SHARPNESS-AWARE MUON: MATRIX-AWARE OPTIMIZATION WITH MOMENTUM-GUIDED PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon is a matrix-aware optimizer that combines momentum with orthogonalization to produce structured update directions with strong empirical behavior. However, it only shapes the geometry of the update itself and does not explicitly control neighborhood worst-case loss or bias training toward flat regions, which are often linked to stronger generalization. To address this limitation, we introduce the sharpness-aware formulation of Muon. We first develop MUSAM, which incorporates Sharpness-Aware Minimization (SAM)-style perturbation into Muon's orthogonalized update and establishes a principled bridge between sharpness-aware training and matrix-aware optimization. We then show that this formulation faces an efficiency–stability tension: constructing the perturbation requires an extra gradient evaluation, while the perturbation itself is generated from a single minibatch gradient and is therefore noisy. To resolve this issue, we propose MOMUSAM, which turns Muon's momentum from an update variable into a sharpness-probing signal. By using normalized momentum as the perturbation direction, MOMUSAM removes the extra clean-gradient computation and replaces the single-batch perturbation proxy with a temporally aggregated one. We prove that both MUSAM and MOMUSAM attain the same asymptotic first-order stationarity order, and further show that MOMUSAM yields a more stable stochastic linearization than MUSAM. Experiments on standard benchmarks demonstrate that sharpness awareness consistently improves Muon.

1 INTRODUCTION

The generalization behavior of modern deep networks depends not only on how well they fit the training data, but also on the geometry of the optimization process that selects the final solution (Hochreiter & Schmidhuber, 1994; Keskar et al., 2016). This observation has increasingly shaped modern optimizer design along two complementary directions. One direction, exemplified by *Sharpness-Aware Minimization* (SAM), improves generalization by explicitly penalizing large neighborhood worst-case loss and thereby biasing training toward flatter regions of the landscape (Foret et al., 2020; Andriushchenko & Flammarion, 2022; Bartlett et al., 2023). The other reshapes the update rule itself through structured transformations of gradient information, exploiting momentum, geometry, or parameter structure to alter the inductive bias of optimization beyond plain first-order descent (Chen et al., 2023; Yuan et al., 2024; Vyas et al., 2024; Pethick et al., 2025). Muon is a particularly compelling instance of this second direction: by coupling momentum with orthogonalization, it produces matrix-aware update directions with distinctive geometry and strong empirical optimization behavior (Jordan et al., 2024).

Despite the success of both directions, however, their interaction remains poorly understood. Sharpness-aware methods primarily control the local loss geometry explored during training, whereas structured optimizers primarily control the geometry of the update used to explore it. This distinction is particularly salient in Muon, where momentum is explicitly orthogonalized into matrix-aware search directions. Such a design gives the optimizer strong control over *how* it moves through parameter space, while leaving unresolved whether those structured movements are guided toward locally flat, generalization-favorable regions. In this sense, Muon does not lack geometry; rather, it leaves unmodeled the part of geometry most directly tied to flatness, namely the local worst-case

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.

loss around the iterate. Existing sharpness-aware extensions of standard momentum methods suggest that internal optimizer states can be useful for constructing perturbations (Becker et al., 2024; Li et al., 2025), but this idea has not been systematically developed for matrix-aware orthogonalized optimization. Muon therefore provides a particularly revealing setting for studying whether update geometry and local loss geometry can be coordinated, which naturally raises the following question:

Can the internal state of a matrix-aware optimizer be made to govern not only how updates are taken, but also which local geometries are favored during training?

Our answer is affirmative, and it leads to a new optimizer family. We begin by formulating MuSAM, a sharpness-aware Muon method that couples SAM-style local worst-case perturbation with Muon's orthogonalized momentum update. This formulation plays a foundational role: it shows that explicit sharpness control is fully compatible with Muon's matrix-aware update geometry, and thereby establishes the correct starting point for sharpness-aware Muon. More importantly, it reveals the exact challenge that such an integration creates. The sharpness perturbation is expensive to construct, because it requires an additional gradient that the base Muon update does not otherwise need (Wen et al., 2018; Zhuang et al., 2022; Zhao et al., 2022). At the same time, it is built from a *single minibatch gradient*, and is therefore an inherently noisy and batch-local proxy for local sharpness. Prior work has shown that such stochastic linearization can be unstable and may fail to faithfully capture the broader local geometry that ultimately matters for generalization (Reddi et al., 2016; Wen et al., 2022; Kim et al., 2022; Li et al., 2025). In this way, MuSAM serves a dual purpose: it provides the first principled sharpness-aware formulation of Muon, and it exposes the precise efficiency–stability tension that the remainder of the paper resolves.

The key technical insight of this paper is that resolving this tension does not require introducing an external perturbation mechanism; instead, it requires identifying the *right* sharpness signal within Muon's own optimization dynamics. In particular, Muon's momentum forms a matrix-valued, temporally aggregated state that carries information across both iterations and minibatches, and is therefore typically smoother, less myopic, and more robust to batch-level noise than an instantaneous stochastic gradient (Li et al., 2021; Cutkosky & Orabona, 2019; Qian et al., 2025). This leads to a principled design criterion for sharpness-aware Muon: the perturbation direction should reuse the temporal information accumulated in Muon's momentum while also serving as a stable proxy for local ascent. Guided by this criterion, we design MoMuSAM, in which the normalized momentum generates the perturbation and the gradient is evaluated only once at the perturbed point. Rather than merely reusing an existing optimizer state, this design connects sharpness probing with Muon's temporal momentum dynamics, turning momentum from a pure update variable into a *history-aware geometric probe* of local sharpness.

This design has consequences at three levels. Algorithmically, MoMuSAM eliminates the extra clean-gradient evaluation required by MuSAM, bringing sharpness-aware Muon much closer to the computational profile of vanilla Muon. Statistically, it replaces a highly local single-batch perturbation with a temporally aggregated one, yielding a more stable slope proxy for stochastic linearization. Theoretically, this distinction is substantive rather than cosmetic: our analysis shows that MuSAM and MoMuSAM attain the same asymptotic first-order stationarity order, while MoMuSAM admits a sharper certified directional-stability bound in the stochastic regime. More broadly, our results reveal a new principle for optimizer design: internal optimizer states can encode reusable geometric information, and exploiting that information can lead to genuinely stronger training algorithms. The contributions of this paper are summarized as follows.

- We introduce, to the best of our knowledge, the first systematic sharpness-aware formulation of Muon. In particular, we establish MuSAM, which unifies Muon's matrix-aware orthogonalized updates with an explicit SAM-style approximation to local worst-case training and makes sharpness control an intrinsic part of the Muon framework.
- We uncover and resolve the fundamental efficiency–stability tension in sharpness-aware Muon. Our proposed MoMuSAM turns Muon's momentum from an update variable into a sharpness-probing signal, thereby reducing extra gradient cost and replacing the noisy single-batch perturbation with a more stable aggregated direction.
- We provide convergence guarantees for both MuSAM and MoMuSAM, showing that they attain the same $\mathcal{O}(T^{-1/4})$ average first-order stationarity order for the original objective at both the clean iterates and the corresponding perturbed query points. Beyond matching this convergence guarantee,

we prove that MoMuSAM improves the certified average directional-linearization stability from $\mathcal{O}(T^{-3/4})$ for MuSAM to $\mathcal{O}(T^{-1})$, while retaining only one gradient evaluation per iteration.

- Through experiments on standard benchmarks, we show that introducing sharpness awareness consistently strengthens Muon, while MoMuSAM delivers a markedly better generalization–efficiency tradeoff than MuSAM. Taken together, these results position sharpness-aware Muon as a new optimizer family that tightly integrates structured matrix updates, local geometry control, and computational efficiency.

2 Preliminaries

2.1 Problem Formulation

We consider stochastic nonconvex optimization over a matrix-valued parameter $W \in \mathbb{R}^{m\times n}$, with objective

$$\min_{W\in\mathbb{R}^{m\times n}} f(W) \;:=\; \mathbb{E}_{\xi}\big[\ell(W;\xi)\big], \tag{1}$$

where ξ denotes a random sample (or minibatch) drawn from the data distribution, and $\ell(W;\xi)$ is the corresponding stochastic loss. At iteration t, given a minibatch $\mathcal{B}_t$, we write the stochastic gradient as $\nabla\ell(W_t;\mathcal{B}_t)$. For simplicity of presentation, the analysis below is developed for a single matrix block W; in a neural network with multiple matrix-valued hidden-layer parameters, the same construction is applied blockwise.

Notation. For a matrix $X \in \mathbb{R}^{m\times n}$, $\|X\|_F$ denotes the Frobenius norm, $\|X\|_*$ denotes the nuclear norm, and $\|X\|_{\mathrm{op}}$ denotes the spectral (operator) norm. We use the standard inequality $\|X\|_* \le \sqrt{r}\,\|X\|_F$, where r is an upper bound on $\mathrm{rank}(X)$. Without loss of generality, we assume $m \ge n$, so the rank of the matrix is at most n. For $X, Y \in \mathbb{R}^{m\times n}$, we define the matrix inner product $\langle X, Y\rangle := \mathrm{Tr}(X^\top Y)$, where $\mathrm{Tr}(\cdot)$ denotes the trace. We write $\mathrm{Polar}(X)$ for the polar factor of X, and use $\mathrm{NS_Orth}(X)$ to denote its practical Newton–Schulz approximation. For notational simplicity, whenever the minibatch $\mathcal{B}_t$ is clear from context, we write $\nabla\ell(W_t)$ as shorthand for $\nabla\ell(W_t;\mathcal{B}_t)$. Whenever a normalized direction $V/\|V\|_F$ appears below, it is understood to be zero when $V = 0$. We reserve $\eta > 0$ for the learning rate, $\mu \in [0, 1)$ for the momentum coefficient, and $\rho > 0$ for the sharpness perturbation radius.

In this paper, the following standard assumptions from stochastic nonconvex optimization (Ghadimi & Lan, 2013; Bottou et al., 2018; Mi et al., 2022; Zhuang et al., 2022; Li et al., 2025; Kim & Oh, 2026) are made for the convergence analysis:

Assumption 2.1 (Lower bounded objective)**.** The objective f is bounded from below, i.e., there exists $f_\star > -\infty$ such that $f(W) \ge f_\star, \forall W \in \mathbb{R}^{m\times n}$.

Assumption 2.2 (Sample-wise Lipschitz smoothness)**.** For almost every sample ξ, the stochastic loss $\ell(\cdot;\xi)$ is continuously differentiable and satisfies

$$\|\nabla\ell(X;\xi) - \nabla\ell(Y;\xi)\|_* \le L\|X - Y\|_{\mathrm{op}}, \qquad \forall X, Y \in \mathbb{R}^{m\times n}.$$

By Jensen's inequality, Assumption 2.2 implies that the same smoothness bound holds almost surely for every sampled minibatch gradient and also for the population gradient.

Assumption 2.3 (Conditionally unbiased stochastic gradients)**.** Let $\mathcal{F}_{t-1}$ denote the history before sampling the minibatch at iteration t. The minibatch $\mathcal{B}_t$, consisting of B independent samples, is sampled independently of $\mathcal{F}_{t-1}$. For every $\mathcal{F}_{t-1}$-measurable query point X_t, we have

$$\mathbb{E}[\nabla\ell(X_t;\mathcal{B}_t) \mid \mathcal{F}_{t-1}] = \nabla f(X_t), \quad \mathbb{E}\Big[\|\nabla\ell(X_t;\mathcal{B}_t) - \nabla f(X_t)\|_F^2 \mid \mathcal{F}_{t-1}\Big] \le \frac{\sigma^2}{B}.$$

Assumption 2.2 is naturally expressed with respect to a norm and its dual norm; in the matrix setting, the operator–nuclear pair is especially compatible with orthogonalized or polar-type updates (Nesterov, 2013; Jaggi, 2013; Beck, 2017; Kim & Oh, 2026). Assumption 2.3 is equally standard and captures the stochastic noise introduced by minibatch sampling, under which the variance of a minibatch estimator decreases proportionally to the batch size (Ghadimi & Lan, 2013; Bottou et al., 2018). Under these assumptions, we use the standard notion of first-order stationarity.

Definition 2.4 (ε-stationary point)**.** A point $W \in \mathbb{R}^{m\times n}$ is called an ε-stationary point if $\|\nabla f(W)\|_* \le \varepsilon$. For an iterative algorithm producing $\{W_t\}_{t=0}^{T-1}$, we say that it attains ε-stationarity in T iterations if $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(W_t)\|_*] \le \varepsilon$.

2.2 Momentum Orthogonalization

Recent work on orthogonalized optimization, exemplified by Muon, advocates exploiting the intrinsic matrix structure of neural network parameters instead of flattening them into long vectors. The central idea is to first construct a matrix-valued momentum term and then orthogonalize it before taking an update step. This procedure preserves the dominant directional subspaces of the momentum while removing singular-value magnitudes, thereby yielding a structured search direction adapted to matrix parameters (Liu et al., 2025a; An et al., 2025; Liu et al., 2025b; Shah et al., 2025).

For exposition, we assume $m \geq n$; when $m < n$, the polar operation is applied to $M_t^\top$, and the resulting direction is transposed back. Denote $M_t \in \mathbb{R}^{m\times n}$ as the momentum matrix and $O_t \in \mathbb{R}^{m\times n}$ as the orthogonalized update direction. A generic Muon-style update takes the form

$$M_t \;=\; \mu M_{t-1} + (1-\mu)\nabla \ell(W_t), \qquad \mu \in [0,1), \tag{2}$$

followed by

$$O_t \;\in\; \arg\min_{O\in\mathbb{R}^{m\times n}} \|O - M_t\|_F \quad \text{s.t.} \quad O^\top O = I_n, \tag{3}$$

and the parameter update

$$W_{t+1} \;=\; W_t - \eta O_t. \tag{4}$$

Thus, Muon updates along the closest column-orthonormal matrix to the momentum.

If $M_t = U_t \Sigma_t V_t^\top$ is the thin singular value decomposition (SVD), then the solution to Eq. (3) is the polar factor $O_t = \mathrm{Polar}(M_t) = U_t V_t^\top$. In practice, this orthogonalization is not computed by exact SVD; Muon instead uses a few Newton–Schulz iterations (Jordan et al., 2024) to obtain an efficient SVD-free approximation, $O_t \approx \mathrm{NS_Orth}(M_t)$, which relies mainly on matrix multiplications and is therefore much cheaper at scale.

