensional subspace rather than perturbing the entire parameter space. Reducing the perturbation dimension helps improve the efficiency of gradient estimation and mitigates the variance associated with full-dimensional ZO optimization.

- **LOZO** (Chen et al., 2025) further exploits low-rank structure by parameterizing zeroth-order perturbations in a low-rank search space. This structured restriction lowers the effective optimization dimension and provides a more efficient alternative to full-space ZO fine-tuning under limited memory budgets.
- **LOZO+** (Fang et al., 2026) extends the low-rank ZO framework of LOZO with an enhanced subspace optimization strategy, while retaining the memory-efficient low-rank parameterization. It provides a stronger low-rank ZO baseline for evaluating whether more informative subspace construction can further improve optimization performance.
- **ZO-Muon** (Lang et al., 2026) integrates zeroth-order gradient estimation with the matrix-structured Muon update. Instead of directly applying an SGD-style update to the estimated gradient, ZO-Muon performs a NS-based spectral transformation in the reduced matrix space and maps the resulting update back to the original parameter space, enabling matrix-geometry-aware ZO optimization without backpropagation.

B.2 The Hyperparameter Settings

Table 3 summarizes the hyperparameter grids used for AHZO and all baselines in our experiments on the Qwen and OPT models. Following the training setup of (Chen et al., 2025; Malladi et al., 2023), we use 30k queries for the ZO methods. For FT and FT-LoRA, we use 625 update steps.

Table 3: The hyperparameter grids for experiments.

Qwen2.5-7B			**OPT-13B**		
Experiment	**Hyperparameters**	**Values**	**Experiment**	**Hyperparameters**	**Values**
DAZM	Batch size	16	DAZM	Batch size	16
	Rank (r)	{128, 256}		Rank (r)	{128, 256}
	Interval (ν)	{50, 100, 200}		Interval (ν)	{50, 100, 200}
	ϵ	1e−3		ϵ	1e−3
	Learning rate	1e−2		Learning rate	1e−2
	p	{32, 64, 128}		p	{32, 64, 128}
ZO-Muon	Batch size	16	ZO-Muon	Batch size	16
	Rank (r)	{64, 128}		Rank (r)	{64, 128}
	Interval (ν)	{50, 100, 200}		Interval (ν)	{50, 100, 200}
	ϵ	1e−3		ϵ	1e−3
	Learning rate	1e−2		Learning rate	1e−2
LOZO+	Batch size	16	LOZO+	Batch size	16
	Rank (r)	{4, 8}		Rank (r)	{2, 4, 8}
	Threshold (ϕ)	0.1		Threshold (ϕ)	0.1
	Interval (ν)	{50, 100, 200, 300}		Interval (ν)	{50, 100, 200, 300}
	ϵ	1e−2		ϵ	1e−2
	Learning rate	{1e−3, 1e−4}		Learning rate	{1e−3, 1e−4}
LOZO	Batch size	16	LOZO	Batch size	16
	Rank (r)	{4, 8}		Rank (r)	{2, 4, 8}
	Interval (ν)	{50, 100, 200, 300}		Interval (ν)	{50, 100, 200, 300}
	ϵ	1e−3		ϵ	1e−3
	Learning rate	{1e−6, 1e−7}		Learning rate	{1e−6, 1e−7}
SubZero	Batch size	16	SubZero	Batch size	16
	Gauss Rank	24		Gauss Rank	24
	ϵ	1e−3		ϵ	1e−3
	Learning rate	{1e−6, 1e−7}		Learning rate	{1e−6, 1e−7}
	Update interval	1000		Update interval	1000
MeZO	Batch size	16	MeZO	Batch size	16
	ϵ	1e−3		ϵ	1e−3
	Learning rate	{1e−7, 5e−7}		Learning rate	{1e−7, 1e−6, 5e−7}
MeZO-LoRA	Batch size	16	MeZO-LoRA	Batch size	16
	(r, α)	$(4, 16)$		(r, α)	$(8, 16)$
	ϵ	1e−2		ϵ	1e−3
	Learning rate	1e−5		Learning rate	{1e−5, 5e−5, 1e−4}
FT-LoRA	Batch size	8	FT-LoRA	Batch size	8
	(r, α)	$(4, 16)$		(r, α)	$(4, 16)$
	Learning rate	1e−5		Learning rate	{1e−5, 5e−5, 8e−5}
FT	Batch size	8	FT	Batch size	8
	Learning rate	1e−5		Learning rate	{1e−5, 5e−5, 8e−5}

B.3 Memory Consumption

As reported in Tables 4 and 5, DAZM maintains a consistently low memory overhead on both Qwen2.5-7B and OPT-13B. Its peak memory usage remains in the same range as memory-efficient ZO methods with batch size 16 and FO baselines with batch size 8, while requiring markedly less memory than full fine-tuning across all evaluated tasks.

Table 4: Estimated peak GPU memory (GiB) on Qwen2.5-7B with rank $r = 128$.

Method	SST-2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	Avg.
FT	80.00	121.80	153.10	123.70	85.30	80.90	116.80	76.20	127.70	139.70	149.60	114.07
FT-LoRA	31.04	68.94	58.44	74.74	33.54	29.64	49.04	19.94	79.24	91.44	99.14	57.74
MeZO	20.05	31.21	39.49	43.77	23.58	20.44	59.48	20.05	28.75	37.73	72.06	36.05
MeZO-LoRA	19.18	31.28	39.56	43.85	23.65	20.51	59.55	16.72	28.83	37.81	72.14	35.73
SubZero	19.79	31.90	36.15	44.47	31.04	30.98	60.17	19.73	29.45	38.43	72.76	37.71
LOZO	19.41	31.52	39.80	44.08	23.89	20.75	59.78	19.34	29.06	38.04	62.96	35.33
LOZO+	21.49	32.43	36.67	44.99	31.57	22.28	60.69	31.47	29.97	38.95	73.28	38.53
ZO-Muon	20.60	32.71	41.00	45.28	25.08	21.94	60.98	19.65	30.26	39.24	60.25	36.09
DAZM	19.81	31.93	36.17	44.50	21.16	21.16	60.21	18.86	29.48	38.47	59.50	34.66

Table 5: Estimated peak GPU memory (GiB) on OPT-13B with rank $r = 128$.

Method	SST-2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	Avg.
FT	103.00	140.60	167.00	203.10	107.40	108.50	198.00	101.20	109.80	219.00	176.10	148.52
FT-LoRA	39.54	79.74	110.84	143.64	42.04	43.04	108.74	31.64	42.74	176.84	111.04	84.53
MeZO	26.80	34.10	40.10	44.50	27.40	27.20	52.60	25.50	33.80	41.00	74.50	38.86
MeZO-LoRA	27.38	34.38	40.58	45.18	27.48	27.68	53.08	25.38	33.78	41.38	75.08	39.22
SubZero	29.83	40.23	49.33	46.03	29.13	28.93	64.03	28.13	37.13	49.03	74.43	43.30
LOZO	27.34	36.34	41.54	44.24	29.04	28.44	52.34	27.34	35.24	47.64	76.54	40.55
LOZO+	28.13	34.54	49.67	39.89	49.67	28.40	50.58	49.67	49.67	39.18	70.87	44.57
ZO-Muon	28.32	35.01	40.63	40.35	30.73	28.86	51.05	27.17	33.54	39.64	71.33	38.79
DAZM	27.10	33.78	36.59	39.12	27.64	27.63	49.81	25.94	32.31	38.41	70.10	37.13

B.4 Wall-clock Time

Tables 6 and 7 report the practical wall-clock runtime of DAZM with $\nu = 100$ on Qwen2.5-7B and OPT-13B. Although DAZM introduces additional matrix decompositions for displacement-aware subspace refresh, these operations are performed only periodically and therefore incur limited computational overhead. As a result, its overall runtime remains close to that of ZO-Muon. More importantly, the Muon-based methods require less training time than non-Muon ZO baselines under the same query budget, indicating higher optimization efficiency and faster practical convergence. Overall, DAZM improves the quality of the search subspace while largely preserving the computational efficiency of ZO-Muon.

Table 6: Pure training runtime on Qwen2.5-7B under a 30k-query equivalent budget.

Method	SST-2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	Avg.
LOZO+	3h 58m	7h 03m	8h 17m	9h 37m	4h 29m	4h 20m	16h 42m	3h 24m	6h 42m	7h 00m	11h 10m	7h 31m
ZO-Muon	1h 57m	4h 53m	6h 50m	7h 20m	3h 08m	2h 18m	14h 06m	1h 25m	4h 33m	4h 50m	8h 23m	5h 26m
DAZM	2h 07m	5h 03m	6h 21m	7h 33m	6h 17m	2h 24m	14h 45m	1h 57m	6h 02m	5h 05m	8h 29m	6h 00m

Table 7: Pure training runtime on OPT-13B under a 30k-query equivalent budget.

Method	SST-2	RTE	CB	BoolQ	WSC	WiC	MultiRC	COPA	ReCoRD	SQuAD	DROP	Avg.
LOZO+	6h 30m	11h 10m	13h 23m	15h 03m	7h 15m	7h 08m	29h 07m	5h 40m	10h 48m	11h 12m	18h 19m	12h 20m
ZO-Muon	3h 53m	9h 58m	11h 06m	10h 49m	5h 03m	3h 40m	23h 15m	2h 23m	8h 38m	7h 46m	13h 53m	9h 08m
DAZM	4h 40m	7h 49m	10h 10m	11h 55m	4h 32m	3h 51m	25h 35m	2h 30m	7h 14m	8h 18m	14h 49m	9h 13m

B.5 Sensitivity Analysis

Figure 3 evaluates the sensitivity of DAZM to the learning rate η, subspace rank r, and refresh interval ν. A moderate learning rate provides the best trade-off between rapid descent and stable convergence. Increasing the subspace rank generally improves optimization by enlarging the accessible search space, although the gain gradually saturates at larger ranks. Meanwhile, smaller refresh

intervals tend to achieve lower final training loss, indicating the benefit of periodically adapting the search subspace to the evolving optimization trajectory. Overall, DAZM remains effective across a broad range of these hyperparameters.

