When Muon Optimizer Meets Adversarial Training: A Theoretical and Empirical Study

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial training (AT) remains one of the most reliable empirical defenses against adversarial attacks. Its robustness critically depends on how the underlying min-max objective is optimized. In practice, Stochastic Gradient Descent (SGD) optimizer remains the default optimization choice for AT, whereas adaptive optimizers often improve standard training but may yield inferior robustness. Recently, the Muon optimizer, which orthogonalizes matrix-valued updates via an approximate polar decomposition, has achieved notable success in large-scale training, motivating the investigation of its behavior beyond standard training settings. Robustness claims in AT can be highly sensitive to the optimizer used in the outer minimization, yet optimizer choice is often treated as an implementation detail. A defense may appear robust under one training dynamic but degrade under heterogeneous or adaptive threat models. This raises a question: *how does orthogonalized outer optimization alter the robustness and training dynamics of AT, and in which regimes is it beneficial?* Focusing on this problem, this paper revisits the choice of optimizer as a security-relevant component of robustness. Theoretically, we show that Muon's polar-factor-based update, together with weight decay, induces a finite worst-case bound on matrix-weight trajectories without explicitly projecting the learned weights. Empirically, across five architectures and multiple ℓ_p threat models on CIFAR-scale settings, Muon often outperforms AdamW and remains competitive with SGD, while SGD remains stronger in some ViT and ImageNet settings. Our results show that outer-optimizer choice can materially change the observed robustness profile under the evaluated AT configurations. These results delineate the regimes in which orthogonalized updates are beneficial. Our findings therefore motivate treating the optimizer as a robustness-relevant design and evaluation variable rather than as an implementation detail.

1 INTRODUCTION

Deep neural networks are increasingly deployed in security-sensitive settings, yet their predictions can be manipulated by small, human-imperceptible adversarial perturbations (Szegedy et al., 2014; Goodfellow et al., 2015; Madry et al., 2018; Carlini and Wagner, 2017). To mitigate such potential threats to neural networks, robust training methods known as adversarial training (AT) have been proposed (Madry et al., 2018; Zhang et al., 2019b; Wong et al., 2020; Croce and Hein, 2022). These methods are based on robust optimization, aiming to maximize adversarial perturbations while minimizing empirical risk.

The outer optimization problem in AT is often noisier and more ill-conditioned than standard empirical risk minimization, because each update is computed on adversarially perturbed inputs produced by an inner maximization procedure. In this regime, optimizer-induced geometry can have a non-negligible effect on both clean accuracy and robust accuracy. Adaptive optimizers such as Adam and AdamW rescale each coordinate of the update by an estimate of its gradient second moment. The Stochastic Gradient Descent (SGD) optimizer and its momentum variants avoid this per-coordinate rescaling, and have been observed to yield greater robustness under the high-variance gradients of AT (Dabouei et al., 2022). Empirical studies have also validated their advantages in AT (Pang et al., 2021). Some studies have also focused on enhancing robustness by exploring the Sharpness-Aware Minimization (SAM) method, which perturbs model weights to seek a flatter loss landscape (Zhang et al., 2024; Zhou et al., 2025). These studies make the parameter dynamics of min-max optimization analytically

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.

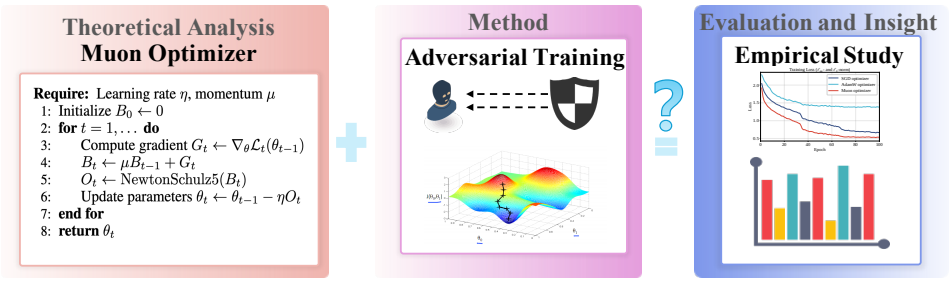


Figure 1: The framework of this study. The left part of the figure is adapted from the work (Jordan et al., 2024).

tractable, allowing robustness gains to be attributed to specific update-rule properties (e.g., implicit regularization of curvature or spectral norms) rather than treated as emergent artifacts.

A prior study (Pang et al., 2021) found Adam (Kingma and Ba, 2015) and AdamW (Loshchilov and Hutter, 2019) to be suboptimal for AT. Recently, the Muon optimizer (Jordan et al., 2024) has been introduced, which avoids expensive high-order projection algorithms while exploiting orthogonalized update geometry and enables a large-batch-friendly update that can be efficiently approximated by a small fixed number of Newton-Schulz iterations. This optimization method constrains the update direction (the polar factor of the momentum-accumulated gradient) to the Stiefel manifold, yielding orthogonalized, spectrally balanced parameter updates (though the weights W_t themselves need not lie on the manifold) (Liu et al., 2025). The Muon optimizer differs from other optimizers in that it orthogonalizes matrix-valued updates. For each 2D weight tensor, the Muon optimizer transforms the momentum update into its approximate polar factor, which can be realized via Newton–Schulz iterations. The update geometry is fundamentally different from SGD (Robbins and Monro, 1951; Bottou, 2012) and AdamW (Loshchilov and Hutter, 2019). The optimizer choice in AT can influence the optimization trajectory and consequently the robustness obtained under a fixed training objective. It changes robustness under heterogeneous threat models, yet this problem under the new geometry-aware optimizer remains unexplored.

To fill this research gap, we carry out a systematic theoretical and experimental study of the Muon optimizer in the AT paradigm, as illustrated in Figure 1. We first characterize an optimizer-side stability mechanism relevant to robustness. Under bounded orthogonalization error and positive weight decay, Muon's orthogonalized update admits a worst-case spectral-norm trajectory bound, which is structural rather than a tight robustness certificate. Each matrix-valued update has a controlled operator norm after orthogonalization. This property does not directly enforce small spectral norms of the learned weights. Instead, it limits worst-case spectral growth in the optimization trajectory. Empirically, we first adapt the Muon optimizer for AT, then conduct systematic experiments across multiple datasets and model architectures, to evaluate the robustness of AT with SGD, AdamW, SAM, and Muon. Across this grid, we observe that the choice of optimizer can induce large and consistent shifts in the robustness frontier: the matched AT settings can yield markedly different robustness profiles depending on the optimization dynamics. Across these settings, Muon generally outperforms AdamW and remains competitive with SGD in several regimes, while SGD is stronger in others. Compared with SAM, Muon should be viewed as an optimizer-geometry alternative rather than a direct replacement. It improves update geometry without introducing an additional weight-perturbation step, but it does not uniformly dominate SAM in robustness. To summarize, our key empirical findings are: **(1)** Muon constrains the update geometry: each matrix-valued update has a bounded operator norm after polar orthogonalization. This does not imply that Muon always produces smaller weight spectral norms than SGD, but it helps avoid the large spectral growth observed with AdamW in several of our settings. **(2)** On small-scale datasets like CIFAR-10 (Krizhevsky, 2009), Muon is on par with or superior to SGD on overparameterized CNNs (Zagoruyko and Komodakis, 2016). On ViTs (Dosovitskiy et al., 2021), Muon lags behind SGD yet avoids the severe degradation observed with AdamW under the tested recipes. However, on ImageNet (Deng et al., 2009), SGD remains stronger under the tested schedule, revealing that orthogonalized AT is sensitive to large-scale learning-rate and weight-decay design. **(3)** Muon is an efficient optimizer-geometry alternative rather than a strict replacement for SAM.

Our contributions are summarized as follows:

- **Optimizer-induced robustness dynamics.** We identify the outer optimizer as a previously underexplored source of variation in adversarial training dynamics. Under matched AT objectives and threat models, we show that different update geometries lead to distinct robustness trajectories and checkpoint behaviors.
- **Mechanistic analysis of Muon's spectral update geometry.** We show that Muon's orthogonalized update admits a trajectory-level spectral-norm stability ceiling and yields a nuclear-norm-aligned descent term for the adversarial objective. We explicitly characterize the scope of this analysis: it explains a stability mechanism, but it is not a certified robustness guarantee.
- **Systematic multi-norm robustness evaluation.** We conduct a systematic empirical comparison of SGD, AdamW, and Muon across multiple architectures, datasets, and threat norms. We further include representative SAM comparisons on WideResNet-34-10 (WRN-34-10) (Zagoruyko and Komodakis, 2016) and ViT-B (Dosovitskiy et al., 2021) to clarify the trade-off between orthogonalized updates and sharpness-aware perturbation-based optimization.
- **Threat model and scope**. We consider standard test-time adversarial robustness under white-box ℓ_p-bounded attacks. The attacker has full access to the deployed classifier and its gradients, but the proposed method changes only the training-time optimizer. We additionally report transfer-based and query-based attacks to cover restricted-information attacker settings.

2 Related Work

Optimizer design is of significant importance in AT, which relies on the min-max optimization mechanism (Madry et al., 2018; Zhang et al., 2019b). These studies postulate that a stronger inner attack yields a more robust defense model. If the attack is too weak, it can lead to gradient masking or an overestimation of robustness.

Most theoretical research on AT focuses on the dynamics of min-max optimization, while less attention is paid to the choice of optimizer itself (Wang et al., 2019; Gao et al., 2019; Deng et al., 2020). Early studies found that SGD optimization can lead to a more robust loss landscape when combined with regularization and data augmentation (Liu et al., 2020). AT with the SGD optimizer and weight decay regularization can achieve feature purification (Allen-Zhu and Li, 2022). Empirical studies have found it more suitable to apply SGD with momentum rather than other adaptive optimizers (Pang et al., 2021).

SAM minimizes both the loss value and its sharpness (Foret et al., 2021). It can help improve the robustness of neural networks (Zhang et al., 2024; Zhou et al., 2025). The implicit bias of SAM yields a better generalization solution than that of standard gradient descent (Andriushchenko and Flammarion, 2022). At the same time, SAM tends to "bounce" on both sides along the direction of maximum curvature. In more general cases, the effect of SAM is linked to the regularization of the Hessian spectrum (especially the maximum eigenvalue) (Bartlett et al., 2023). However, SAM requires two forward-backward passes per step: one for the gradient perturbation and another for the actual update (Foret et al., 2021).

Previous studies have stated that SGD-trained neural networks provide better robustness to input perturbations than those trained with adaptive gradient methods due to smaller Lipschitz constants (Ma et al., 2023). Recently, the Muon optimizer (Jordan et al., 2024; Liu et al., 2025) has been proposed for training new foundation models. When the Hessian satisfies the low-rank property, Muon admits a favorable first-order descent characterization under the nuclear norm geometry (Shen et al., 2025). Using a Lyapunov function, it has been proven that Muon with momentum and decoupled weight decay is stable and converges to the KKT point of the implicitly constrained problem (Chen et al., 2025). The Muon optimizer is built on Newton–Schulz iterative updates, and Bernstein and Newhouse (2025) constructed dual mappings for general network structures. This analytical framework can also be extended to the Muon optimizer. Newhouse et al. (2025) proposed training a Transformer with Lipschitz constraints, employing constraint techniques such as Muon and soft cap. Although Chen et al. (2025) analyze Muon's convergence behavior under spectral norm constraints in a general

nonconvex optimization setting, its effectiveness under AT remains unexplored. Recent work has further analyzed Muon from convergence, spectral, and modular-duality perspectives (Chen et al., 2025; Shen et al., 2025; Bernstein and Newhouse, 2025). These studies primarily concern standard optimization or general spectral geometry. Our scope is complementary: we use the reference Muon optimizer as a fixed representative of orthogonalized updates and study its interaction with adversarial training, rather than comparing or redesigning Muon variants.

3 Theoretical Analysis

3.1 Preliminaries

Adversarial Training (AT). Let $f_\theta : \mathcal{X} \to \mathbb{R}^K$ denote a classifier parameterized by θ, with logits $g_\theta(x)$ and a classification margin $m_\theta(x, y) := g_\theta(x)_y - \max_{y' \neq y} g_\theta(x)_{y'}$. AT solves the min-max problem

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in \Delta} \ell(f_\theta(x + \delta), y) \right], \tag{1}$$

where $\ell(\cdot, \cdot)$ is the classification loss.

Threat model. We consider the union of multiple ℓ_p threat models. Given a set of norms $\mathcal{P} \subseteq \{1, 2, \infty\}$ with per-norm radii $\{\varepsilon_p\}_{p\in\mathcal{P}}$, the perturbation set is

$$\Delta_{\text{union}} := \bigcup_{p\in\mathcal{P}} \{\delta \in \mathbb{R}^d : \|\delta\|_p \leq \varepsilon_p\}. \tag{2}$$

Muon update rule. For each 2D weight block $W_t \in \mathbb{R}^{m\times n}$, let $G_t = \nabla_W \mathcal{L}_{\text{adv}}(\theta_{t-1})$ denote the AT gradient and $M_t = \mu M_{t-1} + G_t$ the momentum-accumulated gradient. Muon applies the orthogonalized update

$$U_t = \text{polar}(M_t), \qquad W_{t+1} = (1 - \eta\lambda)W_t - \eta U_t, \tag{3}$$

where $\eta > 0$ is the learning rate, $\lambda \geq 0$ is weight decay, and $\text{polar}(\cdot)$ is approximated by a fixed number of Newton–Schulz iterations.