Muon is compelling because it is explicitly matrix-aware: it does not merely accumulate gradients, but transforms momentum into a spectrally normalized update direction. However, this geometric normalization is still *sharpness-agnostic*. It shapes the update direction, yet leaves uncontrolled the local worst-case loss around the iterate. As a result, Muon does not explicitly encourage convergence to flat regions, and any improvement in generalization arises only implicitly. This missing link is precisely what makes sharpness-aware Muon both important and nontrivial: it requires enriching a structured matrix optimizer with an explicit mechanism for local geometry control. (Keskar et al., 2016; Foret et al., 2020; Li et al., 2025)

3 Algorithm Design

3.1 A Sharpness-Aware Extension of Muon: MuSAM

We start from the most direct way to make Muon sharpness-aware. Since Muon already provides a structured, orthogonalized update rule for matrix-valued parameters, a natural next step is to explicitly incorporate local worst-case geometry into its update. To this end, we consider a single matrix parameter block $W \in \mathbb{R}^{m\times n}$ and construct a sharpness-aware Muon update at each iteration. The same design is applied blockwise to all matrix-valued parameters in a full network.

Step 1: a local minimax objective. We define the local sharpness-aware objective

$$\min_{W} \max_{\|\delta\|_F \leq \rho} f(W+\delta), \tag{5}$$

where $\rho > 0$ is the perturbation radius. The inner maximization searches for a worst-case perturbation inside a small Frobenius ball, and therefore explicitly measures the local sensitivity of the loss around the current iterate. This is precisely the mechanism through which sharpness-aware training biases optimization toward flatter regions of the loss landscape (Keskar et al., 2016; Li et al., 2021; Foret et al., 2020; Andriushchenko & Flammarion, 2022).

Step 2: stochastic linearization of the inner maximization. Directly solving the inner problem in Eq. (5) is computationally intractable for deep networks. Following the SAM principle (Li et al., 2025; Becker et al., 2024), we approximate it by a first-order stochastic linearization:

$$\begin{aligned}\delta_t &= \arg\max_{\|\delta\|_F\leq\rho} f(W_{t-1}+\delta) \\ &\approx \arg\max_{\|\delta\|_F\leq\rho} \Big(f(W_{t-1}) + \langle \nabla f(W_{t-1}), \delta\rangle\Big) \approx \arg\max_{\|\delta\|_F\leq\rho} \Big(f(W_{t-1}) + \langle \nabla \ell(W_{t-1}), \delta\rangle\Big),\end{aligned}$$

Algorithm 1 Sharpness-Aware Muon (MuSAM)

Require: Loss $\ell(W)$, initial matrix parameter $W_0 \in \mathbb{R}^{m\times n}$, learning rate η, momentum coefficient $\mu \in [0, 1)$, perturbation radius $\rho > 0$

Initialize $M_0 \leftarrow 0$
for $t = 1, 2, \ldots$ **do**
  Compute clean stochastic gradient $\nabla\ell(W_{t-1}; \mathcal{B}_t)$
  Construct sharpness-aware perturbation $\delta_t \leftarrow \rho \frac{\nabla\ell(W_{t-1};\mathcal{B}_t)}{\|\nabla\ell(W_{t-1};\mathcal{B}_t)\|_F}$
  Compute perturbed stochastic gradient $\nabla\ell(W_{t-1} + \delta_t; \mathcal{B}_t)$
  Update momentum matrix $M_t \leftarrow \mu M_{t-1} + (1 - \mu)\nabla\ell(W_{t-1} + \delta_t; \mathcal{B}_t)$
  Orthogonalize the momentum matrix $O_t \in \arg\min_{O\in\mathbb{R}^{m\times n}} \|O - M_t\|_F$, s.t. $O^\top O = I_n$
  Update parameter $W_t \leftarrow W_{t-1} - \eta O_t$
end for

where the first approximation is the local first-order Taylor expansion of $f(W_{t-1} + \delta)$, which is accurate when the perturbation radius ρ is small. The second replaces the full gradient $\nabla f(W_{t-1})$ by the minibatch stochastic gradient $\nabla\ell(W_{t-1})$, yielding a computationally tractable adversarial direction. By Cauchy–Schwarz, the maximizer of Eq. (5) admits the closed form

$$\delta_t = \rho \frac{\nabla\ell(W_{t-1}; \mathcal{B}_t)}{\|\nabla\ell(W_{t-1}; \mathcal{B}_t)\|_F}. \tag{6}$$

Thus, the perturbation is the normalized stochastic ascent direction.

Step 3: Muon update from the perturbed point. Once the perturbation is obtained, we evaluate the gradient at the perturbed iterate: $\nabla\ell(W_{t-1} + \delta_t; \mathcal{B}_t)$, and use it to update the momentum matrix:

$$M_t \;=\; \mu M_{t-1} + (1 - \mu)\nabla\ell(W_{t-1} + \delta_t; \mathcal{B}_t), \tag{7}$$

The orthogonalization and parameter update then follow the standard Muon pipeline introduced in Eqs. (3) and (4). This yields a direct sharpness-aware Muon method, which we call MuSAM, as summarized in Algorithm 1.

Why this construction is the right starting point. The above design is not an arbitrary hybrid. It is the most faithful sharpness-aware extension of Muon. The perturbation δ_t is obtained from a stochastic first-order approximation of the local inner maximization, and thus serves as an explicit SAM-style probe of local worst-case geometry. The outer update, in contrast, remains purely Muon-style: the perturbed gradient is accumulated into a momentum matrix and then orthogonalized before the parameter update. In this way, MuSAM cleanly separates two roles: the perturbation probes local geometry, while the orthogonalized Muon step preserves the matrix-aware update structure. This makes MuSAM a principled baseline for asking whether explicit sharpness control can improve Muon.

Practical considerations. MuSAM already establishes an important point: sharpness awareness can be introduced into Muon in a principled way, without giving up Muon's matrix-aware inductive bias. At the same time, this formulation exposes two practical issues. First, the perturbation in Eq. (6) requires a clean gradient at W_{t-1}, while the actual update uses a second gradient at $W_{t-1} + \delta_t$; this substantially increases the per-step cost, just as in vanilla SAM (Wen et al., 2018; Zhuang et al., 2022; Zhao et al., 2022). Second, the perturbation direction is determined by a single minibatch gradient. In stochastic training, such a direction can be overly batch-local and sensitive to sampling noise, so the resulting adversarial probe may not faithfully reflect the broader local geometry that matters for generalization. Recent analyses describe this phenomenon as instability of stochastic linearization, or equivalently, as the emergence of a friendly adversary that weakens the intended sharpness signal (Becker et al., 2024; Li et al., 2025).

3.2 A Momentum-Guided Refinement: MoMuSAM

We now turn Muon's momentum from a pure update variable into a sharpness-probing signal.

Core idea. The key observation is that Muon already maintains a matrix-valued momentum state M_{t-1} before the current backward pass. This state is available at no extra cost and aggregates gradient

Table 1: Comparison between MUON, MUSAM, and MOMUSAM.

Aspect	MUON	MUSAM	MOMUSAM
Perturbation source	–	$\nabla \ell(W_{t-1})$	M_{t-1}
Per-iteration gradient cost	1	2	1
Directional stability[†]	–	$\mathcal{O}(T^{-3/4})$	$\mathcal{O}(T^{-1})$
First-order stationarity[◇]	$\mathcal{O}(T^{-1/4})$	$\mathcal{O}(T^{-1/4})$	$\mathcal{O}(T^{-1/4})$

[†] The average directional-stability bounds for MUSAM and MOMUSAM are established in Theorem 4.6.
[◇] The stationarity rate of MUON follows from Shen et al. (2025); those of MUSAM and MOMUSAM are established in Theorems 4.1 and 4.2, respectively.

information across previous iterations. We therefore use M_{t-1} as the slope proxy for constructing the local sharpness perturbation, rather than recomputing a fresh clean gradient at W_{t-1}.

Momentum-guided local linearization. We use M_{t-1} as a *proxy* for the local ascent direction when constructing the perturbation. This yields the momentum-guided perturbation

$$\delta_t = \rho \frac{M_{t-1}}{\|M_{t-1}\|_F}. \tag{8}$$

Here the perturbation reuses Muon's matrix-valued momentum, while the matrix-aware geometry is retained by the polar outer update.

Why momentum yields a more stable perturbation proxy. The key reason is that the momentum state aggregates gradient information over time, rather than relying on a single minibatch gradient. Starting from the recursion $M_t = \mu M_{t-1} + (1-\mu)\nabla \ell(W_{t-1} + \delta_t; \mathcal{B}_t)$, one can unroll it from $M_0 = 0$ to obtain

$$M_{t-1} = (1-\mu) \sum_{s=1}^{t-1} \mu^{t-1-s} \nabla \ell(W_{s-1} + \delta_s; \mathcal{B}_s). \tag{9}$$

Thus, M_{t-1} combines perturbed stochastic gradients from multiple iterations with exponentially decaying weights, reducing its sensitivity to the fluctuations of any single minibatch. Because these gradients are evaluated along a changing optimization trajectory, the resulting stability depends on both stochastic-noise averaging and momentum-tracking error. By controlling these two effects jointly, Theorem 4.6 shows that, under the prescribed horizon-dependent hyperparameters, the certified average achieved-linearization gap improves from $\mathcal{O}(T^{-3/4})$ for the single-batch perturbation of MUSAM to $\mathcal{O}(T^{-1})$ for the momentum-guided perturbation of MOMUSAM. Hence, M_{t-1} provides a smoother and more history-aware sharpness probe without being interpreted as an exact solution of the inner maximization (Becker et al., 2024; Li et al., 2025).

Update rule. After constructing the perturbation in Eq. (8), we evaluate the gradient at the perturbed point and then proceed with the standard Muon update, following Eqs. (3), (4) and (7). The resulting method is termed MOMUSAM, short for *Momentum-guided Sharpness-Aware Muon*; its full procedure is summarized in Algorithm 2 in Section B.3.

What MOMUSAM buys us. The key differences among MUON, MUSAM, and MOMUSAM are summarized in Table 1. Relative to the direct sharpness-aware formulation, MOMUSAM provides three complementary advantages.

- First, MOMUSAM eliminates the additional clean-gradient evaluation required by MUSAM. It therefore reduces the per-iteration gradient cost from two evaluations to one, recovering the computational profile of vanilla Muon.
- Second, MOMUSAM replaces the instantaneous single-batch perturbation of MUSAM with the temporally aggregated momentum state. As established in Theorem 4.6, this improves the certified average achieved-linearization gap from $\mathcal{O}(T^{-3/4})$ to $\mathcal{O}(T^{-1})$ under the same horizon-dependent hyperparameters. Importantly, this sharper result controls the complete direction-sensitive stability error, including both stochastic fluctuation and momentum-tracking error, rather than only an isolated variance component.
- Third, MOMUSAM obtains this stability improvement without introducing an auxiliary estimator or optimizer state. At iteration t, the momentum M_t is orthogonalized to produce the current Muon update and is subsequently normalized to construct the next sharpness perturbation δ_{t+1}. The information accumulated for optimization is therefore immediately reused for neighborhood

probing, tightly coupling sharpness awareness with Muon's temporal state evolution without adding another backward pass.

Design philosophy. The key idea of MoMuSAM is to construct the perturbation from a signal that is less sensitive to the randomness of a single minibatch. Prior studies on momentum-guided and stabilized SAM variants suggest that such smoother adversarial directions can better capture the local geometry relevant to generalization under stochastic training (Becker et al., 2024; Li et al., 2025). What has remained missing, however, is an explicit theoretical characterization of this advantage. The next section fills this gap by showing that the momentum-guided perturbation in MoMuSAM yields a more stable stochastic linearization while preserving the same convergence order as MuSAM.

4 Theoretical Analysis

We now show that both MuSAM and MoMuSAM retain favorable convergence guarantees under the stochastic matrix-valued setting in Section 2. Our analysis addresses two complementary questions. First, do the proposed sharpness-aware Muon variants converge to first-order stationary points, both at the clean iterate and at the perturbed iterate? Second, does the momentum-guided perturbation in MoMuSAM enjoy a genuine stability advantage over the raw-gradient perturbation used in MuSAM? The first question is answered by Theorems 4.1 and 4.2, while the second is addressed by Theorem 4.6. Throughout this section, let $r := \min\{m, n\}$. All proofs are provided in the appendix.

Theorem 4.1 (Convergence of MuSAM)**.** *Suppose Assumptions 2.1 to 2.3 hold, and let* $\{W_t\}_{t=0}^{T-1}$ *be generated by* MuSAM. *Then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_* \lesssim \frac{f(W_0)-f_\star}{\eta T}+\frac{\sqrt{r}}{(1-\mu^2)T}\sqrt{\mu^2 L(f(W_0)-f_\star)+(1-\mu)^2L^2\rho^2+(1-\mu)^2\frac{\sigma^2}{B}} + \frac{\sqrt{r(1+\mu^2)}}{1-\mu^2}L\big(\mu\eta+(1-\mu)\rho\big)+\sqrt{\frac{r(1-\mu)}{1+\mu}}\left(\frac{\sigma}{\sqrt{B}}+L\rho\right)+(1+\sqrt{r})L\rho+L\eta. \tag{10}$$

Moreover, if the hyperparameters are chosen as $\eta = \rho = T^{-3/4}$ *and* $1-\mu = T^{-1/2}$*, then* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_* = \mathcal{O}(1/T^{\frac{1}{4}})$ *and* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t+\delta_{t+1})\|_* = \mathcal{O}(1/T^{\frac{1}{4}})$.

Theorem 4.2 (Convergence of MoMuSAM)**.** *Suppose Assumptions 2.1 to 2.3 hold, and let* $\{W_t\}_{t=0}^{T-1}$ *be generated by* MoMuSAM. *Then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_* \lesssim \frac{f(W_0)-f_*}{\eta T}+\frac{\sqrt{r}}{(1-\mu^2)T}\sqrt{\mu^2 L(f(W_0)-f_*)+(1-\mu)^2\frac{\sigma^2}{B}} + \frac{\sqrt{r(1+\mu^2)}}{1-\mu^2}L(\eta+\rho)+\sqrt{\frac{r(1-\mu)}{1+\mu}}\frac{\sigma}{\sqrt{B}}+L\rho+L\eta. \tag{11}$$

Moreover, if the hyperparameters are chosen as $\eta = \rho = T^{-3/4}$ *and* $1-\mu = T^{-1/2}$*, then* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_* = \mathcal{O}(1/T^{\frac{1}{4}})$ *and* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t+\delta_{t+1})\|_* = \mathcal{O}(1/T^{\frac{1}{4}})$.