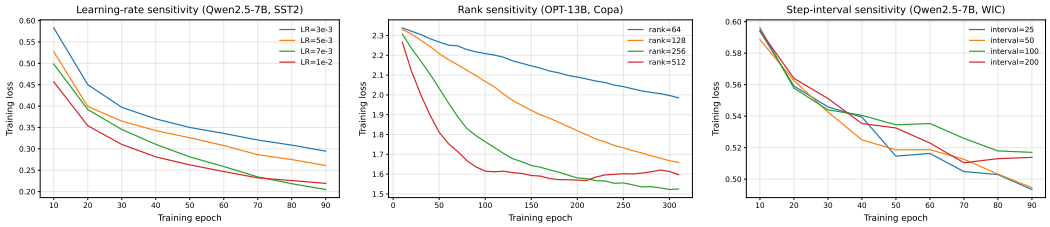


Figure 3: Sensitivity analysis for η, r and ν

B.6 Ablation Study

Table 8 evaluates the effect of the proposed displacement-aware subspace adaptation. *Frozen* keeps the initial search subspace fixed throughout optimization, while *Random* periodically refreshes the search subspace using randomly sampled bases without exploiting the accumulated displacement. Our full method consistently outperforms both variants across all four tasks, achieving an average score of 85.10, compared with 82.71 for Frozen and 82.52 for Random. The degradation of Frozen indicates that a fixed subspace becomes insufficient as the optimization trajectory evolves, while the inferior performance of Random shows that subspace refresh alone is not sufficient. These results support the benefit of using accumulated Muon displacement to adapt the search subspace toward directions that have persistently contributed to recent optimization.

Table 8: Ablation study.

Method	Qwen2.5-7B		OPT-13B		Average
	RTE	WiC	COPA	SST-2	
Ours	**89.17**	**70.69**	**88.00**	**92.55**	**85.10**
Frozen	88.81	67.87	83.00	91.17	82.71
Random	87.73	68.50	82.00	91.86	82.52

B.7 Verification of the Displacement-Aware Capture Bound

Figure 4 empirically examines the lower bound established in Theorem 2, namely $\rho_{\tau_{k+1}} \geq \gamma_{p,k+1} - \varepsilon^2_{\text{exploit},k}$. Across different models, datasets, layers, and refresh steps, the certified lower bound remains consistently stable, indicating that the displacement-retained subspace preserves a non-negligible fraction of the gradient signal available for the subsequent Muon update. For $r = 256$, the expected random-subspace capture benchmark $r/\sqrt{mn}$ is approximately 5.0% for the 5120×5120 attention projections in OPT-13B and 7.14% for the 3584×3584 projections in Qwen2.5-7B. In contrast, the observed DAZM lower bounds are approximately 9.1%–9.8% for OPT-13B and 13.2%–14.5% for Qwen2.5-7B, remaining above the corresponding random-subspace benchmarks throughout training. These results support the theoretical characterization that, when the retained displacement modes remain aligned with the leading gradient structure, displacement-aware adaptation can expose substantially more Muon-accessible gradient signal than an independent random refresh.

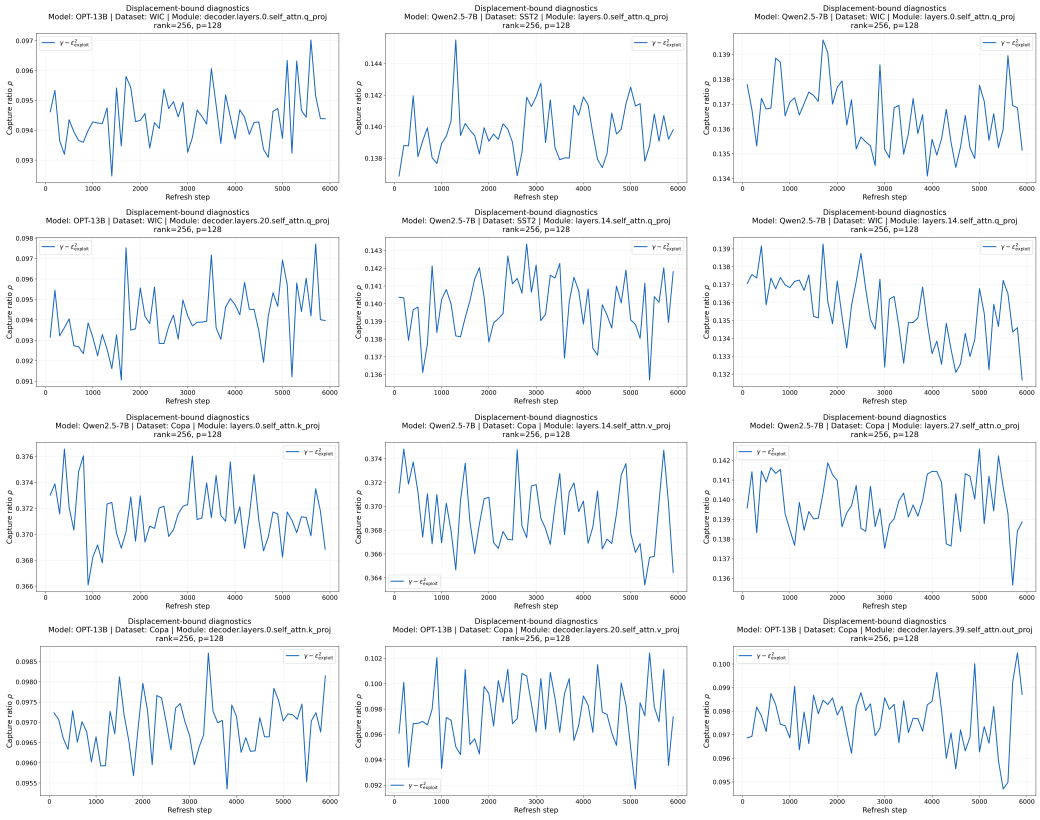


Figure 4: Evolution of the certified displacement-aware capture lower bound $\gamma_{p,k+1} - \varepsilon^2_{\mathrm{exploit},k}$.

C AUXILIARY LEMMAS

We consider a matrix-valued parameter layer $\mathbf{W} \in \mathbb{R}^{m\times n}$ and the objective $\mathcal{F}(\mathbf{W})$. At iteration t, the two-sided rank-r search bases are

$$\mathbf{U}_t \in \mathbb{R}^{m\times r}, \qquad \mathbf{V}_t \in \mathbb{R}^{r\times n}, \tag{24}$$

with

$$\mathbf{U}_t^\top \mathbf{U}_t = \mathbf{I}_r, \qquad \mathbf{V}_t \mathbf{V}_t^\top = \mathbf{I}_r. \tag{25}$$

The true projected gradient is

$$\mathbf{G}_t := \mathbf{U}_t^\top \nabla \mathcal{F}(\mathbf{W}_t) \mathbf{V}_t^\top \in \mathbb{R}^{r\times r}. \tag{26}$$

The practical Muon-style direction is generated from the transported reduced momentum by

$$\mathbf{O}_t = \text{Newton-Schulz}(\mathbf{M}_t, q), \tag{27}$$

and the update is

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \mathbf{U}_t \mathbf{O}_t \mathbf{V}_t. \tag{28}$$

Reduced Gaussian-core estimator. At iteration t, sample $\mathbf{\Sigma}_{t,i} \in \mathbb{R}^{r\times r}$ with $\text{vec}(\mathbf{\Sigma}_{t,i}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{r^2})$ and define

$$\Phi_{t,i} := \frac{\mathcal{F}(\mathbf{W}_t + \mu \mathbf{U}_t \mathbf{\Sigma}_{t,i} \mathbf{V}_t; \mathcal{B}_t) - \mathcal{F}(\mathbf{W}_t; \mathcal{B}_t)}{\mu}, \tag{29}$$

$$\widehat{\mathbf{G}}_t := \frac{1}{N} \sum_{i=1}^{N} \Phi_{t,i} \mathbf{\Sigma}_{t,i}. \tag{30}$$

The effective Gaussian dimension is therefore

$$d := r^2. \tag{31}$$

As in the estimator bound used by the original proof, we write

$$\Theta := \frac{L^2 \mu^2}{2N} d(d+2)(d+4) + \frac{2\sigma^2}{B}, \qquad d = r^2, \tag{32}$$

and use the reduced-space mean-square error condition

$$\mathbb{E}\big[\|\widehat{\mathbf{G}}_t - \mathbf{G}_t\|_F^2\big] \le \Theta. \tag{33}$$

This keeps the same estimator-error premise as the supplied one-sided proof, with the effective dimension changed from rn to r^2.

Momentum transport across subspace refreshes. To retain the original momentum-tracking proof while allowing the two-sided bases to change, define, for $t \ge 1$,

$$\mathcal{T}_t(\mathbf{A}) := \mathbf{U}_t^\top \mathbf{U}_{t-1}\, \mathbf{A}\, \mathbf{V}_{t-1} \mathbf{V}_t^\top. \tag{34}$$

Since both factors are contractions,

$$\|\mathcal{T}_t(\mathbf{A})\|_F \le \|\mathbf{A}\|_F. \tag{35}$$

When the bases are unchanged, $\mathcal{T}_t$ is the identity. The scaled practical and ideal momenta are defined by

$$\mathbf{N}_t = \beta \mathcal{T}_t(\mathbf{N}_{t-1}) + (1-\beta)\widehat{\mathbf{G}}_t, \tag{36}$$

$$\overline{\mathbf{N}}_t = \beta \mathcal{T}_t(\overline{\mathbf{N}}_{t-1}) + (1-\beta)\mathbf{G}_t, \tag{37}$$

with $\mathbf{N}_{-1} = \overline{\mathbf{N}}_{-1} = \mathbf{0}$. Equivalently, the unscaled momentum obeys $\mathbf{M}_t = \beta \mathcal{T}_t(\mathbf{M}_{t-1}) + \widehat{\mathbf{G}}_t$ and $\mathbf{N}_t = (1-\beta)\mathbf{M}_t$.

Define the refresh mismatch

$$\mathbf{R}_t := \mathbf{G}_t - \mathcal{T}_t(\mathbf{G}_{t-1}) - \mathbf{U}_t^\top \big(\nabla \mathcal{F}(\mathbf{W}_t) - \nabla \mathcal{F}(\mathbf{W}_{t-1})\big) \mathbf{V}_t^\top, \tag{38}$$

$$\chi_t := \|\mathbf{R}_t\|_F, \quad \Gamma_{1,T} := \frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E}[\chi_t], \quad \Gamma_{2,T} := \frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E}[\chi_t^2]. \tag{39}$$

If the bases are fixed over the analyzed horizon, then $\chi_t = \Gamma_{1,T} = \Gamma_{2,T} = 0$.