Notation. We use $\|\cdot\|_2$ for the spectral (operator-2) norm, $\|\cdot\|_F$ for the Frobenius norm, $\|\cdot\|_*$ for the nuclear norm, and $\sigma_i(\cdot)$ for singular values. $J_f(x)$ denotes the Jacobian of f at x. Per-layer weights are W_ℓ with spectral norm upper bound B_ℓ (to be established in Theorem 3.1). We analyze how the update rule of Eq. (3) governs $\|W_\ell\|_2$, the Lipschitz constant of m_θ, and the per-step descent on the AT objective defined in Eq. (1).

3.2 Spectral-Norm Stability Ceiling

This section proves that Muon's orthogonalized update admits a spectral-norm stability ceiling (Theorem 3.1). Theoretically, we show that Muon's orthogonalized update induces a spectral-norm stability ceiling on the weight trajectory, preventing unbounded spectral growth under the stated assumptions.

Proposition 3.1 (Robustness governed by the dual norm of input gradients)**.** *Under a first-order Taylor expansion of m_θ,*

$$m_\theta(x + \delta, y) \approx m_\theta(x, y) + \langle \nabla_x m_\theta(x, y), \delta \rangle.$$

Hence, by Hölder's inequality with $1/p + 1/q = 1$,

$$\max_{\|\delta\|_p \leq \varepsilon} |m_\theta(x + \delta, y) - m_\theta(x, y)| \approx \varepsilon \, \|\nabla_x m_\theta(x, y)\|_q,$$

so, under this first-order approximation, the linearized margin remains positive whenever $m_\theta(x, y) > \varepsilon \, \|\nabla_x m_\theta(x, y)\|_q$.

Proposition 3.2 (Bounding the input Jacobian with layer spectral norms)**.** *For a feedforward network $f(x) = W_L \phi(\cdots \phi(W_1 x))$ with L layers, if the activations ϕ are 1-Lipschitz (e.g., ReLU, leaky-ReLU), then*

$$\|J_f(x)\|_2 \leq \prod_{\ell=1}^{L} \|W_\ell\|_2 . \tag{4}$$

This gives $\|\nabla_x g_\theta(x)\|_2 \le C \cdot \prod_\ell \|W_\ell\|_2$*, where* C *absorbs activation Lipschitz constants and output projections.*

The Muon optimizer uses the Newton–Schulz approximation for polar decomposition (Jordan et al., 2024), whose update rule is defined in Eq. (3). The exact polar factor is orthogonal (or semi-orthogonal in the rectangular case) and has operator norm one. The practical finite-step Newton–Schulz update need not coincide with the exact polar factor. We therefore analyze the ideal polar update first and treat the Newton–Schulz approximation as a perturbation captured by the orthogonalization error. The following lemma extends prior analysis (Jordan et al., 2024; Higham, 2008).

Lemma 3.1. *Let* $A \in \mathbb{R}^{m\times n}$ *be full rank with compact SVD* $A = P\Sigma V^\top$*, and define* $\mathrm{Ortho}(A) = PV^\top$*. Then* $\|\mathrm{Ortho}(A)\|_2 = 1$*. Moreover, if* $m \ge n$*,*

$$\mathrm{Ortho}(A)^\top \mathrm{Ortho}(A) = I_n, \qquad A = \mathrm{Ortho}(A)(A^\top A)^{1/2}; \tag{5}$$

if $m < n$*, we have* $\mathrm{Ortho}(A)\,\mathrm{Ortho}(A)^\top = I_m$*,* $A = (AA^\top)^{1/2}\,\mathrm{Ortho}(A)$*. Equivalently,* $\mathrm{Ortho}(A)$ *is the Frobenius projection of* A *onto the corresponding rectangular Stiefel set.*

The proof of Lemma 3.1 is given in Appendix C.1.

Theorem 3.1. *Assume there is a quantifiable orthogonalization error* $\varepsilon_t \ge 0$*, such that*

$$\left\|\widetilde{U}_t\right\|_2 \le 1 + \varepsilon_t. \tag{6}$$

The exact-polar case corresponds to $\varepsilon_t = 0$*, whereas the finite-step Newton–Schulz implementation is covered by any quantifiable* $\varepsilon_t \ge 0$ *satisfying Eq. (6). The theorem does not require this approximation error to vanish.*

Under the assumption of Eq. (6) and $0 \le \eta\lambda \le 1$*, for any* $t \ge 0$*, the spectral norm satisfies the following recursive upper bound*

$$\|W_{t+1}\|_2 \le (1-\eta\lambda)\,\|W_t\|_2 + \eta\,(1+\varepsilon_t)\,. \tag{7}$$

Furthermore, if $\eta\lambda \in (0,1)$*, the following explicit trajectory upper bound holds*

$$\begin{aligned}\|W_t\|_2 &\le (1-\eta\lambda)^t\,\|W_0\|_2 + \frac{1-(1-\eta\lambda)^t}{\lambda} \\ &\quad + \eta\sum_{k=0}^{t-1}(1-\eta\lambda)^{t-1-k}\varepsilon_k.\end{aligned} \tag{8}$$

Specifically, if $\varepsilon_t \le \bar{\varepsilon}$ *for all* t*, then*

$$\begin{aligned}\|W_t\|_2 &\le (1-\eta\lambda)^t\,\|W_0\|_2 + \frac{1+\bar{\varepsilon}}{\lambda}\left(1-(1-\eta\lambda)^t\right) \\ &\le \max\left\{\|W_0\|_2\,, \frac{1+\bar{\varepsilon}}{\lambda}\right\}.\end{aligned} \tag{9}$$

The proof of Theorem 3.1 is given in Appendix C.2.

Theorem 3.2. *Fix* $\mathcal{P} \subseteq \{1,2,\infty\}$ *with radii* $\{\varepsilon_p\}_{p\in\mathcal{P}}$ *and let* $\Delta_{\mathrm{union}} = \bigcup_{p\in\mathcal{P}}\{\delta \in \mathbb{R}^d : \|\delta\|_p \le \varepsilon_p\}$*. Suppose every layer satisfies* $\|W_\ell\|_2 \le B_\ell$ *(as obtained from Theorem 3.1 under its stated assumptions), and define*

$$\hat{L}_p(\theta) := c_p C_{\mathrm{marg}} \prod_{\ell=1}^{L} B_\ell,\ C_{\mathrm{marg}} := \max_{j\ne y}\|e_y - e_j\|_2 = \sqrt{2}. \tag{10}$$

where c_p *is the norm-equivalence constant satisfying* $\|\delta\|_2 \le c_p\|\delta\|_p$ *for all* $\delta \in \mathbb{R}^d$ *(concretely,* $c_\infty = \sqrt{d}$*,* $c_2 = 1$*, and* $c_1 = 1$*, since* $\|\delta\|_2 \le \sqrt{d}\|\delta\|_\infty$ *and* $\|\delta\|_2 \le \|\delta\|_1$*). Furthermore,* C_{marg} *absorbs the output-layer projection from logits to margin. If*

$$m_\theta(x,y) \;>\; \max_{p\in\mathcal{P}} \hat{L}_p(\theta)\,\varepsilon_p, \tag{11}$$

then $m_\theta(x+\delta, y) > 0$ *for all* $\delta \in \Delta_{\mathrm{union}}$*.*

The proof of Theorem 3.2 is given in Appendix C.3. This is a structural sufficient condition rather than a numerically tight certificate, since the product of layerwise worst-case bounds can be highly conservative.

3.3 MUON'S GRADIENT DYNAMICS UNDER MIN-MAX OPTIMIZATION

Our main result (Theorem 3.3) shows that Muon descends the adversarial loss along the nuclear-norm direction with the per-step guarantee. The polar direction yields a nuclear-norm-aligned descent term.

Assumption 3.1. At a given outer iteration, let x^{adv} denote the adversarial example returned by the finite-step inner solver and held fixed during the parameter update. We view the adversarial loss L_{adv} as a function $F(W)$ of a particular weight block W. Assume that F is locally β-smooth, i.e., the following inequality holds

$$F(W+\Delta) \le F(W) + \langle \nabla F(W), \Delta \rangle + \frac{\beta}{2}\|\Delta\|_F^2, \tag{12}$$

where Δ represents any perturbation or increment to the parameter block W.

Theorem 3.3 (Muon optimizer on the adversarial loss for the min-max dynamics)**.** *In AT, assume F is β-smooth w.r.t. $\|\cdot\|_F$. Let $G = \nabla_W F(W)$ be the block gradient and let M be the momentum-accumulated gradient used by Muon. Let $\widetilde{U}$ be the inexact polar direction used in practice, satisfying*

$$\|\widetilde{U} - \mathrm{Ortho}(M)\|_F \le \delta_{\mathrm{orth}},\ \|\widetilde{U}\|_2 \le 1+\varepsilon,\ \|\widetilde{U}\|_F^2 \le r(1+\varepsilon)^2, \tag{13}$$

where r is $\mathrm{rank}(M)$ or any upper bound on it. Consider the update $W^+ = W - \eta\,\widetilde{U}$ with $\eta > 0$. Then the following inequality holds:

$$\begin{aligned} F(W^+) \le F(W) &- \eta\|M\|_* + \eta(1+\varepsilon)\|G - M\|_* \\ &+ \eta\sqrt{r}\,\|M\|_2\,\delta_{\mathrm{orth}} + \frac{\beta}{2}\eta^2\, r(1+\varepsilon)^2. \end{aligned} \tag{14}$$

The proof of Theorem 3.3 with orthogonalized component analysis is given in Appendix C.4.

4 EMPIRICAL RESULTS AND INSIGHTS

4.1 ADAPTING MUON FOR AT

We follow the multi-norm AT procedure (Croce and Hein, 2022). We instantiate the AT objective defined in Eq. (1) with APGD as the inner attack and Muon as the outer optimizer. The attack method used in our study is Auto Projected Gradient Descent (APGD) (Croce and Hein, 2020b). It replaces the fixed step size of PGD with an adaptive schedule and an automatic restart mechanism. For ℓ_∞ and ℓ_2 attacks, it uses $\mathrm{sign}(\nabla_x \ell)$ and $\nabla_x \ell / \|\nabla_x \ell\|_2$, respectively. For the ℓ_1 attack, the step ascends along the top-k coordinates of $|\nabla_x \ell|$ (a sparse steepest-ascent direction) and projects the iterate back onto the ℓ_1 ball of radius ε_1 after each step (Croce and Hein, 2022).

Algorithm 1 outlines the AT procedure. It uses a sampled-norm surrogate during training rather than solving the exact inner maximization over Δ_{union} at every minibatch. The worst-case union robustness is reported at evaluation. Various optimizers can be integrated as plugins within this AT framework. Algorithm 2 (Appendix B) is applied independently to each eligible matrix-valued parameter block, with a separate momentum buffer for each block. These blockwise updates together constitute one model-level optimizer step. The experimental protocol can be referred to Appendix D.1.

4.2 MAIN EVALUATION RESULTS ON CIFAR-10

The models selected for our experiment include PreActResNet-18 (He et al., 2016b), WRN-34-10 (Zagoruyko and Komodakis, 2016), WRN-34-20 (Zagoruyko and Komodakis, 2016), ViT-B (Dosovitskiy et al., 2021), and ViT-L (Dosovitskiy et al., 2021). Best-checkpoint selection is based on their ℓ_∞ robustness in the training-time evaluation. In our study, the best checkpoint is selected using a small validation subset during training, and the final reported performance is evaluated on the full test set. This protocol may introduce checkpoint-selection bias. The evaluation is performed using APGD-CE together with APGD-T (targeted APGD with DLR loss). The robust accuracy reported for each norm is the worst-case result over the cascade of two attacks. We compare Muon (Jordan et al., 2024), AdamW (Loshchilov and Hutter, 2019), and SGD (Robbins and Monro, 1951; Bottou,

Algorithm 1 Multi-Norm APGD Adversarial Training

Require: Model f_θ, data loader $\mathcal{D}$, threat norms $\mathcal{P} = \{p_1, \ldots, p_M\}$, radii $\{\varepsilon_i\}_{i=1}^M$, attack steps $\{T_i\}_{i=1}^M$, optimizer Opt (SGD / AdamW / Muon), adversarial loss $\ell(\cdot, \cdot)$, optional scheduler Scheduler

Initialize parameters θ (random or pretrained)
for epoch $= 1$ to K **do**
  for each minibatch $(x, y) \sim \mathcal{D}$ **do**
    $i \sim \mathrm{Unif}\{1, \ldots, M\}$
    $p \leftarrow p_i,\ \varepsilon \leftarrow \varepsilon_i,\ T \leftarrow T_i$
    $x_{\mathrm{adv}} \leftarrow \mathrm{APGD}(f_\theta, x, y;\ p, \varepsilon, T, \ell)$
    Opt.zero_grad()
    $L_{\mathrm{adv}} \leftarrow \ell(f_\theta(x_{\mathrm{adv}}), y)$
    Backprop(L_{adv})
    Opt.step() ▷ *If Muon: see Alg. 2 for the expanded step*
  end for
  if Scheduler is used **then**
    Scheduler.step()
  end if
end for

Table 1: AT results on PreActResNet-18 (He et al., 2016b).

