SPECTRAL ALLOCATION: WHY MUON OUTPERFORMS ADAM, AND HOW TO IMPROVE MUON

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon accelerates language-model pretraining, but why its uniform spectral scaling works and where it falls short remain unclear. We probe singular directions of momentum buffers at training checkpoints and estimate their per-rank loss-optimal step sizes on held-out batches. Across models and training stages, we observe a stable spectral profile. The loss-optimal step size is relatively flat across the step-size-tolerant bulk but drops sharply at the leading direction, termed the step-size-sensitive head. This profile provides a spectral-allocation account of the performance ordering from SGD through Adam to Muon. It also suggests that Muon still underuses the tolerant bulk. We introduce **Spectral-Aware Muon (SAMuon)**, which holds the head at the same scale as Muon while increasing the scale of the bulk using a static spectral prior. SAMuon follows the measured profile, whereas SAMuon-lite uses a simpler two-level approximation. Neither variant adds persistent optimiser state. Under idealised exact whitening, both variants retain the asymptotic convergence rates of Muon. Across evaluated language-model settings with 124M to 1B parameters, both variants achieve lower final validation loss than tuned AdamW and Muon (Scion implementation). SAMuon reaches the same validation loss as Muon using 13.3% to 24.0% fewer training tokens. SAMuon-lite retains most of this improvement and runs at nearly the same wall-clock speed as Muon in the measured 1B-model setting.

1 INTRODUCTION

Orthogonal optimisers such as Muon (Jordan et al., 2024b; Liu et al., 2025) have accelerated Transformer-based large language model (LLM) pretraining relative to Adam (Kingma & Ba, 2015). Muon whitens each momentum matrix, assigning approximately equal update coefficients to its singular directions. Existing analyses based on spectral-norm (Bernstein & Newhouse, 2024; Chen et al., 2026; Bernstein & Newhouse, 2025; Kovalev, 2025) and curvature (Wang et al., 2026; Su, 2025) offer explanations for Muon's success. However, the loss-landscape structure that makes rescaling along the momentum buffer's singular directions effective remains insufficiently understood. This motivates a more fine-grained question: **how does the loss geometry vary across these singular directions in real Transformer pretraining, and what does this imply for optimiser design?**

To distinguish progress on unseen data from refitting the gradient batch, we introduce *out-of-sample spectral probing*, an offline analysis of saved training checkpoints that estimates the loss-optimal step along each singular direction of the momentum buffer using held-out batches. After an initial burn-in stage, the profile becomes highly anisotropic yet has a consistent shape across the probed batches, training stages and model scales. Loss-optimal step sizes are large and approximately flat across the bulk, which we term the **Tolerant Bulk**. Over the leading ranks, they decrease approximately linearly with log rank towards the single leading singular direction, which has the smallest loss-optimal step and the highest measured directional curvature. We term this step-size-sensitive direction the **Sensitive Head**. Viewed through this spectral profile (Figure 2), momentum-based stochastic gradient descent (SGD) favours the least tolerant directions by scaling them in proportion to their singular values; Adam's coordinate-wise rescaling reduces this imbalance in our measurements only partially; and Muon's uniform whitening increases the relative allocation to the bulk, bringing its allocation closer to the measured profile. This yields a unified **Spectral Allocation** account of *why Muon outperforms Adam, which in turn outperforms SGD* in Transformer pretraining.

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.

The same profile also reveals room to improve Muon, as its uniform allocation still leaves the bulk below its measured loss-optimal scale. We therefore propose *Spectral-Aware Muon (SAMuon)*, which amplifies the bulk while keeping the head at the Muon scale. Unlike SOAP (Vyas et al., 2025), which maintains additional state for adaptive rescaling in Shampoo's (Gupta et al., 2018) eigenbasis, SAMuon exploits the observed stability through a static, measurement-derived prior on spectral allocation. The full SAMuon follows the log-rank-linear profile using a low-rank randomised SVD, whereas SAMuon-lite uses power iteration to implement a simpler two-level allocation. Neither variant adds persistent optimiser state beyond Muon's momentum buffer, and both introduce only one additional tuned scale parameter, γ. By extending the convergence analysis framework in Shen et al. (2026), we also show that their idealised exact-whitening forms retain Muon's asymptotic convergence rates.

We validate both SAMuon variants on "modded-nanogpt" (Jordan et al., 2024a) pretraining following the protocol in Pethick et al. (2025), across models from 124M to 1B parameters and batch sizes from 1024 to 4096 and token budgets ranging from $1\times$ to $4\times$ the Chinchilla-optimal allocation (Hoffmann et al., 2022). Hyperparameters are tuned per batch size at 124M and transferred across model scales. Specifically, the 124M-sized model with $\sim 4\times$ the Chinchilla-optimal token budget tests whether gains persist in the over-trained regime (Wen et al., 2026). By better exploiting the tolerant bulk, the full SAMuon improves on AdamW and the Muon (Scion implementation) baseline in each trained model and batch size combination, reaching the same final validation loss as Muon with 13.3%–24.0% fewer tokens. SAMuon-lite recovers most of this improvement (13.3%–22.1%) with nearly unchanged wall-clock speed as Muon.

Our contributions are: **1) Out-of-sample spectral analysis:** We introduce offline spectral probing on held-out batches and identify a stable loss-optimal step-size profile with a *Sensitive Head* and *Tolerant Bulk*, yielding a unified *Spectral Allocation* account of the SGD–Adam–Muon performance ordering and revealing the limitations of Muon's uniform spectral allocation. **2) SAMuon optimisers:** We develop SAMuon and SAMuon-lite, which distil the measured structure into static, head-anchored spectral allocations without adding persistent optimiser state. **3) Empirical and theoretical validation:** Across evaluated 124M–1B parameter pretraining settings, SAMuon and SAMuon-lite match tuned Muon's final validation loss using 13.3%–24.0% and 13.3%–22.1% fewer tokens, respectively. Both variants also retain Muon's asymptotic convergence rates under idealised exact whitening.

2 Related work[1]

Theoretical Interpretations of Muon's Success. Norm-geometric analyses characterise Muon's orthogonalised direction through spectral-norm steepest descent and trust-region formulations (Bernstein & Newhouse, 2024; Kovalev, 2025), but do not resolve how loss-optimal update scales vary across momentum singular directions. Su (2025) analyses spectral homogenisation under an isotropic curvature model, with orthogonalisation becoming optimal in a limiting regime. Our curvature measurements reveal anisotropy across the probed directions. The closely related work of Wang et al. (2026) links the advantage of Muon to a smaller normalised directional-sharpness penalty and reports alignment between the gradient and the dominant *in-sample* Hessian subspace. We extend this curvature perspective by estimating *out-of-sample* loss-optimal step sizes along individual singular directions of momentum buffers at training checkpoints. Our analysis shows why spectral allocation in the momentum buffer's singular basis matters for optimiser design, and the measured loss-optimal step-size profile provides a practical guide for allocating update scales across singular directions.