Lemma 1 (Descent estimate) *Under Assumption 1, consider*

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta \mathbf{U}_t \mathbf{O}_t \mathbf{V}_t. \tag{40}$$

Then

$$\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \leq -\eta \left\langle \nabla \mathcal{F}(\mathbf{W}_t), \mathbf{U}_t \mathbf{O}_t \mathbf{V}_t \right\rangle_F + \frac{L\eta^2}{2} \|\mathbf{U}_t \mathbf{O}_t \mathbf{V}_t\|_{\mathrm{op}}^2. \tag{41}$$

Moreover,

$$\|\mathbf{U}_t \mathbf{O}_t \mathbf{V}_t\|_{\mathrm{op}} = \|\mathbf{O}_t\|_{\mathrm{op}}. \tag{42}$$

Proof 1 *Let $\mathbf{D}_t := \mathbf{W}_{t+1} - \mathbf{W}_t$ and define $h(s) := \mathcal{F}(\mathbf{W}_t + s\mathbf{D}_t)$ for $s \in [0, 1]$. By the fundamental theorem of calculus,*

$$\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) = \int_0^1 \left\langle \nabla \mathcal{F}(\mathbf{W}_t + s\mathbf{D}_t), \mathbf{D}_t \right\rangle_F ds. \tag{43}$$

Separating the gradient at $\mathbf{W}_t$ gives

$$\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \left\langle \nabla \mathcal{F}(\mathbf{W}_t), \mathbf{D}_t \right\rangle_F + \int_0^1 \left\langle \nabla \mathcal{F}(\mathbf{W}_t + s\mathbf{D}_t) - \nabla \mathcal{F}(\mathbf{W}_t), \mathbf{D}_t \right\rangle_F ds. \tag{44}$$

By nuclear/operator norm duality and Assumption 1,

$$\left\langle \nabla \mathcal{F}(\mathbf{W}_t + s\mathbf{D}_t) - \nabla \mathcal{F}(\mathbf{W}_t), \mathbf{D}_t \right\rangle_F \leq \|\nabla \mathcal{F}(\mathbf{W}_t + s\mathbf{D}_t) - \nabla \mathcal{F}(\mathbf{W}_t)\|_* \|\mathbf{D}_t\|_{\mathrm{op}} \leq Ls \|\mathbf{D}_t\|_{\mathrm{op}}^2. \tag{45}$$

Integrating over s yields

$$\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \leq \left\langle \nabla \mathcal{F}(\mathbf{W}_t), \mathbf{D}_t \right\rangle_F + \frac{L}{2} \|\mathbf{D}_t\|_{\mathrm{op}}^2. \tag{46}$$

Substituting $\mathbf{D}_t = -\eta \mathbf{U}_t \mathbf{O}_t \mathbf{V}_t$ proves Equation 41. Finally, because $\mathbf{U}_t$ has orthonormal columns and $\mathbf{V}_t$ has orthonormal rows, multiplication by these partial isometries preserves the nonzero singular values of $\mathbf{O}_t$, hence Equation 42 follows.

Lemma 2 (Mini-batch variance) *Let*

$$\nabla \mathcal{F}(\mathbf{W}_t; \mathcal{B}_t) = \frac{1}{B} \sum_{b=1}^{B} \nabla \mathcal{F}(\mathbf{W}_t; \xi_{t,b}). \tag{47}$$

Then

$$\mathbb{E}\Big[\|\nabla \mathcal{F}(\mathbf{W}_t; \mathcal{B}_t) - \nabla \mathcal{F}(\mathbf{W}_t)\|_F^2 \Big] \leq \frac{\sigma^2}{B}, \tag{48}$$

and

$$\mathbb{E}\Big[\|\mathbf{U}_t^\top (\nabla \mathcal{F}(\mathbf{W}_t; \mathcal{B}_t) - \nabla \mathcal{F}(\mathbf{W}_t)) \mathbf{V}_t^\top\|_F^2 \Big] \leq \frac{\sigma^2}{B}. \tag{49}$$

Proof 2 *The first inequality follows from independence, unbiasedness, and Assumption 2. For the second inequality, for any matrix $\mathbf{A}$,*

$$\|\mathbf{U}_t^\top \mathbf{A} \mathbf{V}_t^\top\|_F \leq \|\mathbf{A}\mathbf{V}_t^\top\|_F \leq \|\mathbf{A}\|_F, \tag{50}$$

where Equation 25 is used twice. Applying this to the mini-batch error proves Equation 49.

Lemma 3 (Gradient-gap inequality) *Under Assumption 1, for any $\mathbf{W}$,*

$$\|\nabla \mathcal{F}(\mathbf{W})\|_F^2 \leq 2L\big(\mathcal{F}(\mathbf{W}) - \mathcal{F}(\mathbf{W}^\star)\big). \tag{51}$$

In particular,

$$\|\nabla \mathcal{F}(\mathbf{W}_0)\|_F^2 \leq 2LH. \tag{52}$$

Proof 3 *The descent inequality induced by Assumption 1 gives, for any* $\mathbf{Y}$*,*

$$\mathcal{F}(\mathbf{Y}) \le \mathcal{F}(\mathbf{W}) + \langle \nabla \mathcal{F}(\mathbf{W}), \mathbf{Y} - \mathbf{W} \rangle_F + \frac{L}{2} \|\mathbf{Y} - \mathbf{W}\|_{\mathrm{op}}^2. \tag{53}$$

Choose $\mathbf{Y} = \mathbf{W} - L^{-1} \nabla \mathcal{F}(\mathbf{W})$*. Then*

$$\begin{aligned} \mathcal{F}\left(\mathbf{W} - \frac{1}{L} \nabla \mathcal{F}(\mathbf{W})\right) &\le \mathcal{F}(\mathbf{W}) - \frac{1}{L} \|\nabla \mathcal{F}(\mathbf{W})\|_F^2 + \frac{1}{2L} \|\nabla \mathcal{F}(\mathbf{W})\|_{\mathrm{op}}^2 \\ &\le \mathcal{F}(\mathbf{W}) - \frac{1}{2L} \|\nabla \mathcal{F}(\mathbf{W})\|_F^2, \end{aligned} \tag{54}$$

where $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$ *is used. Since* $\mathcal{F}(\mathbf{W}^\star)$ *is no larger than the left-hand side, rearrangement proves Equation 51. Equation Equation 52 follows from Assumption 3.*

Lemma 4 (Projected polar alignment) *For* $\mathbf{G}_t$ *in Equation 26,*

$$\langle \nabla \mathcal{F}(\mathbf{W}_t), \mathbf{U}_t \operatorname{Polar}(\mathbf{G}_t) \mathbf{V}_t \rangle_F = \|\mathbf{G}_t\|_*. \tag{55}$$

Moreover, for any reduced matrix $\mathbf{M}_t$*,*

$$\langle \mathbf{M}_t, \operatorname{Polar}(\mathbf{M}_t) \rangle_F = \|\mathbf{M}_t\|_*. \tag{56}$$

Proof 4 *Using the definition of* $\mathbf{G}_t$*,*

$$\langle \nabla \mathcal{F}(\mathbf{W}_t), \mathbf{U}_t \operatorname{Polar}(\mathbf{G}_t) \mathbf{V}_t \rangle_F = \langle \mathbf{G}_t, \operatorname{Polar}(\mathbf{G}_t) \rangle_F. \tag{57}$$

Let the compact SVD be $\mathbf{G}_t = \mathbf{A}_t \mathbf{\Lambda}_t \mathbf{B}_t^\top$*. Then* $\operatorname{Polar}(\mathbf{G}_t) = \mathbf{A}_t \mathbf{B}_t^\top$ *and*

$$\langle \mathbf{G}_t, \operatorname{Polar}(\mathbf{G}_t) \rangle_F = \operatorname{tr}\left[(\mathbf{A}_t \mathbf{\Lambda}_t \mathbf{B}_t^\top)^\top \mathbf{A}_t \mathbf{B}_t^\top \right] = \operatorname{tr}(\mathbf{B}_t \mathbf{\Lambda}_t \mathbf{B}_t^\top) = \operatorname{tr}(\mathbf{\Lambda}_t) = \|\mathbf{G}_t\|_*. \tag{58}$$

The second identity follows identically by replacing $\mathbf{G}_t$ *with* $\mathbf{M}_t$*.*

Lemma 5 (Momentum error bound for projected polar directions) *Let* $\mathbf{N}_t = (1-\beta)\mathbf{M}_t$*. Since positive scaling does not change the polar factor,* $\operatorname{Polar}(\mathbf{M}_t) = \operatorname{Polar}(\mathbf{N}_t)$*. Then*

$$|\langle \mathbf{G}_t, \operatorname{Polar}(\mathbf{G}_t) - \operatorname{Polar}(\mathbf{M}_t) \rangle_F| \le 2 \|\mathbf{G}_t - \mathbf{N}_t\|_* \le 2\sqrt{r} \|\mathbf{G}_t - \mathbf{N}_t\|_F. \tag{59}$$

Proof 5 *Let*

$$\delta_t := \langle \mathbf{G}_t, \operatorname{Polar}(\mathbf{G}_t) - \operatorname{Polar}(\mathbf{N}_t) \rangle_F. \tag{60}$$

By Lemma 4,

$$\delta_t = \|\mathbf{G}_t\|_* - \langle \mathbf{G}_t, \operatorname{Polar}(\mathbf{N}_t) \rangle_F. \tag{61}$$

Adding and subtracting $\mathbf{N}_t$ *gives*

$$\delta_t = \|\mathbf{G}_t\|_* - \|\mathbf{N}_t\|_* - \langle \mathbf{G}_t - \mathbf{N}_t, \operatorname{Polar}(\mathbf{N}_t) \rangle_F. \tag{62}$$

Therefore,

$$\begin{aligned} |\delta_t| &\le \left| \|\mathbf{G}_t\|_* - \|\mathbf{N}_t\|_* \right| + \left| \langle \mathbf{G}_t - \mathbf{N}_t, \operatorname{Polar}(\mathbf{N}_t) \rangle_F \right| \\ &\le \|\mathbf{G}_t - \mathbf{N}_t\|_* + \|\mathbf{G}_t - \mathbf{N}_t\|_* \|\operatorname{Polar}(\mathbf{N}_t)\|_{\mathrm{op}} \\ &\le 2 \|\mathbf{G}_t - \mathbf{N}_t\|_*. \end{aligned} \tag{63}$$

Since $\mathbf{G}_t - \mathbf{N}_t \in \mathbb{R}^{r \times r}$*,*

$$\|\mathbf{G}_t - \mathbf{N}_t\|_* \le \sqrt{\operatorname{rank}(\mathbf{G}_t - \mathbf{N}_t)} \|\mathbf{G}_t - \mathbf{N}_t\|_F \le \sqrt{r} \|\mathbf{G}_t - \mathbf{N}_t\|_F. \tag{64}$$

This proves Equation 59.