Norm	Opt.	Best					Last				
		Clean	ℓ_∞	ℓ_2	ℓ_1	Union	Clean	ℓ_∞	ℓ_2	ℓ_1	Union
ℓ_∞+ℓ_1	SGD	**79.21**	**40.84**	**64.86**	**52.15**	**40.37**	81.22	35.77	64.16	48.00	35.07
	AdamW	66.26	36.83	55.19	45.59	36.53	69.98	32.78	55.36	43.76	32.38
	Muon	77.48	39.67	63.08	51.03	39.27	**84.81**	**38.58**	**67.24**	**50.90**	**37.83**
ℓ_∞	SGD	**78.60**	**45.81**	58.20	8.05	8.05	82.37	41.73	57.17	5.80	5.79
	AdamW	65.18	39.64	47.76	10.07	10.07	69.22	38.63	48.96	**7.87**	**7.87**
	Muon	78.04	43.86	**58.58**	**10.16**	**10.16**	**85.74**	**43.68**	**60.84**	6.00	6.00
ℓ_1	SGD	**84.08**	17.10	**59.15**	**53.05**	17.10	84.08	16.98	59.34	53.20	16.98
	AdamW	71.50	21.83	54.64	49.62	21.83	79.36	**19.09**	54.36	49.68	**19.09**
	Muon	76.30	**23.63**	57.75	51.71	**23.63**	**86.68**	18.19	**63.50**	**57.57**	18.19
ℓ_2	SGD	87.58	25.91	**65.53**	25.00	20.52	87.74	22.81	63.46	22.73	17.82
	AdamW	80.74	24.58	58.84	25.42	19.97	82.48	**23.31**	58.92	24.11	18.66
	Muon	**85.10**	**28.52**	64.09	**27.75**	**22.85**	**89.66**	23.82	**66.94**	**25.75**	**19.63**

2012) on NVIDIA RTX 4080 SUPER. We report both best and last checkpoints, and we treat the gap as evidence of robust overfitting rather than as a hidden tuning advantage. Appendix D.2 shows the experimental results on ViT-B and ViT-L.

PreActResNet-18. Table 1 shows the results on PreActResNet-18 (11.2 million parameters) (He et al., 2016b). Under the multi-norm training ($\ell_\infty + \ell_1$), Muon's best checkpoint achieves the union robustness of 39.27, compared to SGD's 40.37 and AdamW's 36.53. On shallower CNNs, Muon is not necessarily superior to SGD. Muon's advantage lies in its balanced performance across norms.

WideResNet. Table 2 illustrates the robustness evaluation results on WRN-34-10 (48.2 million parameters) (Zagoruyko and Komodakis, 2016). Muon achieves the strongest union robustness under AT. With ℓ_∞-only training, Muon attains a high ℓ_∞ robust accuracy (46.71), but its ℓ_1 robustness nearly collapses (4.88). This reflects a general phenomenon where training against a single threat model tends to produce robustness mainly against that specific threat. Table 3 further shows robustness evaluation results on WRN-34-20 (192.9 million parameters) (Zagoruyko and Komodakis, 2016). Muon's best checkpoint achieves a union robustness of 39.15. Furthermore, the Muon optimizer attains higher ℓ_∞/ℓ_2 robustness (39.99/61.22). Under the multi-norm $\ell_\infty + \ell_1$ setting, Muon's ℓ_1 robust accuracy drops sharply from 47.68% at the best checkpoint to 11.10% at the final checkpoint, indicating pronounced robust overfitting.

Table 2: AT results on WRN-34-10 (Zagoruyko and Komodakis, 2016).

Norm	Opt.	Best					Last				
		Clean	ℓ_∞	ℓ_2	ℓ_1	Union	Clean	ℓ_∞	ℓ_2	ℓ_1	Union
$\ell_\infty + \ell_1$	SGD	**78.19**	38.18	**63.29**	**50.51**	37.74	81.52	32.03	61.43	43.71	31.36
	AdamW	75.43	37.80	60.42	47.49	37.15	78.90	35.72	61.48	46.57	34.99
	Muon	73.27	**38.87**	58.44	45.73	**37.89**	**85.21**	**38.14**	**65.13**	**46.58**	**37.09**
ℓ_∞	SGD	80.02	44.11	57.73	**7.95**	**7.95**	82.50	37.23	53.21	**5.69**	**5.69**
	AdamW	69.92	42.47	49.69	4.41	4.41	72.39	41.38	48.25	3.18	3.18
	Muon	**83.16**	**46.71**	**58.11**	4.88	4.88	**85.96**	**43.81**	**54.05**	4.34	4.34
ℓ_1	SGD	47.59	19.39	36.99	33.27	19.39	83.90	16.80	58.03	51.04	16.80
	AdamW	65.26	23.49	50.97	46.03	23.49	75.77	**20.91**	55.29	49.41	**20.91**
	Muon	**78.81**	**24.89**	**58.87**	**53.56**	**24.89**	**87.83**	19.60	**63.03**	**58.15**	19.60
ℓ_2	SGD	76.59	22.72	54.04	23.72	18.87	87.51	21.49	62.31	24.65	18.18
	AdamW	73.40	24.07	51.93	24.18	19.90	86.17	21.65	61.15	19.38	15.75
	Muon	**89.97**	**26.85**	**68.16**	**25.96**	**21.63**	**90.18**	**26.44**	**68.56**	**26.59**	**21.53**

Table 3: AT results on WRN-34-20 (Zagoruyko and Komodakis, 2016).

Norm	Opt.	Best					Last				
		Clean	ℓ_∞	ℓ_2	ℓ_1	Union	Clean	ℓ_∞	ℓ_2	ℓ_1	Union
$\ell_\infty + \ell_1$	SGD	75.35	36.84	59.99	**48.18**	36.51	82.58	34.14	**62.16**	**44.79**	33.26
	AdamW	72.39	35.44	57.51	43.99	34.65	72.39	35.43	57.51	44.04	**34.63**
	Muon	**76.41**	**39.99**	**61.22**	47.68	**39.15**	**86.51**	**37.87**	61.69	11.10	11.10
ℓ_∞	SGD	75.98	41.12	55.09	**8.30**	**8.30**	83.62	39.17	54.40	**5.34**	**5.34**
	AdamW	73.37	42.59	48.31	2.48	2.48	75.17	40.91	47.35	2.22	2.22
	Muon	**82.13**	**46.58**	**57.98**	5.30	5.30	**87.44**	**46.76**	**56.99**	5.15	5.15
ℓ_1	SGD	56.49	21.24	41.95	37.74	21.24	85.31	17.94	60.45	54.45	17.94
	AdamW	68.56	21.79	52.11	47.44	21.79	82.03	17.37	57.17	50.30	17.37
	Muon	**88.53**	**21.45**	**65.71**	**60.19**	**21.45**	**88.61**	**20.83**	**65.16**	**59.95**	**20.83**
ℓ_2	SGD	88.33	24.37	64.29	25.26	20.11	88.48	23.93	64.84	25.66	20.05
	AdamW	85.55	22.73	61.20	22.15	17.92	85.45	22.59	60.76	21.53	17.76
	Muon	**91.09**	**28.15**	**70.14**	**27.26**	**22.64**	**91.16**	**27.84**	**70.38**	**27.25**	**22.44**

In summary, AT with the Muon optimizer achieves considerable robustness on overparameterized CNNs. Muon achieves union robustness on par with or better than SGD on WRN-34-10/20, while outperforming AdamW by a large margin. Additional ViT results are provided in Appendix D.2.

4.3 Gradient-norm Trajectory Analysis

Figure 2 provides a controlled diagnostic under a shared nominal learning rate. We use this experiment only to visualize the resulting optimizer trajectories. Because identical numerical learning rates do not imply matched effective update magnitudes, we do not interpret the comparison as a causal isolation of update direction. Although Muon exhibits faster adversarial-loss descent despite smaller gradient norms, different optimizers induce different effective update magnitudes. It is qualitatively consistent with the polar-alignment geometry characterized in Theorem 3.3.

4.4 Results on ImageNet

We additionally evaluate multi-norm AT on ImageNet (Deng et al., 2009) with ResNet-50 (He et al., 2016a) as a large-scale stress test. Under the reported configurations, the initial learning rate is set to 1.0 for SGD and 10^{-4} for Muon and AdamW.

Table 4 shows that, under the evaluated training recipe and compute budget, SGD attains higher selected-checkpoint and final robustness than Muon on the full ImageNet validation set. Because

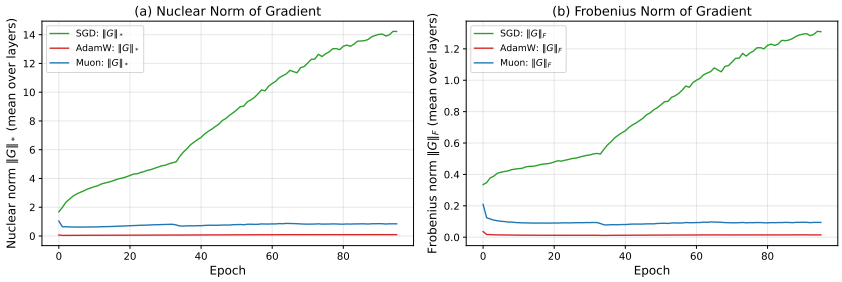


Figure 2: Gradient-norm trajectories during AT: (a) nuclear norm and (b) Frobenius norm

Table 4: Results on the ImageNet validation set (Deng et al., 2009).

Opt.	Best					Last				
	Clean	ℓ_∞	ℓ_2	ℓ_1	Union	Clean	ℓ_∞	ℓ_2	ℓ_1	Union
SGD	**53.58**	**21.50**	**35.77**	**23.21**	**21.49**	48.07	8.21	**22.07**	**12.24**	**8.13**
AdamW	40.52	6.99	17.83	2.88	2.25	40.16	6.77	17.93	9.50	6.74
Muon	47.68	18.91	31.13	18.69	18.21	**48.80**	**8.39**	16.77	3.11	2.35

we do not conduct an exhaustive optimizer-specific hyperparameter search at ImageNet scale, we do not interpret this gap as establishing an intrinsic large-scale ranking between the optimizers. Instead, the result shows that the optimizer behavior observed on CIFAR does not automatically transfer to a substantially larger training regime. We therefore view ImageNet as a scale stress test that identifies a concrete scope boundary of the CIFAR-scale observations, rather than as evidence for a universal optimizer ordering.

5 Conclusion

This paper studies optimizer geometry as a security-relevant factor in adversarial training. Using Muon as a representative orthogonalized optimizer, we characterize two properties of its update geometry: bounded operator-norm updates that, together with weight decay, yield a uniform upper bound on weight norms throughout training, and a nuclear-norm-aligned first-order descent interpretation. These results characterize optimization mechanisms rather than a universal robustness guarantee. Empirically, Muon generally outperforms AdamW and is competitive with SGD in several CNN regimes, while SGD remains stronger in some ViT and ImageNet settings. Muon also exhibits robust overfitting in some configurations, so we do not claim improved last-checkpoint stability or a universal replacement for SGD. Overall, our results show that outer-optimizer choice can materially change the robustness obtained from the same AT objective and should therefore be reported, tuned, and stress-tested as part of adversarial-defense evaluation.

AI use statement

The core idea of our research is generated through human discussions. AI tools were used for language refinement and limited assistance in code-related tasks. All scientific decisions, experiments, and conclusions were made by the authors.

Ethics statement

Our research is focused on adversarial attacks. Nevertheless, we have chosen to conduct our study using the CIFAR-10 and ImageNet datasets. This does not pose any ethical risks in practical scenarios.

Reproducibility statement

Our implementation builds on `https://github.com/fra31/robust-finetuning`. We extend it to support ImageNet-scale adversarial training, modern outer optimizers including Muon, and the architectures and threat models evaluated in this work.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *IEEE 62nd annual symposium on foundations of computer science (FOCS)*, pages 977–988, 2022.

Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020.

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

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pages 484–501, 2020.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283, 2018a.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293, 2018b.

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.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. In *International Conference on Machine Learning*, pages 3920–3930, 2025.

Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700 of *Lecture Notes in Computer Science*, pages 421–436. Springer, 2012.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 39–57, 2017.

Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1739–1747, 2020.

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

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320, 2019.

Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International conference on machine learning*, pages 2196–2205, 2020a.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216, 2020b.

Francesco Croce and Matthias Hein. Adversarial robustness against multiple and single l_p-threat models via quick fine-tuning of robust classifiers. In *International Conference on Machine Learning*, pages 4436–4454, 2022.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Conference on Neural Information Processing Systems*, 2021.

Ali Dabouei, Fariborz Taherkhani, Sobhan Soleymani, and Nasser M Nasrabadi. Revisiting outer optimization in adversarial training. In *European Conference on Computer Vision*, pages 244–261, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

Zhun Deng, Hangfeng He, Jiaoyang Huang, and Weijie Su. Towards understanding the dynamics of the first-order adversaries. In *International Conference on Machine Learning*, pages 2484–2493, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, pages 1–13, 2021.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2021.

Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D Lee. Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016b.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Advances in neural information processing systems*, 30, 2017.

Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

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

Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. Bad global minima exist and sgd can reach them. *Advances in Neural Information Processing Systems*, 33:8543–8552, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Avery Ma, Yangchen Pan, and Amir-massoud Farahmand. Understanding the robustness difference between stochastic gradient descent and adaptive gradient methods. *Transactions on Machine Learning Research*, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Laker Newhouse, R Preston Hess, Franz Cesista, Andrii Zahorodnii, Jeremy Bernstein, and Phillip Isola. Training transformers with enforced lipschitz constants. *arXiv preprint arXiv:2507.13338*, 2025.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *ICLR*, 2021.