Remark 4.3 (Same asymptotic rate, different algorithmic role). Theorems 4.1 and 4.2 show that the momentum-guided refinement in MoMuSAM does *not* pay an optimization penalty in terms of asymptotic first-order stationarity: both MuSAM and MoMuSAM attain the same order at the clean iterate W_t and at the perturbed iterate $W_t + \delta_{t+1}$. Thus, replacing the raw-gradient perturbation by a momentum-guided one is not a heuristic shortcut that sacrifices convergence guarantees.

Remark 4.4 (Where MoMuSAM improves). While MuSAM and MoMuSAM are of the same asymptotic order, MoMuSAM enjoys two additional advantages. First, it removes the extra clean-gradient evaluation required to construct the perturbation in MuSAM, thereby reducing the per-iteration gradient cost. Second, as shown in Theorem 4.6, it uses a more stable, temporally aggregated slope proxy for stochastic linearization. Hence, MoMuSAM preserves the same stationarity guarantee as MuSAM while offering a more efficient and statistically more stable perturbation mechanism.

Stability of stochastic linearization. To evaluate the perturbation actually induced by a slope proxy, we use a direction-sensitive population linearization criterion. For every nonzero $V \in \mathbb{R}^{m\times n}$, let $\mathsf{n}(V) := V/\|V\|_F$, and set $\mathsf{n}(0) := 0$.

Definition 4.5 (Directional linearization quality)**.** Let $g_t := \nabla f(W_{t-1})$. Define

$$L_t(V) := f(W_{t-1}) + \rho\langle g_t, \mathsf{n}(V)\rangle, \qquad \mathcal{E}_t(V) := L_t(g_t) - L_t(V). \tag{12}$$

A slope proxy V_t is said to be Δ_t-stable if $\mathbb{E}[\mathcal{E}_t(V_t)] \le \Delta_t$.

The gap $\mathcal{E}_t(V)$ measures the loss in population linearization caused by using the perturbation direction induced by V instead of the population-gradient direction. In particular, $\mathcal{E}_t(V) \ge 0$ and

$$\mathcal{E}_t(V) \le 2\rho\|V - \nabla f(W_{t-1})\|_F. \tag{13}$$

Theorem 4.6 (Stability of stochastic linearization)**.** *Suppose Assumptions 2.1 to 2.3 hold.*

(i) MuSAM. *For the raw-gradient perturbation used by* MuSAM*,*

$$\mathbb{E}[\mathcal{E}_t(\nabla\ell(W_{t-1}; \mathcal{B}_t))] \le \frac{2\rho\sigma}{\sqrt{B}}. \tag{14}$$

(ii) MoMuSAM. *For every $t \ge 2$, the momentum-guided perturbation used by* MoMuSAM *satisfies*

$$\mathbb{E}[\mathcal{E}_t(M_{t-1})] \le 2\rho\left[L\left(\rho + \frac{\eta}{1-\mu}\right) + \sqrt{\frac{1-\mu}{1+\mu}\frac{1+\mu^{t-1}}{1-\mu^{t-1}}}\frac{\sigma}{\sqrt{B}}\right]. \tag{15}$$

Moreover, if $\eta = \rho = T^{-3/4}$ and $1-\mu = T^{-1/2}$, then

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{E}_t(\nabla\ell(W_{t-1}; \mathcal{B}_t))] = \mathcal{O}\left(\frac{\sigma}{\sqrt{B}\,T^{3/4}}\right), \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{E}_t(M_{t-1})] = \mathcal{O}\left(\frac{L + \sigma/\sqrt{B}}{T}\right). \tag{16}$$

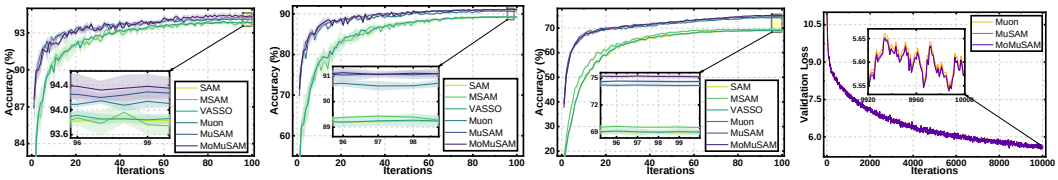
Remark 4.7 (Sharper certified stochastic stability). In the stochastic regime with minibatch noise, Theorem 4.6 establishes a sharper certified asymptotic upper-bound order for the stochastic-linearization error of MoMuSAM. Under the same horizon-dependent hyperparameters, the single-batch perturbation of MuSAM admits an average stability bound of order $\mathcal{O}(T^{-3/4})$, whereas the temporally aggregated perturbation of MoMuSAM improves this order to $\mathcal{O}(T^{-1})$. The improvement follows from averaging stochastic gradient noise across iterations, while the accompanying momentum-tracking error remains controlled at the same $\mathcal{O}(T^{-1})$ scale. Thus, the advantage is obtained for the complete stability bound, rather than only for an isolated variance term.

Remark 4.8 (Sharper stability without an optimization penalty). Taken together, Theorems 4.2 and 4.6 show that MoMuSAM preserves the $\mathcal{O}(T^{-1/4})$ first-order stationarity order of MuSAM, while improving the certified average stochastic-linearization stability from $\mathcal{O}(T^{-3/4})$ to $\mathcal{O}(T^{-1})$. Moreover, MoMuSAM achieves this sharper stability guarantee using only one gradient evaluation per iteration, thereby retaining the per-iteration gradient cost of vanilla Muon.

5 Experiments

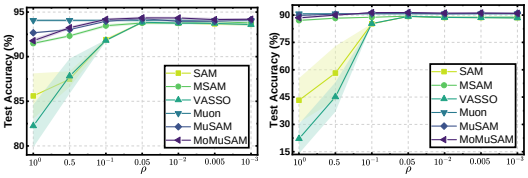
We evaluate the proposed MuSAM and MoMuSAM against a diverse set of baselines spanning both sharpness-aware and matrix-aware optimization. On the sharpness-aware side, we compare with standard SAM (Foret et al., 2020), MSAM (Becker et al., 2024), and VASSO (Li et al., 2025), which represent the canonical SAM formulation, a momentum-guided sharpness-aware variant, and a variance-suppressed stabilization approach, respectively. On the matrix-aware side, we compare with Muon (Jordan et al., 2024), the most relevant baseline for assessing whether explicit sharpness control can enhance matrix-aware orthogonalized optimization. Our experiments cover both image classification and language modeling tasks. For image classification, we evaluate on Fashion-MNIST (FMNIST) (Xiao et al., 2017) using a multilayer perceptron (MLP), and on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) using CifarNet and ResNet-18, respectively. For language modeling, we adopt the NanoGPT framework (Karpathy, 2022) to reproduce a 124M-parameter GPT-2 model (Radford et al., 2019) trained on OpenWebText (Gokaslan & Cohen, 2019).

In Figure 1, we compare the convergence performance of different optimizers across FMNIST, CIFAR-10, CIFAR-100, and OpenWebText. The results show that introducing sharpness awareness into Muon consistently improves the final performance, with MuSAM already outperforming the



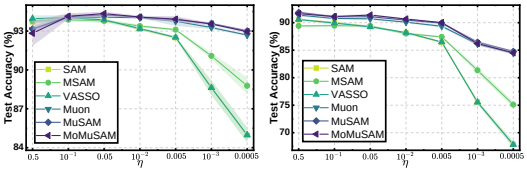
(a) FMNIST (b) CIFAR-10 (c) CIFAR-100 (d) OpenWebText

Figure 1: Performance comparison of different baselines across multiple datasets, showing test accuracy on FMNIST, CIFAR-10, and CIFAR-100, as well as validation loss on OpenWebText.



(a) FMNIST (b) CIFAR-10

Figure 2: Sensitivity analysis on ρ.



(a) FMNIST (b) CIFAR-10

Figure 3: Sensitivity analysis on η.

baselines in most cases and MoMuSAM delivering the strongest overall results. This trend can be understood from the complementary roles of the two components: Muon improves the *update geometry* through matrix-aware momentum orthogonalization, whereas sharpness-aware perturbation improves the *local loss geometry* explored during training by biasing optimization toward regions with smaller neighborhood worst-case loss. The further gain of MoMuSAM over MuSAM shows that the quality of the perturbation signal is also crucial. MuSAM builds its perturbation from a single minibatch gradient, so it remains sensitive to stochastic noise and batch-local fluctuations. MoMuSAM, in contrast, uses the momentum state, which aggregates information across iterations and minibatches. This makes the perturbation direction smoother, more stable, and more informed by Muon's accumulated optimization history, leading to a consistently better tradeoff between convergence quality, generalization, and computational efficiency.

Figures 2 and 3 examine the sensitivity of different optimizers to the perturbation radius ρ and learning rate η. The results reveal a clear robustness advantage for our proposed methods. With respect to ρ, MuSAM and MoMuSAM remain competitive over a substantially wider range of perturbation magnitudes than the SAM-style baselines. On both FMNIST and CIFAR-10, the performance of SAM, MSAM, and VASSO deteriorates rapidly when ρ becomes too large, suggesting that their perturbations can easily become overly aggressive and destabilize training. By contrast, MuSAM and MoMuSAM maintain strong accuracy across a much broader range of ρ, with MoMuSAM consistently achieving the best or near-best performance. This indicates that the matrix-aware Muon update already provides a stabilizing optimization geometry, and that our sharpness-aware methods preserve this robustness while further enhancing generalization. A similar pattern is observed for the learning rate in Figure 3. As η decreases, SAM, MSAM, and VASSO suffer much more pronounced degradation, whereas MuSAM and MoMuSAM remain comparatively stable and continue to deliver the strongest final accuracy over a broad range of stepsizes. This robustness is especially meaningful because it shows that our gains do not rely on fragile hyperparameter tuning. Overall, these sensitivity results show that the proposed methods are not only stronger in peak accuracy, but also substantially easier to tune in practice. Additional experimental results and setups are provided in Section C.

6 Conclusion

In this paper, we introduced the first sharpness-aware formulation of Muon and established a new optimizer family that unifies Muon's matrix-aware orthogonalized updates with explicit local geometry control. Our proposed MuSAM incorporates explicit local geometry control into Muon's matrix-aware updates, while MoMuSAM further turns momentum into a sharpness-probing signal, removing the extra clean-gradient computation and stabilizing the perturbation direction. Theoretically, both methods attain the same asymptotic first-order stationarity order, while MoMuSAM yields a more stable stochastic linearization. Empirically, sharpness awareness consistently strengthens Muon, and MoMuSAM delivers a markedly better generalization–efficiency tradeoff.

AI USE STATEMENT

In this work, we used generative AI tools for grammar checking and language polishing. The authors reviewed all AI-assisted edits and take full responsibility for the final content of the paper.

REFERENCES

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. Asgo: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International conference on machine learning*, pp. 639–668. PMLR, 2022.

Peter L Bartlett, Philip M Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *Journal of Machine Learning Research*, 24(316):1–36, 2023.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Marlon Becker, Frederick Altrock, and Benjamin Risse. Momentum-sam: Sharpness aware minimization without computational overhead. *arXiv preprint arXiv:2401.12033*, 2024.

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

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Da Chang, Yongxiang Liu, and Ganzhao Yuan. On the convergence of muon and beyond. *arXiv preprint arXiv:2509.15816*, 2025.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36:49205–49233, 2023.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent Tan, and Joey Tianyi Zhou. Sharpness-aware training for free. *Advances in Neural Information Processing Systems*, 35:23439–23451, 2022.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. *arXiv preprint arXiv:2502.04664*, 2025.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

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

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. *Advances in neural information processing systems*, 7, 1994.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pp. 427–435. PMLR, 2013.

Weisen Jiang, Hansi Yang, Yu Zhang, and James Kwok. An adaptive policy to employ sharpness-aware minimization. *arXiv preprint arXiv:2304.14647*, 2023.

Keller Jordan, Yiding Jin, Victor Boza, Yang Jiacheng, Franciszek Cecista, Landon Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon.

Andrej Karpathy. nanogpt, 2022. URL https://github.com/karpathy/nanoGPT.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

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

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher sam: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, pp. 11148–11161. PMLR, 2022.

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International conference on machine learning*, pp. 5905–5914. PMLR, 2021.

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.

Bingcong Li, Alireza Sadeghi, and Georgios Giannakis. Heavy ball momentum for conditional gradient. *Advances in Neural Information Processing Systems*, 34:21244–21255, 2021.

Bingcong Li, Yilang Zhang, and Georgios B Giannakis. Vasso: Variance suppression for sharpness-aware minimization. *arXiv preprint arXiv:2509.02433*, 2025.

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

Jiaxiang Li and Mingyi Hong. A note on the convergence of muon. *arXiv preprint arXiv:2502.02900*, 2025.

Chen Liu, Mathieu Salzmann, Tao Lin, Ryota Tomioka, and Sabine Süsstrunk. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. *Advances in Neural Information Processing Systems*, 33:21476–21487, 2020.

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*, 2025a.

Liming Liu, Zhenghao Xu, Zixuan Zhang, Hao Kang, Zichong Li, Chen Liang, Weizhu Chen, and Tuo Zhao. Cosmos: A hybrid adaptive optimizer for memory-efficient training of llms. *arXiv preprint arXiv:2502.17410*, 2025b.

Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12360–12370, 2022.

Peng Mi, Li Shen, Tianhe Ren, Yiyi Zhou, Xiaoshuai Sun, Rongrong Ji, and Dacheng Tao. Make sharpness-aware minimization stronger: A sparsified perturbation approach. *Advances in Neural Information Processing Systems*, 35:30950–30962, 2022.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

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

Xun Qian, Hussein Rammal, Dmitry Kovalev, and Peter Richtarik. Muon is provably faster with momentum variance reduction. *arXiv preprint arXiv:2512.16598*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic frank-wolfe methods for nonconvex optimization. In *2016 54th annual Allerton conference on communication, control, and computing (Allerton)*, pp. 1244–1251. IEEE, 2016.