Spectral Analysis of Deep Neural Networks. Prior work examines Hessian anisotropy (Sagun et al., 2018), inter-block curvature heterogeneity in Transformers (Zhang et al., 2024), and gradient spectral anisotropy (Refael et al., 2025; Cao et al., 2026). We complement these studies by measuring the loss response along singular directions of momentum buffers within parameter blocks. Magakyan et al. (2026) show that retaining only a small fraction of leading singular directions degrades Muon, which is consistent with our finding that Muon's advantage lies in better utilising the tolerant spectral bulk.

Adaptive Methods and Practical Muon Variants. Beyond convergence speed, a central appeal of Muon is its low memory overhead. Adaptive preconditioners such as SOAP (Vyas et al., 2025) and several Muon variants (Si et al., 2025; Du & Su, 2026; Zhang et al., 2026b) require additional

[1]We present a concise overview of the related work here. See Appendix A for the extended discussion.

persistent optimiser state relative to Muon. COSMOS (Liu et al., 2026) reduces this memory overhead relative to SOAP by limiting adaptiveness to the leading moment subspace. LiMuon (Huang et al., 2026) uses low-rank randomised SVD to compress stored momentum, whereas SAMuon uses it to shape spectral allocation. NorMuon (Li et al., 2026) applies per-neuron rescaling which is potentially complementary to SAMuon. SAMuon preserves the memory advantage of Muon through a static, measured spectral prior without additional persistent optimiser state. We build on Scion (Pethick et al., 2025) for its normalisation scheme, which supports hyperparameter transfer across model sizes.

3 PRELIMINARIES

The Muon optimiser (Jordan et al., 2024b) operates independently on each 2D weight matrix $\mathbf{W}^{(t)} \in \mathbb{R}^{m\times n}$ of the target model, where we take $m \le n$ without loss of generality. At step t, given the mini-batch gradient $\mathbf{G}^{(t)} = \nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W}^{(t)}; \mathcal{B}_t)$, Muon first integrates the gradient into a momentum buffer $\mathbf{M}^{(t)}$, and it is this buffer that is orthogonalised to produce the update:

$$\mathbf{M}^{(t)} = \mu\,\mathbf{M}^{(t-1)} + (1-\mu)\,\mathbf{G}^{(t)}, \tag{1}$$

$$\mathbf{O}^{(t)} = \mathrm{NS}\big(\mathbf{M}^{(t)}\big) \;\approx\; \big(\mathbf{M}^{(t)}\mathbf{M}^{(t)\top}\big)^{-1/2}\mathbf{M}^{(t)} \;=\; \mathbf{U}\mathbf{V}^{\top}, \tag{2}$$

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta\,\kappa\,\mathbf{O}^{(t)}, \tag{3}$$

where μ is the momentum coefficient and η is the learning rate. For the Scion implementation of Muon used in our paper, a fixed scaling factor $\kappa = \sqrt{d_{\text{out}}/d_{\text{in}}}$ is applied along with η, where d_{out} and d_{in} are the matrix's original output and input dimensions, respectively (Pethick et al., 2025).

For simplicity, we also assume $\mathbf{M}^{(t)}$ to have full row rank (with a rank of m). We write the (reduced) singular value decomposition (SVD) of $\mathbf{M}^{(t)}$ in Equation (2) as $\mathbf{M}^{(t)} = \mathbf{U}\mathbf{S}\mathbf{V}^{\top} = \sum_i \sigma_i\,\boldsymbol{u}_i\boldsymbol{v}_i^{\top}$, where $\mathbf{S}$ is the diagonal matrix of singular values $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_m \ge 0$, $\boldsymbol{u}_i, \boldsymbol{v}_i$ are the corresponding left and right singular vectors, and the index i is the *spectral rank index* of the singular component $\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^{\top}$. The operator $\mathrm{NS}(\cdot)$ in Equation (2) denotes the Newton–Schulz iteration, which approximates the polar factor $\mathbf{U}\mathbf{V}^{\top}$ of its input using only a small number of matrix products, avoiding an explicit SVD (see Appendix J.4 for details of the Newton–Schulz scheme adopted from LaCT (Zhang et al., 2026a)). Exact whitening maps every non-zero singular value to 1, so ideally all m singular directions of each momentum buffer receive the same update coefficient $\eta\kappa$.

4 SPECTRAL PROBING FOR THE LOSS-OPTIMAL STEP SIZE PROFILE

To understand the loss geometry underlying the benefits and limitations of uniform whitening in Muon (Jordan et al., 2024b), we conduct a fine-grained spectral analysis of the Transformer pretraining loss landscape. Using *out-of-sample spectral probing*, we estimate loss-optimal step sizes along singular directions of momentum buffers on held-out batches at checkpoints saved at different training stages.

4.1 METHODOLOGY: SPECTRAL PROBING

Our *spectral probing* procedure builds on the offline directional-probing methodology of Wu et al. (2024), where we estimate the loss-optimal step sizes along singular directions of momentum buffers grouped by rank across the model.

Per-Rank Probe Construction. Let $\mathcal{L}(\boldsymbol{\theta}_t; \mathcal{B})$ be the loss on a batch $\mathcal{B}$ at the checkpoint parameter state $\boldsymbol{\theta}_t$, where t denotes the training iteration of the checkpoint. We probe the spectrum of the updated momentum buffer for the $N_{\mathrm{W}} = 72$ two-dimensional weight matrices in the model, indexed by j. For each matrix, we integrate the gradient $\mathbf{G}_j$ of a *gradient batch* $\mathcal{B}_{\mathrm{g}}$ into the momentum buffer $\mathbf{M}_j$ stored at the checkpoint, $\mathbf{M}_j \leftarrow \mu\,\mathbf{M}_j + (1-\mu)\,\mathbf{G}_j$, following the same momentum-update rule used during training. Crucially, the out-of-sample response on a *held-out batch* $\mathcal{B}_{\mathrm{o}} \neq \mathcal{B}_{\mathrm{g}}$ is measured, rather than on the gradient batch that was used to form the update. Our comparison shows that standard in-sample probing reports a far more permissive spectral tail (the small-singular-value end of the spectrum), largely reflecting refitting of the gradient batch (Figure 8 in Appendix F).

The SVD of the updated momentum buffer is $\mathbf{M}_j = \sum_d \sigma_{j,d}\,\boldsymbol{u}_{j,d}\boldsymbol{v}_{j,d}^{\top}$. The singular components are indexed by their *rank* d, *i.e.* the position of $\sigma_{j,d}$ in the descending spectrum, and grouped across weight

matrices by rank. This grouping is well-defined for GPT-2-style architectures because every probed matrix shares the same minimum dimension, the model width d_{model}, so each matrix carries exactly d_{model} singular components and the d-th component can be aggregated model-wide. Our analysis therefore focuses on within-block spectral directions rather than resolving differences between blocks. The *rank-d probe* $\Delta\boldsymbol{\theta}^{(d)}$ combines the Scion-scaled rank-d direction $\mathbf{P}_{j,d}$ from every probed matrix into a pseudo parameter update

$$\mathbf{P}_{j,d} = -\kappa_j \, \boldsymbol{u}_{j,d}\boldsymbol{v}_{j,d}^{\top} , \qquad \Delta\boldsymbol{\theta}^{(d)} = \operatorname{vec}\left(\{\mathbf{P}_{j,d}\}_{j=1}^{N_{\mathrm{W}}}\right) \quad (d = 1, \ldots, d_{\text{model}}), \tag{4}$$

where $\operatorname{vec}(\cdot)$ concatenates the entries of all N_{W} matrices into a single vector acting on the parameter state $\boldsymbol{\theta}_t$, and κ_j is the fixed per-weight-matrix scaling of Scion for the j-th matrix. Note that $\mathbf{P}_{j,d}$ has the same orientation and scaling as the rank-d component that the actual Muon (Scion implementation) optimiser applies to the j-th momentum buffer, which makes the probe's loss-optimal step size directly comparable with the effective learning rate used in the Muon (Scion) optimiser's trajectory (the red dashed line in the middle panel of Figure 1). Also, every matrix contributes exactly one rank-one component to each probe, which prevents the curvature measurements from being confounded by unequal numbers of aggregated components from different matrices.