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

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in neural information processing systems*, 32, 2019.

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *International Conference on Machine Learning*, pages 6586–6595, 2019.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *ICLR*, 2020.

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

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings 2018 Network and Distributed System Security Symposium*, 2018.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, pages 1–15, 2016.

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. *Advances in neural information processing systems*, 32, 2019a.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482, 2019b.

Yihao Zhang, Hangzhou He, Jingyu Zhu, Huanran Chen, Yifei Wang, and Zeming Wei. On the duality between sharpness-aware minimization and adversarial training. In *International Conference on Machine Learning*, pages 59024–59041, 2024.

Zhanpeng Zhou, Mingze Wang, Yuchen Mao, Bingrui Li, and Junchi Yan. Sharpness-aware minimization efficiently selects flatter minima late in training. In *ICLR*, 2025.

A SUPPLEMENTAL REMARKS OF THEORETICAL ANALYSIS

Remark A.1 (Interpretation of Theorem 3.2). The quantity $\hat{L}_p(\theta)$ is a computable upper bound on the true ℓ_p-Lipschitz constant $L_p(\theta)$, obtained by chaining (i) the Jacobian–spectral bound of Proposition 3.2, (ii) the per-layer spectral bound $\|W_\ell\|_2 \le B_\ell$ from Theorem 3.1, and (iii) the norm-equivalence $\|\delta\|_2 \le c_p\|\delta\|_p$. Theorem 3.2 should be read as a structural statement rather than a usable robustness certificate. Specifically, Muon controls the operator norm of each matrix-valued update after polar orthogonalization. When combined with weight decay, this yields a worst-case trajectory bound on the spectral norm of the corresponding weight block. The bound is generally loose and does not imply that Muon-trained models always have smaller spectral norms than SGD-trained models. Instead, it identifies a sufficient condition under which the update rule avoids uncontrolled spectral growth.

Remark A.2 (Interpretation of Proposition 3.1). We acknowledge that true adversarial robustness depends on the global geometry of the model's decision boundary, not merely on local gradient magnitudes. The first-order characterization adopted here reveals robustness to the dual norm of the input gradient. It is a standard proxy used throughout the certification literature (Hein and Andriushchenko, 2017; Tsuzuku et al., 2018). It is exact for linear models and for perturbations that remain within a fixed linear region of a piecewise-linear network. Although it does not capture higher-order curvature effects (e.g., decision boundary curvature exploited by second-order attacks), it remains predictive in practice.

Remark A.3 (Interpretation of Proposition 3.2). For residual architectures, Proposition 3.2 can be applied blockwise. A block $h_{\ell+1} = h_\ell + F_\ell(h_\ell)$ satisfies $\|Jh_{\ell+1}\|_2 \le 1 + \|JF_\ell(h_\ell)\|_2$, and hence the network Jacobian is bounded by the product of such residual-block factors. Therefore, the sequential feedforward bound is used as a clean base case, while ResNet/WRN-style architectures follow by replacing each layer factor with its corresponding residual-block Lipschitz factor.

Remark A.4 (Interpretation of Theorem 3.1). Theorem 3.1 gives a worst-case spectral-stability bound, $\|W_t\|_2 \le \max\{\|W_0\|_2, \frac{1+\bar{\varepsilon}}{\lambda}\}$. The uniform ceiling in Eq. (7) requires $\lambda > 0$. When $\lambda = 0$, Eq. (9) only provides finite-horizon growth control. The underlying recursion applies to any optimizer with uniformly bounded operator-norm updates, and the Muon-specific ingredient is that polar orthogonalization provides this control directly, $\|\operatorname{Ortho}(M_t)\|_2 = 1$ in the exact case, without projecting the learned weights. Thus, the theorem characterizes trajectory stability rather than predicting tight spectral norms or a universal advantage over SGD. SGD can yield smaller spectral norms in practice, whereas AdamW lacks an analogous update-level bound and exhibits substantially larger spectral growth in several settings (Appendix D.9). Theorem 3.1 should not be interpreted as a certified robustness guarantee for Muon-trained networks. It provides a sufficient condition connecting spectral stability to a first-order robustness proxy. The resulting bound can be loose and does not imply improved empirical robustness.

Remark A.5 (Interpretation of Theorem 3.3). By spectral–nuclear norm duality, $\|\mathbf{M}\|_* = \max_{\|\mathbf{U}\|_2 \le 1} \langle \mathbf{M}, \mathbf{U} \rangle$, and the polar factor $\operatorname{Ortho}(\mathbf{M})$ attains this maximum. Thus, in the exact-polar case, Muon selects the direction that maximizes the first-order decrease predicted by the momentum surrogate among matrix updates with unit operator norm. Theorem 3.3 quantifies how momentum mismatch, finite-step orthogonalization, and curvature perturb this idealized geometry. This characterization does not imply uniformly faster convergence than SGD or AdamW. Theorem 3.3 provides a one-step, first-order characterization of the outer-update geometry, rather than a guarantee of min–max convergence, robust generalization, or uniformly faster optimization.

B MAPPING OF THEORY AND ALGORITHM

Algorithm 2 illustrates the inline mechanism of `Opt.step()` described in Algorithm 1.

Table 5 maps theoretical results with algorithmic operations of the Newton–Schulz subroutine defined in Algorithm 2.

Algorithm 2 One Muon Step under AT (expands line 10 of Alg. 1 when $\mathrm{Opt} = \mathrm{Muon}$; Newton–Schulz iteration inlined as sub-steps of S2).

Require: Weight block $W \in \mathbb{R}^{m\times n}$, momentum buffer M (init. $\mathbf{0}$), gradient $G = \nabla_W L_{\mathrm{adv}}$, learning rate η, momentum μ, weight decay λ, Newton-Schulz steps s

S1 $M \leftarrow \mu M + G$ ▷ *momentum update (Assumption 3.1: $G = \nabla_W F(W)$; F is assumed β-smooth in Theorem 3.3)*

S2 *// Approximate the ideal polar factor characterized in Lemma 3.1*

$X \leftarrow M \,/\, \|M\|_F$ ▷ *scale normalization before Newton–Schulz iterations*

for $k = 1$ **to** s **do**

$X \leftarrow \frac{3}{2}X - \frac{1}{2}XX^\top X$ ▷ *Newton–Schulz polar iteration*

end for

$\widetilde{U} \leftarrow X$ ▷ *as in Theorem 3.1*

S3 $W \leftarrow (1-\eta\lambda)\, W - \eta\, \widetilde{U}$ ▷ *orthogonalized weight update (Theorem 3.1: $\|W_t\|_2 \le \max(\|W_0\|_2,\ \frac{1+\bar{\varepsilon}}{\lambda})$, for all $t \ge 0$)*

return updated W, updated M

Table 5: Mapping of algorithmic operations and theoretical results.

Step	Operation Theory	Implication	
S1	$M \leftarrow \mu M + G$	Assumption 3.1	β-smooth adversarial loss
S2 (init)	$X \leftarrow M/\|M\|_F$	Lemma 3.1 (pre-cond.)	Normalizes scales
S2 (loop)	$X \leftarrow \frac{3}{2}X - \frac{1}{2}XX^\top X$	Theorem 3.1 (Newton–Schulz approximation and tolerance condition)	$\|\widetilde{U}\|_2 \approx 1$
S3	$W \leftarrow (1-\eta\lambda)W - \eta\widetilde{U}$	Theorem 3.1	Bounded $\|W_t\|_2$ trajectory
S3 (all layers)	Multi-norm Δ_{union}	Theorem 3.2	Union robustness analysis
S3	AT loss descent	Theorem 3.3	Nuclear-norm descent $\|M\|_*$

C Proof of Theoretical Analysis

C.1 Proof of Lemma 3.1

Proof. Let $U = \mathrm{Ortho}(A) = PV^\top$, where $A = P\Sigma V^\top$ is the compact SVD of A. If $m \ge n$, then $P^\top P = I_n$ and V is orthogonal. Hence

$$U^\top U = VP^\top PV^\top = I_n.$$

Moreover, since $A^\top A = V\Sigma^2 V^\top$, we have

$$U(A^\top A)^{1/2} = PV^\top(V\Sigma V^\top) = P\Sigma V^\top = A.$$

If $m < n$, then P is orthogonal and $V^\top V = I_m$. Hence

$$UU^\top = PV^\top VP^\top = I_m.$$

Moreover, since $AA^\top = P\Sigma^2 P^\top$, we have

$$(AA^\top)^{1/2}U = (P\Sigma P^\top)(PV^\top) = P\Sigma V^\top = A.$$

Thus, all nonzero singular values of U are equal to one, and therefore $\|U\|_2 = 1$.

It remains to prove the projection statement. Let $\mathcal{S}$ denote the corresponding rectangular Stiefel set, i.e., $\mathcal{S} = \{Q : Q^\top Q = I_n\}$ if $m \ge n$, and $\mathcal{S} = \{Q : QQ^\top = I_m\}$ if $m < n$. For any $Q \in \mathcal{S}$,

$$\|A - Q\|_F^2 = \|A\|_F^2 + \|Q\|_F^2 - 2\langle A, Q\rangle.$$

Since $\|Q\|_F^2 = \min\{m, n\}$ is fixed on $\mathcal{S}$, minimizing $\|A - Q\|_F$ is equivalent to maximizing $\langle A, Q\rangle$. By von Neumann's trace inequality,

$$\langle A, Q\rangle \le \sum_i \sigma_i(A)\sigma_i(Q) = \sum_i \sigma_i(A),$$

because every $Q \in \mathcal{S}$ has all nonzero singular values equal to one. For $Q = U = PV^\top$, equality holds:

$$\langle A, U\rangle = \mathrm{tr}(A^\top U) = \mathrm{tr}(\Sigma) = \sum_i \sigma_i(A).$$

Hence U is a minimizer of $\min_{Q\in\mathcal{S}} \|A - Q\|_F$. □

C.2 Proof of Theorem 3.1

Proof. Using the triangle inequality and absolute homogeneity of the spectral norm, we have

$$\begin{aligned} \|W_{t+1}\|_2 &= \left\|(1-\eta\lambda)W_t - \eta\widetilde{U}_t\right\|_2 \\ &\le (1-\eta\lambda)\,\|W_t\|_2 + \eta\left\|\widetilde{U}_t\right\|_2. \end{aligned} \tag{15}$$

Applying the assumption of Eq. (6) with $a_t = \|W_t\|_2$, we obtain

$$a_{t+1} \le (1-\eta\lambda)a_t + \eta + \eta\varepsilon_t. \tag{16}$$

Unrolling the recursion gives

$$a_t \le (1-\eta\lambda)^t a_0 + \eta\sum_{k=0}^{t-1}(1-\eta\lambda)^{t-1-k} + \eta\sum_{k=0}^{t-1}(1-\eta\lambda)^{t-1-k}\varepsilon_k. \tag{17}$$

The second term is a geometric series

$$\begin{aligned} \eta\sum_{k=0}^{t-1}(1-\eta\lambda)^{t-1-k} &= \eta\sum_{j=0}^{t-1}(1-\eta\lambda)^j \\ &= \eta\cdot\frac{1-(1-\eta\lambda)^t}{\eta\lambda} \\ &= \frac{1-(1-\eta\lambda)^t}{\lambda}. \end{aligned} \tag{18}$$

Substituting back yields Eq. (8). If $\varepsilon_k \le \bar{\varepsilon}$, then

$$\eta\sum_{k=0}^{t-1}(1-\eta\lambda)^{t-1-k}\varepsilon_k \le \eta\bar{\varepsilon}\sum_{j=0}^{t-1}(1-\eta\lambda)^j = \bar{\varepsilon}\cdot\frac{1-(1-\eta\lambda)^t}{\lambda}, \tag{19}$$

which gives Eq. (9). □

C.3 Proof of Theorem 3.2

Proof. **Step 1: From Jacobian spectral norms to an ℓ_2-Lipschitz bound on the margin.**

For each competing class $j \ne y$, define the pairwise margin

$$m_{y,j}(x) := g_\theta(x)_y - g_\theta(x)_j = \langle e_y - e_j,\, g_\theta(x)\rangle. \tag{20}$$

Here $e_y - e_j$ is fixed and satisfies $\|e_y - e_j\|_2 = \sqrt{2}$. Hence

$$\nabla_x m_{y,j}(x) = J_{g_\theta}(x)^\top(e_y - e_j), \tag{21}$$

and therefore, it satisfies

$$\|\nabla_x m_{y,j}(x)\|_2 \le \sqrt{2}\,\|J_{g_\theta}(x)\|_2 \le \sqrt{2}\prod_{\ell=1}^{L}\|W_\ell\|_2, \tag{22}$$

where the last inequality follows from Proposition 3.2 assuming 1-Lipschitz activations. Otherwise, the product is replaced by $L_\phi^{L-1}\prod_\ell\|W_\ell\|_2$.

Since the multiclass margin can be written as

$$m_\theta(x,y) = g_\theta(x)_y - \max_{j\ne y} g_\theta(x)_j = \min_{j\ne y} m_{y,j}(x), \tag{23}$$

and the pointwise minimum of finitely many L-Lipschitz functions is still L-Lipschitz, $m_\theta(\cdot,y)$ is $L_2(\theta)$-Lipschitz with respect to $\|\cdot\|_2$, with

$$L_2(\theta) \le C_{\mathrm{marg}}\prod_{\ell=1}^{L}\|W_\ell\|_2,\ C_{\mathrm{marg}} := \max_{j\ne y}\|e_y - e_j\|_2 = \sqrt{2}. \tag{24}$$

Step 2: Spectral-norm control from Muon.