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

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence bound and critical batch size of muon optimizer. *arXiv preprint arXiv:2507.01598*, 2025.

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and muons: Optimization via stochastic frank-wolfe. *arXiv preprint arXiv:2506.04192*, 2025.

Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical efficiency of muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of muon. *arXiv preprint arXiv:2505.23737*, 2025.

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

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How does sharpness-aware minimization minimize sharpness? *arXiv preprint arXiv:2211.05729*, 2022.

Wei Wen, Yandan Wang, Feng Yan, Cong Xu, Chunpeng Wu, Yiran Chen, and Hai Li. Smoothout: Smoothing out sharp minima to improve generalization in deep learning. *arXiv preprint arXiv:1805.07898*, 2018.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in neural information processing systems*, 33:2958–2969, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

Huizhuo Yuan, Yifeng Liu, Shuang Wu, Xun Zhou, and Quanquan Gu. Mars: Unleashing the power of variance reduction for training large models. *arXiv preprint arXiv:2411.10438*, 2024.

Minxin Zhang, Yuxuan Liu, and Hayden Schaeffer. Adagrad meets muon: Adaptive stepsizes for orthogonal updates. *arXiv preprint arXiv:2509.02981*, 2025.

Yang Zhao, Hao Zhang, and Xiuyuan Hu Ss-sam. Stochastic scheduled sharpness-aware minimization for efficiently training deep neural networks. *arXiv preprint arXiv:2203.09962*, 8:18, 2022.

Yaowei Zheng, Richong Zhang, and Yongyi Mao. Regularizing neural networks via adversarial model perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8156–8165, 2021.

Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. *arXiv preprint arXiv:2203.08065*, 2022.

A Related Works

Muon and matrix-aware orthogonalized optimization. Recent work has increasingly explored optimizers that exploit the intrinsic matrix structure of neural network parameters rather than treating each layer as a long vector. Within this line, Muon (Jordan et al., 2024) is a representative matrix-aware optimizer that orthogonalizes the momentum matrix before updating the weights, yielding structured search directions with strong empirical performance, especially at large scale (Liu et al., 2025a; Shah et al., 2025). In practice, Muon replaces exact SVD-based polar decomposition with a few Newton–Schulz iterations, resulting in an efficient SVD-free implementation based primarily on matrix multiplications (Kim & Oh, 2026). Around Muon, a growing ecosystem of related methods has emerged, including AdaMuon, COSMOS, ASGO, PolarGrad, AdaGO, Gluon, and other matrix-aware or LMO-based optimizers (Si et al., 2025; Liu et al., 2025b; An et al., 2025; Lau et al., 2025; Zhang et al., 2025; Riabinin et al., 2025; Pethick et al., 2025). On the theoretical side, Muon has been interpreted through non-Euclidean trust-region geometry, stochastic Frank–Wolfe viewpoints, spectral-norm constrained optimization, and implicit-bias analyses (Bernstein & Newhouse, 2024; Kovalev, 2025; Sfyraki & Wang, 2025; Chen et al., 2025; Fan et al., 2025). Nevertheless, existing Muon literature remains centered on optimization geometry, convergence, and efficiency, while exposing a persistent gap between practical finite-step Newton–Schulz implementations and analyses based on exact polar factors, as well as relatively slow or restrictive guarantees for several Muon variants (Shen et al., 2025; Sato et al., 2025; Li & Hong, 2025; Chang et al., 2025). Crucially for our setting, these works do not establish any explicit mechanism by which Muon controls local sharpness or steers optimization toward flat regions: Muon reshapes the update direction, but any generalization benefit from flatness can only arise implicitly. This makes sharpness-aware Muon especially compelling, because Muon's matrix-aware inductive bias shapes update geometry, while explicit sharpness control governs the local loss geometry most relevant to generalization.

Sharpness-aware optimization. Sharpness-aware optimization is motivated by the long-standing observation that flatter minima often correlate with stronger generalization, although this connection is subtle and not universal (Hochreiter & Schmidhuber, 1994; Keskar et al., 2016; Li et al., 2018; Yao et al., 2018; Liu et al., 2020; Damian et al., 2021; Dinh et al., 2017). Early efforts such as Entropy-SGD and adversarial weight perturbation already sought to bias optimization toward flatter regions (Chaudhari et al., 2019; Wu et al., 2020; Zheng et al., 2021), but SAM has become the dominant formulation by explicitly minimizing a local worst-case loss through adversarial perturbations in parameter space (Foret et al., 2020). Its empirical success has motivated a broad family of variants, including ASAM, Fisher-SAM, GSAM, and a range of efficiency-oriented methods such as LookSAM, SS-SAM, adaptive SAM policies, and SAF (Kwon et al., 2021; Kim et al., 2022; Zhuang et al., 2022; Liu et al., 2022; Zhao et al., 2022; Jiang et al., 2023; Du et al., 2022). At the same time, this line of work has exposed two persistent limitations. First, vanilla SAM typically requires an additional forward–backward pass to construct the perturbation, which substantially increases training cost (Foret et al., 2020; Wen et al., 2022; Becker et al., 2024). Second, the perturbation is produced by stochastic linearization from a single minibatch gradient, making it a batch-local and noisy proxy for sharpness; recent work shows that such adversaries can become unstable or overly friendly, thereby weakening the intended sharpness signal and limiting the generalization–computation tradeoff (Andriushchenko & Flammarion, 2022; Becker et al., 2024; Li et al., 2025). These limitations are especially relevant in our setting: once sharpness-aware training is brought into Muon, the key challenge is no longer only how to inject local geometry control, but how to do so in a way that respects Muon's structured update geometry while avoiding the overhead and instability of single-batch perturbations. This is precisely the gap addressed by our momentum-guided sharpness-aware Muon design.

B Proof

B.1 Useful lemmas

Lemma B.1. *Under Assumptions 2.1 to 2.3, the Muon+SAM update guarantees that*

$$f(W_t) \le f(W_{t-1}) - \eta\|\nabla f(W_{t-1})\|_* + 2\eta\|M_t - \nabla f(W_{t-1}+\delta_t)\|_* + 2\eta L\rho + \frac{L\eta^2}{2}. \quad (17)$$

Proof. Let the exact SVD of M_t be $M_t = Q_t \Sigma_t R_t^\top$, and define $U_t := \mathrm{Polar}(M_t) = Q_t R_t^\top$. By the descent lemma implied by Assumption 2.2, applied to $W_t = W_{t-1} - \eta U_t$, we have

$$\begin{aligned} f(W_t) &\le f(W_{t-1}) + \langle \nabla f(W_{t-1}), W_t - W_{t-1} \rangle_F + \frac{L}{2}\|W_t - W_{t-1}\|_{\mathrm{op}}^2 \\ &= f(W_{t-1}) - \eta \langle \nabla f(W_{t-1}), U_t \rangle_F + \frac{L\eta^2}{2}\|U_t\|_{\mathrm{op}}^2 \\ &\le f(W_{t-1}) - \eta \langle \nabla f(W_{t-1}), U_t \rangle_F + \frac{L\eta^2}{2}. \end{aligned} \tag{18}$$

Next,

$$\begin{aligned} \langle \nabla f(W_{t-1}), U_t \rangle_F &= \langle M_t, U_t \rangle_F + \langle \nabla f(W_{t-1}) - M_t, U_t \rangle_F \\ &= \|M_t\|_* + \langle \nabla f(W_{t-1}) - M_t, U_t \rangle_F \\ &\ge \|M_t\|_* - \|\nabla f(W_{t-1}) - M_t\|_* \, \|U_t\|_{\mathrm{op}} \\ &\ge \|M_t\|_* - \|\nabla f(W_{t-1}) - M_t\|_*. \end{aligned} \tag{19}$$

By the reverse triangle inequality for the nuclear norm,

$$\|M_t\|_* \ge \|\nabla f(W_{t-1})\|_* - \|\nabla f(W_{t-1}) - M_t\|_*. \tag{20}$$

Combining Eq. (19) and Eq. (20), we obtain

$$\langle \nabla f(W_{t-1}), U_t \rangle_F \ge \|\nabla f(W_{t-1})\|_* - 2\|\nabla f(W_{t-1}) - M_t\|_*. \tag{21}$$

Substituting Eq. (21) into Eq. (18) yields

$$f(W_t) \le f(W_{t-1}) - \eta \|\nabla f(W_{t-1})\|_* + 2\eta \|\nabla f(W_{t-1}) - M_t\|_* + \frac{L\eta^2}{2}. \tag{22}$$

Now add and subtract $\nabla f(W_{t-1} + \delta_t)$:

$$\|\nabla f(W_{t-1}) - M_t\|_* \le \|\nabla f(W_{t-1}) - \nabla f(W_{t-1} + \delta_t)\|_* + \|\nabla f(W_{t-1} + \delta_t) - M_t\|_*. \tag{23}$$

By Assumption 2.2,

$$\|\nabla f(W_{t-1}) - \nabla f(W_{t-1} + \delta_t)\|_* \le L\|\delta_t\|_{\mathrm{op}} \le L\|\delta_t\|_F \le L\rho. \tag{24}$$

Therefore,

$$\|\nabla f(W_{t-1}) - M_t\|_* \le \|M_t - \nabla f(W_{t-1} + \delta_t)\|_* + L\rho. \tag{25}$$

Substituting the above into Eq. (22) proves Eq. (17). □

Lemma B.2. *Under Assumptions 2.1 to 2.3, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f(W_{t-1})\|_* \le \frac{f(W_0) - f_\star}{\eta T} + \frac{2}{T}\sum_{t=1}^{T} \mathbb{E}\|M_t - \nabla f(W_{t-1} + \delta_t)\|_* + 2L\rho + \frac{L\eta}{2}, \tag{26}$$

where $f_\star$ is a lower bound of f.

Proof. Taking expectation in Eq. (17), summing over $t = 1, \ldots, T$, and using the telescoping identity

$$\sum_{t=1}^{T} \big(f(W_t) - f(W_{t-1})\big) = f(W_T) - f(W_0),$$

we obtain

$$\begin{aligned} &\mathbb{E}[f(W_T)] - \mathbb{E}[f(W_0)] \\ &\le -\eta \sum_{t=1}^{T} \mathbb{E}\|\nabla f(W_{t-1})\|_* + 2\eta \sum_{t=1}^{T} \mathbb{E}\|M_t - \nabla f(W_{t-1} + \delta_t)\|_* + 2\eta L T \rho + \frac{TL\eta^2}{2}. \end{aligned} \tag{27}$$

Using $\mathbb{E}[f(W_T)] \ge f_\star$ and rearranging gives Eq. (26). □

B.2 Theoretical Analysis for Algorithm 1

Lemma B.3 (Same-minibatch estimator decomposition)**.** *According to Assumptions 2.2 and 2.3, we have*

$$\begin{aligned}&\mathbb{E}[\nabla\ell(W_{t-1};\mathcal{B}_t)-\nabla f(W_{t-1})+\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\\&-\mathbb{E}[\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\mid\mathcal{F}_{t-1}]\mid\mathcal{F}_{t-1}]=0\end{aligned}\tag{28}$$

and

$$\begin{aligned}&\mathbb{E}[\|\nabla\ell(W_{t-1};\mathcal{B}_t)-\nabla f(W_{t-1})+\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\\&-\mathbb{E}[\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\mid\mathcal{F}_{t-1}]\|_F^2\mid\mathcal{F}_{t-1}]\le 2\left(\frac{\sigma^2}{B}+L^2\rho^2\right).\end{aligned}\tag{29}$$

Define $C_t := M_t - \nabla f(W_{t-1})$, for every $t \ge 2$,

$$\begin{aligned}C_t=\mu C_{t-1}+\widetilde{D}_t+(1-\mu)(&\nabla\ell(W_{t-1};\mathcal{B}_t)-\nabla f(W_{t-1})+\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\\&-\mathbb{E}[\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\mid\mathcal{F}_{t-1}]),\end{aligned}\tag{30}$$

where

$$\widetilde{D}_t := \mu\left(\nabla f(W_{t-2})-\nabla f(W_{t-1})\right)+(1-\mu)\mathbb{E}[\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\mid\mathcal{F}_{t-1}]\tag{31}$$

is $\mathcal{F}_{t-1}$-measurable and satisfies

$$\|\widetilde{D}_t\|_F \le L\big(\mu\eta+(1-\mu)\rho\big).\tag{32}$$

Proof. Because $\nabla\ell(W_{t-1};\mathcal{B}_t)$ is evaluated at the $\mathcal{F}_{t-1}$-measurable point W_{t-1}, Assumption 2.3 gives $\mathbb{E}[\nabla\ell(W_{t-1};\mathcal{B}_t)-\nabla f(W_{t-1})\mid\mathcal{F}_{t-1}]=0$. Moreover, we have

$$\mathbb{E}[\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)-\mathbb{E}[\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\mid\mathcal{F}_{t-1}]\mid\mathcal{F}_{t-1}]=0.$$

Adding these identities proves Eq. (28).

By Assumption 2.2 and $\|\delta_t\|_F \le \rho$,

$$\begin{aligned}\|\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\|_F &\le \|\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\|_*\\&\le L\|\delta_t\|_{\mathrm{op}}\le L\rho.\end{aligned}\tag{33}$$

Consequently,

$$\begin{aligned}&\|\mathbb{E}[\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\mid\mathcal{F}_{t-1}]\|_F\\&\le\mathbb{E}[\|\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\|_F\mid\mathcal{F}_{t-1}]\le L\rho,\end{aligned}\tag{34}$$

and the conditional variance identity gives

$$\begin{aligned}&\mathbb{E}\big[\|\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)-\mathbb{E}[\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\mid\mathcal{F}_{t-1}]\|_F^2\mid\mathcal{F}_{t-1}\big]\\&\le\mathbb{E}\big[\|\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\|_F^2\mid\mathcal{F}_{t-1}\big]\le L^2\rho^2.\end{aligned}\tag{35}$$

Using $\|A+B\|_F^2\le 2\|A\|_F^2+2\|B\|_F^2$ and Assumption 2.3, we obtain

$$\begin{aligned}&\mathbb{E}[\|\nabla\ell(W_{t-1};\mathcal{B}_t)-\nabla f(W_{t-1})+\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\\&\quad-\mathbb{E}[\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\mid\mathcal{F}_{t-1}]\|_F^2\mid\mathcal{F}_{t-1}]\\&\le 2\,\mathbb{E}\big[\|\nabla\ell(W_{t-1};\mathcal{B}_t)-\nabla f(W_{t-1})\|_F^2\mid\mathcal{F}_{t-1}\big]\\&\quad+2\mathbb{E}[\|\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)-\mathbb{E}[\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla\ell(W_{t-1};\mathcal{B}_t)\mid\mathcal{F}_{t-1}]\|_F^2\mid\mathcal{F}_{t-1}]\\&\le 2\left(\frac{\sigma^2}{B}+L^2\rho^2\right),\end{aligned}\tag{36}$$

which proves Eq. (29).