Lemma 6 (Newton–Schulz approximation error) *Let* $\mathbf{M}_t \neq \mathbf{0}$ *and initialize*

$$\mathbf{X}_{t,0} = \frac{\mathbf{M}_t}{\|\mathbf{M}_t\|_{\mathrm{op}}}. \tag{65}$$

Run $\mathbf{X}_{t,k+1} = \frac{1}{2} \mathbf{X}_{t,k} \left(3\mathbf{I} - \mathbf{X}_{t,k}^\top \mathbf{X}_{t,k} \right)$*,* $k = 0, \ldots, q-1$*, and set* $\mathbf{O}_t := \mathbf{X}_{t,q}$*. Define* $\varepsilon_{t,q} := \|\operatorname{Polar}(\mathbf{M}_t) - \mathbf{O}_t\|_{\mathrm{op}}$*. Then* $\|\mathbf{O}_t\|_{\mathrm{op}} \le 1 + \varepsilon_{t,q}$*, and if* $\varepsilon_{t,0} < 2/3$*, we can get:*

$$\varepsilon_{t,q} \le \frac{2}{3} \left(\frac{3}{2} \varepsilon_{t,0} \right)^{2^q}. \tag{66}$$

Proof 6 *Let the compact SVD be*

$$\mathbf{M}_t = \mathbf{A}_t \widehat{\mathbf{\Sigma}}_t \mathbf{B}_t^\top, \qquad \mathrm{Polar}(\mathbf{M}_t) = \mathbf{A}_t \mathbf{B}_t^\top. \tag{67}$$

Then

$$\mathbf{X}_{t,0} = \mathbf{A}_t \mathbf{\Sigma}_{t,0} \mathbf{B}_t^\top, \qquad \mathbf{\Sigma}_{t,0} = \frac{\widehat{\mathbf{\Sigma}}_t}{\|\mathbf{M}_t\|_{\mathrm{op}}}, \tag{68}$$

so all singular values on the compact support lie in $(0, 1]$. *Suppose* $\mathbf{X}_{t,k} = \mathbf{A}_t \mathbf{\Sigma}_{t,k} \mathbf{B}_t^\top$. *Then*

$$\mathbf{X}_{t,k}^\top \mathbf{X}_{t,k} = \mathbf{B}_t \mathbf{\Sigma}_{t,k}^2 \mathbf{B}_t^\top, \tag{69}$$

and substitution into $\mathbf{X}_{t,k+1}$ *gives*

$$\mathbf{X}_{t,k+1} = \mathbf{A}_t \mathbf{\Sigma}_{t,k+1} \mathbf{B}_t^\top, \tag{70}$$

where each singular value evolves as

$$\sigma_{t,i}^{(k+1)} = g(\sigma_{t,i}^{(k)}), \qquad g(x) := \frac{x(3 - x^2)}{2}. \tag{71}$$

Thus Newton–Schulz preserves the left and right singular vectors and only transforms the singular values. Since

$$\mathrm{Polar}(\mathbf{M}_t) - \mathbf{O}_t = \mathbf{A}_t (\mathbf{I} - \mathbf{\Sigma}_{t,q}) \mathbf{B}_t^\top, \tag{72}$$

we have

$$\varepsilon_{t,q} = \max_i |1 - \sigma_{t,i}^{(q)}|. \tag{73}$$

Also,

$$\|\mathbf{O}_t\|_{\mathrm{op}} \le \|\mathrm{Polar}(\mathbf{M}_t)\|_{\mathrm{op}} + \|\mathbf{O}_t - \mathrm{Polar}(\mathbf{M}_t)\|_{\mathrm{op}} \le 1 + \varepsilon_{t,q}. \tag{74}$$

For $x \in [0, 1]$,

$$1 - g(x) = \frac{1}{2}(1 - x)^2(2 + x) \le \frac{3}{2}(1 - x)^2. \tag{75}$$

Let $e_{t,k} := \max_i (1 - \sigma_{t,i}^{(k)})$. *Then*

$$e_{t,k+1} \le \frac{3}{2} e_{t,k}^2. \tag{76}$$

Setting $u_{t,k} := \frac{3}{2} e_{t,k}$ *gives* $u_{t,k+1} \le u_{t,k}^2$. *If* $e_{t,0} < 2/3$, *then* $u_{t,0} < 1$ *and* $u_{t,q} \le u_{t,0}^{2^q}$. *Therefore,*

$$e_{t,q} \le \frac{2}{3} \left(\frac{3}{2} e_{t,0} \right)^{2^q}. \tag{77}$$

Since $e_{t,q} = \varepsilon_{t,q}$, *Equation 66 follows.*

Lemma 7 (Transported momentum-estimation error for ZO-Muon) *Suppose the ZO-Muon update is*

$$\mathbf{W}_t = \mathbf{W}_{t-1} - \eta \mathbf{U}_{t-1} \mathbf{O}_{t-1} \mathbf{V}_{t-1}, \qquad \varepsilon_q := \sup_t \varepsilon_{t,q} < 1. \tag{78}$$

Then

$$\mathbb{E}\|\mathbf{G}_t - \mathbf{N}_t\|_F \le \beta^{t+1} \sqrt{2LH} + \sum_{i=1}^{t} \beta^{t-i+1} \Big(L\eta(1 + \varepsilon_q) + \mathbb{E}[\chi_i] \Big) + (1 - \beta^{t+1}) \sqrt{\Theta}. \tag{79}$$

Consequently,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{G}_t - \mathbf{N}_t\|_F \le \frac{\beta}{T(1-\beta)} \sqrt{2LH} + \frac{\beta}{1-\beta} L\eta(1 + \varepsilon_q) + \frac{\beta}{1-\beta} \Gamma_{1,T} + \sqrt{\Theta}. \tag{80}$$

Proof 7 *By the triangle inequality,*

$$\|\mathbf{G}_t - \mathbf{N}_t\|_F \le \underbrace{\|\mathbf{G}_t - \overline{\mathbf{N}}_t\|_F}_{\textit{momentum lag}} + \underbrace{\|\overline{\mathbf{N}}_t - \mathbf{N}_t\|_F}_{\textit{ZO estimation error}}. \tag{81}$$

For the first term, define $\mathbf{e}_t := \mathbf{G}_t - \overline{\mathbf{N}}_t$*. From Equation 37,*

$$\mathbf{e}_t = \beta\Big(\mathbf{G}_t - \mathcal{T}_t(\overline{\mathbf{N}}_{t-1})\Big) = \beta\mathcal{T}_t(\mathbf{e}_{t-1}) + \beta\Big(\mathbf{G}_t - \mathcal{T}_t(\mathbf{G}_{t-1})\Big). \tag{82}$$

Using Equation 35,

$$\|\mathbf{e}_t\|_F \le \beta\|\mathbf{e}_{t-1}\|_F + \beta\|\mathbf{G}_t - \mathcal{T}_t(\mathbf{G}_{t-1})\|_F. \tag{83}$$

By the definition of $\mathbf{R}_t$*,*

$$\begin{aligned}\|\mathbf{G}_t - \mathcal{T}_t(\mathbf{G}_{t-1})\|_F &\le \big\|\mathbf{U}_t^\top(\nabla\mathcal{F}(\mathbf{W}_t) - \nabla\mathcal{F}(\mathbf{W}_{t-1}))\mathbf{V}_t^\top\big\|_F + \chi_t \\ &\le \|\nabla\mathcal{F}(\mathbf{W}_t) - \nabla\mathcal{F}(\mathbf{W}_{t-1})\|_* + \chi_t \\ &\le L\|\mathbf{W}_t - \mathbf{W}_{t-1}\|_{\mathrm{op}} + \chi_t. \end{aligned}\tag{84}$$

From Equation 78, Lemma 6, and Equation 42,

$$\|\mathbf{W}_t - \mathbf{W}_{t-1}\|_{\mathrm{op}} = \eta\|\mathbf{U}_{t-1}\mathbf{O}_{t-1}\mathbf{V}_{t-1}\|_{\mathrm{op}} \le \eta(1+\varepsilon_q). \tag{85}$$

Hence

$$\|\mathbf{e}_t\|_F \le \beta\|\mathbf{e}_{t-1}\|_F + \beta\big(L\eta(1+\varepsilon_q) + \chi_t\big). \tag{86}$$

Because $\overline{\mathbf{N}}_0 = (1-\beta)\mathbf{G}_0$*,*

$$\|\mathbf{e}_0\|_F = \beta\|\mathbf{G}_0\|_F \le \beta\|\nabla\mathcal{F}(\mathbf{W}_0)\|_F \le \beta\sqrt{2LH}. \tag{87}$$

Unrolling Equation 86 gives

$$\mathbb{E}\|\mathbf{e}_t\|_F \le \beta^{t+1}\sqrt{2LH} + \sum_{i=1}^{t}\beta^{t-i+1}\Big(L\eta(1+\varepsilon_q) + \mathbb{E}[\chi_i]\Big). \tag{88}$$

For the second term, let $\mathbf{z}_t := \mathbf{N}_t - \overline{\mathbf{N}}_t$*. Subtracting Equation 37 from Equation 36 gives*

$$\mathbf{z}_t = \beta\mathcal{T}_t(\mathbf{z}_{t-1}) + (1-\beta)(\widehat{\mathbf{G}}_t - \mathbf{G}_t). \tag{89}$$

Repeated use of Equation 35, Jensen's inequality, and Equation 33 gives

$$\mathbb{E}\|\mathbf{z}_t\|_F \le (1-\beta)\sum_{i=0}^{t}\beta^{t-i}\mathbb{E}\|\widehat{\mathbf{G}}_i - \mathbf{G}_i\|_F \le (1-\beta)\sum_{i=0}^{t}\beta^{t-i}\sqrt{\Theta} = (1-\beta^{t+1})\sqrt{\Theta}. \tag{90}$$

Combining the last inequality with Equation 88 proves Equation 79.

Finally, average over $t = 0, \dots, T-1$*. The initial term is bounded by*

$$\frac{1}{T}\sum_{t=0}^{T-1}\beta^{t+1}\sqrt{2LH} \le \frac{\beta}{T(1-\beta)}\sqrt{2LH}. \tag{91}$$

For the constant drift term,

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{t}\beta^{t-i+1}L\eta(1+\varepsilon_q) \le \frac{\beta}{1-\beta}L\eta(1+\varepsilon_q). \tag{92}$$

For the refresh term, exchanging the order of summation gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{t}\beta^{t-i+1}\mathbb{E}[\chi_i] = \frac{1}{T}\sum_{i=1}^{T-1}\mathbb{E}[\chi_i]\sum_{t=i}^{T-1}\beta^{t-i+1} \le \frac{\beta}{1-\beta}\Gamma_{1,T}. \tag{93}$$

The estimation term is at most $\sqrt{\Theta}$*. This proves Equation 80.*

Lemma 8 (One-step progress for DAZM) *The DAZM update satisfies*

$$\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \le \eta(\varepsilon_{t,q} - 1)\|\mathbf{G}_t\|_* + 2\eta\sqrt{r}\|\mathbf{G}_t - \mathbf{N}_t\|_F + \frac{L\eta^2}{2}(1 + \varepsilon_{t,q})^2. \tag{94}$$

Proof 8 *By Lemma 1,*

$$\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \le -\eta\langle \nabla\mathcal{F}(\mathbf{W}_t), \mathbf{U}_t\mathbf{O}_t\mathbf{V}_t\rangle_F + \frac{L\eta^2}{2}\|\mathbf{O}_t\|_{\mathrm{op}}^2. \tag{95}$$