By Theorem 3.1, running Muon with learning rate η and weight decay $\lambda > 0$ guarantees

$$\|W_\ell\|_2 \;\le\; B_\ell \text{ for all } \ell \in \{1, \dots, L\} \text{ and all training steps,} \tag{25}$$

where $B_\ell = \max\big(\|W_\ell^{(0)}\|_2,\, (1+\bar{\varepsilon})/\lambda\big)$. Combining with Eq. (24) yields

$$L_2(\theta) \le C_{\text{marg}} \prod_{\ell=1}^{L} B_\ell. \tag{26}$$

Step 3: From ℓ_2-Lipschitzness to ℓ_p-Lipschitzness via norm equivalence.

For each $p \in \{1, 2, \infty\}$, let c_p denote the norm-equivalence constant satisfying

$$\|\delta\|_2 \;\le\; c_p \, \|\delta\|_p, \qquad \text{for all } \delta \in \mathbb{R}^d, \tag{27}$$

with $c_\infty = \sqrt{d}$, $c_2 = 1$, and $c_1 = 1$. Then for any $\delta \in \mathbb{R}^d$ and any $p \in \mathcal{P}$,

$$\begin{aligned} \big|m_\theta(x+\delta, y) - m_\theta(x, y)\big| &\le L_2(\theta)\, \|\delta\|_2 \\ &\le L_2(\theta)\, c_p\, \|\delta\|_p \\ &\le \hat{L}_p(\theta)\, \|\delta\|_p, \end{aligned} \tag{28}$$

where

$$\hat{L}_p(\theta) := c_p \, C_{\text{marg}} \prod_{\ell=1}^{L} B_\ell$$

is the computable upper bound from Eq. (10), and the last inequality uses Eq. (26).

Step 4: Union robustness.

Take any $\delta \in \Delta_{\text{union}} = \bigcup_{p \in \mathcal{P}} \{\delta : \|\delta\|_p \le \varepsilon_p\}$. Then there exists some $p \in \mathcal{P}$ with $\|\delta\|_p \le \varepsilon_p$. By Eq. (28),

$$\begin{aligned} m_\theta(x+\delta, y) &\ge m_\theta(x, y) - \hat{L}_p(\theta)\, \|\delta\|_p \\ &\ge m_\theta(x, y) - \hat{L}_p(\theta)\, \varepsilon_p \\ &\ge m_\theta(x, y) - \max_{p' \in \mathcal{P}} \hat{L}_{p'}(\theta)\, \varepsilon_{p'}. \end{aligned} \tag{29}$$

Under the assumption of Theorem 3.2, $m_\theta(x, y) > \max_{p' \in \mathcal{P}} \hat{L}_{p'}(\theta)\, \varepsilon_{p'}$, hence $m_\theta(x+\delta, y) > 0$. Since $\delta \in \Delta_{\text{union}}$ was arbitrary, $m_\theta(x+\delta, y) > 0$ for all $\delta \in \Delta_{\text{union}}$. □

C.4 Proof of Theorem 3.3

Proof. Let M denote the momentum-accumulated gradient used by Muon, and let $\widetilde{U}$ denote the inexact Newton–Schulz output satisfying

$$\|\widetilde{U} - \text{Ortho}(M)\|_F \le \delta_{\text{orth}},\; \|\widetilde{U}\|_2 \le 1 + \varepsilon,\; \|\widetilde{U}\|_F^2 \le r(1+\varepsilon)^2, \tag{30}$$

where $r = \text{rank}(M)$, or any upper bound on it. Consider the actual update $W^+ = W - \eta\, \widetilde{U}$. By Assumption 3.1 ($\beta$-smoothness of F w.r.t. $\|\cdot\|_F$), we have

$$F(W^+) \le F(W) - \eta \langle G, \widetilde{U} \rangle + \frac{\beta}{2} \eta^2 \|\widetilde{U}\|_F^2. \tag{31}$$

We decompose the inner product as

$$\begin{aligned} \langle G, \widetilde{U} \rangle = &\underbrace{\langle M, \text{Ortho}(M) \rangle}_{\text{(i)}} \\ &+ \underbrace{\langle G - M, \widetilde{U} \rangle}_{\text{(ii)}} \\ &+ \underbrace{\langle M, \widetilde{U} - \text{Ortho}(M) \rangle}_{\text{(iii)}}. \end{aligned} \tag{32}$$

(i) Ideal inner-product term. By the same SVD argument as in the idealized case (Lemma 3.1), the polar factor satisfies

$$\langle M, \mathrm{Ortho}(M)\rangle = \|M\|_*, \tag{33}$$

where $\|M\|_* = \sum_{i=1}^{r} \sigma_i(M)$ is the nuclear norm.

(ii) Gradient–momentum mismatch. By Hölder's inequality for Schatten norms, i.e., the duality between the nuclear norm and the spectral norm, we have

$$\langle G - M, \widetilde{U}\rangle \geq -\|G - M\|_* \, \|\widetilde{U}\|_2 \geq -(1+\varepsilon)\|G - M\|_*. \tag{34}$$

(iii) Orthogonalization error. By the Cauchy–Schwarz inequality for the Frobenius inner product,

$$\begin{aligned}\Big\langle M, \tilde{U} - \mathrm{Ortho}(M)\Big\rangle &\geq -\|M\|_F \Big\|\tilde{U} - \mathrm{Ortho}(M)\Big\|_F \\ &\geq -\|M\|_F \delta_{\mathrm{orth}} \geq -\sqrt{r}\,\|M\|_2 \delta_{\mathrm{orth}},\end{aligned} \tag{35}$$

where the last step uses $\|M\|_F \leq \sqrt{r}\,\|M\|_2$.

Combining (i)–(iii). Substituting Eq. (33), Eq. (34), and Eq. (35) into Eq. (32) gives

$$\langle G, \widetilde{U}\rangle \geq \|M\|_* - (1+\varepsilon)\|G - M\|_* - \sqrt{r}\,\|M\|_2\,\delta_{\mathrm{orth}}. \tag{36}$$

Plugging Eq. (36) into Eq. (31) and using $\|\widetilde{U}\|_F^2 \leq r(1+\varepsilon)^2$, we obtain

$$\begin{aligned}F(W^+) \leq\ & F(W) - \eta\|M\|_* + \eta(1+\varepsilon)\|G - M\|_* \\ &+ \eta\sqrt{r}\,\|M\|_2\,\delta_{\mathrm{orth}} + \frac{\beta}{2}\eta^2 r(1+\varepsilon)^2.\end{aligned} \tag{37}$$

This is exactly the descent inequality of Theorem 3.3. □

D Experimental Setup and Additional Results

D.1 Experiment Setup

Datasets. The CIFAR-10 dataset (Krizhevsky, 2009) is a commonly used dataset in computer vision. It consists of 60,000 32×32 color images in 10 classes with 50,000 training images and 10,000 test images. It is a standard benchmark in adversarial robustness research. ImageNet (Deng et al., 2009) is a large-scale image database for visual recognition research. It contains over 14 million human-annotated images spanning more than 20,000 categories, and around 1 million images include object bounding boxes. In research practice, the dataset is often used in its 1,000-class subset, which includes 1,281,167 training samples and 50,000 validation samples.

Training Details. On CIFAR-10 (Krizhevsky, 2009), the training process lasts for 100 epochs. The APGD method (Croce and Hein, 2020b) is used as the default attack method with 10 iteration steps. For the ℓ_∞-attack, the perturbation budget is $\varepsilon_\infty = \frac{8.0}{255.0}$. For the ℓ_2-attack, the perturbation budget is $\varepsilon_2 = 0.5$. For the ℓ_1-attack, the perturbation budget is $\varepsilon_1 = 12$. The batch size is set to 128. For CNNs (He et al., 2016b; Zagoruyko and Komodakis, 2016), the initial learning rate is 1.0. Unless otherwise noted, the initial learning rate is 0.001 for ViTs (Dosovitskiy et al., 2021). The learning rate is set in a stepwise manner, decreasing to 0.1 and 0.01 of the initial learning rate at 1/3 and 2/3 of the entire training period, respectively. The momentum coefficient μ is 0.9, and the weight decay λ is 0.0005. On ImageNet (Deng et al., 2009), the training process lasts for 100 epochs. Due to computing resource constraints, the defense is only deployed on ResNet-50 (He et al., 2016a). We implement the multi-norm AT. The batch size is still set to 128. For the ℓ_∞-attack, the perturbation budget is $\varepsilon_\infty = \frac{4.0}{255.0}$, and the attack step number is 5. For the ℓ_2-attack, the perturbation budget is $\varepsilon_2 = 2$, and the attack step number is 10. For the ℓ_1-attack, the perturbation budget is $\varepsilon_1 = 255$, and the attack step number is 15. We keep the AT objective, threat model, data pipeline, batch size, training horizon, and scheduler form fixed across optimizers. Learning rate and weight decay are treated as optimizer-specific configuration parameters, since identical numerical values need not induce comparable update scales across fundamentally different update geometries. We therefore use stable practical configurations rather than enforcing a single numerical learning rate or weight decay across optimizers. We do not perform exhaustive optimizer-specific hyperparameter searches.

Table 6: AT results on ViT-B (Dosovitskiy et al., 2021).

Norm	Opt.	Best					Last				
		Clean	ℓ_∞	ℓ_2	ℓ_1	Union	Clean	ℓ_∞	ℓ_2	ℓ_1	Union
$\ell_\infty + \ell_1$	SGD	**66.35**	**29.64**	**52.69**	**44.78**	**29.62**	**71.96**	**24.88**	**52.94**	**40.32**	**24.80**
	AdamW	27.22	12.10	18.97	16.54	12.02	32.77	13.62	23.11	19.88	13.53
	Muon	61.53	27.03	48.63	42.61	27.03	65.49	19.17	45.79	35.42	19.15
ℓ_∞	SGD	**64.65**	**30.97**	**50.42**	**32.59**	**28.10**	**71.64**	**26.15**	**52.58**	**27.86**	**27.71**
	AdamW	17.93	12.12	11.60	4.48	4.48	17.25	12.93	12.31	5.77	5.77
	Muon	60.66	30.36	47.20	28.00	25.73	66.26	19.93	44.20	26.23	19.30
ℓ_1	SGD	**68.48**	**22.63**	**51.30**	**45.27**	**22.63**	**72.24**	**17.19**	**49.72**	**43.06**	**17.19**
	AdamW	25.57	8.99	17.08	15.94	8.97	28.26	6.47	16.98	15.66	6.44
	Muon	58.58	21.15	44.41	40.93	21.15	64.64	12.15	40.61	34.79	12.15
ℓ_2	SGD	57.08	17.49	40.32	29.44	17.48	**76.50**	**18.61**	**52.99**	**33.55**	**18.46**
	AdamW	18.51	11.53	13.95	13.01	11.49	35.10	8.81	20.54	16.13	8.71
	Muon	**58.49**	**17.66**	**41.01**	**31.77**	**17.62**	67.89	12.70	42.60	30.96	12.70

Accordingly, the reported comparisons should be interpreted as optimizer-dependent robustness under the stated configurations, rather than as hyperparameter-optimal rankings. Following the reference Muon implementation (Jordan et al., 2024), we fix the default Newton–Schulz iteration count to $s = 5$ in all experiments and do not tune this internal approximation parameter. The optimizer-specific configuration choices considered in this study therefore concern the learning rate and weight decay. Our analysis accounts for the resulting finite-step errors ε_t and δ_{orth}, without assuming an exact polar decomposition. We do not introduce additional ViT-specific pretraining, gradient clipping, or optimizer-specific warm-up beyond the training recipe described above. Accordingly, the ViT results should not be interpreted as best-tuned robustness for each optimizer.

D.2 Experimental Results on Vision Transformers

Table 6 shows the robustness of ViT-B (85.7 million parameters) (Dosovitskiy et al., 2021) under different optimizers. On ViT-B, Muon substantially outperforms AdamW for adversarial training, although it does not surpass SGD. While SGD can also achieve a certain level of robustness on ViT-B, Muon's advantage is that it remains relatively stable across AT scenarios. Under the tested AT configurations, Muon remains substantially more stable than AdamW on the evaluated ViTs. We emphasize that these results characterize the stated training recipe rather than a hyperparameter-independent ranking of Transformer optimizers. Table 7 shows the robustness of ViT-L (303.1 million parameters) (Dosovitskiy et al., 2021) under different optimizers. The experimental observations on ViT-L are similar to those on ViT-B. On ViT-L, AdamW exhibits partial training failure, analogous to what happens when training ViT-B under the ℓ_∞-norm. To stabilize AT with AdamW, we set the learning rate to 0.00005 and the weight decay to 0.02. We find that AdamW's final model under ℓ_∞-only training on ViT-L is close to random performance (e.g., the union robust accuracy is low), whereas Muon still maintains considerable robustness.

Overall, on ViT-B and ViT-L, both Muon and SGD avoid the severe degradation observed with AdamW under the tested AT recipes. Muon remains competitive in ℓ_∞ robustness, whereas SGD is stronger in several union-robustness settings.