Next, the momentum recursion gives

$$C_t=\mu M_{t-1}+(1-\mu)\nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t)-\nabla f(W_{t-1})$$

$$
\begin{aligned}
&= \mu\,(M_{t-1} - \nabla f(W_{t-2})) + \mu\,(\nabla f(W_{t-2}) - \nabla f(W_{t-1})) \\
&\quad + (1-\mu)\mathbb{E}[\nabla \ell(W_{t-1}+\delta_t;\mathcal{B}_t) - \nabla \ell(W_{t-1};\mathcal{B}_t) \mid \mathcal{F}_{t-1}] \\
&\quad + (1-\mu)(\nabla \ell(W_{t-1};\mathcal{B}_t) - \nabla f(W_{t-1}) + \nabla \ell(W_{t-1}+\delta_t;\mathcal{B}_t) - \nabla \ell(W_{t-1};\mathcal{B}_t) \\
&\qquad\qquad - \mathbb{E}[\nabla \ell(W_{t-1}+\delta_t;\mathcal{B}_t) - \nabla \ell(W_{t-1};\mathcal{B}_t) \mid \mathcal{F}_{t-1}]) \\
&= \mu C_{t-1} + \widetilde{D}_t + (1-\mu)(\nabla \ell(W_{t-1};\mathcal{B}_t) - \nabla f(W_{t-1}) + \nabla \ell(W_{t-1}+\delta_t;\mathcal{B}_t) - \nabla \ell(W_{t-1};\mathcal{B}_t) \\
&\quad - \mathbb{E}[\nabla \ell(W_{t-1}+\delta_t;\mathcal{B}_t) - \nabla \ell(W_{t-1};\mathcal{B}_t) \mid \mathcal{F}_{t-1}]). \qquad (37)
\end{aligned}
$$

This proves Eq. (30).

Finally, Assumption 2.2 and the exact-polar update imply

$$\|\nabla f(W_{t-2}) - \nabla f(W_{t-1})\|_F \le L\|W_{t-2} - W_{t-1}\|_{\mathrm{op}} \le L\eta.$$

Combining this inequality with $\|\mathbb{E}[\nabla \ell(W_{t-1}+\delta_t;\mathcal{B}_t) - \nabla \ell(W_{t-1};\mathcal{B}_t) \mid \mathcal{F}_{t-1}]\|_F \le L\rho$ gives $\|\widetilde{D}_t\|_F \le L\big(\mu\eta + (1-\mu)\rho\big)$, which proves Eq. (32). □

Theorem B.4. *Suppose Assumptions 2.1 to 2.3 hold.*

(i) Recursion for the clean-point estimator error. *For every $t \ge 2$,*

$$
\begin{aligned}
\mathbb{E}\|C_t\|_F^2 \le \frac{1+\mu^2}{2}\mathbb{E}\|C_{t-1}\|_F^2 + \frac{1+\mu^2}{1-\mu^2}L^2\big(\mu\eta + (1-\mu)\rho\big)^2 \\
+ 2(1-\mu)^2\left(\frac{\sigma^2}{B} + L^2\rho^2\right). \qquad (38)
\end{aligned}
$$

(ii) Unrolled bound. *For every $t \ge 1$,*

$$
\begin{aligned}
\mathbb{E}\|C_t\|_F^2 \le \left(\frac{1+\mu^2}{2}\right)^{t-1}\mathbb{E}\|C_1\|_F^2 + \frac{2(1+\mu^2)}{(1-\mu^2)^2}L^2\big(\mu\eta + (1-\mu)\rho\big)^2 \\
+ \frac{4(1-\mu)}{1+\mu}\left(\frac{\sigma^2}{B} + L^2\rho^2\right). \qquad (39)
\end{aligned}
$$

(iii) Pointwise mismatch bound. *Let $E_t := M_t - \nabla f(W_{t-1}+\delta_t)$. Then, for every $t \ge 1$,*

$$
\begin{aligned}
\mathbb{E}\|E_t\|_F^2 \le 2\left(\frac{1+\mu^2}{2}\right)^{t-1}\mathbb{E}\|C_1\|_F^2 + \frac{4(1+\mu^2)}{(1-\mu^2)^2}L^2\big(\mu\eta + (1-\mu)\rho\big)^2 \\
+ \frac{8(1-\mu)}{1+\mu}\left(\frac{\sigma^2}{B} + L^2\rho^2\right) + 2L^2\rho^2. \qquad (40)
\end{aligned}
$$

Writing $r := \min\{m,n\}$, we further have

$$
\begin{aligned}
\mathbb{E}\|E_t\|_* \le \sqrt{2r}\left(\frac{1+\mu^2}{2}\right)^{(t-1)/2}\sqrt{\mathbb{E}\|C_1\|_F^2} + \frac{2\sqrt{r(1+\mu^2)}}{1-\mu^2}L\big(\mu\eta + (1-\mu)\rho\big) \\
+ 2\sqrt{\frac{2r(1-\mu)}{1+\mu}}\sqrt{\frac{\sigma^2}{B} + L^2\rho^2} + \sqrt{2r}\,L\rho. \qquad (41)
\end{aligned}
$$

(iv) Averaged mismatch bound. *For every $T \ge 1$,*

$$
\begin{aligned}
\frac{2}{T}\sum_{t=1}^{T}\mathbb{E}\|E_t\|_* \le \frac{8\sqrt{2r\mathbb{E}\|C_1\|_F^2}}{(1-\mu^2)T} + \frac{4\sqrt{r(1+\mu^2)}}{1-\mu^2}L\big(\mu\eta + (1-\mu)\rho\big) \\
+ 4\sqrt{\frac{2r(1-\mu)}{1+\mu}}\sqrt{\frac{\sigma^2}{B} + L^2\rho^2} + 2\sqrt{2r}\,L\rho. \qquad (42)
\end{aligned}
$$

(v) Averaged stationarity bound. *We have*

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_* \\ &\le \frac{f(W_0)-f_\star}{\eta T}+\frac{8\sqrt{2r}}{(1-\mu^2)T}\sqrt{4\mu^2 L\big(f(W_0)-f_\star\big)+4(1-\mu)^2L^2\rho^2+2(1-\mu)^2\frac{\sigma^2}{B}} \\ &\quad+\frac{4\sqrt{r(1+\mu^2)}}{1-\mu^2}L\big(\mu\eta+(1-\mu)\rho\big)+4\sqrt{\frac{2r(1-\mu)}{1+\mu}}\sqrt{\frac{\sigma^2}{B}+L^2\rho^2}+(2+2\sqrt{2r})L\rho+\frac{L\eta}{2}.\end{aligned} \tag{43}$$

In particular, if $T\ge 2$, $\eta=\rho=T^{-3/4}$, *and* $1-\mu=T^{-1/2}$, *then*

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_* \\ &=\mathcal{O}\left(\frac{f(W_0)-f_\star+\sqrt{r}\,L}{T^{1/4}}+\frac{\sqrt{r}\,\sigma}{\sqrt{B}\,T^{1/4}}+\frac{\sqrt{rL(f(W_0)-f_\star)}}{T^{1/2}}+\frac{(1+\sqrt{r})L}{T^{3/4}}\right),\end{aligned} \tag{44}$$

and

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t+\delta_{t+1})\|_* \\ &=\mathcal{O}\left(\frac{f(W_0)-f_\star+\sqrt{r}\,L}{T^{1/4}}+\frac{\sqrt{r}\,\sigma}{\sqrt{B}\,T^{1/4}}+\frac{\sqrt{rL(f(W_0)-f_\star)}}{T^{1/2}}+\frac{(1+\sqrt{r})L}{T^{3/4}}\right).\end{aligned} \tag{45}$$

Proof. We prove the statements in order.

Proof of (i). By Eqs. (28) and (30), conditioned on $\mathcal{F}_{t-1}$, we have

$$\mathbb{E}\big[\|C_t\|_F^2\mid\mathcal{F}_{t-1}\big]=\Big\|\mu C_{t-1}+\widetilde{D}_t\Big\|_F^2+(1-\mu)^2\mathbb{E}\big[\|\zeta_t\|_F^2\mid\mathcal{F}_{t-1}\big]. \tag{46}$$

For any $\alpha>0$, Young's inequality gives $\|A+B\|_F^2\le(1+\alpha)\|A\|_F^2+\big(1+\frac{1}{\alpha}\big)\|B\|_F^2$. For $\mu\in(0,1)$, apply this inequality with $A=\mu C_{t-1}$, $B=\widetilde{D}_t$, and $\alpha=\frac{1-\mu^2}{2\mu^2}$. Then $\mu^2(1+\alpha)=\frac{1+\mu^2}{2}$ and $1+\frac{1}{\alpha}=\frac{1+\mu^2}{1-\mu^2}$. Using Eqs. (29) and (32) in Eq. (46), and then taking total expectation, gives

$$\mathbb{E}\|C_t\|_F^2\le\frac{1+\mu^2}{2}\mathbb{E}\|C_{t-1}\|_F^2+\frac{1+\mu^2}{1-\mu^2}L^2\big(\mu\eta+(1-\mu)\rho\big)^2+2(1-\mu)^2\left(\frac{\sigma^2}{B}+L^2\rho^2\right). \tag{47}$$

This proves Eq. (38). For $\mu=0$, Eq. (30) and conditional orthogonality give $\mathbb{E}\|C_t\|_F^2\le L^2\rho^2+2\left(\frac{\sigma^2}{B}+L^2\rho^2\right)$, which is bounded by the right-hand side of Eq. (38) evaluated at $\mu=0$. Hence, the same recursion also holds for $\mu=0$.

Proof of (ii). Let $q:=\frac{1+\mu^2}{2}$. Unrolling Eq. (38) from C_1 gives

$$\mathbb{E}\|C_t\|_F^2\le q^{t-1}\mathbb{E}\|C_1\|_F^2+\frac{1-q^{t-1}}{1-q}\left[\frac{1+\mu^2}{1-\mu^2}L^2\big(\mu\eta+(1-\mu)\rho\big)^2+2(1-\mu)^2\left(\frac{\sigma^2}{B}+L^2\rho^2\right)\right]. \tag{48}$$

Because $1-q=\frac{1-\mu^2}{2}$, we obtain

$$\mathbb{E}\|C_t\|_F^2\le q^{t-1}\mathbb{E}\|C_1\|_F^2+\frac{2(1+\mu^2)}{(1-\mu^2)^2}L^2\big(\mu\eta+(1-\mu)\rho\big)^2+\frac{4(1-\mu)}{1+\mu}\left(\frac{\sigma^2}{B}+L^2\rho^2\right), \tag{49}$$

which proves Eq. (39).

Proof of (iii). By definition, $E_t = C_t - (\nabla f(W_{t-1} + \delta_t) - \nabla f(W_{t-1}))$. Therefore,

$$\|E_t\|_F^2 \le 2\|C_t\|_F^2 + 2\left\|\nabla f(W_{t-1} + \delta_t) - \nabla f(W_{t-1})\right\|_F^2 \le 2\|C_t\|_F^2 + 2L^2\rho^2, \tag{50}$$

where we used Assumption 2.2 and $\|\delta_t\|_F \le \rho$. Combining this inequality with Eq. (39) proves Eq. (40).

Furthermore, $\|E_t\|_* \le \sqrt{r}\,\|E_t\|_F$. Jensen's inequality and $\sqrt{x_1 + \cdots + x_k} \le \sum_{i=1}^k \sqrt{x_i}$ then give Eq. (41).

Proof of (iv). Summing Eq. (41) over $t = 1, \ldots, T$, and using

$$\sum_{t=1}^{T} \left(\frac{1+\mu^2}{2}\right)^{(t-1)/2} \le \frac{1}{1 - \sqrt{(1+\mu^2)/2}} = \frac{2\left(1 + \sqrt{(1+\mu^2)/2}\right)}{1-\mu^2} \le \frac{4}{1-\mu^2}, \tag{51}$$

proves Eq. (42).