Insert $\mathrm{Polar}(\mathbf{G}_t)$ *and* $\mathrm{Polar}(\mathbf{M}_t)$*:*

$$\begin{aligned} &\eta\langle \mathbf{G}_t, \mathbf{O}_t\rangle_F \\ &= -\eta\langle \mathbf{G}_t, \mathrm{Polar}(\mathbf{G}_t)\rangle_F + \eta\langle \mathbf{G}_t, \mathrm{Polar}(\mathbf{G}_t) - \mathrm{Polar}(\mathbf{M}_t)\rangle_F + \eta\langle \mathbf{G}_t, \mathrm{Polar}(\mathbf{M}_t) - \mathbf{O}_t\rangle_F. \end{aligned} \tag{96}$$

Lemma 4 gives the first term as $-\eta\|\mathbf{G}_t\|_*$*. Nuclear/operator duality and Lemma 6 yield*

$$\left|\langle \mathbf{G}_t, \mathrm{Polar}(\mathbf{M}_t) - \mathbf{O}_t\rangle_F\right| \le \varepsilon_{t,q}\|\mathbf{G}_t\|_*. \tag{97}$$

Lemma 5 bounds the middle term by $2\sqrt{r}\|\mathbf{G}_t - \mathbf{N}_t\|_F$*, while Lemma 6 gives* $\|\mathbf{O}_t\|_{\mathrm{op}} \le 1 + \varepsilon_{t,q}$*. Combining the bounds proves Equation 94.*

Lemma 9 (One-step progress for two-sided ZO-SGD with momentum) *Let*

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \alpha\mathbf{U}_t\mathbf{N}_t\mathbf{V}_t, \qquad \alpha := \frac{\eta}{1-\beta}. \tag{98}$$

Then

$$\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \le -\frac{\alpha}{2}\|\mathbf{G}_t\|_F^2 + \frac{\alpha}{2}\|\mathbf{G}_t - \mathbf{N}_t\|_F^2 + \frac{(L\alpha - 1)\alpha}{2}\|\mathbf{N}_t\|_F^2. \tag{99}$$

Proof 9 *Lemma 1 gives*

$$\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \le -\alpha\langle \mathbf{G}_t, \mathbf{N}_t\rangle_F + \frac{L\alpha^2}{2}\|\mathbf{U}_t\mathbf{N}_t\mathbf{V}_t\|_{\mathrm{op}}^2 \le -\alpha\langle \mathbf{G}_t, \mathbf{N}_t\rangle_F + \frac{L\alpha^2}{2}\|\mathbf{N}_t\|_F^2. \tag{100}$$

By the polarization identity,

$$\langle \mathbf{G}_t, \mathbf{N}_t\rangle_F = \frac{1}{2}\left(\|\mathbf{G}_t\|_F^2 + \|\mathbf{N}_t\|_F^2 - \|\mathbf{G}_t - \mathbf{N}_t\|_F^2\right). \tag{101}$$

Substitution gives Equation 99.

Lemma 10 (Exact reduced representation of accumulated displacement) *Consider a refresh interval* $\mathcal{I}_k = \{\tau_k, \dots, \tau_{k+1} - 1\}$ *during which* $\mathbf{U}_k, \mathbf{V}_k$ *are fixed. Let*

$$\mathbf{\Delta}_k := \mathbf{W}_{\tau_k} - \mathbf{W}_{\tau_{k+1}} = \sum_{t\in\mathcal{I}_k} \eta_t\mathbf{U}_k\mathbf{O}_t\mathbf{V}_k = \mathbf{U}_k\mathbf{\Psi}_k\mathbf{V}_k, \quad \mathbf{\Psi}_k := \sum_{t\in\mathcal{I}_k} \eta_t\mathbf{O}_t. \tag{102}$$

If $\mathbf{\Psi}_k = \mathbf{A}_k\mathbf{\Lambda}_k\mathbf{B}_k^\top$ *is an SVD, then*

$$\mathbf{\Delta}_k = (\mathbf{U}_k\mathbf{A}_k)\mathbf{\Lambda}_k(\mathbf{B}_k^\top\mathbf{V}_k) \tag{103}$$

is a thin SVD of $\mathbf{\Delta}_k$*. Hence, the leading* p *left and right singular bases of* $\mathbf{\Delta}_k$ *are exactly*

$$\mathbf{U}_{k+1}^{\mathrm{exploit}} = \mathbf{U}_k\mathbf{A}_{k,1:p}, \qquad \mathbf{R}_{k+1}^{\mathrm{exploit}} := \left(\mathbf{V}_{k+1}^{\mathrm{exploit}}\right)^\top = \mathbf{V}_k^\top\mathbf{B}_{k,1:p}.$$

Proof 10 *For every* $t \in \mathcal{I}_k$*, the two-sided Muon-style update is*

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t\mathbf{U}_k\mathbf{O}_t\mathbf{V}_k, \tag{104}$$

where $\mathbf{U}_k$ *and* $\mathbf{V}_k$ *remain fixed throughout the refresh interval* $\mathcal{I}_k$*. Hence,*

$$\mathbf{W}_t - \mathbf{W}_{t+1} = \eta_t\mathbf{U}_k\mathbf{O}_t\mathbf{V}_k. \tag{105}$$

Summing over $t \in \mathcal{I}_k$ and using the telescoping identity gives

$$\boldsymbol{\Delta}_k := \mathbf{W}_{\tau_k} - \mathbf{W}_{\tau_{k+1}} = \sum_{t \in \mathcal{I}_k} (\mathbf{W}_t - \mathbf{W}_{t+1}) = \sum_{t \in \mathcal{I}_k} \eta_t \mathbf{U}_k \mathbf{O}_t \mathbf{V}_k = \mathbf{U}_k \boldsymbol{\Psi}_k \mathbf{V}_k, \tag{106}$$

where $\boldsymbol{\Psi}_k := \sum_{t \in \mathcal{I}_k} \eta_t \mathbf{O}_t$.

Let $\boldsymbol{\Psi}_k = \mathbf{A}_k \boldsymbol{\Lambda}_k \mathbf{B}_k^\top$ be an SVD of $\boldsymbol{\Psi}_k$, where $\mathbf{A}_k^\top \mathbf{A}_k = \mathbf{I}_r$, $\mathbf{B}_k^\top \mathbf{B}_k = \mathbf{I}_r$, and $\boldsymbol{\Lambda}_k$ contains the singular values of $\boldsymbol{\Psi}_k$ in nonincreasing order. Substituting this decomposition into the accumulated displacement gives

$$\boldsymbol{\Delta}_k = \mathbf{U}_k \mathbf{A}_k \boldsymbol{\Lambda}_k \mathbf{B}_k^\top \mathbf{V}_k = (\mathbf{U}_k \mathbf{A}_k) \boldsymbol{\Lambda}_k (\mathbf{B}_k^\top \mathbf{V}_k). \tag{107}$$

We next verify that the two outer factors form orthonormal singular bases. Since $\mathbf{U}_k^\top \mathbf{U}_k = \mathbf{I}_r$,

$$(\mathbf{U}_k \mathbf{A}_k)^\top (\mathbf{U}_k \mathbf{A}_k) = \mathbf{A}_k^\top \mathbf{U}_k^\top \mathbf{U}_k \mathbf{A}_k = \mathbf{A}_k^\top \mathbf{A}_k = \mathbf{I}_r. \tag{108}$$

Thus, $\mathbf{U}_k \mathbf{A}_k$ has orthonormal columns.

Similarly, using $\mathbf{V}_k \mathbf{V}_k^\top = \mathbf{I}_r$, we can get:

$$(\mathbf{B}_k^\top \mathbf{V}_k)(\mathbf{B}_k^\top \mathbf{V}_k)^\top = \mathbf{B}_k^\top \mathbf{V}_k \mathbf{V}_k^\top \mathbf{B}_k = \mathbf{B}_k^\top \mathbf{B}_k = \mathbf{I}_r. \tag{109}$$

Hence, $\mathbf{B}_k^\top \mathbf{V}_k$ has orthonormal rows. Equivalently, defining $\mathbf{R}_k := \mathbf{V}_k^\top \mathbf{B}_k$, we have

$$\mathbf{R}_k^\top \mathbf{R}_k = \mathbf{I}_r. \tag{110}$$

Therefore,

$$\boldsymbol{\Delta}_k = (\mathbf{U}_k \mathbf{A}_k) \boldsymbol{\Lambda}_k \mathbf{R}_k^\top \tag{111}$$

is a reduced SVD of $\boldsymbol{\Delta}_k$. In particular, $\boldsymbol{\Delta}_k$ and $\boldsymbol{\Psi}_k$ have the same nonzero singular values.

Since the singular values in $\boldsymbol{\Lambda}_k$ are ordered nonincreasingly, the leading p left singular vectors of $\boldsymbol{\Delta}_k$ are

$$\widehat{\mathbf{U}}_k = \mathbf{U}_k \mathbf{A}_{k,1:p}, \tag{112}$$

while the leading p right singular vectors are

$$\widehat{\mathbf{R}}_k = \mathbf{V}_k^\top \mathbf{B}_{k,1:p}. \tag{113}$$

Equivalently, if the right basis is stored in row form,

$$\widehat{\mathbf{V}}_k = \mathbf{B}_{k,1:p}^\top \mathbf{V}_k, \tag{114}$$

then

$$\widehat{\mathbf{R}}_k = (\widehat{\mathbf{V}}_k)^\top = \mathbf{V}_k^\top \mathbf{B}_{k,1:p}. \tag{115}$$

Lemma 11 (Expected capture ratio of a random two-sided subspace) *Let $\mathbf{H} \in \mathbb{R}^{m \times n}$ be fixed, and let $\mathbf{U}_{\mathrm{rnd}} \in \mathbb{R}^{m \times r}$ and $\mathbf{R}_{\mathrm{rnd}} \in \mathbb{R}^{n \times r}$ be independent Haar-random orthonormal bases sampled independently of $\mathbf{H}$. Then*

$$\mathbb{E}\left[\frac{\|\mathbf{U}_{\mathrm{rnd}}^\top \mathbf{H} \mathbf{R}_{\mathrm{rnd}}\|_*}{\|\mathbf{H}\|_*} \right] \le \frac{r}{\sqrt{mn}}. \tag{116}$$

Proof 11 *Let $\mathbf{H} = \sum_{j=1}^{q} \sigma_j \mathbf{a}_j \mathbf{b}_j^\top$ be a compact singular value decomposition of $\mathbf{H}$, where $q = \mathrm{rank}(\mathbf{H})$ and*

$$\|\mathbf{a}_j\|_2 = \|\mathbf{b}_j\|_2 = 1. \tag{117}$$

Then

$$\mathbf{U}_{\mathrm{rnd}}^\top \mathbf{H} \mathbf{R}_{\mathrm{rnd}} = \sum_{j=1}^{q} \sigma_j \mathbf{U}_{\mathrm{rnd}}^\top \mathbf{a}_j \mathbf{b}_j^\top \mathbf{R}_{\mathrm{rnd}} = \sum_{j=1}^{q} \sigma_j \left(\mathbf{U}_{\mathrm{rnd}}^\top \mathbf{a}_j \right) \left(\mathbf{R}_{\mathrm{rnd}}^\top \mathbf{b}_j \right)^\top. \tag{118}$$