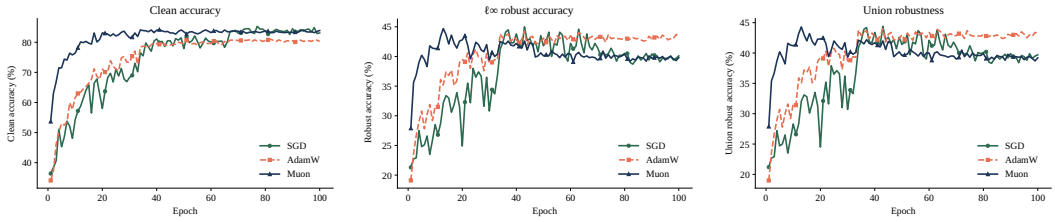
D.3 Robust Overfitting Diagnostics

The "robust overfitting" phenomenon is a concern in adversarial learning. We visualize the clean accuracy, ℓ_∞ robust accuracy, and union robust accuracy of PreActResNet-18 (He et al., 2016b). On CIFAR-10, Figure 3 illustrates the transition process. As shown in the figures, Muon converges substantially faster than both SGD and AdamW in the early training phase: by epoch 10–15, Muon already achieves about 40–44% union robust accuracy, while SGD is still around 25–35% and AdamW around 30–38%. This rapid early convergence is consistent with the nuclear-norm-aligned descent geometry characterized in Theorem 3.3. We emphasize that the theorem does not establish a matched-step convergence advantage over SGD; the empirical curves reflect the complete optimizer

Table 7: AT results on ViT-L (Dosovitskiy et al., 2021).

Norm	Opt.	Best					Last				
		Clean	ℓ_∞	ℓ_2	ℓ_1	Union	Clean	ℓ_∞	ℓ_2	ℓ_1	Union
$\ell_\infty + \ell_1$	SGD	**69.87**	**32.51**	**55.86**	**47.39**	**32.50**	**73.10**	**26.69**	**55.09**	**43.03**	**26.64**
	AdamW	26.52	11.43	19.14	16.28	11.23	25.10	10.62	17.12	15.19	10.59
	Muon	63.85	28.18	50.85	43.94	28.18	68.02	21.65	47.79	37.25	21.64
ℓ_∞	SGD	**70.15**	**34.26**	**55.30**	**34.95**	**30.25**	**72.32**	**25.61**	**51.95**	**27.67**	**22.68**
	AdamW	16.98	12.85	14.12	10.41	10.32	18.05	6.13	8.14	0.79	0.65
	Muon	60.42	31.79	47.76	28.72	26.61	68.35	23.03	48.16	27.52	21.95
ℓ_1	SGD	24.68	9.22	18.22	18.58	9.22	40.60	13.48	29.27	28.23	13.48
	AdamW	17.84	11.12	14.90	14.83	11.12	24.27	5.62	13.27	13.02	5.62
	Muon	**58.00**	**21.13**	**43.75**	**39.80**	**21.13**	**67.47**	**14.05**	**44.00**	**39.09**	**14.05**
ℓ_2	SGD	**77.03**	18.37	55.32	33.31	18.15	**77.22**	**19.01**	**54.40**	**34.42**	**18.78**
	AdamW	21.31	10.24	16.63	14.63	10.24	26.38	5.77	14.80	11.62	5.73
	Muon	67.40	**21.40**	**48.68**	**35.54**	**21.33**	71.17	15.32	46.74	31.90	15.26

dynamics. Empirically, Muon reaches SGD-level robustness substantially earlier in this setting, making it attractive under early-stopped or compute-constrained training. In the later phase (epoch 40–100), Muon stabilizes at approximately 40% (ℓ_∞) and 40% (union), SGD converges to a similar level (about 40%), while AdamW continues to improve and ultimately reaches about 36%. AdamW continues to narrow the gap later in training. We also observe larger short-horizon oscillations for SGD in this run. These curves illustrate distinct optimizer trajectories, but we do not interpret them as a causal isolation of update geometry.



(a) Clean accuracy (b) ℓ_∞ robust accuracy (c) Union robust accuracy

Figure 3: Robustness evaluation during AT process.

We observe robust overfitting in several Muon-based AT runs: a strong selected checkpoint can coexist with substantial degradation at the end of training. This does not contradict our theoretical analysis, which characterizes update-level trajectory control and local descent geometry rather than robust generalization across training epochs. The rapid early progress observed in Figure 3 is qualitatively consistent with Theorem 3.3, but the experiment does not isolate this mechanism causally. We therefore do not claim that Muon mitigates robust overfitting or uniformly improves last-checkpoint stability. Rather, its empirical advantage lies in favorable early- and middle-stage optimization and in the regimes where this translates into improved selected-checkpoint robustness.

For practical use, we recommend robust-validation-based early stopping. Specifically, for multi-norm AT, checkpoints should be selected according to validation union robust accuracy rather than clean accuracy, final-epoch performance, or a single-norm robust metric. This criterion introduces no additional training cost and directly targets the evaluation metric of interest. Other alleviation methods, such as Adversarial Weight Perturbation (AWP) (Wu et al., 2020) and label smoothing (Szegedy et al., 2016), are complementary to Muon and can be combined with its orthogonalized update geometry. It is shown in Appendix D.10.

Table 8: AT results on WRN-34-10 and ViT-B (SAM vs. Muon).

Model	Norm	Opt.	Best					Last				
			Clean	ℓ_∞	ℓ_2	ℓ_1	Union	Clean	ℓ_∞	ℓ_2	ℓ_1	Union
WRN-34-10	$\ell_\infty + \ell_1$	SAM	**78.75**	**40.98**	**64.64**	**52.77**	**40.64**	80.93	**40.93**	**65.84**	**52.71**	**40.52**
		Muon	73.27	38.87	58.44	45.73	37.89	**85.21**	38.14	65.13	46.58	37.09
	ℓ_∞	SAM	80.53	46.08	**60.30**	**10.43**	**10.43**	81.84	**46.20**	**60.76**	**9.74**	**9.73**
		Muon	**83.16**	**46.71**	58.11	4.88	4.88	**85.96**	43.81	54.06	4.34	4.34
	ℓ_1	SAM	**83.93**	24.87	**63.78**	**57.57**	24.86	85.04	**22.34**	**63.92**	57.66	**22.34**
		Muon	78.81	**24.89**	58.87	53.56	**24.89**	**87.83**	19.60	63.03	**58.15**	19.60
	ℓ_2	SAM	87.95	**29.27**	67.50	**29.34**	**24.23**	89.28	25.76	67.49	**26.62**	21.21
		Muon	**89.97**	26.85	**68.16**	25.96	21.63	**90.18**	**26.44**	**68.56**	26.59	**21.53**
ViT-B	$\ell_\infty + \ell_1$	SAM	56.48	**28.05**	45.24	40.01	**28.04**	57.83	**28.99**	**46.35**	**41.19**	**28.99**
		Muon	**61.53**	27.03	**48.63**	**42.61**	27.03	**65.49**	19.17	45.79	35.42	19.15
	ℓ_∞	SAM	51.70	28.08	40.76	24.79	23.44	51.88	**28.52**	40.99	24.83	**23.66**
		Muon	**60.66**	**30.36**	**47.20**	**28.00**	**25.73**	**66.26**	19.93	**44.20**	**26.23**	19.30
	ℓ_1	SAM	**66.03**	**24.65**	**50.57**	**46.24**	**24.65**	**67.63**	**25.26**	**52.11**	**47.21**	**25.26**
		Muon	58.58	21.15	44.41	40.93	21.15	64.64	12.15	40.61	34.79	12.15
	ℓ_2	SAM	**75.30**	**24.98**	**55.97**	**37.28**	**24.57**	**75.45**	**24.91**	**56.33**	**37.17**	**24.49**
		Muon	58.49	17.66	41.01	31.77	17.62	67.89	12.70	42.60	30.96	12.70

D.4 Comparison with SAM

We conduct a comparison experiment between Muon and SAM ($\rho = 0.05$) on WRN-34-10 (Zagoruyko and Komodakis, 2016) and ViT-B (Dosovitskiy et al., 2021). SAM is not treated as a mutually exclusive update geometry. Rather, it is a sharpness-aware wrapper on SGD. Table 8 demonstrates the comparison. SAM remains a strong optimizer, while Muon provides an optimizer-geometry alternative that does not require SAM's additional weight-perturbation gradient evaluation. Two optimizers address different mechanisms. SAM perturbs weights to regularize sharpness with two backpropagations, while Muon orthogonalizes updates to stabilize spectral geometry. Nevertheless, SAM requires two backpropagations. Taking ℓ_2-norm AT on WRN-34-10 as an example, SAM takes 67,741.1 s, while Muon requires only 60,415.0 s.

D.5 AutoAttack Evaluation

We evaluate robustness using the complete AutoAttack suite (Croce and Hein, 2020b) for reliability. Besides APGD-CE and APGD-T, the evaluation includes FAB (Croce and Hein, 2020a) and Square Attack (Andriushchenko et al., 2020) to form the complete AutoAttack evaluation suite. This study compares the defense methods (Croce and Hein, 2022) on three optimizers with other representative methods, such as AT (Madry et al., 2018), FastAT (Wong et al., 2020) and its variant (Andriushchenko and Flammarion, 2020), FreeAT (Shafahi et al., 2019), and YOPO (Zhang et al., 2019a). The WRN-34-10 models (Zagoruyko and Komodakis, 2016) are trained with AT, FreeAT, and YOPO. The PreActResNet-18 (He et al., 2016b) models are trained with FastAT methods (Wong et al., 2020; Andriushchenko and Flammarion, 2020). Table 9 reports the complete AutoAttack suite under the standard ℓ_∞ threat model ($\varepsilon = 8/255$) for the ℓ_∞-trained checkpoints of all five evaluated architectures: PreActResNet-18, WRN-34-10, WRN-34-20, ViT-B, and ViT-L. These evaluations provide a standardized sanity check against attack underfitting across both CNN and Transformer architectures.

D.6 Robustness under Additional White-box Attacks

To rule out attack underfitting and loss-specific overestimation, we evaluate the trained models under other white-box attacks, including PGD-100 and C&W-100 under both ℓ_∞ and ℓ_2 threat models. These evaluations complement AutoAttack/RobustBench-style testing and provide additional evidence that the observed robustness is not an artifact of the default APGD evaluation. Since Muon

Table 9: The robustness comparison (RobustBench (Croce et al., 2021)). "Defense" denotes the baseline method (Croce and Hein, 2022).

Models	Method	Clean	AA
WRN-34-10 (Zagoruyko and Komodakis, 2016)	Defense + Muon (best)	83.16	**46.71**
	Defense + SGD (best)	80.02	44.11
	Defense + AdamW (best)	69.92	42.46
	Defense + Muon (final)	85.96	43.80
	Defense + SGD (final)	82.50	37.23
	Defense + AdamW (final)	72.39	41.37
	AT (Madry et al., 2018)	87.14	44.04
	YOPO (Zhang et al., 2019a)	**87.20**	44.83
	FreeAT (Shafahi et al., 2019)	86.11	41.47
PreActResNet-18 (He et al., 2016b)	Defense + Muon (best)	78.04	43.85
	Defense + SGD (best)	78.60	**45.81**
	Defense + AdamW (best)	65.18	39.65
	Defense + Muon (final)	**85.74**	43.67
	Defense + SGD (final)	82.37	41.71
	Defense + AdamW (final)	69.22	38.64
	FastAT (Wong et al., 2020)	83.34	43.21
	FastAT's variant (Andriushchenko and Flammarion, 2020)	79.84	43.93
WRN-34-20 (Zagoruyko and Komodakis, 2016)	Defense + Muon (best)	82.14	46.58
	Defense + SGD (best)	75.98	41.10
	Defense + AdamW (best)	73.37	42.56
	Defense + Muon (final)	**87.44**	**46.79**
	Defense + SGD (final)	83.64	39.16
	Defense + AdamW (final)	75.17	40.89
ViT-B (Dosovitskiy et al., 2021)	Defense + Muon (best)	60.66	30.36
	Defense + SGD (best)	64.65	**30.97**
	Defense + AdamW (best)	17.93	12.12
	Defense + Muon (final)	66.26	19.93
	Defense + SGD (final)	**71.64**	26.15
	Defense + AdamW (final)	17.24	12.93
ViT-L (Dosovitskiy et al., 2021)	Defense + Muon (best)	60.42	31.79
	Defense + SGD (best)	70.15	**34.26**
	Defense + AdamW (best)	16.98	12.88
	Defense + Muon (final)	68.35	23.03
	Defense + SGD (final)	**72.32**	25.60
	Defense + AdamW (final)	18.05	6.05

affects only the optimization trajectory during training, the deployed classifier remains a standard deterministic differentiable model. Hence, there is no inference-time stochastic transformation over which EOT must average, and no non-differentiable or gradient-obfuscating module for BPDA to approximate (Athalye et al., 2018a;b). This distinguishes our setting from preprocessing defenses (Xu et al., 2018) or randomization methods (Cohen et al., 2019), where adaptive attacks must be tailored to the defense mechanism (Tramer et al., 2020). Next, we evaluate the robustness under additional white-box attacks, transfer attacks, and query-based attacks.