Proof of (v). The definitions in Lemma B.3 also apply at $t = 1$. Since $M_0 = 0$, we have

$$\begin{aligned} C_1 &= (1-\mu)\nabla\ell(W_0 + \delta_1; \mathcal{B}_1) - \nabla f(W_0) \\ &= -\mu\nabla f(W_0) + (1-\mu)\mathbb{E}[\nabla\ell(W_0 + \delta_1; \mathcal{B}_1) - \nabla\ell(W_0; \mathcal{B}_1) \mid \mathcal{F}_0] \\ &\quad + (1-\mu)(\nabla\ell(W_0; \mathcal{B}_1) - \nabla f(W_0) + \nabla\ell(W_0 + \delta_1; \mathcal{B}_1) - \nabla\ell(W_0; \mathcal{B}_1) \\ &\qquad - \mathbb{E}[\nabla\ell(W_0 + \delta_1; \mathcal{B}_1) - \nabla\ell(W_0; \mathcal{B}_1) \mid \mathcal{F}_0]). \end{aligned} \tag{52}$$

Because $\mathbb{E}[\nabla\ell(W_0; \mathcal{B}_1) - \nabla f(W_0) + \nabla\ell(W_0 + \delta_1; \mathcal{B}_1) - \nabla\ell(W_0; \mathcal{B}_1) - \mathbb{E}[\nabla\ell(W_0 + \delta_1; \mathcal{B}_1) - \nabla\ell(W_0; \mathcal{B}_1) \mid \mathcal{F}_0] \mid \mathcal{F}_0] = 0$, the cross term between $-\mu\nabla f(W_0) + (1-\mu)\mathbb{E}[\nabla\ell(W_0 + \delta_1; \mathcal{B}_1) - \nabla\ell(W_0; \mathcal{B}_1) \mid \mathcal{F}_0]$ and $\nabla\ell(W_0; \mathcal{B}_1) - \nabla f(W_0) + \nabla\ell(W_0 + \delta_1; \mathcal{B}_1) - \nabla\ell(W_0; \mathcal{B}_1) - \mathbb{E}[\nabla\ell(W_0 + \delta_1; \mathcal{B}_1) - \nabla\ell(W_0; \mathcal{B}_1) \mid \mathcal{F}_0]$ vanishes. Hence,

$$\begin{aligned} \mathbb{E}\|C_1\|_F^2 &= \|-\mu\nabla f(W_0) + (1-\mu)\mathbb{E}[\nabla\ell(W_0 + \delta_1; \mathcal{B}_1) - \nabla\ell(W_0; \mathcal{B}_1) \mid \mathcal{F}_0]\|_F^2 \\ &\quad + (1-\mu)^2\mathbb{E}\|\nabla\ell(W_0; \mathcal{B}_1) - \nabla f(W_0) + \nabla\ell(W_0 + \delta_1; \mathcal{B}_1) - \nabla\ell(W_0; \mathcal{B}_1) \\ &\qquad - \mathbb{E}[\nabla\ell(W_0 + \delta_1; \mathcal{B}_1) - \nabla\ell(W_0; \mathcal{B}_1) \mid \mathcal{F}_0]\|_F^2 \\ &\le 2\mu^2\|\nabla f(W_0)\|_F^2 + 2(1-\mu)^2L^2\rho^2 + 2(1-\mu)^2\left(\frac{\sigma^2}{B} + L^2\rho^2\right) \\ &= 2\mu^2\|\nabla f(W_0)\|_F^2 + 4(1-\mu)^2L^2\rho^2 + 2(1-\mu)^2\frac{\sigma^2}{B}. \end{aligned} \tag{53}$$

Under Assumptions 2.1 and 2.2, the standard gradient-gap inequality yields $\|\nabla f(W_0)\|_F^2 \le 2L\big(f(W_0) - f_\star\big)$. Consequently,

$$\mathbb{E}\|C_1\|_F^2 \le 4\mu^2L\big(f(W_0) - f_\star\big) + 4(1-\mu)^2L^2\rho^2 + 2(1-\mu)^2\frac{\sigma^2}{B}. \tag{54}$$

Substituting Eq. (54) into Eq. (42), and then applying Eq. (26), gives Eq. (43).

To recover the simplified main-text bound, use $\sqrt{\frac{\sigma^2}{B} + L^2\rho^2} \le \frac{\sigma}{\sqrt{B}} + L\rho$ and $\sqrt{\frac{1-\mu}{1+\mu}} \le 1$. This yields

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_* &\lesssim \frac{f(W_0) - f_\star}{\eta T} + \frac{\sqrt{r(1+\mu^2)}}{1-\mu^2}L\big(\mu\eta + (1-\mu)\rho\big) \\ &\quad + \frac{\sqrt{r}}{(1-\mu^2)T}\sqrt{\mu^2L\big(f(W_0) - f_\star\big) + (1-\mu)^2L^2\rho^2 + (1-\mu)^2\frac{\sigma^2}{B}} \\ &\quad + \sqrt{\frac{r(1-\mu)}{1+\mu}}\left(\frac{\sigma}{\sqrt{B}} + L\rho\right) + (1+\sqrt{r})L\rho + L\eta. \end{aligned} \tag{55}$$

Now set $\eta = \rho = T^{-3/4}$ and $1 - \mu = T^{-1/2}$, then Eq. (55) implies

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_* \\
&= \mathcal{O}\left(\frac{f(W_0) - f_\star + \sqrt{r}\,L}{T^{1/4}} + \frac{\sqrt{r}\,\sigma}{\sqrt{B}\,T^{1/4}} + \frac{\sqrt{rL(f(W_0) - f_\star)}}{T^{1/2}} + \frac{(1+\sqrt{r})L}{T^{3/4}}\right). \qquad (56)
\end{aligned}$$

Finally, for every t,

$$\begin{aligned}
\|\nabla f(W_t + \delta_{t+1})\|_* &\le \|\nabla f(W_t)\|_* + \|\nabla f(W_t + \delta_{t+1}) - \nabla f(W_t)\|_* \\
&\le \|\nabla f(W_t)\|_* + L\|\delta_{t+1}\|_{\mathrm{op}} \\
&\le \|\nabla f(W_t)\|_* + L\rho. \qquad (57)
\end{aligned}$$

Taking expectation and averaging gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t + \delta_{t+1})\|_* \le \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_* + L\rho. \qquad (58)$$

By combining Eq. (58) and Eq. (55), and further setting $\eta = \rho = T^{-3/4}$ and $1 - \mu = T^{-1/2}$, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t + \delta_{t+1})\|_* \\
&\lesssim \frac{f(W_0) - f_\star}{\eta T} + \frac{\sqrt{r(1+\mu^2)}}{1-\mu^2}L\big(\mu\eta + (1-\mu)\rho\big) \\
&\quad + \frac{\sqrt{r}}{(1-\mu^2)T}\sqrt{\mu^2 L\big(f(W_0) - f_\star\big) + (1-\mu)^2L^2\rho^2 + (1-\mu)^2\frac{\sigma^2}{B}} \\
&\quad + \sqrt{\frac{r(1-\mu)}{1+\mu}}\left(\frac{\sigma}{\sqrt{B}} + L\rho\right) + (1+\sqrt{r})L\rho + L\eta \\
&= \mathcal{O}\left(\frac{f(W_0) - f_\star + \sqrt{r}\,L}{T^{1/4}} + \frac{\sqrt{r}\,\sigma}{\sqrt{B}\,T^{1/4}} + \frac{\sqrt{rL(f(W_0) - f_\star)}}{T^{1/2}} + \frac{(1+\sqrt{r})L}{T^{3/4}}\right). \qquad (59)
\end{aligned}$$

□

Algorithm 2 Momentum-guided Sharpness-Aware Muon (MoMuSAM)

Require: Loss $\ell(W)$, initial matrix parameter $W_0 \in \mathbb{R}^{m\times n}$, learning rate η, momentum coefficient $\mu \in [0, 1)$, perturbation radius $\rho > 0$

1: Initialize $M_0 \leftarrow 0$ and $\delta_1 \leftarrow 0$
2: **for** $t = 1, 2, \ldots$ **do**
3: Compute perturbed stochastic gradient $\nabla\ell(W_{t-1} + \delta_t; \mathcal{B}_t)$
4: Update momentum matrix $M_t \leftarrow \mu M_{t-1} + (1-\mu)\nabla\ell(W_{t-1} + \delta_t; \mathcal{B}_t)$
5: Orthogonalize the momentum matrix $O_t \in \arg\min_{O\in\mathbb{R}^{m\times n}} \|O - M_t\|_F$, s.t. $O^\top O = I_n$
6: Update parameter $W_t \leftarrow W_{t-1} - \eta O_t$
7: Construct momentum-guided perturbation $\delta_{t+1} \leftarrow \rho \frac{M_t}{\|M_t\|_F}$
8: **end for**

B.3 Theoretical Analysis for Algorithm 2

Let $E_t = M_t - \nabla f(W_{t-1} + \delta_t)$. Using $M_t = \mu M_{t-1} + (1-\mu)\nabla\ell(W_{t-1} + \delta_t)$, we have

$$\begin{aligned} E_t &= \mu M_{t-1} + (1-\mu)\nabla\ell(W_{t-1} + \delta_t) - \nabla f(W_{t-1} + \delta_t) \\ &= \mu(M_{t-1} - \nabla f(W_{t-2} + \delta_{t-1})) + \mu(\nabla f(W_{t-2} + \delta_{t-1}) - \nabla f(W_{t-1} + \delta_t)) \\ &\quad + (1-\mu)(\nabla\ell(W_{t-1} + \delta_t) - \nabla f(W_{t-1} + \delta_t)) \\ &= \mu E_{t-1} + \mu(\nabla f(W_{t-2} + \delta_{t-1}) - \nabla f(W_{t-1} + \delta_t)) \\ &\quad + (1-\mu)(\nabla\ell(W_{t-1} + \delta_t) - \nabla f(W_{t-1} + \delta_t)). \end{aligned} \tag{60}$$

Hence

$$\begin{aligned} \|E_t\|_* \le \mu\|E_{t-1}\|_* &+ \mu\|\nabla f(W_{t-2} + \delta_{t-1}) - \nabla f(W_{t-1} + \delta_t)\|_* \\ &+ (1-\mu)\|\nabla\ell(W_{t-1} + \delta_t) - \nabla f(W_{t-1} + \delta_t)\|_*. \end{aligned} \tag{61}$$

Conditioned on $\mathcal{F}_{t-1}$, both E_{t-1} and $\nabla f(W_{t-2} + \delta_{t-1}) - \nabla f(W_{t-1} + \delta_t)$ are deterministic. Therefore, the cross term vanishes:

$$\begin{aligned} \mathbb{E}\big[\|E_t\|_F^2 \mid \mathcal{F}_{t-1}\big] &= \mathbb{E}[\|\mu(E_{t-1} + \nabla f(W_{t-2} + \delta_{t-1}) - \nabla f(W_{t-1} + \delta_t)) \\ &\qquad +(1-\mu)(\nabla\ell(W_{t-1} + \delta_t) - \nabla f(W_{t-1} + \delta_t))\|_F^2 \mid \mathcal{F}_{t-1}\big] \\ &= \mu^2\|E_{t-1} + \nabla f(W_{t-2} + \delta_{t-1}) - \nabla f(W_{t-1} + \delta_t)\|_F^2 \\ &\quad + (1-\mu)^2\mathbb{E}\big[\|(\nabla\ell(W_{t-1} + \delta_t) - \nabla f(W_{t-1} + \delta_t))\|_F^2 \mid \mathcal{F}_{t-1}\big]. \end{aligned} \tag{62}$$

By Assumption 2.2, we have

$$\begin{aligned} \|\nabla f(W_{t-2} + \delta_{t-1}) - \nabla f(W_{t-1} + \delta_t)\|_F &\le \|\nabla f(W_{t-2} + \delta_{t-1}) - \nabla f(W_{t-1} + \delta_t)\|_* \\ &\le L\|(W_{t-2} + \delta_{t-1}) - (W_{t-1} + \delta_t)\|_{\mathrm{op}} \\ &\le L\|W_{t-1} - W_{t-2}\|_{\mathrm{op}} + L\|\delta_t - \delta_{t-1}\|_{\mathrm{op}}. \end{aligned} \tag{63}$$

Since $W_{t-1} = W_{t-2} - \eta U_{t-1}$ and $\|U_{t-1}\|_{\mathrm{op}} \le 1$,

$$\|W_{t-1} - W_{t-2}\|_{\mathrm{op}} \le \eta.$$

Also $\|\delta_t\|_{\mathrm{op}} \le \|\delta_t\|_F = \rho$, hence

$$\|\delta_t - \delta_{t-1}\|_{\mathrm{op}} \le 2\rho.$$

Therefore,

$$\|\nabla f(W_{t-2} + \delta_{t-1}) - \nabla f(W_{t-1} + \delta_t)\|_F \le L(\eta + 2\rho). \tag{64}$$

For any $\alpha > 0$, we have

$$\|A + B\|_F^2 \le (1+\alpha)\|A\|_F^2 + \Big(1 + \frac{1}{\alpha}\Big)\|B\|_F^2.$$

Applying the inequality above together with Assumption 2.3 to $A = E_{t-1}$, $B = \nabla f(W_{t-2} + \delta_{t-1}) - \nabla f(W_{t-1} + \delta_t)$ in Eq. (62), we obtain

$$\mathbb{E}\big[\|E_t\|_F^2 \mid \mathcal{F}_{t-1}\big]$$

$$\leq \mu^2(1+\alpha)\|E_{t-1}\|_F^2 + \mu^2\Big(1+\frac{1}{\alpha}\Big)\|\nabla f(W_{t-2}+\delta_{t-1}) - \nabla f(W_{t-1}+\delta_t)\|_F^2 + (1-\mu)^2\frac{\sigma^2}{B}. \tag{65}$$

For $\mu \in (0,1)$, choose $\alpha := \frac{1-\mu^2}{2\mu^2}$. Then $\mu^2(1+\alpha) = \frac{1+\mu^2}{2}$ and $\mu^2\Big(1+\frac{1}{\alpha}\Big) = \mu^2\frac{1+\mu^2}{1-\mu^2}$. Using Eq. (64), we get

$$\mathbb{E}\big[\|E_t\|_F^2 \mid \mathcal{F}_{t-1}\big] \leq \frac{1+\mu^2}{2}\|E_{t-1}\|_F^2 + \frac{\mu^2(1+\mu^2)}{1-\mu^2}L^2(\eta+2\rho)^2 + (1-\mu)^2\frac{\sigma^2}{B}. \tag{66}$$

For $\mu = 0$, we have $E_t = \nabla\ell(W_{t-1}+\delta_t;\mathcal{B}_t) - \nabla f(W_{t-1}+\delta_t)$, and hence $\mathbb{E}[\|E_t\|_F^2 \mid \mathcal{F}_{t-1}] \leq \frac{\sigma^2}{B}$ by Assumption 2.3. This is bounded by the right-hand side of Eq. (66) evaluated at $\mu = 0$, so the same recursion remains valid.