Using the triangle inequality for the nuclear norm, we obtain

$$\left\|\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{H}\mathbf{R}_{\mathrm{rnd}}\right\|_{*} \leq \sum_{j=1}^{q} \sigma_j \left\| \left(\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{a}_j\right)\left(\mathbf{R}_{\mathrm{rnd}}^{\top}\mathbf{b}_j\right)^{\top} \right\|_{*}. \tag{119}$$

For any two vectors $\mathbf{x}$ *and* $\mathbf{y}$*, the rank-one matrix* $\mathbf{x}\mathbf{y}^{\top}$ *has exactly one nonzero singular value, equal to* $\|\mathbf{x}\|_2\|\mathbf{y}\|_2$*. Hence,*

$$\|\mathbf{x}\mathbf{y}^{\top}\|_{*} = \|\mathbf{x}\|_2\|\mathbf{y}\|_2. \tag{120}$$

Applying this identity to (119) *gives*

$$\left\|\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{H}\mathbf{R}_{\mathrm{rnd}}\right\|_{*} \leq \sum_{j=1}^{q} \sigma_j \left\|\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{a}_j\right\|_2 \left\|\mathbf{R}_{\mathrm{rnd}}^{\top}\mathbf{b}_j\right\|_2. \tag{121}$$

We next bound the expected overlap of a fixed unit vector with a Haar-uniform random subspace. Since $\mathbf{U}_{\mathrm{rnd}}$ *is Haar-uniform on the Stiefel manifold, rotational invariance implies*

$$\mathbb{E}\left[\mathbf{U}_{\mathrm{rnd}}\mathbf{U}_{\mathrm{rnd}}^{\top}\right] = c_U\mathbf{I}_m \tag{122}$$

for some scalar c_U*. Taking traces on both sides yields*

$$r = \mathbb{E}\left[\mathrm{tr}\left(\mathbf{U}_{\mathrm{rnd}}\mathbf{U}_{\mathrm{rnd}}^{\top}\right)\right] = c_U m, \tag{123}$$

and therefore

$$c_U = \frac{r}{m}. \tag{124}$$

Hence,

$$\mathbb{E}\left[\mathbf{U}_{\mathrm{rnd}}\mathbf{U}_{\mathrm{rnd}}^{\top}\right] = \frac{r}{m}\mathbf{I}_m. \tag{125}$$

For each left singular vector $\mathbf{a}_j$*,*

$$\mathbb{E}\left[\left\|\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{a}_j\right\|_2^2\right] = \mathbb{E}\left[\mathbf{a}_j^{\top}\mathbf{U}_{\mathrm{rnd}}\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{a}_j\right] = \mathbf{a}_j^{\top}\mathbb{E}\left[\mathbf{U}_{\mathrm{rnd}}\mathbf{U}_{\mathrm{rnd}}^{\top}\right]\mathbf{a}_j = \frac{r}{m}\|\mathbf{a}_j\|_2^2 = \frac{r}{m}. \tag{126}$$

By Jensen's inequality,

$$\mathbb{E}\left[\left\|\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{a}_j\right\|_2\right] \leq \sqrt{\mathbb{E}\left[\left\|\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{a}_j\right\|_2^2\right]} = \sqrt{\frac{r}{m}}. \tag{127}$$

Similarly, Haar isotropy of the right random basis gives

$$\mathbb{E}\left[\mathbf{R}_{\mathrm{rnd}}\mathbf{R}_{\mathrm{rnd}}^{\top}\right] = \frac{r}{n}\mathbf{I}_n, \tag{128}$$

and therefore

$$\mathbb{E}\left[\left\|\mathbf{R}_{\mathrm{rnd}}^{\top}\mathbf{b}_j\right\|_2\right] \leq \sqrt{\frac{r}{n}}. \tag{129}$$

Since $\mathbf{U}_{\mathrm{rnd}}$ *and* $\mathbf{R}_{\mathrm{rnd}}$ *are sampled independently, the two projected lengths are independent. Consequently,*

$$\mathbb{E}\left[\left\|\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{a}_j\right\|_2\left\|\mathbf{R}_{\mathrm{rnd}}^{\top}\mathbf{b}_j\right\|_2\right] = \mathbb{E}\left[\left\|\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{a}_j\right\|_2\right]\mathbb{E}\left[\left\|\mathbf{R}_{\mathrm{rnd}}^{\top}\mathbf{b}_j\right\|_2\right] \leq \sqrt{\frac{r}{m}}\sqrt{\frac{r}{n}} = \frac{r}{\sqrt{mn}}. \tag{130}$$

Taking expectations in (121) *and using* (130)*, we obtain*

$$\mathbb{E}\left[\left\|\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{H}\mathbf{R}_{\mathrm{rnd}}\right\|_{*}\right] \leq \sum_{j=1}^{q}\sigma_j\mathbb{E}\left[\left\|\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{a}_j\right\|_2\left\|\mathbf{R}_{\mathrm{rnd}}^{\top}\mathbf{b}_j\right\|_2\right] \leq \frac{r}{\sqrt{mn}}\sum_{j=1}^{q}\sigma_j. \tag{131}$$

Finally, since

$$\sum_{j=1}^{q}\sigma_j = \|\mathbf{H}\|_{*}, \tag{132}$$

we conclude that

$$\mathbb{E}\left[\left\|\mathbf{U}_{\mathrm{rnd}}^{\top}\mathbf{H}\mathbf{R}_{\mathrm{rnd}}\right\|_{*}\right] \leq \frac{r}{\sqrt{mn}}\|\mathbf{H}\|_{*}. \tag{133}$$

By dividing both sides by $\|\mathbf{H}\|_{*}$*, we can obtain the result.*

D CONVERGENCE THEOREMS

D.1 ZO-MUON

Theorem 1 (Convergence of ZO-Muon) Under Assumptions 1–3 and the reduced estimator condition Equation 33, let $\mathbf{G}_t$ be defined by Equation 26, and let $\mathbf{N}_t = (1-\beta)\mathbf{M}_t$ follow the transported momentum recursion Equation 36. Suppose

$$\varepsilon_q := \sup_t \varepsilon_{t,q} < 1. \tag{134}$$

Then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{G}_t\|_* \le \frac{1}{1-\varepsilon_q}\Bigg[\frac{H}{\eta T} + \frac{2\beta\sqrt{2rLH}}{T(1-\beta)} + \frac{2\beta\sqrt{r}}{1-\beta}L\eta(1+\varepsilon_q) + \frac{2\beta\sqrt{r}}{1-\beta}\Gamma_{1,T} + 2\sqrt{r\Theta} + \frac{L\eta}{2}(1+\varepsilon_q)^2\Bigg]. \tag{135}$$

In particular, choosing

$$\eta = \sqrt{\frac{H}{LT(1+\varepsilon_q)^2}}, \qquad \beta = 1 - \min\{T^{-1/4}, 1\}, \tag{136}$$

gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{G}_t\|_* = \mathcal{O}\Bigg(\frac{1}{1-\varepsilon_q}\Big[(1+\varepsilon_q)\sqrt{LH}\,T^{-1/2} + \sqrt{rLH}\,T^{-3/4} + \sqrt{rLH}\,T^{-1/4} + \sqrt{r\Theta} + \sqrt{r}\,T^{1/4}\Gamma_{1,T}\Big]\Bigg). \tag{137}$$

Proof 12 *From Lemma 8, the one-step progress of the DAZM update satisfies*

$$\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \le \eta(\varepsilon_{t,q}-1)\|\mathbf{G}_t\|_* + 2\eta\sqrt{r}\|\mathbf{G}_t - \mathbf{N}_t\|_F + \frac{L\eta^2}{2}(1+\varepsilon_{t,q})^2. \tag{138}$$

Since $\varepsilon_{t,q} \le \varepsilon_q := \sup_t \varepsilon_{t,q} < 1$, we have

$$\varepsilon_{t,q} - 1 \le -(1-\varepsilon_q), \tag{139}$$

$$(1+\varepsilon_{t,q})^2 \le (1+\varepsilon_q)^2. \tag{140}$$

Therefore,

$$\mathcal{F}(\mathbf{W}_{t+1}) - \mathcal{F}(\mathbf{W}_t) \le -\eta(1-\varepsilon_q)\|\mathbf{G}_t\|_* + 2\eta\sqrt{r}\|\mathbf{G}_t - \mathbf{N}_t\|_F + \frac{L\eta^2}{2}(1+\varepsilon_q)^2. \tag{141}$$

Taking expectations on both sides gives

$$\mathbb{E}[\mathcal{F}(\mathbf{W}_{t+1})] - \mathbb{E}[\mathcal{F}(\mathbf{W}_t)] \le -\eta(1-\varepsilon_q)\mathbb{E}\|\mathbf{G}_t\|_* + 2\eta\sqrt{r}\mathbb{E}\|\mathbf{G}_t - \mathbf{N}_t\|_F + \frac{L\eta^2}{2}(1+\varepsilon_q)^2. \tag{142}$$

Rearranging yields

$$\eta(1-\varepsilon_q)\mathbb{E}\|\mathbf{G}_t\|_* \le \mathbb{E}[\mathcal{F}(\mathbf{W}_t)] - \mathbb{E}[\mathcal{F}(\mathbf{W}_{t+1})] + 2\eta\sqrt{r}\mathbb{E}\|\mathbf{G}_t - \mathbf{N}_t\|_F + \frac{L\eta^2}{2}(1+\varepsilon_q)^2. \tag{143}$$

Summing over $t = 0, \dots, T-1$ and using the telescoping identity gives

$$\eta(1-\varepsilon_q)\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{G}_t\|_* \le \mathcal{F}(\mathbf{W}_0) - \mathbb{E}[\mathcal{F}(\mathbf{W}_T)] + 2\eta\sqrt{r}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{G}_t - \mathbf{N}_t\|_F + \frac{L\eta^2 T}{2}(1+\varepsilon_q)^2. \tag{144}$$

Since $\mathcal{F}(\mathbf{W}_T) \geq \mathcal{F}(\mathbf{W}^\star)$ and Assumption 3 gives

$$\mathcal{F}(\mathbf{W}_0) - \mathcal{F}(\mathbf{W}^\star) \leq H, \tag{145}$$

we obtain

$$\eta(1-\varepsilon_q) \sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{G}_t\|_* \leq H + 2\eta\sqrt{r} \sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{G}_t - \mathbf{N}_t\|_F + \frac{L\eta^2 T}{2}(1+\varepsilon_q)^2. \tag{146}$$