Table 10 checks the stronger white-box attack evaluation results on PreActResNet-18 (He et al., 2016b). The perturbation budgets are set to $\varepsilon_\infty = 8.0/255.0$ and $\varepsilon_2 = 0.5$ for ℓ_∞ and ℓ_2 attacks, respectively. For each norm, SGD, AdamW, and Muon are compared using best and last checkpoints across PGD-100 (Madry et al., 2018) and C&W-100 (Carlini and Wagner, 2017) attacks under ℓ_∞ and ℓ_2 settings. Overall, these evaluations provide complementary white-box stress tests for the reported checkpoints.

Table 11 reports white-box robustness on WRN-34-10 (Zagoruyko and Komodakis, 2016) under PGD-100 and C&W-100 attacks for both ℓ_∞ and ℓ_2 threat models. Across the four training norms, Muon attains competitive white-box robustness in many of the reported settings, particularly under ℓ_1 and ℓ_2 training.

Table 12 presents white-box robustness results on WRN-34-20 (Zagoruyko and Komodakis, 2016) under PGD and C&W attacks with ℓ_∞ and ℓ_2 constraints. Muon remains competitive under these

Table 10: The white-box attack evaluation results on PreActResNet-18 (He et al., 2016b).

Norm	Opt.	Best				Last			
		PGD (ℓ_∞)	C&W (ℓ_∞)	PGD (ℓ_2)	C&W (ℓ_2)	PGD (ℓ_∞)	C&W (ℓ_∞)	PGD (ℓ_2)	C&W (ℓ_2)
$\ell_\infty + \ell_1$	SGD	**43.44**	**43.25**	**66.29**	**65.32**	37.30	38.23	64.96	64.67
	AdamW	39.24	39.55	56.56	55.67	24.74	24.89	57.87	56.12
	Muon	42.76	42.10	65.25	63.56	**39.97**	**40.89**	**68.09**	**67.78**
ℓ_∞	SGD	**48.07**	**48.11**	60.00	**59.19**	43.78	44.19	58.54	58.44
	AdamW	41.76	40.95	49.55	48.53	36.70	35.48	51.64	50.15
	Muon	47.03	45.66	**61.15**	**59.19**	**45.12**	**45.78**	**62.07**	**61.97**
ℓ_1	SGD	19.02	19.51	**59.44**	**59.95**	19.11	19.62	59.53	60.01
	AdamW	24.28	26.39	55.62	55.19	**21.27**	**23.46**	56.04	55.82
	Muon	**25.97**	**27.03**	59.39	58.32	19.76	20.81	**63.67**	**64.23**
ℓ_2	SGD	27.15	28.94	**65.95**	**66.32**	21.27	23.46	63.58	64.02
	AdamW	25.99	28.48	59.71	59.97	24.84	**27.62**	59.82	60.38
	Muon	**29.88**	**31.71**	65.23	64.75	**25.12**	25.80	**67.06**	**67.40**

Table 11: The white-box attack evaluation results on WRN-34-10 (Zagoruyko and Komodakis, 2016).

Norm	Opt.	Best				Last			
		PGD (ℓ_∞)	C&W (ℓ_∞)	PGD (ℓ_2)	C&W (ℓ_2)	PGD (ℓ_∞)	C&W (ℓ_∞)	PGD (ℓ_2)	C&W (ℓ_2)
$\ell_\infty + \ell_1$	SGD	40.48	40.60	**64.59**	**63.73**	33.30	33.89	61.91	61.98
	AdamW	39.47	40.64	61.38	61.05	37.89	39.02	62.80	62.44
	Muon	**41.98**	**41.14**	60.19	58.93	**39.43**	**39.69**	**65.35**	**65.48**
ℓ_∞	SGD	46.53	46.35	59.66	58.65	38.68	39.34	54.03	54.25
	AdamW	44.56	43.93	51.65	50.63	44.21	43.92	50.87	49.89
	Muon	**49.41**	**48.80**	**60.48**	**59.53**	**44.81**	**45.05**	**54.72**	**55.12**
ℓ_1	SGD	22.24	21.26	38.47	37.33	18.85	18.89	58.27	58.55
	AdamW	25.42	26.10	52.07	51.42	**23.29**	**26.12**	56.51	56.57
	Muon	**27.20**	**28.89**	**60.53**	**59.87**	22.50	22.77	**63.32**	**63.48**
ℓ_2	SGD	24.10	25.48	55.21	54.72	23.47	23.44	62.58	62.77
	AdamW	25.54	27.19	53.40	53.03	23.75	25.18	61.53	62.36
	Muon	**28.77**	**28.96**	**68.33**	**68.59**	**28.85**	**29.04**	**68.76**	**68.99**

Table 12: The white-box attack evaluation results on WRN-34-20 (Zagoruyko and Komodakis, 2016).

Norm	Opt.	Best				Last			
		PGD (ℓ_∞)	C&W (ℓ_∞)	PGD (ℓ_2)	C&W (ℓ_2)	PGD (ℓ_∞)	C&W (ℓ_∞)	PGD (ℓ_2)	C&W (ℓ_2)
$\ell_\infty + \ell_1$	SGD	40.58	39.12	62.14	60.47	35.23	35.53	62.25	62.42
	AdamW	38.34	37.86	59.50	58.12	38.34	37.86	59.49	58.12
	Muon	**43.03**	**42.41**	**63.19**	**61.66**	**41.78**	**42.07**	**62.82**	**62.96**
ℓ_∞	SGD	45.13	43.08	57.92	55.81	40.13	40.47	54.90	55.36
	AdamW	44.81	45.05	50.96	50.22	44.17	43.17	50.60	49.25
	Muon	**49.71**	**48.56**	**60.39**	**59.07**	**47.71**	**48.01**	**57.83**	**58.08**
ℓ_1	SGD	23.91	23.80	43.44	42.51	20.04	20.20	60.65	60.79
	AdamW	23.79	**26.23**	53.03	52.74	21.26	22.64	57.92	59.11
	Muon	**24.20**	24.43	**65.88**	**66.05**	**23.65**	**23.74**	**65.36**	**65.44**
ℓ_2	SGD	26.23	26.38	64.42	64.66	25.67	25.78	65.06	65.22
	AdamW	26.83	26.58	61.62	62.95	27.53	26.52	61.33	62.56
	Muon	**30.71**	**30.88**	**70.37**	**70.49**	**30.29**	**30.50**	**70.55**	**70.69**

additional attacks, particularly after ℓ_2 training, but we do not interpret this as improved last-checkpoint stability.

Table 13: The black-box transfer attack results on WRN-34-10 (Zagoruyko and Komodakis, 2016).

Norm	Opt.	Best				Last			
		PGD (ℓ_∞)	C&W (ℓ_∞)	PGD (ℓ_2)	C&W (ℓ_2)	PGD (ℓ_∞)	C&W (ℓ_∞)	PGD (ℓ_2)	C&W (ℓ_2)
$\ell_\infty + \ell_1$	SGD	**74.56**	**74.32**	**77.06**	**76.97**	77.97	78.08	80.20	80.19
	AdamW	71.76	71.58	74.17	74.05	74.39	74.34	77.35	77.30
	Muon	68.68	68.45	71.80	71.72	**82.06**	**81.96**	**84.14**	**84.03**
ℓ_∞	SGD	76.38	76.27	78.52	78.43	79.48	79.32	81.23	81.29
	AdamW	66.64	66.58	68.37	68.40	68.44	68.63	70.66	70.78
	Muon	**79.83**	**79.66**	**81.72**	**81.65**	**83.40**	**83.44**	**84.96**	**85.06**
ℓ_1	SGD	41.90	41.69	45.52	45.57	79.50	79.22	82.61	82.46
	AdamW	59.98	59.85	63.61	63.66	68.60	68.50	73.46	73.60
	Muon	**72.50**	**71.71**	**76.91**	**76.90**	**83.30**	**82.90**	**86.66**	**86.60**
ℓ_2	SGD	70.98	70.94	74.47	74.29	83.82	83.93	86.29	86.21
	AdamW	66.00	65.94	70.75	70.88	81.81	81.66	84.65	84.61
	Muon	**86.86**	**86.74**	**89.07**	**88.98**	**86.86**	**86.71**	**89.19**	**89.10**

Table 14: The black-box transfer attack results on WRN-34-20 (Zagoruyko and Komodakis, 2016).

Norm	Opt.	Best				Last			
		PGD (ℓ_∞)	C&W (ℓ_∞)	PGD (ℓ_2)	C&W (ℓ_2)	PGD (ℓ_∞)	C&W (ℓ_∞)	PGD (ℓ_2)	C&W (ℓ_2)
$\ell_\infty + \ell_1$	SGD	70.67	70.51	73.52	73.61	79.19	78.90	81.37	81.38
	AdamW	68.34	68.23	70.95	70.98	68.34	68.23	70.95	70.97
	Muon	**72.05**	**71.79**	**75.00**	**74.98**	**84.35**	**84.13**	**85.73**	**85.54**
ℓ_∞	SGD	72.19	71.94	74.39	74.30	80.72	80.61	82.30	82.38
	AdamW	69.47	69.45	71.50	71.47	71.23	71.31	73.35	73.39
	Muon	**78.82**	**78.48**	**80.66**	**80.56**	**85.10**	**84.96**	**86.48**	**86.42**
ℓ_1	SGD	49.58	49.40	54.09	54.06	80.49	80.19	83.99	83.96
	AdamW	62.73	62.49	66.79	66.85	76.56	76.20	80.46	80.38
	Muon	**84.49**	**84.15**	**87.64**	**87.64**	**84.15**	**83.76**	**87.49**	**87.52**
ℓ_2	SGD	84.78	84.56	87.20	87.03	85.13	84.70	87.29	87.36
	AdamW	81.56	81.56	84.02	83.92	81.11	81.03	83.82	83.67
	Muon	**88.19**	**88.11**	**90.26**	**90.21**	**88.30**	**88.15**	**90.35**	**90.32**

D.7 Robustness under Black-box Attacks

D.7.1 Robustness under Transfer-based Attacks

Black-box transfer attacks provide a complementary stress test under restricted attacker information. We do not use transfer robustness as evidence against gradient obfuscation; that role is served by the strong white-box evaluations above. Instead, these experiments characterize robustness when the attacker accesses only a surrogate model.

We first evaluate a naturally trained PreActResNet-18 (He et al., 2016b) under adversarial attacks. Its robust accuracies under PGD-100 (ℓ_∞), C&W-100 (ℓ_∞), PGD-100 (ℓ_2), and C&W-100 (ℓ_2) are 2.34%, 2.85%, 16.73%, and 16.22%, respectively. Then, we transfer the adversarial examples to WRN-34-10 (Zagoruyko and Komodakis, 2016). Table 13 shows that Muon remains competitive under transferred attacks.

Similar settings are implemented on WRN-34-20 (Zagoruyko and Komodakis, 2016). Table 14 reports black-box transfer robustness on WRN-34-20 (PreActResNet-18 as the source model). Muon consistently outperforms SGD and AdamW in nearly all settings, with especially large margins under ℓ_1 and ℓ_2 training. Notably, Muon maintains strong last-checkpoint robustness, achieving 85.10/84.96 under ℓ_∞ transfer attacks for ℓ_∞ training and 90.35/90.32 under ℓ_2 transfer attacks for ℓ_2 training. These results indicate that Muon's gains are not limited to white-box optimization paths, but generalize to transferred adversarial examples, suggesting stronger cross-model robustness and better resistance to real black-box threat models.

Table 15: The black-box query attack results on PreActResNet-18 (He et al., 2016b).

		Best				Last			
		Square		RayS		Square		RayS	
Norm	Opt.	RA	Avg. Q	RA	Avg. Q	RA	Avg. Q	RA	Avg. Q
$\ell_\infty + \ell_1$	SGD	**50.16**	916.9	**54.53**	2040.6	47.27	888.2	51.84	2081.4
	AdamW	45.51	911.3	47.48	1910.9	42.71	875.4	46.17	1972.2
	Muon	48.96	894.7	53.41	2073.7	**50.82**	890.3	**55.93**	2091.8
ℓ_∞	SGD	**52.83**	837.3	**57.71**	1808.1	50.61	847.5	57.11	2020.7
	AdamW	44.74	805.7	48.56	1685.9	45.67	770.5	49.64	1750.3
	Muon	50.73	880.5	57.09	1942.0	**53.67**	848.4	**59.83**	2080.5
ℓ_1	SGD	30.75	853.0	39.07	1998.0	30.99	865.2	39.97	2031.6
	AdamW	**36.15**	896.5	40.89	2207.1	**34.10**	888.3	39.93	2074.1
	Muon	35.38	909.3	**42.03**	2006.6	33.13	935.6	**43.80**	2244.1
ℓ_2	SGD	41.61	1042.6	53.04	2389.3	37.34	1035.7	49.36	2414.7
	AdamW	40.07	963.8	47.51	2167.1	39.79	982.6	48.03	2225.2
	Muon	**42.94**	998.9	**54.05**	2473.7	**39.96**	1065.9	**53.58**	2547.7

Table 16: Training time for WRN-34-20 (Zagoruyko and Komodakis, 2016) under different threat norms. Times are converted from seconds to hours (h) and minutes (m), rounded to the nearest minute.