Then Eq. (66) implies

$$\mathbb{E}\|E_t\|_F^2 \leq \frac{1+\mu^2}{2}\mathbb{E}\|E_{t-1}\|_F^2 + \frac{\mu^2(1+\mu^2)}{1-\mu^2}L^2(\eta+2\rho)^2 + (1-\mu)^2\frac{\sigma^2}{B}. \tag{67}$$

Unrolling gives

$$\begin{aligned}
\mathbb{E}\|E_t\|_F^2 &\leq \left(\frac{1+\mu^2}{2}\right)^{t-1}\mathbb{E}\|E_1\|_F^2 + \frac{1-\left(\frac{1+\mu^2}{2}\right)^{t-1}}{1-\left(\frac{1+\mu^2}{2}\right)}\left(\frac{\mu^2(1+\mu^2)}{1-\mu^2}L^2(\eta+2\rho)^2 + (1-\mu)^2\frac{\sigma^2}{B}\right) \\
&\leq \left(\frac{1+\mu^2}{2}\right)^{t-1}\mathbb{E}\|E_1\|_F^2 + \frac{2}{1-\mu^2}\left(\frac{\mu^2(1+\mu^2)}{1-\mu^2}L^2(\eta+2\rho)^2 + (1-\mu)^2\frac{\sigma^2}{B}\right) \\
&= \left(\frac{1+\mu^2}{2}\right)^{t-1}\mathbb{E}\|E_1\|_F^2 + \frac{2\mu^2(1+\mu^2)}{(1-\mu^2)^2}L^2(\eta+2\rho)^2 + \frac{2(1-\mu)}{1+\mu}\frac{\sigma^2}{B}.
\end{aligned} \tag{68}$$

Let $r = \min\{m,n\}$. Since $\|A\|_* \leq \sqrt{r}\,\|A\|_F$, Jensen's inequality gives

$$\mathbb{E}\|E_t\|_* \leq \sqrt{r}\,\mathbb{E}\|E_t\|_F \leq \sqrt{r}\,\sqrt{\mathbb{E}\|E_t\|_F^2}. \tag{69}$$

Using Eq. (68) and $\sqrt{x+y+z} \leq \sqrt{x}+\sqrt{y}+\sqrt{z}$, we obtain

$$\mathbb{E}\|E_t\|_* \leq \sqrt{r}\left(\frac{1+\mu^2}{2}\right)^{\frac{t-1}{2}}\sqrt{\mathbb{E}\|E_1\|_F^2} + \frac{\sqrt{2r}\,\mu\sqrt{1+\mu^2}}{1-\mu^2}L(\eta+2\rho) + \sqrt{\frac{2r(1-\mu)}{1+\mu}}\,\frac{\sigma}{\sqrt{B}}. \tag{70}$$

Summing Eq. (70) over $t = 1,\ldots,T$ gives

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\|E_t\|_* \\
&\leq \sqrt{r}\,\sqrt{\mathbb{E}\|E_1\|_F^2}\sum_{t=1}^{T}\left(\frac{1+\mu^2}{2}\right)^{\frac{t-1}{2}} + \frac{\sqrt{2r}\,\mu\sqrt{1+\mu^2}}{1-\mu^2}TL(\eta+2\rho) + \sqrt{\frac{2r(1-\mu)}{1+\mu}}\,\frac{T\sigma}{\sqrt{B}}.
\end{aligned} \tag{71}$$

Since $\sum_{t=1}^{T}\left(\frac{1+\mu^2}{2}\right)^{\frac{t-1}{2}} = \sum_{s=0}^{T-1}\left(\sqrt{\frac{1+\mu^2}{2}}\right)^s = \frac{1-\left(\sqrt{\frac{1+\mu^2}{2}}\right)^T}{1-\sqrt{\frac{1+\mu^2}{2}}}$, we obtain

$$\begin{aligned}
\frac{2}{T}\sum_{t=1}^{T}\mathbb{E}\|E_t\|_* \leq{}& \frac{2\sqrt{r}\,\sqrt{\mathbb{E}\|E_1\|_F^2}}{T}\cdot\frac{1-\left(\sqrt{\frac{1+\mu^2}{2}}\right)^T}{1-\sqrt{\frac{1+\mu^2}{2}}} + \frac{2\sqrt{2r}\,\mu\sqrt{1+\mu^2}}{1-\mu^2}L(\eta+2\rho) \\
&+ 2\sqrt{\frac{2r(1-\mu)}{1+\mu}}\,\frac{\sigma}{\sqrt{B}}.
\end{aligned} \tag{72}$$

Using $\frac{1-\left(\sqrt{\frac{1+\mu^2}{2}}\right)^T}{1-\sqrt{\frac{1+\mu^2}{2}}} \leq \frac{4}{1-\mu^2}$, we further have

$$\frac{2}{T}\sum_{t=1}^{T}\mathbb{E}\|E_t\|_* \leq \frac{8\sqrt{r}\,\sqrt{\mathbb{E}\|E_1\|_F^2}}{(1-\mu^2)T} + \frac{2\sqrt{2r}\,\mu\sqrt{1+\mu^2}}{1-\mu^2}\,L(\eta+2\rho) + 2\sqrt{\frac{2r(1-\mu)}{1+\mu}}\,\frac{\sigma}{\sqrt{B}}. \tag{73}$$

By combining equation Eq. (26) and Eq. (72), we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_* \\
&\leq \frac{f(W_0)-f_*}{\eta T} + \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}\|E_t\|_* + 2L\rho + \frac{L\eta}{2} \\
&\leq \frac{f(W_0)-f_*}{\eta T} + \frac{8\sqrt{r}}{(1-\mu^2)T}\sqrt{2\mu^2L(f(W_0)-f_*) + (1-\mu)^2\frac{\sigma^2}{B}} \\
&\quad + \frac{2\sqrt{2r}\,\mu\sqrt{1+\mu^2}}{1-\mu^2}\,L(\eta+2\rho) + 2\sqrt{\frac{2r(1-\mu)}{1+\mu}}\,\frac{\sigma}{\sqrt{B}} + 2L\rho + \frac{L\eta}{2} \\
&\lesssim \frac{f(W_0)-f_*}{\eta T} + \frac{\sqrt{r}}{(1-\mu^2)T}\sqrt{\mu^2L(f(W_0)-f_*) + (1-\mu)^2\frac{\sigma^2}{B}} \\
&\quad + \frac{\sqrt{r}\,\mu\sqrt{1+\mu^2}}{1-\mu^2}\,L(\eta+\rho) + \sqrt{\frac{r(1-\mu)}{1+\mu}}\,\frac{\sigma}{\sqrt{B}} + L\rho + L\eta.
\end{aligned} \tag{74}$$

Let $\eta = \rho = \frac{1}{T^{\frac{3}{4}}}$ and $1-\mu = \frac{1}{T^{\frac{1}{2}}}$, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_* \\
&\leq \frac{f(W_0)-f_*+\sqrt{r}L}{T^{\frac{1}{4}}} + \frac{\sqrt{r}\sigma}{\sqrt{B}T^{\frac{1}{4}}} + \frac{\sqrt{rL(f(W_0)-f_*)}}{T^{\frac{1}{2}}} + \frac{L}{T^{\frac{3}{4}}} \\
&= \mathcal{O}\left(\frac{f(W_0)-f_\star+\sqrt{r}\,L}{T^{1/4}} + \frac{\sqrt{r}\,\sigma}{\sqrt{B}\,T^{1/4}} + \frac{\sqrt{rL(f(W_0)-f_\star)}}{T^{1/2}} + \frac{L}{T^{3/4}}\right).
\end{aligned} \tag{75}$$

Moreover, according to Eq. (57), if we set $\eta = \rho = \frac{1}{T^{\frac{3}{4}}}$ and $1-\mu = \frac{1}{T^{\frac{1}{2}}}$, we further have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t+\delta_{t+1})\|_* \\
&= \mathcal{O}\left(\frac{f(W_0)-f_\star+\sqrt{r}\,L}{T^{1/4}} + \frac{\sqrt{r}\,\sigma}{\sqrt{B}\,T^{1/4}} + \frac{\sqrt{rL(f(W_0)-f_\star)}}{T^{1/2}} + \frac{L}{T^{3/4}}\right).
\end{aligned} \tag{76}$$

B.4 Proof of Theorem 4.6

We first establish a direction-sensitive comparison between a slope proxy and the population-gradient perturbation.

Lemma B.5 (Achieved-linearization deviation)**.** *For every $V \in \mathbb{R}^{m\times n}$,*

$$0 \le \mathcal{E}_t(V) \le 2\rho \left\| V - \nabla f(W_{t-1}) \right\|_F . \tag{77}$$

Proof. Let $g_t := \nabla f(W_{t-1})$. By Eq. (12),

$$\mathcal{E}_t(V) = \rho \left\langle g_t, \mathsf{n}(g_t) - \mathsf{n}(V) \right\rangle . \tag{78}$$

Since $\mathsf{n}(g_t)$ maximizes $\langle g_t, U\rangle$ over $\|U\|_F \le 1$, it follows immediately that $\mathcal{E}_t(V) \ge 0$.

Furthermore,

$$\mathcal{E}_t(V) = \rho \left\langle g_t - V, \mathsf{n}(g_t) - \mathsf{n}(V) \right\rangle + \rho \left\langle V, \mathsf{n}(g_t) - \mathsf{n}(V) \right\rangle . \tag{79}$$

Because $\mathsf{n}(V)$ maximizes $\langle V, U\rangle$ over $\|U\|_F \le 1$,

$$\left\langle V, \mathsf{n}(g_t) - \mathsf{n}(V) \right\rangle \le 0. \tag{80}$$

Therefore, by Cauchy–Schwarz,

$$\begin{aligned} \mathcal{E}_t(V) &\le \rho \left\| g_t - V \right\|_F \left\| \mathsf{n}(g_t) - \mathsf{n}(V) \right\|_F \\ &\le 2\rho \left\| g_t - V \right\|_F , \end{aligned} \tag{81}$$

where the last inequality uses $\|\mathsf{n}(g_t)\|_F \le 1$ and $\|\mathsf{n}(V)\|_F \le 1$. The argument also covers $g_t = 0$ or $V = 0$ under the convention $\mathsf{n}(0) = 0$. □

B.4.1 Stability of the MuSAM perturbation

Let $G_t^{\mathrm{clean}} := \nabla \ell(W_{t-1}; \mathcal{B}_t)$ denote the raw stochastic gradient used to construct the MuSAM perturbation. Applying Lemma B.5 gives

$$\mathcal{E}_t(G_t^{\mathrm{clean}}) \le 2\rho \left\| G_t^{\mathrm{clean}} - \nabla f(W_{t-1}) \right\|_F . \tag{82}$$

Taking expectation and applying Jensen's inequality together with Assumption 2.3, we obtain

$$\begin{aligned} \mathbb{E}\left[\mathcal{E}_t(G_t^{\mathrm{clean}})\right] &\le 2\rho\, \mathbb{E}\left[\left\| G_t^{\mathrm{clean}} - \nabla f(W_{t-1}) \right\|_F\right] \\ &\le 2\rho \sqrt{\mathbb{E}\left[\left\| G_t^{\mathrm{clean}} - \nabla f(W_{t-1}) \right\|_F^2\right]} \\ &\le \frac{2\rho\sigma}{\sqrt{B}}. \end{aligned} \tag{83}$$

This proves Eq. (14).

B.4.2 Stability of the MoMuSAM perturbation

For $s \ge 1$, define $X_s := W_{s-1} + \delta_s$ and $G_s := \nabla \ell(X_s; \mathcal{B}_s)$. Unrolling the momentum recursion from $M_0 = 0$ gives, for every $t \ge 2$,

$$M_{t-1} = (1-\mu) \sum_{s=1}^{t-1} \mu^{t-1-s} G_s. \tag{84}$$

Define the bias-corrected momentum $\overline{M}_{t-1} := \frac{M_{t-1}}{1-\mu^{t-1}} = \sum_{s=1}^{t-1} w_{t,s} G_s$, where $w_{t,s} := \frac{(1-\mu)\mu^{t-1-s}}{1-\mu^{t-1}}$. These weights are nonnegative and satisfy $\sum_{s=1}^{t-1} w_{t,s} = 1$.

Since $1 - \mu^{t-1} > 0$, positive rescaling does not change the normalized direction: $\mathsf{n}(\overline{M}_{t-1}) = \mathsf{n}(M_{t-1})$. Consequently, $\mathcal{E}_t(\overline{M}_{t-1}) = \mathcal{E}_t(M_{t-1})$. Thus, $\overline{M}_{t-1}$ is used only for the analysis and induces exactly the same perturbation direction as the momentum appearing in MoMuSAM.

Define the stochastic-gradient noise $Z_s := G_s - \nabla f(X_s)$. We decompose the bias-corrected momentum error as

$$\overline{M}_{t-1} - \nabla f(W_{t-1}) = \underbrace{\sum_{s=1}^{t-1} w_{t,s} \left(\nabla f(X_s) - \nabla f(W_{t-1})\right)}_{\mathsf{B}_t} + \underbrace{\sum_{s=1}^{t-1} w_{t,s} Z_s}_{\mathsf{N}_t}. \tag{85}$$

Here, B_t is the tracking component and N_t is the martingale-noise component.

Bounding the tracking component. The exact-polar update satisfies $W_j = W_{j-1} - \eta O_j$ and $\|O_j\|_{\mathrm{op}} \le 1$. Hence, for every $s \le t-1$,

$$\|W_{t-1} - W_{s-1}\|_{\mathrm{op}} \le \sum_{j=s}^{t-1} \|W_j - W_{j-1}\|_{\mathrm{op}} \le (t-s)\eta. \tag{86}$$

Since $\|\delta_s\|_{\mathrm{op}} \le \|\delta_s\|_F \le \rho$, we have

$$\|X_s - W_{t-1}\|_{\mathrm{op}} \le (t-s)\eta + \rho. \tag{87}$$

Using $\|A\|_F \le \|A\|_*$, Assumption 2.2, and $\sum_{s=1}^{t-1} w_{t,s} = 1$, we obtain

$$\begin{aligned}
\|\mathsf{B}_t\|_F &\le \sum_{s=1}^{t-1} w_{t,s} \left\|\nabla f(X_s) - \nabla f(W_{t-1})\right\|_F \\
&\le \sum_{s=1}^{t-1} w_{t,s} \left\|\nabla f(X_s) - \nabla f(W_{t-1})\right\|_* \\
&\le L \sum_{s=1}^{t-1} w_{t,s} \|X_s - W_{t-1}\|_{\mathrm{op}} \\
&\le L \left(\rho + \eta \sum_{s=1}^{t-1} w_{t,s}(t-s)\right).
\end{aligned} \tag{88}$$

The weighted delay admits the exact expression

$$\begin{aligned}
A_t(\mu) &:= \sum_{s=1}^{t-1} w_{t,s}(t-s) \\
&= \frac{1-\mu}{1-\mu^{t-1}} \sum_{k=0}^{t-2} (k+1)\mu^k \\
&= \frac{1 - t\mu^{t-1} + (t-1)\mu^t}{(1-\mu)(1-\mu^{t-1})}.
\end{aligned} \tag{89}$$

Moreover, $1 - t\mu^{t-1} + (t-1)\mu^t \le 1 - \mu^{t-1}$, because the difference between the right- and left-hand sides is $(t-1)\mu^{t-1}(1-\mu) \ge 0$. Therefore, $A_t(\mu) \le \frac{1}{1-\mu}$. Substituting this into Eq. (88) yields

$$\|\mathsf{B}_t\|_F \le L\big(\rho + \eta A_t(\mu)\big) \le L\left(\rho + \frac{\eta}{1-\mu}\right). \tag{90}$$