Dividing both sides by $\eta T(1-\varepsilon_q)$ gives

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{G}_t\|_* \leq \frac{1}{1-\varepsilon_q}\left[\frac{H}{\eta T} + \frac{2\sqrt{r}}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{G}_t - \mathbf{N}_t\|_F + \frac{L\eta}{2}(1+\varepsilon_q)^2\right]. \tag{147}$$

By Lemma 7,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{G}_t - \mathbf{N}_t\|_F \leq \frac{\beta\sqrt{2LH}}{T(1-\beta)} + \frac{\beta}{1-\beta} L\eta(1+\varepsilon_q) + \frac{\beta}{1-\beta}\Gamma_{1,T} + \sqrt{\Theta}. \tag{148}$$

Multiplying (148) *by $2\sqrt{r}$ yields*

$$\frac{2\sqrt{r}}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{G}_t - \mathbf{N}_t\|_F \leq \frac{2\beta\sqrt{2rLH}}{T(1-\beta)} + \frac{2\beta\sqrt{r}}{1-\beta} L\eta(1+\varepsilon_q) + \frac{2\beta\sqrt{r}}{1-\beta}\Gamma_{1,T} + 2\sqrt{r\Theta}. \tag{149}$$

Substituting this inequality into (147) *gives*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{G}_t\|_* \leq \frac{1}{1-\varepsilon_q}\Bigg[&\frac{H}{\eta T} + \frac{2\beta\sqrt{2rLH}}{T(1-\beta)} + \frac{2\beta\sqrt{r}}{1-\beta} L\eta(1+\varepsilon_q) \\ &+ \frac{2\beta\sqrt{r}}{1-\beta}\Gamma_{1,T} + 2\sqrt{r\Theta} + \frac{L\eta}{2}(1+\varepsilon_q)^2\Bigg],\end{aligned} \tag{150}$$

which proves Equation 135.

We next consider the parameter schedule

$$\eta = \sqrt{\frac{H}{LT(1+\varepsilon_q)^2}}, \qquad \beta = 1 - \min\{T^{-1/4}, 1\}. \tag{151}$$

For the asymptotic regime $T > 1$,

$$1 - \beta = T^{-1/4}. \tag{152}$$

Moreover,

$$\frac{H}{\eta T} = (1+\varepsilon_q)\sqrt{LH}\, T^{-1/2}, \tag{153}$$

$$\frac{1}{T(1-\beta)} = T^{-3/4}, \tag{154}$$

$$L\eta(1+\varepsilon_q) = \sqrt{LH}\, T^{-1/2}, \tag{155}$$

$$\frac{\eta}{1-\beta} = \frac{\sqrt{H/L}}{1+\varepsilon_q} T^{-1/4}, \tag{156}$$

$$\frac{\Gamma_{1,T}}{1-\beta} = T^{1/4}\Gamma_{1,T}, \tag{157}$$

$$L\eta(1+\varepsilon_q)^2 = (1+\varepsilon_q)\sqrt{LH}\, T^{-1/2}. \tag{158}$$

Consequently,

$$\frac{2\beta\sqrt{2rLH}}{T(1-\beta)} = \mathcal{O}\left(\sqrt{rLH}\,T^{-3/4}\right), \tag{159}$$

$$\frac{2\beta\sqrt{r}}{1-\beta}L\eta(1+\varepsilon_q) = \mathcal{O}\left(\sqrt{rLH}\,T^{-1/4}\right), \tag{160}$$

$$\frac{2\beta\sqrt{r}}{1-\beta}\Gamma_{1,T} = \mathcal{O}\left(\sqrt{r}\,T^{1/4}\Gamma_{1,T}\right). \tag{161}$$

Collecting all terms gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{G}_t\|_* = \mathcal{O}\Bigg(\frac{1}{1-\varepsilon_q}\left[(1+\varepsilon_q)\sqrt{LH}\,T^{-1/2} + \sqrt{rLH}\,T^{-3/4} + \sqrt{rLH}\,T^{-1/4} + \sqrt{r\Theta} + \sqrt{r}\,T^{1/4}\Gamma_{1,T}\right]\Bigg), \tag{162}$$

which proves Equation 137.

D.2 Displacement-Aware Subspace Advantage

Theorem 2 (Lower bound of displacement-aware gradient capture) Consider the DAZM subspace obtained after refresh interval $\mathcal{I}_k = \{\tau_k, \dots, \tau_{k+1} - 1\}$. Let $\mathbf{H}_{\tau_{k+1}} := \nabla F(\mathbf{W}_{\tau_{k+1}})$ denote the gradient at the beginning of the next refresh interval, with leading p left and right singular bases $\mathbf{A}_{p,k+1}$ and $\mathbf{B}_{p,k+1}$. Let $\mathbf{U}_{k+1}^{\text{exploit}}$ and $\mathbf{V}_{k+1}^{\text{exploit}}$ be the displacement-based exploitation bases. Let $\varepsilon_{\text{exploit},k} := \max\{\|\sin\Theta(\mathbf{U}_{k+1}^{\text{exploit}}, \mathbf{A}_{p,k+1})\|_{\text{op}}, \|\sin\Theta(\mathbf{R}_{k+1}^{\text{exploit}}, \mathbf{B}_{p,k+1})\|_{\text{op}}\} < 1$ measure the alignment error of the retained displacement modes, and define $\gamma_{p,k+1} := \sum_{j=1}^{p}\sigma_j(\mathbf{H}_{\tau_{k+1}})/\|\mathbf{H}_{\tau_{k+1}}\|_*$ as the fraction of nuclear-norm mass contained in the leading p singular modes. Then the capture ratio of the refreshed DAZM subspace satisfies

$$\rho_{\tau_{k+1}} := \frac{\left\|\mathbf{U}_{k+1}^\top\mathbf{H}_{\tau_{k+1}}\mathbf{V}_{k+1}^\top\right\|_*}{\|\mathbf{H}_{\tau_{k+1}}\|_*} \geq \gamma_{p,k+1} - \varepsilon_{\text{exploit},k}^2. \tag{163}$$

Proof 13 *For brevity, denote* $\mathbf{U}_e := \mathbf{U}_{k+1}^{\text{exploit}}$ *and* $\mathbf{R}_e := \mathbf{R}_{k+1}^{\text{exploit}} = (\mathbf{V}_{k+1}^{\text{exploit}})^\top$. *Partition the singular value decomposition of* $\mathbf{H}_{\tau_{k+1}}$ *into its leading* p *and remaining singular components as* $\mathbf{H}_{\tau_{k+1}} = \mathbf{A}_{p,k+1}\mathbf{\Sigma}_{p,k+1}\mathbf{B}_{p,k+1}^\top + \mathbf{A}_{\perp,k+1}\mathbf{\Sigma}_{\perp,k+1}\mathbf{B}_{\perp,k+1}^\top$.

Projecting this decomposition onto the exploitation subspace and using the reverse triangle inequality for the nuclear norm gives

$$\|\mathbf{U}_e^\top\mathbf{H}_{\tau_{k+1}}\mathbf{R}_e\|_* \geq \|\mathbf{U}_e^\top\mathbf{A}_{p,k+1}\mathbf{\Sigma}_{p,k+1}\mathbf{B}_{p,k+1}^\top\mathbf{R}_e\|_* - \|\mathbf{U}_e^\top\mathbf{A}_{\perp,k+1}\mathbf{\Sigma}_{\perp,k+1}\mathbf{B}_{\perp,k+1}^\top\mathbf{R}_e\|_*. \tag{164}$$

We first lower bound the contribution from the leading p *singular modes. By the definition of* $\varepsilon_{\text{exploit},k}$*, the largest principal-angle sine between* $\mathbf{U}_e$ *and* $\mathbf{A}_{p,k+1}$*, as well as that between* $\mathbf{R}_e$ *and* $\mathbf{B}_{p,k+1}$*, is at most* $\varepsilon_{\text{exploit},k}$*. Therefore, the smallest singular values of* $\mathbf{U}_e^\top\mathbf{A}_{p,k+1}$ *and* $\mathbf{B}_{p,k+1}^\top\mathbf{R}_e$ *satisfy* $\sigma_{\min}(\mathbf{U}_e^\top\mathbf{A}_{p,k+1}) \geq \sqrt{1-\varepsilon_{\text{exploit},k}^2}$ *and* $\sigma_{\min}(\mathbf{B}_{p,k+1}^\top\mathbf{R}_e) \geq \sqrt{1-\varepsilon_{\text{exploit},k}^2}$*. Using* $\|\mathbf{AXB}\|_* \geq \sigma_{\min}(\mathbf{A})\sigma_{\min}(\mathbf{B})\|\mathbf{X}\|_*$*, we obtain*

$$\|\mathbf{U}_e^\top\mathbf{A}_{p,k+1}\mathbf{\Sigma}_{p,k+1}\mathbf{B}_{p,k+1}^\top\mathbf{R}_e\|_* \geq (1-\varepsilon_{\text{exploit},k}^2)\|\mathbf{\Sigma}_{p,k+1}\|_* = (1-\varepsilon_{\text{exploit},k}^2)\sum_{j=1}^{p}\sigma_j(\mathbf{H}_{\tau_{k+1}}). \tag{165}$$

We next bound the contribution from the remaining singular modes. The principal-angle definition also implies $\|\mathbf{U}_e^\top\mathbf{A}_{\perp,k+1}\|_{\text{op}} \leq \varepsilon_{\text{exploit},k}$ *and* $\|\mathbf{B}_{\perp,k+1}^\top\mathbf{R}_e\|_{\text{op}} \leq \varepsilon_{\text{exploit},k}$*. Using the nuclear-norm inequality* $\|\mathbf{AXB}\|_* \leq \|\mathbf{A}\|_{\text{op}}\|\mathbf{X}\|_*\|\mathbf{B}\|_{\text{op}}$*, we have*

$$\|\mathbf{U}_e^\top\mathbf{A}_{\perp,k+1}\mathbf{\Sigma}_{\perp,k+1}\mathbf{B}_{\perp,k+1}^\top\mathbf{R}_e\|_* \leq \varepsilon_{\text{exploit},k}^2\|\mathbf{\Sigma}_{\perp,k+1}\|_* = \varepsilon_{\text{exploit},k}^2\sum_{j>p}\sigma_j(\mathbf{H}_{\tau_{k+1}}). \tag{166}$$

Substituting (165) *and* (166) *into* (164) *yields*

$$\|\mathbf{U}_e^\top \mathbf{H}_{\tau_{k+1}} \mathbf{R}_e\|_* \geq (1-\varepsilon_{\text{exploit},k}^2) \sum_{j=1}^{p} \sigma_j(\mathbf{H}_{\tau_{k+1}}) - \varepsilon_{\text{exploit},k}^2 \sum_{j>p} \sigma_j(\mathbf{H}_{\tau_{k+1}})$$

$$= \sum_{j=1}^{p} \sigma_j(\mathbf{H}_{\tau_{k+1}}) - \varepsilon_{\text{exploit},k}^2 \|\mathbf{H}_{\tau_{k+1}}\|_*. \tag{167}$$