Measurement via Local-Quadratic Approximation. For each probe, we primarily measure the *loss-optimal step size*, defined as the step size that minimises the *out-of-sample* loss along the probe direction, $\eta_d^* = \arg\min_{\eta \geq 0} \mathcal{L}(\boldsymbol{\theta}_t + \eta\, \Delta\boldsymbol{\theta}^{(d)}; \mathcal{B}_{\mathrm{o}})$. Running an exact line search for all ranks of every checkpoint is prohibitively expensive. Therefore, we estimate η_d^* from the local-quadratic approximation (LQA) of the loss along the probe direction,

$$\mathcal{L}(\boldsymbol{\theta}_t + \eta\, \Delta\boldsymbol{\theta}; \mathcal{B}_{\mathrm{o}}) \;\approx\; \mathcal{L}(\boldsymbol{\theta}_t; \mathcal{B}_{\mathrm{o}}) \;+\; \eta\, \boldsymbol{g}^{\top} \Delta\boldsymbol{\theta} \;+\; \tfrac{1}{2}\, \eta^2\, \Delta\boldsymbol{\theta}^{\top} \mathbf{H}\, \Delta\boldsymbol{\theta}, \tag{5}$$

where $\boldsymbol{g}$ and $\mathbf{H}$ denote the full-parameter gradient and Hessian of $\mathcal{L}(\cdot\,; \mathcal{B}_{\mathrm{o}})$ at $\boldsymbol{\theta}_t$. For a descent direction with positive directional curvature, minimising Equation (5) over η yields

$$\eta_d^* \;\approx\; -\frac{\boldsymbol{g}^{\top} \Delta\boldsymbol{\theta}^{(d)}}{\Delta\boldsymbol{\theta}^{(d)\top} \mathbf{H}\, \Delta\boldsymbol{\theta}^{(d)}}. \tag{6}$$

The directional curvature in the denominator does not require an explicit Hessian, since rearranging Equation (5) at a small displacement ε gives the finite-difference estimate

$$\Delta\boldsymbol{\theta}^{\top} \mathbf{H}\, \Delta\boldsymbol{\theta} \;\approx\; \frac{2}{\varepsilon^2}\Big[\, \mathcal{L}(\boldsymbol{\theta}_t + \varepsilon\, \Delta\boldsymbol{\theta}; \mathcal{B}_{\mathrm{o}}) \;-\; \mathcal{L}(\boldsymbol{\theta}_t; \mathcal{B}_{\mathrm{o}}) \;-\; \varepsilon\, \boldsymbol{g}^{\top} \Delta\boldsymbol{\theta} \Big], \tag{7}$$

which reuses the loss and gradient and consumes only *one additional forward pass* per probe. The resulting collection $\{\eta_d^*\}_{d=1}^{d_{\text{model}}}$ constitutes the *out-of-sample loss-optimal spectral step size profile* of the loss landscape (the *measured profile* for short) at checkpoint $\boldsymbol{\theta}_t$, which Figure 1 tracks across training stages. We exclude probes that fail the descent or positive-curvature condition, accounting for only 0.52% of all probes. Appendix C details these conditions and validates the accuracy of the local-quadratic approximation in the probed Transformer loss landscapes.

4.2 Observations: a stable, anisotropic loss-optimal step size profile

Figure 1 shows the loss-optimal spectral step size profile along a Muon (Scion) trajectory of a 64M "modded-nanogpt" model (Jordan et al., 2024a; Pethick et al., 2025). At each checkpoint, we probe 32 ranks on 20 held-out batches and interpolate between sampled ranks for visualisation. After early training, the profile shows a highly anisotropic yet stable shape across checkpoints and batches: a relatively flat region of larger loss-optimal step sizes, the *Tolerant Bulk*, and a sharp drop towards the single leading singular direction, the *Sensitive Head*. The same qualitative structure also appears along an AdamW trajectory and at width 768 (Appendices D and E). This observed stability motivates using the measured profile to construct a static spectral prior for optimiser design.

A sensitive head constrains Muon's uniform scale. The leading ranks dominate the buffer's spectral energy, yet their loss-optimal step sizes are substantially smaller than those of the bulk. The leading direction's loss-optimal step size approximately matches Muon's applied scale (red dashed line in Figure 1). This suggests that the head operates near the Edge-of-Stability (Cohen et al., 2021) and constrains the shared step scale that Muon applies to all singular directions.

The tolerant bulk is under-exploited. The largest increase in loss-optimal step sizes occurs between ranks 1 and 2, followed by a slower rise towards a relatively flat bulk at several times the head's

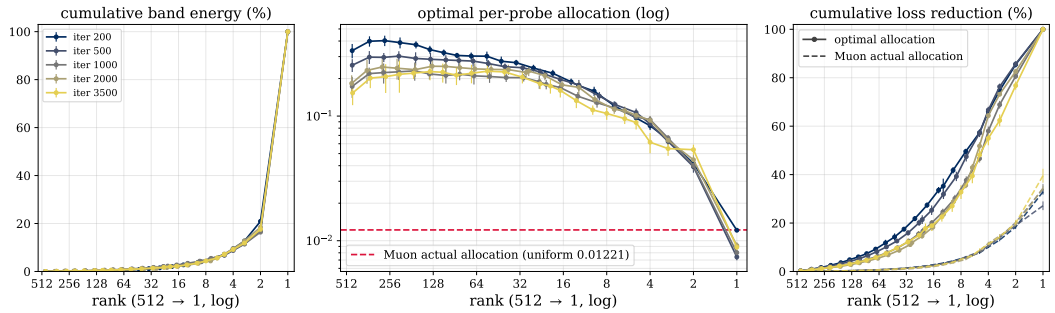

Figure 1: **Out-of-sample loss-optimal step size profile across training.** We probe momentum buffers along a Muon (Scion) trajectory of a 64M-parameter "modded-nanogpt" model (Pethick et al., 2025), at checkpoints from iterations 200–3500 in the constant-learning-rate stage. Points show means over 20 held-out batches and error bars indicate standard deviations. The horizontal axis runs from the tail (rank 512) to the head (rank 1) on a log scale. **Left:** cumulative spectral energy. **Middle:** loss-optimal step size η_d^* estimated by the local-quadratic approximation, shown on a log scale. The red dashed line marks Muon's uniform applied step size during training. **Right:** cumulative predicted loss reduction when each rank uses its own loss-optimal step size (solid), compared with using Muon's uniform step (dashed). For each checkpoint, both curves are normalised by the summed loss reduction under per-rank optimal steps, so the solid curve reaches 100% by construction. These sums omit cross-rank curvature interactions and therefore illustrate idealised room for improvement. Section 4.2 analyses the head–bulk structure and its evolution during training depicted in this figure.

scale. Muon's head-limited uniform allocation therefore leaves substantial room for larger bulk steps. The right panel of Figure 1 illustrates the potential gain by summing the predicted loss reductions when each probe uses its own optimal step. Note that this is an optimistic idealisation that assumes Hessian-conjugate probes. This idealised sum omits cross-rank curvature interactions and therefore indicates potential headroom, not a guaranteed achievable loss reduction from a joint update.