Norm	SGD	AdamW	Muon
$\ell_\infty + \ell_1$	51h59m	52h06m	52h45m
ℓ_∞	51h47m	51h52m	52h29m
ℓ_1	52h18m	52h21m	53h04m
ℓ_2	51h49m	51h56m	52h32m

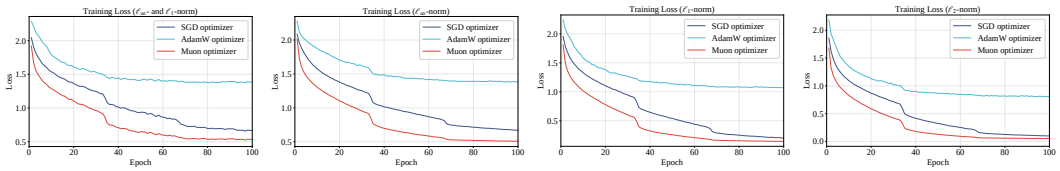
D.7.2 Robustness under Query-based Attacks

Query-based attacks iteratively optimize adversarial examples via direct target-model feedback. In real-world security scenarios, attackers may query neural network services through Application Programming Interfaces (APIs), where the model's internal structure remains opaque and only confidence scores or final predictions are returned. The evaluation results on query-based attacks can provide additional complementary black-box stress tests. Thus, we assess PreActResNet-18 under two query-based black-box attacks on the full CIFAR-10 test set: Square Attack (Andriushchenko et al., 2020), a score-based random-search method (query budget 5,000), and RayS (Chen and Gu, 2020), a decision-based hard-label method (query budget 10,000). Both operate under the ℓ_∞ threat model with $\varepsilon_\infty = 8.0/255.0$. For each model, we report the robust accuracy (RA) and the average number of queries over successful attacks (Avg. Q), for both the best and last checkpoints. Square is score-based (5000-query budget), and RayS is decision-based hard-label (10000-query budget). Both under ℓ_∞, $\varepsilon = 8.0/255.0$. Table 15 indicates that models trained with Muon show relatively stable performance under various training-norm settings.

D.8 Training Time and Loss of Different Optimizers

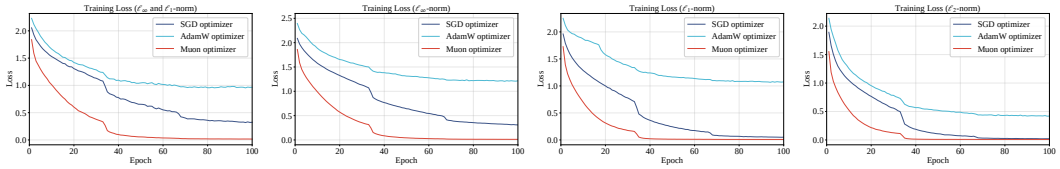
Table 16 reports the AT time across different optimizers. We train WRN-34-20 for 100 epochs. The training time is comparable across the three optimizers.

Figures 4, 5, 6, 7, and 8 depict the training loss curves during AT on CIFAR-10 (Krizhevsky, 2009) with different optimizers on PreActResNet-18 (He et al., 2016b), WRN-34-10 (Zagoruyko and Komodakis, 2016), WRN-34-20 (Zagoruyko and Komodakis, 2016), ViT-B (Dosovitskiy et al., 2021),



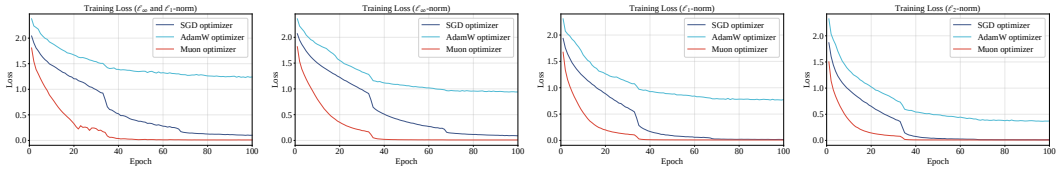
(a) $\ell_\infty + \ell_1$ (b) ℓ_∞ (c) ℓ_1 (d) ℓ_2

Figure 4: Training loss on PreActResNet-18 (He et al., 2016b).



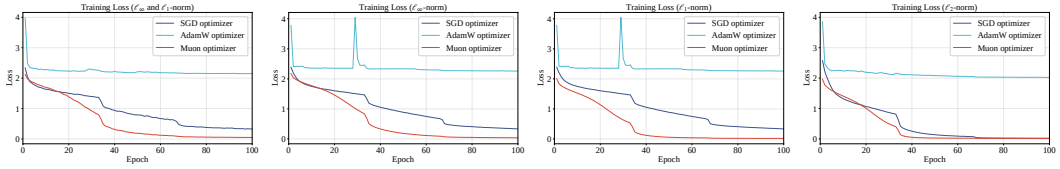
(a) $\ell_\infty + \ell_1$ (b) ℓ_∞ (c) ℓ_1 (d) ℓ_2

Figure 5: Training loss on WRN-34-10 (Zagoruyko and Komodakis, 2016).



(a) $\ell_\infty + \ell_1$ (b) ℓ_∞ (c) ℓ_1 (d) ℓ_2

Figure 6: Training loss on WRN-34-20 (Zagoruyko and Komodakis, 2016).



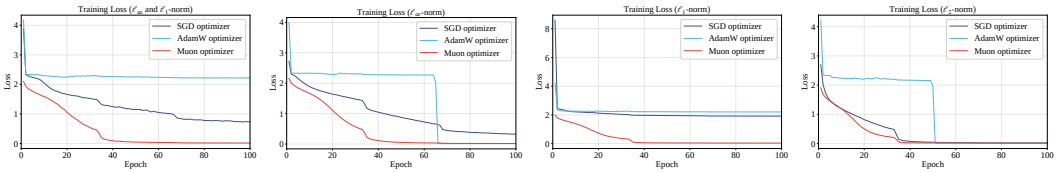
(a) $\ell_\infty + \ell_1$ (b) ℓ_∞ (c) ℓ_1 (d) ℓ_2

Figure 7: Training loss on ViT-B (Dosovitskiy et al., 2021).

and ViT-L (Dosovitskiy et al., 2021). In the reported runs, Muon often reaches lower adversarial training loss earlier than SGD and AdamW and exhibits comparatively smooth loss trajectories. These empirical observations are consistent with the favorable nuclear-norm-aligned term in Theorem 3.3. In contrast, on the ViT architecture (Dosovitskiy et al., 2021), it is difficult to use an adaptive optimizer such as AdamW for AT. When AdamW exhibits numerical instability on ViTs, we report the resulting evaluation accuracies in the main tables rather than marking the runs as "failed". Such cases are described as severe degradation or near-random performance when the reported robust accuracies are close to chance level.

D.9 Spectral Analysis: Spectral Norms and Condition Numbers

As a lightweight diagnostic of optimizer-induced spectral geometry, we track the mean per-layer condition number $\kappa = \sigma_{\max}/\sigma_{\min}$ every 5 epochs throughout AT on PreActResNet-18 (Figure 9). Muon maintains a low and stable $\kappa \approx 5$ throughout training, comparable to SGD, while AdamW's κ grows rapidly to ≈ 50 within the first 30 epochs. Correlating κ with robust accuracy (Figure 10) reveals two distinct regimes: SGD and Muon concentrate in the low κ, high-robustness region, while AdamW occupies the high κ, lower-robustness region. This correlation suggests that matrix



(a) $\ell_\infty + \ell_1$ (b) ℓ_∞ (c) ℓ_1 (d) ℓ_2

Figure 8: Training loss on ViT-L (Dosovitskiy et al., 2021).

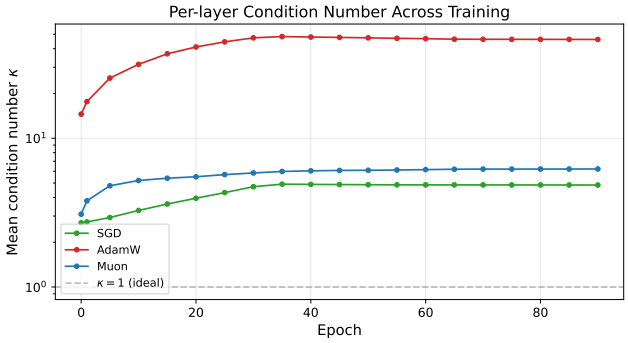


Figure 9: Per-layer condition number change during training.

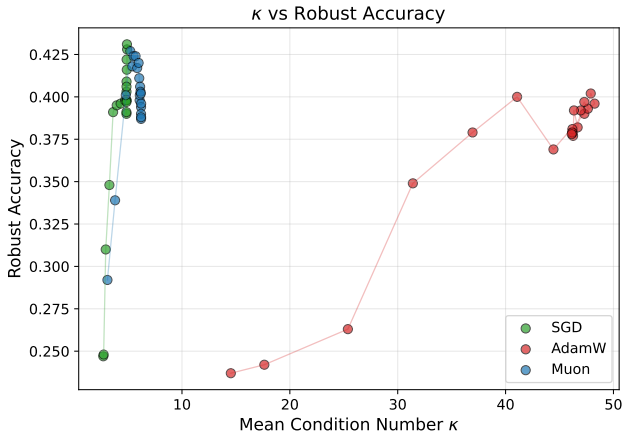
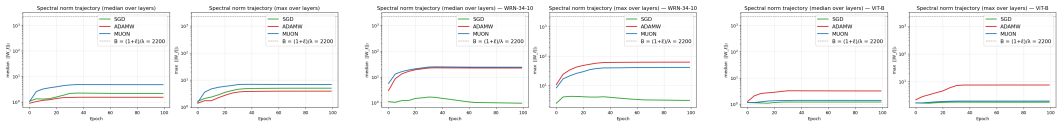


Figure 10: Spectral conditioning and robustness.

anisotropy may serve as a useful descriptive diagnostic of optimizer-induced trajectories: SGD and Muon occupy a similar low-κ, high-robustness regime, whereas AdamW exhibits substantially larger κ. We therefore interpret conditioning as a diagnostic of optimizer-induced geometry rather than as a causal predictor of robustness. SGD often attains the smallest weight norms, while Muon does not uniformly dominate it in magnitude. Instead, Muon's weight-matrix behavior is generally closer to SGD than to AdamW and avoids the large growth observed with AdamW in several settings. Our claim therefore concerns spectral behavior of the optimization trajectory rather than strict norm minimization, with Muon remaining closer to SGD than to AdamW in these diagnostics.

Figure 11 compares weight-norm trajectories on three architectures. For Muon, the observed norms remain far below the conservative bound of Theorem 3.1, confirming that the theorem should be interpreted as a stability guarantee rather than a tight predictor. Muon and SGD plateau near the first learning-rate decay, whereas AdamW exhibits architecture-dependent behavior, attaining smaller norms on PreActResNet-18 but larger norms on WRN-34-10 and ViT-B. We interpret this comparison as an empirical diagnostic rather than a causal explanation of robustness.



(a) PreActResNet-18 (b) WRN-34-10 (c) ViT-B

Figure 11: Weight spectral-norm trajectories during AT.

Table 17: Robustness of TRADES-trained models using different optimizers.

	SGD		AdamW		Muon	
Metric	Best	Last	Best	Last	Best	Last
Clean Accuracy	80.54	80.54	80.13	83.19	81.70	**84.64**
FGSM	53.47	53.47	55.40	56.44	**57.60**	57.17
PGD-20	49.01	48.98	51.24	47.63	**52.48**	47.36
PGD-100	48.74	48.81	51.14	47.09	**52.29**	46.71
CW-20	46.86	46.92	48.97	48.33	**51.14**	48.50
CW-100	46.79	46.74	48.87	48.00	**50.91**	48.13
AA	46.80	46.80	47.80	46.30	**49.00**	46.20

Table 18: Robustness of TRADES+AWP models with different optimizers.

	SGD		AdamW		Muon	
Metric	Best	Last	Best	Last	Best	Last
Clean Accuracy	81.28	81.26	80.08	78.18	**82.34**	82.13
FGSM	57.68	57.75	56.68	52.83	**59.09**	57.38
PGD-20	54.78	54.70	52.78	46.74	**56.00**	52.65
PGD-100	54.75	54.65	52.63	46.53	**56.02**	52.37
CW-20	51.19	51.10	49.96	45.76	**53.30**	51.98
CW-100	51.14	50.99	49.82	45.60	**53.25**	51.89
AA	50.90	50.10	49.80	45.70	**52.20**	50.50

D.10 Does the Optimizer Effect Transfer Across Other AT Objectives?

We further evaluate Muon under TRADES (Zhang et al., 2019b) on CIFAR-10 Krizhevsky (2009). To accommodate AdamW, all models are trained for 74 epochs at a learning rate of 0.01 and for the final two epochs at 0.001. Table 17 reports the best and final checkpoints for SGD, AdamW, and Muon. Muon achieves the strongest best-checkpoint robustness, attaining the highest accuracy under all evaluated attacks, including 49.00% under AutoAttack. Its final checkpoint also yields the highest clean accuracy (84.64%), but with a noticeable robustness drop. By contrast, SGD changes little between the best and final checkpoints, indicating greater training stability.

Table 18 compares SGD, AdamW, and Muon under TRADES+AWP (Wu et al., 2020) training with a learning rate of 0.01 for epochs 1–100, 10^{-3} for 101–150, and 10^{-4} for 151–200. For TRADES+AWP, Muon's selected checkpoint attains 82.34% clean accuracy and 52.20% AutoAttack accuracy, corresponding to point differences of $+1.06$ and $+1.30$ percentage points relative to SGD, respectively. AdamW degrades late in training, whereas SGD remains stable and Muon retains strong final performance.