Bounding the martingale-noise component. Let $\mathcal{F}_{s-1}$ denote the history before sampling $\mathcal{B}_s$. Since X_s is $\mathcal{F}_{s-1}$-measurable, the stochastic-oracle assumption gives $\mathbb{E}[Z_s \mid \mathcal{F}_{s-1}] = 0$ and $\mathbb{E}[\|Z_s\|_F^2 \mid \mathcal{F}_{s-1}] \le \frac{\sigma^2}{B}$. For $i < j$, Z_i is $\mathcal{F}_{j-1}$-measurable. Thus,

$$\mathbb{E}\langle Z_i, Z_j\rangle = \mathbb{E}[\mathbb{E}[\langle Z_i, Z_j\rangle \mid \mathcal{F}_{j-1}]] = \mathbb{E}[\langle Z_i, \mathbb{E}[Z_j \mid \mathcal{F}_{j-1}]\rangle] = 0. \tag{91}$$

It follows that

$$\mathbb{E}\|\mathsf{N}_t\|_F^2 = \sum_{s=1}^{t-1} w_{t,s}^2 \mathbb{E}\|Z_s\|_F^2$$

$$\leq \frac{\sigma^2}{B}\sum_{s=1}^{t-1} w_{t,s}^2$$
$$= \frac{\sigma^2}{B}\frac{(1-\mu)^2}{(1-\mu^{t-1})^2}\sum_{k=0}^{t-2}\mu^{2k}$$
$$= \frac{\sigma^2}{B}\frac{1-\mu}{1+\mu}\frac{1+\mu^{t-1}}{1-\mu^{t-1}}. \tag{92}$$

Define $\Gamma_t(\mu) := \frac{1-\mu}{1+\mu}\frac{1+\mu^{t-1}}{1-\mu^{t-1}}$. By Jensen's inequality,

$$\mathbb{E}\|\mathsf{N}_t\|_F \leq \sqrt{\Gamma_t(\mu)}\frac{\sigma}{\sqrt{B}}. \tag{93}$$

Combining Eqs. (85), (90) and (93) with the triangle inequality gives

$$\mathbb{E}\left\|\overline{M}_{t-1} - \nabla f(W_{t-1})\right\|_F \leq L\left(\rho + \frac{\eta}{1-\mu}\right) + \sqrt{\Gamma_t(\mu)}\frac{\sigma}{\sqrt{B}}. \tag{94}$$

Applying Lemma B.5 and using $\mathcal{E}_t(\overline{M}_{t-1}) = \mathcal{E}_t(M_{t-1})$, we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{E}_t(M_{t-1})] &= \mathbb{E}[\mathcal{E}_t(\overline{M}_{t-1})] \\ &\leq 2\rho\,\mathbb{E}\left\|\overline{M}_{t-1} - \nabla f(W_{t-1})\right\|_F \\ &\leq 2\rho\left[L\left(\rho + \frac{\eta}{1-\mu}\right) + \sqrt{\frac{1-\mu}{1+\mu}\frac{1+\mu^{t-1}}{1-\mu^{t-1}}}\frac{\sigma}{\sqrt{B}}\right].\end{aligned} \tag{95}$$

This proves the finite-time bound Eq. (15).

B.4.3 Average stability under the prescribed schedule

We first bound the average of $\sqrt{\Gamma_t(\mu)}$. Let $a := 1-\mu$. For $k = t-1 \geq 1$,

$$\Gamma_{k+1}(\mu) = \frac{a}{2-a}\frac{1+(1-a)^k}{1-(1-a)^k}. \tag{96}$$

For $a \in (0,1)$, Bernoulli's inequality gives $(1-a)^{-k} \geq (1+a)^k \geq 1+ak$, and therefore $(1-a)^k \leq \frac{1}{1+ak}$. Consequently, $1-(1-a)^k \geq \frac{ak}{1+ak}$. Using $1+(1-a)^k \leq 2$ and $2-a \geq 1$, we obtain

$$\Gamma_{k+1}(\mu) \leq a\frac{2(1+ak)}{ak} = 2a + \frac{2}{k}. \tag{97}$$

Hence, $\sqrt{\Gamma_{k+1}(\mu)} \leq \sqrt{2a} + \sqrt{\frac{2}{k}}$. Since $\sum_{k=1}^{N} k^{-1/2} \leq 2\sqrt{N}$, we have

$$\frac{1}{T-1}\sum_{t=2}^{T}\sqrt{\Gamma_t(\mu)} \leq \sqrt{2(1-\mu)} + \frac{2\sqrt{2}}{\sqrt{T-1}}. \tag{98}$$

It remains to include the initial perturbation. Since $M_0 = 0$, $\mathsf{n}(M_0) = 0$, and $\delta_1 = 0$,

$$\mathcal{E}_1(M_0) = L_1(\nabla f(W_0)) - L_1(M_0) = \rho\|\nabla f(W_0)\|_F. \tag{99}$$

For completeness, the gradient-gap inequality used below follows directly from Assumptions 2.1 and 2.2. The corresponding descent lemma gives

$$\begin{aligned}f\left(W_0 - \frac{1}{L}\nabla f(W_0)\right) &\leq f(W_0) - \frac{1}{L}\|\nabla f(W_0)\|_F^2 + \frac{1}{2L}\|\nabla f(W_0)\|_{\mathrm{op}}^2 \\ &\leq f(W_0) - \frac{1}{2L}\|\nabla f(W_0)\|_F^2.\end{aligned} \tag{100}$$

Since $f \geq f_\star$, this implies $\|\nabla f(W_0)\|_F \leq \sqrt{2L\big(f(W_0) - f_\star\big)}$. Therefore, $\frac{1}{T}\mathbb{E}[\mathcal{E}_1(M_0)] \leq \frac{\rho}{T}\sqrt{2L\big(f(W_0) - f_\star\big)}$.

Averaging Eq. (95) over $t = 2, \ldots, T$, using Eq. (98), and adding the initial term gives

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{E}_t(M_{t-1})] \\
&\leq \frac{\rho}{T}\sqrt{2L\big(f(W_0) - f_\star\big)} + \frac{2\rho(T-1)}{T}\left[L\left(\rho + \frac{\eta}{1-\mu}\right) + \frac{\sigma}{\sqrt{B}}\left(\sqrt{2(1-\mu)} + \frac{2\sqrt{2}}{\sqrt{T-1}}\right)\right].
\end{aligned} \tag{101}$$

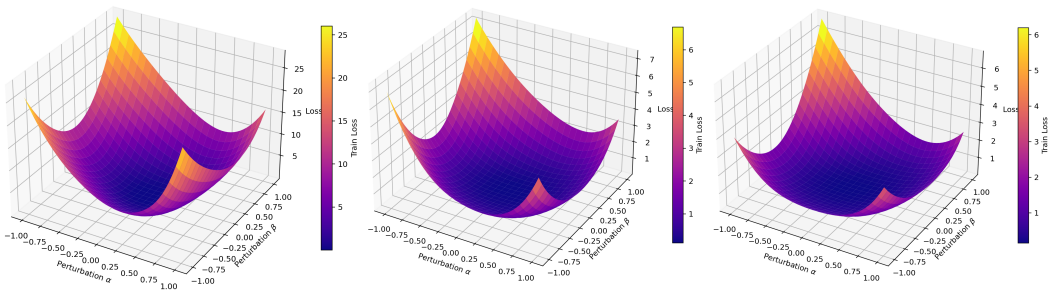
Now choose $\eta = \rho = T^{-3/4}$ and $1 - \mu = T^{-1/2}$. It follows that

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{E}_t(M_{t-1})] &= \mathcal{O}\left(\frac{L + \sigma/\sqrt{B}}{T} + \frac{\sigma}{\sqrt{B}\,T^{5/4}} + \frac{L}{T^{3/2}} + \frac{\sqrt{L(f(W_0) - f_\star)}}{T^{7/4}}\right) \\
&= \mathcal{O}\left(\frac{L + \sigma/\sqrt{B}}{T}\right).
\end{aligned} \tag{102}$$

Finally, averaging Eq. (83) over $t = 1, \ldots, T$ gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\mathcal{E}_t(G_t^{\text{clean}})\big] \leq \frac{2\rho\sigma}{\sqrt{B}} = \mathcal{O}\left(\frac{\sigma}{\sqrt{B}\,T^{3/4}}\right), \tag{103}$$

where the last equality uses $\rho = T^{-3/4}$. Together, Eqs. (102) and (103) prove Eq. (16). □



(a) MUON (b) MUSAM (c) MOMUSAM

Figure 4: 3D loss landscapes for different algorithms.

Table 2: Test Accuracy under different datasets.

Dataset	SAM	MSAM	VASSO	MUON	MUSAM	MOMUSAM
FMNIST	$93.83_{\pm 0.11}$	$93.74_{\pm 0.14}$	$93.86_{\pm 0.08}$	$94.10_{\pm 0.06}$	$94.27_{\pm 0.08}$	$94.35_{\pm 0.17}$
CIFAR-10	$89.27_{\pm 0.05}$	$89.40_{\pm 0.07}$	$89.26_{\pm 0.25}$	$90.72_{\pm 0.23}$	$91.11_{\pm 0.18}$	$91.21_{\pm 0.19}$
CIFAR-100	$68.98_{\pm 0.29}$	$68.48_{\pm 0.14}$	$68.96_{\pm 0.33}$	$74.18_{\pm 0.20}$	$74.62_{\pm 0.17}$	$75.14_{\pm 0.16}$

C ADDITIONAL EXPERIMENTS

Experimental setup. For the baseline optimizers, we tune the hyperparameters by grid search and report the best-performing configuration for each method. In contrast, for the proposed MUSAM and MOMUSAM, the hyperparameter choices are derived directly from the theoretical prescriptions in Theorems 4.1 and 4.2. In particular, the resulting closed-form settings depend only on the total number of iterations T, and are used throughout our experiments without additional task-specific search. This highlights an important practical advantage of our proposed methods: their hyperparameters are not only theoretically grounded, but also straightforward to instantiate in practice. All image-classification experiments related to convergence and sensitivity analysis are repeated over 10 random seeds, and we report the mean and standard deviation. For loss-landscape visualization, we follow Li et al. (2018) and plot the landscape along two random filter-normalized perturbation directions. The experiments were conducted on a host machine equipped with an Intel(R) Xeon(R) W9-3475X CPU running at 2.20 GHz (maximum turbo frequency 4.80 GHz), with 36 physical cores and 72 threads, 256 GB DDR5 ECC RAM, and a single NVIDIA(R) RTX(TM) A6000 GPU with 48 GB of memory.

Geometry analysis. The results in Figure 4, Tables 2 and 3 consistently support the central claim of our paper: introducing explicit sharpness awareness substantially strengthens Muon, and the momentum-guided design yields the strongest overall behavior. From Table 2, MUON already outperforms SAM-style baselines on all three image-classification benchmarks, confirming the benefit of matrix-aware orthogonalized updates; however, both MUSAM and MOMUSAM further improve upon MUON consistently, showing that matrix-aware update geometry and explicit sharpness control are indeed complementary. This gain becomes more pronounced on the harder dataset CIFAR-100, where MOMUSAM achieves the largest margin, suggesting that explicit local geometry control becomes increasingly valuable in more challenging and noisier optimization regimes. The geometry diagnostics in Table 3 and Figure 4 explain this trend. Relative to MUON, MUSAM sharply reduces the top Hessian eigenvalues, and MOMUSAM reduces them even further, indicating substantially flatter local curvature around the converged solution. This is also visually reflected in the 3D loss landscapes: while MUON already reaches a relatively smooth basin, MUSAM produces a visibly flatter neighborhood, and MOMUSAM yields the broadest and smoothest basin among the three. Together, these results suggest that the accuracy gains are not merely optimization artifacts, but are tightly coupled with improved local flatness. Moreover, the consistent advantage of MOMUSAM over MUSAM indicates that the quality of the perturbation signal is critical: by constructing the perturbation from the momentum state rather than a single minibatch gradient, MOMUSAM probes sharpness with a direction that is temporally aggregated, less noisy, and more informed by

Table 3: Top 5 Hessian eigenvalues.

Algorithm	λ_1	λ_2	λ_3	λ_4	λ_5
MUON	30.21	17.89	9.50	7.59	6.08
MUSAM	8.71	6.42	5.42	3.48	2.71
MOMUSAM	5.55	3.18	1.78	1.41	1.06

Table 4: Wall-clock time test accuracy of CIFAR-10 for different algorithms.

	Wall-clock time test accuracy (%)						
Algorithm	**200 sec**	**500 sec**	**800 sec**	**1100 sec**	**1400 sec**	**1700 sec**	**2000 sec**
SAM	68.68	74.26	78.27	81.94	84.65	85.84	86.53
MSAM	75.02	77.78	82.50	84.23	85.15	87.20	87.86
VASSO	68.40	76.17	79.17	85.26	85.35	86.33	87.64
Muon	83.74	85.67	88.38	88.96	89.34	89.82	90.02
MuSAM	83.44	86.61	87.57	89.10	89.49	89.60	89.91
MoMuSAM	84.30	87.27	88.45	89.22	89.67	89.93	90.23

Muon's accumulated optimization history, which in turn leads to a flatter solution and stronger final generalization.

Wall-clock time comparison. Table 4 provides a more practical view of the efficiency–performance tradeoff by comparing test accuracy under the same wall-clock budget. The key observation is that MOMUSAM consistently matches or exceeds MUON across almost the entire time range, and achieves the best final accuracy overall. This is particularly significant because, unlike MUSAM, MOMUSAM does not require an additional gradient computation to construct the perturbation; instead, it reuses the momentum state already maintained by Muon, and therefore preserves essentially the same per-iteration computational profile as MUON. In other words, the gain of MOMUSAM is obtained without paying an additional wall-clock overhead relative to the base matrix-aware optimizer. By contrast, although MUSAM also improves generalization, its extra clean-gradient evaluation makes its advantage less pronounced under a fixed time budget. These results therefore highlight a central strength of our design: MOMUSAM is not only more accurate than MUON, but also achieves this improvement at nearly the same computational cost, yielding a distinctly better accuracy–efficiency tradeoff in practical training.