The near-head transition is approximately log-rank-linear. At both model widths 512 and 768, loss-optimal step sizes rise approximately *linearly with log rank* from the head to the bulk plateau (Figure 6 in Appendix E). This shape motivates the allocation for the full SAMuon in Section 5.

The head gap emerges gradually. The separation between the loss-optimal step sizes of head and bulk widens over the first few hundred iterations, alongside increasing concentration of spectral energy and directional curvature in the head (Figure 11 in Appendix J). This gradual emergence motivates warming up SAMuon's anisotropic spectral allocation from Muon's uniform profile, rather than trying to exploit the full head–bulk gap before it has developed (Section 5).

4.3 Why Muon Can Outperform Adam, and How to Improve Muon

The measured profile establishes the importance of considering not only the overall step size, but also how update scale is distributed across singular directions of the momentum buffer. We refer to this distribution as **Spectral Allocation**. The measured loss-optimal step size profile provides a direct empirical reference for assessing an optimiser's spectral allocation and identifying where it can be improved. Momentum SGD assigns scales proportional to the singular values, whereas Muon assigns uniform scales under exact whitening. For Adam, we examine the actual update projected onto its buffer's singular directions. We normalise the measured profile and each optimiser's allocation by their respective head values to compare their relative allocation across ranks independently of the overall step size. This comparison (Figure 2) provides a unified account of why Muon improves on Adam and SGD and identifies the remaining mismatch that motivates SAMuon.

SGD's head-heavy allocation and Adam's partial correction. SGD's singular-value-proportional allocation concentrates the update in the leading directions (Cao et al., 2026; Refael et al., 2025), where the measured loss-optimal steps are smallest. Its allocation therefore runs counter to the measured profile. Adam's coordinate-wise scaling partially corrects this mismatch. Projected onto its momentum buffer's singular basis, the actual Adam update raises the bulk's scale relative to the

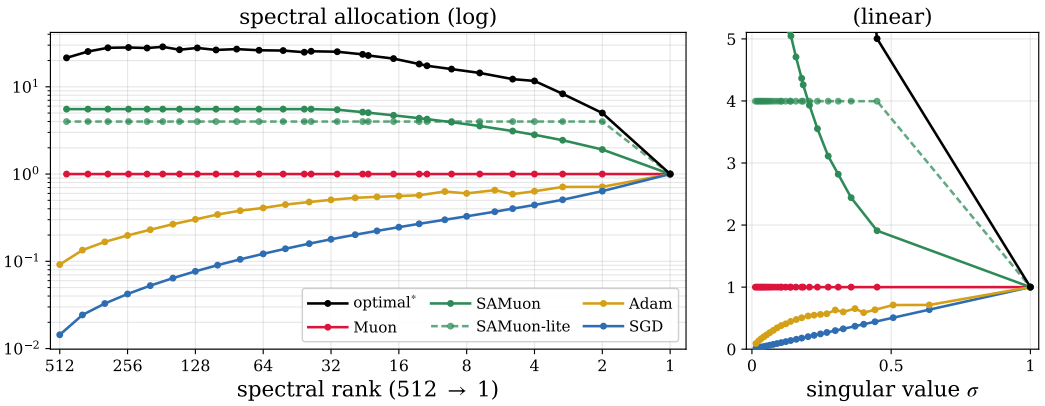

Figure 2: **Spectral allocations compared with the measured loss-optimal step size profile.** The optimiser curves show projections of updates onto each singular direction of their momentum buffers. The reference curve *optimal** is the out-of-sample loss-optimal step size profile at iteration 1000 of the Muon trajectory (Figure 1, middle panel). All profiles are normalised to 1 at the head for comparison. **Left:** allocation against spectral rank, with logarithmic axes. **Right:** allocation against normalised singular value σ, with linear axes. SGD assigns coefficients proportional to the buffer's singular values, while Muon assigns uniform coefficients (assuming exact whitening). Adam's curve shows the projection of an actual update at iteration 1000 of the AdamW trajectory. The two SAMuon profiles preserve the head's scale while amplifying the bulk: SAMuon-lite uses a uniform boost ($\gamma = 5$), whereas SAMuon follows the measured log-rank-linear transition ($\gamma = 7.07$) (both are illustrative). Section 4.3 discusses the implications and Appendix G details the curve construction.

head by approximately an order of magnitude compared with SGD, although the head still dominates (Figure 2). This provides a within-block account of Adam's advantage over SGD, complementary to the inter-block curvature heterogeneity identified by Zhang et al. (2024) and consistent with the curvature-based perspective of Wang et al. (2026). Nevertheless, Adam still allocates too little scale to the bulk relative to the head compared with the measured profile.

The advantage and limit of Muon's uniform allocation. Muon corrects this imbalance through uniform spectral scaling. With the head scales matched, exact whitening increases the coefficient at rank d by σ_1/σ_d relative to momentum SGD, shifting relative update scale towards the tolerant bulk (Figure 2). This brings the allocation closer to the measured profile and suggests why whitening accelerates convergence: it increases the relative update scale of bulk directions that admit larger loss-reducing steps. However, Muon's shared scale remains constrained by the head, while the bulk still tolerates steps several times larger (Figure 1).

Improving Muon's spectral allocation. The remaining mismatch suggests maintaining Muon's head scale while amplifying the bulk towards its measured tolerance. Because the profile's shape is stable across batches and training stages, this corrected allocation can be encoded as a static spectral prior without maintaining additional state for online adaptation. Figure 2 illustrates two choices: a simple uniform bulk boost and a finer allocation following the measured log-rank-linear transition. Both allocations more closely match the measured loss-optimal profile than Muon's uniform allocation. These motivate SAMuon-lite and SAMuon, respectively, whose construction is detailed in Section 5.

Taken together, Figure 2 shows progressively closer alignment with the measured loss-optimal profile from SGD through Adam and Muon to SAMuon. This offers a unified account of their performance ordering and highlights spectral allocation as an actionable principle for optimiser design.

5 Spectral-Aware Muon (SAMuon)

Guided by the measured spectral profile (Section 4), we introduce **Spectral-Aware Muon (SAMuon)**, which improves the uniform spectral allocation of Muon by amplifying the bulk while keeping the head at the original Muon scale. We provide two variants: the full SAMuon follows the measured

log-rank-linear profile, whereas the simplified SAMuon-lite uses a two-level approximation. Neither variant adds persistent optimiser state beyond the momentum buffer used by Muon.