It remains to relate the capture of the exploitation subspace to that of the full rank-r DAZM subspace. By construction, the first p directions of $\mathbf{U}_{k+1}$ and $\mathbf{V}_{k+1}$ are precisely the retained exploitation directions. Let $\mathbf{E}_p = [\mathbf{I}_p, \mathbf{0}]^\top \in \mathbb{R}^{r\times p}$. Then $\mathbf{U}_e = \mathbf{U}_{k+1}\mathbf{E}_p$ and $\mathbf{R}_e = \mathbf{V}_{k+1}^\top \mathbf{E}_p$, with $\mathbf{E}_p^\top \mathbf{E}_p = \mathbf{I}_p$ and $\|\mathbf{E}_p\|_{\text{op}} = 1$. Consequently,

$$\|\mathbf{U}_e^\top \mathbf{H}_{\tau_{k+1}} \mathbf{R}_e\|_* = \left\|\mathbf{E}_p^\top \left(\mathbf{U}_{k+1}^\top \mathbf{H}_{\tau_{k+1}} \mathbf{V}_{k+1}^\top\right) \mathbf{E}_p\right\|_* \leq \left\|\mathbf{U}_{k+1}^\top \mathbf{H}_{\tau_{k+1}} \mathbf{V}_{k+1}^\top\right\|_*, \tag{168}$$

where the inequality follows from the submultiplicativity of the nuclear norm with respect to the operator norm.

Combining (167) *and* (168)*, and dividing by $\|\mathbf{H}_{\tau_{k+1}}\|_*$, gives*

$$\rho_{\tau_{k+1}} = \frac{\left\|\mathbf{U}_{k+1}^\top \mathbf{H}_{\tau_{k+1}} \mathbf{V}_{k+1}^\top\right\|_*}{\|\mathbf{H}_{\tau_{k+1}}\|_*} \geq \frac{\sum_{j=1}^{p} \sigma_j(\mathbf{H}_{\tau_{k+1}})}{\|\mathbf{H}_{\tau_{k+1}}\|_*} - \varepsilon_{\text{exploit},k}^2 = \gamma_{p,k+1} - \varepsilon_{\text{exploit},k}^2. \tag{169}$$

D.3 ZO-SGD WITH MOMENTUM

The transported reduced momentum is again

$$\mathbf{N}_t = \beta \mathcal{T}_t(\mathbf{N}_{t-1}) + (1-\beta)\widehat{\mathbf{G}}_t, \tag{170}$$

and the update is

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \alpha \mathbf{U}_t \mathbf{N}_t \mathbf{V}_t, \qquad \alpha = \frac{\eta}{1-\beta}. \tag{171}$$

Theorem 3 (Convergence of ZO-SGD with momentum) *Under Assumptions 1–3 and Equation 33, fix any $\zeta > 0$. Suppose $\alpha \leq 1/L$ and*

$$\frac{4(1+\zeta)\beta^2 L^2 \alpha^2}{(1-\beta)^2} \leq \frac{1}{2}. \tag{172}$$

Then

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{G}_t\|_F^2 \leq \frac{4H}{\alpha T} + \frac{8LH}{T(1-\beta)} + \frac{8(1+\zeta)\beta^2 L^2 \alpha^2}{(1-\beta)^2}\Theta + 4\Theta + \frac{4(1+\zeta^{-1})\beta^2}{(1-\beta)^2}\Gamma_{2,T}. \tag{173}$$

Choosing the same schedule as in the original momentum-SGD proof,

$$\beta = 1 - \min\{T^{-1/4}, 1\}, \qquad \alpha = \min\left\{\frac{1-\beta}{2\sqrt{2(1+\zeta)}\,\beta L}, \sqrt{\frac{H}{LT(4LH+2\Theta)}}\right\}, \tag{174}$$

yields, up to constants depending only on ζ,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{G}_t\|_* = \mathcal{O}\left(\sqrt{r}\left[\sqrt{LH}\,T^{-1/4} + (LH\Theta)^{1/4}T^{-1/4} + \sqrt{LH}\,T^{-3/8} + \sqrt{\Theta}\right]\right). \tag{175}$$

Proof 14 *From Lemma 9, if $\alpha \leq 1/L$, then the final term in Equation 99 is non-positive. Hence*

$$\frac{\alpha}{2}\|\mathbf{G}_t\|_F^2 \leq \mathcal{F}(\mathbf{W}_t) - \mathcal{F}(\mathbf{W}_{t+1}) + \frac{\alpha}{2}\|\mathbf{G}_t - \mathbf{N}_t\|_F^2. \tag{176}$$

Taking expectations, summing, and using Assumption 3 gives

$$S_T := \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{G}_t\|_F^2 \leq \frac{2H}{\alpha T} + \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{G}_t - \mathbf{N}_t\|_F^2. \tag{177}$$

Introduce the ideal transported momentum $\overline{\mathbf{N}}_t$ from Equation 37. Then

$$\|\mathbf{G}_t - \mathbf{N}_t\|_F^2 \le 2\|\mathbf{G}_t - \overline{\mathbf{N}}_t\|_F^2 + 2\|\overline{\mathbf{N}}_t - \mathbf{N}_t\|_F^2. \tag{178}$$

For the ideal momentum lag, let $\mathbf{e}_t := \mathbf{G}_t - \overline{\mathbf{N}}_t$. As in Lemma 7,

$$\mathbf{e}_t = \beta\mathcal{T}_t(\mathbf{e}_{t-1}) + \beta\big(\mathbf{G}_t - \mathcal{T}_t(\mathbf{G}_{t-1})\big). \tag{179}$$

Using Young's inequality with $c = (1-\beta)/\beta$ and Equation 35,

$$\|\mathbf{e}_t\|_F^2 \le \beta\|\mathbf{e}_{t-1}\|_F^2 + \frac{\beta^2}{1-\beta}\|\mathbf{G}_t - \mathcal{T}_t(\mathbf{G}_{t-1})\|_F^2. \tag{180}$$

The projected-gradient drift satisfies

$$\|\mathbf{G}_t - \mathcal{T}_t(\mathbf{G}_{t-1})\|_F \le L\|\mathbf{W}_t - \mathbf{W}_{t-1}\|_{\mathrm{op}} + \chi_t \le L\alpha\|\mathbf{N}_{t-1}\|_F + \chi_t. \tag{181}$$

For any $\zeta > 0$,

$$(a+b)^2 \le (1+\zeta)a^2 + (1+\zeta^{-1})b^2, \tag{182}$$

so

$$\|\mathbf{G}_t - \mathcal{T}_t(\mathbf{G}_{t-1})\|_F^2 \le (1+\zeta)L^2\alpha^2\|\mathbf{N}_{t-1}\|_F^2 + (1+\zeta^{-1})\chi_t^2. \tag{183}$$

Summing Equation 180 recursively and using $\mathbb{E}\|\mathbf{e}_0\|_F^2 \le 2LH$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{e}_t\|_F^2 \le \frac{2LH}{T(1-\beta)} + \frac{(1+\zeta)\beta^2L^2\alpha^2}{(1-\beta)^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{N}_t\|_F^2 + \frac{(1+\zeta^{-1})\beta^2}{(1-\beta)^2}\Gamma_{2,T}. \tag{184}$$

For the estimation component $\mathbf{z}_t := \mathbf{N}_t - \overline{\mathbf{N}}_t$, Minkowski's inequality, transport contraction, and Equation 33 give

$$\big(\mathbb{E}\|\mathbf{z}_t\|_F^2\big)^{1/2} \le (1-\beta)\sum_{i=0}^{t}\beta^{t-i}\sqrt{\Theta} \le \sqrt{\Theta}, \tag{185}$$

therefore

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{z}_t\|_F^2 \le \Theta. \tag{186}$$

Combining Equation 178–Equation 186,

$$\frac{1}{T}\sum_{t}\mathbb{E}\|\mathbf{G}_t - \mathbf{N}_t\|_F^2 \le \frac{4LH}{T(1-\beta)} + \frac{2(1+\zeta)\beta^2L^2\alpha^2}{(1-\beta)^2}\frac{1}{T}\sum_{t}\mathbb{E}\|\mathbf{N}_t\|_F^2 + \frac{2(1+\zeta^{-1})\beta^2}{(1-\beta)^2}\Gamma_{2,T} + 2\Theta. \tag{187}$$

Next, from the transported momentum recursion and convexity of $\|\cdot\|_F^2$,

$$\mathbb{E}\|\mathbf{N}_t\|_F^2 \le (1-\beta)\sum_{i=0}^{t}\beta^{t-i}\mathbb{E}\|\widehat{\mathbf{G}}_i\|_F^2 \le 2(1-\beta)\sum_{i=0}^{t}\beta^{t-i}\left(\mathbb{E}\|\mathbf{G}_i\|_F^2 + \Theta\right). \tag{188}$$

Averaging over t and exchanging the order of summation yields

$$\frac{1}{T}\sum_{t}\mathbb{E}\|\mathbf{N}_t\|_F^2 \le 2S_T + 2\Theta. \tag{189}$$

Substitute Equation 189 into Equation 187, and then into Equation 177:

$$\left(1 - \frac{4(1+\zeta)\beta^2L^2\alpha^2}{(1-\beta)^2}\right)S_T \le \frac{2H}{\alpha T} + \frac{4LH}{T(1-\beta)} + \frac{4(1+\zeta)\beta^2L^2\alpha^2}{(1-\beta)^2}\Theta + 2\Theta + \frac{2(1+\zeta^{-1})\beta^2}{(1-\beta)^2}\Gamma_{2,T}. \tag{190}$$

Under Equation 172, the coefficient on the left is at least $1/2$. Multiplying the right-hand side by two proves Equation 173. Finally, $\|\mathbf{G}_t\|_* \le \sqrt{r}\|\mathbf{G}_t\|_F$ and Jensen's inequality give Equation 175.

Learning-rate comparison. The two convergence certificates retain the same qualitative separation as in the supplied proof. Under Equation 136, the ZO-Muon step size satisfies

$$\eta_{\mathrm{M}} = \mathcal{O}(T^{-1/2}). \tag{191}$$

For ZO-SGD with momentum, $\eta = (1-\beta)\alpha$ and $1-\beta = \Theta(T^{-1/4})$; with the second branch of Equation 174, $\alpha = \mathcal{O}(T^{-1/2})$, giving

$$\eta_{\mathrm{S}} = \mathcal{O}(T^{-3/4}). \tag{192}$$

Thus the smoothness-based convergence certificate permits a larger-order Muon step. This is a comparison of sufficient certified schedules.