5.1 HEAD-ANCHORED SPECTRAL ALLOCATION

SAMuon and SAMuon-lite distil the measured profile of Section 4 (mostly flat across the bulk, log-rank-linear near the head, see Figure 6) into a fixed target allocation that anchors the head and rescales the rest of the spectrum. Let $(\boldsymbol{u}_i, \boldsymbol{v}_i)_{i \le k}$ be the leading k singular pairs of the buffer $\mathbf{M}^{(t)}$. For each weight matrix, the update direction $\mathbf{O}_{\text{SA}}^{(t)}$ scales the whitened momentum buffer by γ_t and then corrects the leading k directions to their target scales $s_i(t)$:

$$\mathbf{O}_{\text{SA}}^{(t)} \;=\; \gamma_t \,\mathrm{NS}\big(\mathbf{M}^{(t)}\big) \;-\; \sum_{i=1}^{k} \big(\gamma_t - s_i(t)\big)\, \boldsymbol{u}_i \boldsymbol{v}_i^{\top}, \tag{8}$$

where $\mathbf{O}_{\text{SA}}^{(t)}$ is then used to update the model weights $\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta\,\kappa\,\mathbf{O}_{\text{SA}}^{(t)}$. Under exact whitening, this allocates rank $i \le k$ the coefficient $s_i(t)$ and every other direction the coefficient γ_t. The target profile anchors the head at Muon's unit scale, $s_1(t) = 1$, avoiding further amplification of the direction that operates near the Edge-of-Stability. The scales rise monotonically to the bulk scale, γ_t, with $1 = s_1(t) \le s_i(t) \le \gamma_t$.

The bulk boost $\gamma \ge 1$ sets the target amplification of the bulk relative to Muon, while the head remains at unit scale. Both variants warm up gradually to the target profile. A warmup weight $w_t \in [0, 1]$, rising from 0 to 1 over the warmup horizon (cosine schedule, see Appendix J), sets $\gamma_t = 1 + (\gamma - 1)\, w_t$ and $s_i(t) = 1 + (s_i - 1)\, w_t$ for target scales s_i. Training begins at Muon's uniform allocation and gradually approaches the target profile as the head–bulk gap in the spectral profile emerges (Appendix J). Setting $\gamma = 1$ recovers the standard Muon update exactly for SAMuon / SAMuon-lite. Note that both SAMuon variants maintain a single momentum buffer per 2D weight matrix, using the same momentum update rule as Muon (Equation (1)). The spectral modification changes only how this buffer is mapped to the parameter update.

SAMuon: the measured log-rank-linear profile. The full SAMuon follows the empirical shape of Figure 6 directly so that the target scales increase linearly in log rank from the head to the bulk scale,

$$s_i \;=\; 1 + (\gamma - 1)\,\frac{\log i}{\log k}, \quad 1 \le i \le k \;\; (k \ge 2), \tag{9}$$

and all ranks beyond k receive the fixed bulk scale γ. The value of k is set by the measured onset of the plateau, $k = 32$ on the width-512 probing model, and transfers across widths by a square-root rule inspired by Figure 6, *i.e.* $k(d_{\text{model}}) = \left\lfloor 32\sqrt{d_{\text{model}}/512} \right\rceil$. The leading k singular pairs are obtained from a randomised low-rank SVD of the buffer at each update step (`torch.svd_lowrank`; Halko et al., 2011). We emphasise that this shape and scaling rule are an *empirical* distillation of the measurement and not a theoretically derived optimum. A theoretical explanation of the profile shape and width dependence remains future work.

SAMuon-lite: the efficient two-level simplification. SAMuon-lite sets $k = 1$ and $s_1 = 1$ in Equation (8), yielding a two-level allocation that preserves the head at Muon's scale while amplifying all remaining directions by γ_t. The leading pair $(\boldsymbol{u}_1, \boldsymbol{v}_1)$ is estimated using a few steps of power iteration on $\mathbf{M}^{(t)}$, avoiding the low-rank SVD used by the full SAMuon. The complete update procedure for both variants is provided in Algorithm 1 in Appendix H.

5.2 ALGORITHM ANALYSIS: COMPLEXITY AND CONVERGENCE GUARANTEE

Memory and computational cost. Both variants retain Muon's single momentum buffer per $m \times n$ weight matrix, requiring mn persistent optimiser-state entries, compared with Adam's $2mn$. The estimated singular-vector pairs require only $k(m+n)$ transient entries but add no persistent state. For fixed iteration counts, the head estimation costs $O(mn)$ for SAMuon-lite and $O(mnk)$ for SAMuon, compared with $O(m^2 n)$ for Newton–Schulz whitening. Thus, the head estimation has a much lower arithmetic cost than whitening when $k \ll m$, which is true given our square-root scaling rule for k and when m is large.

Measured runtime. The lower arithmetic cost does not translate equally into wall-clock efficiency for both variants. In our 1B-model timing benchmark (Appendix J.7), the power iteration used by SAMuon-lite adds 0.5% of the Muon iteration time, whereas the `torch.svd_lowrank` call used by the full SAMuon adds 7.4%. Thus, SAMuon-lite retains near-Muon per-iteration cost, while the full SAMuon variant incurs a larger overhead under the current implementation.

Convergence guarantee. Under idealised exact whitening and exact singular-vector estimation, both variants retain Muon's asymptotic convergence rates for the iteration-averaged expected gradient nuclear norm, up to γ-dependent constants: $\mathcal{O}(T^{-1/4})$ in the stochastic setting and $\mathcal{O}(T^{-1/2})$ in the deterministic setting (Shen et al., 2026). The guarantee applies to any allocation with scales in $[1, \gamma]$ along the exact singular vectors of the momentum buffer, covering both proposed profiles. The assumptions and proof are provided in Theorem I.4, Appendix I.

6 Experiments

We evaluate the SAMuon and SAMuon-lite variants on language-model pretraining, following the experimental protocol established in Pethick et al. (2025).

6.1 Experimental setup

Architecture and training protocol. We train "modded-nanogpt" models (Jordan et al., 2024a; Pethick et al., 2025) at three scales: 124M, 300M, and 1B parameters, using the FineWeb dataset (Penedo et al., 2024) (100B-token sample). The 124M and 300M models use a 10B-token budget and the 1B model uses a 20B-token budget, corresponding to approximately $4\times$, $1.7\times$ and $1\times$ Chinchilla-optimal token budget (Hoffmann et al., 2022), respectively. We evaluate batch sizes of 1024, 2048, and 4096 sequences of length 1024, with the evaluated model-scale and batch-size combinations reported in Table 1. Iteration counts are adjusted to preserve the token budget across batch sizes. Following Pethick et al. (2025), we use a trapezoidal learning-rate schedule with a constant phase followed by linear decay over the final 28.5% of training. AdamW uses an additional 5% linear learning-rate warmup, while the SAMuon variants use a 30% spectral warmup as described in Section 5.1. Each configuration is evaluated with a single seed. Within each model-scale and batch-size setting, all optimisers use the same model initialisation and identical data ordering. Architectural specifications, hardware, and complete training details are provided in Appendix J.

Baselines. We compare against AdamW (Loshchilov & Hutter, 2019), with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, and Muon using the unconstrained Scion implementation of Pethick et al. (2025). We select Scion for its simpler hyperparameter tuning and reported performance advantages over the original Muon implementation (Appendix J.1). Both SAMuon variants are built directly on this codebase. Muon and both SAMuon variants use a momentum coefficient of $\mu = 0.9$.

Hyperparameter tuning protocol. We tune hyperparameters separately for each batch size on the 124M model. For AdamW, we sweep the learning rate with 2-spaced grids. For Muon (Scion implementation), we fix the learning rate at the Scion default (2^{-12}) and sweep the spectral radius that controls the update scale for 2D weight matrices with $\sqrt{2}$-spaced grids. For the SAMuon variants, we fix the learning rate and the spectral radius at the Scion default (2^{-12} and 50, respectively) and tune only the bulk boost γ with $\sqrt{2}$-spaced grids. The profile rank k follows the width rule in Section 5.1 and is not tuned. For larger models, we follow the transfer convention of Pethick et al. (2025): Muon's learning rate and spectral radius are transferred directly, while AdamW's learning rate is halved at each model-scale step. For the SAMuon variants, we also transfer the learning rate, spectral radius and bulk boost without further tuning, and only k changes according to the width rule. Note that the SAMuon variants adopt the learning-rate schedule and model-scale transfer convention of Muon without method-specific tuning, leaving potential gains from further tuning unexplored (Appendix B). Search grids, selected hyperparameters, and full sweep results are provided in Appendix J.

6.2 Main results

Table 1 reports final validation losses and token-efficiency improvements across the seven evaluated model-scale and batch-size settings. Token savings are measured and interpolated using shortened training runs that reach the tuned Muon baseline's final validation loss, as detailed in Appendix K.

Table 1: **Final validation loss and token-efficiency improvement across model scales and batch sizes.** Batch size is the number of sequences in a batch, and validation losses are reported after training on the token budgets listed in the table. Hyperparameters are tuned on the 124M model separately for each batch size and transferred to larger models following Section 6.1. The "improv." columns report the estimated percentage of training tokens saved relative to tuned Muon at matched final validation loss. The measurement procedure is detailed in Appendix K. The best unrounded loss and token-efficiency improvement in each row are highlighted in bold.

Model	Batch	Tokens	AdamW	Muon (Scion)	SAMuon		SAMuon-lite	
			loss	loss	loss	improv.	loss	improv.
124M	1024	10B	3.2105	3.1622	**3.1435**	**20.3%**	3.1438	20.3%
	2048	10B	3.2410	3.1726	**3.1484**	**23.8%**	3.1511	21.5%
	4096	10B	3.3050	3.1953	**3.1658**	**24.0%**	3.1678	22.1%
300M	1024	10B	3.0297	2.9836	**2.9640**	**17.7%**	2.9663	15.9%
	2048	10B	3.0658	2.9941	**2.9713**	**18.8%**	2.9738	17.1%
1B	1024	20B	2.7599	2.7251	**2.7114**	**13.3%**	2.7117	13.3%
	4096	20B	2.8067	2.7413	**2.7222**	**15.6%**	2.7240	14.4%

Both variants outperform the baselines across evaluated settings. Both SAMuon and SAMuon-lite achieve lower final validation loss than AdamW and tuned Muon in every evaluated setting. In the matched-loss experiments, SAMuon requires 13.3%–24.0% fewer tokens than Muon, while SAMuon-lite requires 13.3%–22.1% fewer tokens. These improvements span all three model scales, with training budgets ranging from approximately Chinchilla-optimal to substantially beyond that allocation. These results support the practical value of amplifying the tolerant spectral bulk while preserving the head direction's scale.

Token-efficiency gains increase with batch size. Within each model scale, both variants achieve larger token savings over Muon at the larger evaluated batch sizes. On the 124M model, for example, the savings increase from 20.3% for both variants at batch size 1024 to 24.0% for SAMuon and 22.1% for SAMuon-lite at batch size 4096. The partial batch-size comparisons at 300M and 1B show the same trend. One possible explanation is that larger batches provide less noisy estimates of the momentum buffer's spectral structure.

SAMuon-lite retains most of the full profile's benefit. The full SAMuon achieves slightly lower final validation loss than SAMuon-lite in every evaluated setting, with a token-saving advantage of up to 2.3 percentage points. This suggests that matching the measured log-rank-linear profile more closely can improve optimisation beyond the gains obtained by the two-level approximation. However, the small difference between the variants relative to their gains over Muon suggests that separating the sensitive head from the tolerant bulk captures the main practical benefit of the measured profile. With its near-Muon per-iteration cost (Section 5.2), SAMuon-lite offers a practical alternative to the full SAMuon that retains most of the token-efficiency benefit.

7 Conclusion

In this paper, we contribute an *out-of-sample* spectral probing analysis that identifies a stable loss-optimal step size profile with a *Sensitive Head* and a *Tolerant Bulk*. This fine-grained profile provides a *Spectral Allocation* account of the SGD–Adam–Muon performance ordering and reveals that Muon's uniform allocation still underuses the bulk. Guided by this finding, SAMuon and SAMuon-lite amplify the bulk while preserving the head scale, without additional persistent optimiser state. Across evaluated 124M–1B-parameter language-model settings, SAMuon reaches Muon's final validation loss with 13.3%–24.0% fewer tokens; SAMuon-lite retains most gains (13.3%–22.1%) at near-Muon per-iteration cost. Both retain Muon's asymptotic convergence rates under idealised exact whitening. These results support *Spectral Allocation* as an actionable principle for general optimiser design. A further discussion of future work and limitations is provided in Appendix B.

AI USE STATEMENT

Generative AI tools assisted with literature retrieval and review, developing experimental and plotting code, preparing figures, and formalising, writing, and checking mathematical arguments and proofs. They also assisted with drafting, editing, and organising the manuscript and provided feedback on the methodology, presentation, and interpretation of results. All AI-assisted code, figures, mathematical arguments, and text were carefully reviewed for technical accuracy and consistency before use, with suggestions corrected and refined where necessary. Checks included assessing the internal consistency of experimental results and comparing them with relevant published findings.

ETHICS STATEMENT

This work studies optimisation methods for language-model pretraining and does not involve human-subject experiments or the collection of new personal data. We do not identify additional ethical concerns specific to the proposed methods beyond those generally associated with language-model research.

REPRODUCIBILITY STATEMENT

The optimiser definitions and implementation-level pseudocode are provided in Section 5 and Algorithm 1. The spectral-probing procedure is specified in Section 4, with its assumptions and validation in Appendix C. Model architecture, data, training schedules, optimiser hyperparameters, Newton–Schulz coefficients, and wall-clock measurements are documented in Section 6.1 and appendices J, J.4, J.5 and J.7. The convergence assumptions and complete proof are given in Appendix I, and the token-efficiency measurement procedure is given in Appendix K. We will publicly release the code upon acceptance of the paper.

REFERENCES

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *16th Annual Workshop on Optimization for Machine Learning (OPT 2024)*, 2024. URL https://opt-ml.org/papers/2024/paper93.pdf.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=hErdffTsLu.

Hengjie Cao, Mengyi Chen, Yifeng Yang, Fang Dong, Ruijun Huang, Jixian Zhou, Anrui Chen, Mingzhi Dong, Yujiang Wang, Jinlong Hou, Yuan Cheng, Fan Wu, Fan Yang, Tun Lu, Ning Gu, and Li Shang. Metis: Training LLMs with FP4 quantization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=I2ZrCi5O84.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=Blz4hjxLwU.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Zhehang Du and Weijie Su. The Newton–Muon optimizer. *arXiv preprint arXiv:2604.01472*, 2026. URL https://arxiv.org/abs/2604.01472.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and Muon on multiclass separable data. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2502.04664. Spotlight; arXiv:2502.04664.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*,

pp. 1842–1850. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53:217–288, 2011. doi: 10.1137/090771806.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2176. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html.

Feihu Huang, Yuning Luo, and Songcan Chen. LiMuon: Light and fast Muon optimizer for large models. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2509.14562.

Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024a. URL https://github.com/KellerJordan/modded-nanogpt.

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

Kimi Team. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. URL https://arxiv.org/abs/2507.20534.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization, 2025. URL https://arxiv.org/abs/2503.12645.

Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm. In *Advances in Neural Information Processing Systems*, volume 37, pp. 73501–73548, 2024. URL https://openreview.net/forum?id=SFxAjB7UXx. arXiv:2405.14813.

Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon. *arXiv preprint arXiv:2502.02900*, 2025. URL https://arxiv.org/abs/2502.02900.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. NorMuon: Making Muon more efficient and scalable. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2510.05491.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025. URL https://arxiv.org/abs/2502.16982.

Liming Liu, Zhenghao Xu, Zixuan Zhang, Hao Kang, Zichong Li, Chen Liang, Weizhu Chen, and Tuo Zhao. COSMOS: A hybrid adaptive optimizer for efficient training of large language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=j2QTOOtM8R.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Gagik Magakyan, Pablo Parrilo, and Asuman Ozdaglar. Spectral scaling laws of Muon, 2026. URL https://arxiv.org/abs/2606.04058.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417. PMLR, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2406.17557.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49069–49104. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/pethick25a.html.

Yehonathan Refael, Guy Smorodinsky, Tom Tirer, and Ofir Lindenbaum. SUMO: Subspace-aware moment-orthogonalization for accelerating memory-efficient LLM training. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=DIjRvEKOeG.

Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. ICLR Workshop, 2018. URL https://openreview.net/forum?id=rJrTwxbCb.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=4nH4CulGaP.

Chongjie Si, Debing Zhang, and Wei Shen. AdaMuon: Adaptive Muon optimizer. *arXiv preprint arXiv:2507.11005*, 2025. URL https://arxiv.org/abs/2507.11005.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Weijie Su. Isotropic curvature model for understanding deep learning optimization: Is gradient orthogonalization optimal? *arXiv preprint arXiv:2511.00674*, 2025. URL https://arxiv.org/abs/2511.00674.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M. Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IDxZhXrpNf.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why Muon outperforms Adam: A curvature perspective, 2026. URL https://arxiv.org/abs/2606.04662.

Kaiyue Wen, David Leo Wright Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2J51qUZ0iG.

Xiaodong Wu, Wenyi Yu, Chao Zhang, and Philip Woodland. An improved empirical Fisher approximation for natural gradient descent. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NeurIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f23098fa0cfcdef0e743b134d380eeb9-Abstract-Conference.html.

Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T. Freeman, and Hao Tan. Test-time training done right. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ffbfff81fb78e4bb558273b91e12e318-Abstract-Conference.html.

Yechen Zhang, Shuhao Xing, Junhao Huang, Kai Lv, Yunhua Zhou, Xipeng Qiu, Qipeng Guo, and Kai Chen. Mousse: Rectifying the geometry of Muon with curvature-aware preconditioning. *arXiv preprint arXiv:2603.09697*, 2026b. URL https://arxiv.org/abs/2603.09697.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need Adam: A Hessian perspective. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NeurIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ee0e45ff4de76cbfdf07015a7839f339-Abstract-Conference.html.

A Extended Related Work

Theoretical Interpretations of Muon's Success. Norm-geometric interpretations derive Muon's orthogonalised direction from spectral-norm steepest descent and norm duality (Bernstein & Newhouse, 2024; 2025; Jordan et al., 2024b), or formulate it as a linear-minimisation-oracle or non-Euclidean trust-region step (Pethick et al., 2025; Kovalev, 2025). Related work establishes convergence guarantees (Li & Hong, 2025), while modular norm constructions and Scion connect norm choice and normalisation to update scaling and learning-rate transfer (Large et al., 2024; Pethick et al., 2025). Distinct from a constraint on the update, Chen et al. (2026) analyse spectral-norm constraints on the weights induced by Muon with weight decay. Fan et al. (2025) establish a spectral-norm max-margin bias in separable multiclass linear classification. Norm-based analyses incorporate curvature through norm-dependent smoothness bounds or surrogate loss models (Bernstein & Newhouse, 2024; Large et al., 2024; Kovalev, 2025), but do not resolve the measured variation in loss-optimal update scales across momentum singular directions. Su (2025) studies an isotropic curvature model and shows that, under suitable curvature-growth conditions, the optimal update has a more homogeneous spectrum; orthogonalisation becomes optimal in a phase-transition limit. Our measurements instead show strongly nonuniform directional curvature across equally normalised probes, indicating a limitation of an isotropic description of the observed geometry. The closely related work of Wang et al. (2026) links Muon's advantage to a smaller normalised directional-sharpness penalty and analyses how redistributing update energy across curvature groups can improve descent. This provides a curvature-based explanation alongside the norm-geometric interpretations; they further report substantial alignment between the gradient and the dominant in-sample Hessian subspace, motivating examination at finer spectral granularity. Our spectral probing analysis extends this curvature perspective by estimating *out-of-sample* loss-optimal step sizes along individual singular directions of momentum buffers on held-out batches at checkpoints along Transformer training trajectories. The resulting anisotropic profile shows why spectral allocation in the momentum buffer's singular basis matters for optimiser design and provides a practical guide for assigning update scales across singular directions. The empirical improvements achieved by SAMuon further support the practical value of this approach.

Spectral Analysis of Deep Neural Networks. Empirical studies of the Hessian eigenspectrum have long established that neural network loss landscapes are highly anisotropic, the spectrum splitting into a bulk concentrated near zero and a small number of outlier eigenvalues orders of magnitude larger (Sagun et al., 2018). For Transformers, Zhang et al. (2024) analyse this structure at the granularity of parameter blocks, attributing the failure of single-learning-rate SGD, relative to coordinate-wise Adam, to this inter-block curvature disparity, *i.e.* "block heterogeneity". Complementary to their

work, we argue that the distribution of curvature *within* each block across the singular directions of the update is another key aspect that separates orthogonal optimisers from Adam (Section 4.3). There is also a body of work discussing the spectral anisotropy of gradients and momentum buffers (Refael et al., 2025; Cao et al., 2026; Magakyan et al., 2026), instead of the curvature behaviour that is our focus. Specifically, Magakyan et al. (2026) show that retaining only a small fraction of leading singular directions markedly degrades Muon's performance, providing evidence consistent with our interpretation that Muon's advantage involves better utilising the spectral bulk.

Adaptive Methods and Practical Muon Variants. Beyond convergence speed, a central appeal of Muon (Jordan et al., 2024b) is its low memory overhead. Muon maintains only a single momentum buffer, rather than the two moments used by Adam (Kingma & Ba, 2015) and the even heavier memory requirement of most second-order methods. Kronecker-Factored Approximate Curvature (K-FAC) (Martens & Grosse, 2015), Shampoo (Gupta et al., 2018), and SOAP (Vyas et al., 2025) maintain structured matrix statistics for preconditioning, and some recent Muon variants likewise add persistent state in pursuit of adaptivity: AdaMuon (Si et al., 2025) adds an element-wise second moment, Newton–Muon (Du & Su, 2026) a right preconditioner based on uncentred activation second moments, and Mousse (Zhang et al., 2026b) Shampoo-style matrix factors, with a single-sided variant reducing this overhead. These methods report improvements over Muon while carrying additional persistent state; the resulting training-memory overhead depends on the variant and implementation. A complementary hybrid, COSMOS (Liu et al., 2026), instead confines adaptive (SOAP-style) preconditioning to the leading eigen-subspace of the second-moment matrix and applies a low-cost orthogonalised update to the remainder. COSMOS adapts its subspace preconditioner online, whereas our method shapes the momentum spectrum using a fixed, measurement-derived allocation without additional persistent state. LiMuon (Huang et al., 2026) likewise applies a randomised low-rank SVD to the momentum buffer, but to compress the stored momentum rather than shape its spectral allocation, *i.e.* the same computational primitive as SAMuon but with a different purpose. Two lines of research instead keep the footprint close to the single buffer used by Muon. NorMuon (Li et al., 2026) adds only a per-neuron second-moment vector to rescale the orthogonalised update; this rescaling is potentially complementary to our spectral allocation, although their combination remains untested. Scion (Pethick et al., 2025) is a renormalised Muon which also replaces the auxiliary Adam optimiser used for embeddings with a lightweight signed update. SAMuon retains this single-buffer footprint while exploiting the within-block spectral structure through a fixed, measurement-derived shaping of the update spectrum, at negligible additional FLOPs and with a single *tuned* extra hyperparameter (bulk boost γ; the profile rank of SAMuon follows a fixed width rule and is not tuned). Note that we build our work directly on Scion, which is our primary baseline and is referred to as Muon in Section 6.

B Discussion and Limitations

Spectral probing as a tool for optimiser design. Our out-of-sample spectral probing turns saved training checkpoints into a practical guide to spectral allocation. The gains of SAMuon and SAMuon-lite show how this measured profile can inform effective updates through a static prior, without additional persistent optimiser state. More broadly, such measurements could guide future optimiser design by identifying underused subspaces before committing to extensive full-training searches.

Further tuning for SAMuon variants. The SAMuon variants inherit the learning-rate schedule and model-scale transfer convention developed for Muon in Pethick et al. (2025). Tuning these choices specifically for the SAMuon variants may yield further gains and remains future work.

Is there a better spectral profile? The measured profile provides an effective spectral allocation, but the probes evaluate directions separately and do not capture curvature interactions between them. Accounting for these interactions could further improve the allocation. The static prior also leaves open whether adaptation to temporal changes beyond early-stage warmup would improve training. Finally, the profile aggregates measurements across weight matrices, while the per-module analysis in Appendix L reveals clear module-dependent differences in directional alignment. This motivates investigating module-specific spectral allocations. These directions are left for future work.

Limitations. Our optimiser comparisons cover 124M–1B-parameter models from one decoder-only Transformer family trained on FineWeb. Whether the measured spectral profile and optimisation gains extend to other architectures, datasets, or larger models remains to be established.

C Spectral Probing: Assumptions and Validation

This section states the assumptions and retention conditions required to interpret the local-quadratic estimate in Section 4.1, reports the filtering of probes that fail these conditions, and validates the local-quadratic approximation against an exact line-search reference.

Assumption C.1 (Local Quadraticity)**.** Along each probe direction $\{\boldsymbol{\theta}_t + \eta\,\Delta\boldsymbol{\theta}^{(d)} : \eta \geq 0\}$, the batch loss is well-approximated by its second-order Taylor expansion (Equation (5)) over the step range containing the loss minimum.

Assumption C.2 (Held-Out Descent Direction)**.** For every retained rank-d probe, the directional derivative of the held-out loss is negative, $\boldsymbol{g}^\top \Delta\boldsymbol{\theta}^{(d)} < 0$, where $\boldsymbol{g}$ is the gradient of the held-out batch loss at $\boldsymbol{\theta}_t$. Thus, a positive step along the probe direction initially decreases the held-out loss.

Assumption C.3 (Positive Directional Curvature)**.** The probed curvature $\Delta\boldsymbol{\theta}^{(d)\top}\mathbf{H}\,\Delta\boldsymbol{\theta}^{(d)}$ is positive. Together with assumption C.2, this ensures that the quadratic in Equation (5) has a positive interior minimum. All probes retained in our analysis satisfy this condition.

Filtering ill-behaved probes. The LQA yields a positive interior optimum under assumptions C.2 and C.3. We therefore discard any probe that fails either check before aggregating or plotting the results. Across the 64M and 124M out-of-sample spectral probing analyses, 31 of the $6{,}400$ sampled rank probes (0.48%) fail the descent-direction check, while two additional rank probes (0.03%) have negative curvature, giving 33 filtered rank probes (0.52%) in total. The two negative-curvature cases occur at the deepest ranks late in training and are numerically negligible, with curvature magnitudes no greater than 4.0×10^{-7}; the non-descent cases are likewise concentrated late in training. Thus, every probe included in the analysis satisfies both conditions, while the filtering affects only a tiny fraction of the measurements.

Validation against an exact reference: golden-section line search. For every LQA probe, the finite-difference curvature estimate of Equation (7) uses the fixed displacement $\varepsilon = 0.3$. To further test assumption C.1 directly, we measure the exact optimal step of a probe by a one-dimensional line search along the probe direction, free from any quadratic assumptions. An exponential coarse search first brackets the loss minimum, and then a golden-section refinement in $\log_{10}$-space locates it to a target precision. The search operates purely on forward evaluations of the batch loss but requires tens of evaluations per probe, compared with the single additional forward evaluation required by the LQA estimate of Equation (7). We report the loss-optimal step in update-norm units as $s_d^* = \|\eta_d^* \Delta\boldsymbol{\theta}^{(d)}\|_2 = \eta_d^* \|\Delta\boldsymbol{\theta}^{(d)}\|_2$, where the Scion matrix scales κ_j are already included in $\Delta\boldsymbol{\theta}^{(d)}$. Figure 3 compares s_d^* estimated by the LQA with the line-search reference over 24 checkpoint–batch analyses spanning training stages and both Muon (Scion) and AdamW trajectories (773 probes in total). The two agree closely across more than four orders of magnitude of update norm: the median relative deviation remains near 5% throughout the operationally relevant range ($s_d^* \lesssim 10$), which covers the head and most of the bulk. Deviations grow for the flattest directions in the deep tail ($s_d^* \gtrsim 10^2$), where the loss basin is extremely shallow, so the precise location of the minimum is both harder for the reference search to pin down and less consequential for the resulting profile. This close agreement confirms that the pretraining loss is locally near-quadratic along the probed directions (across training stages and update orientations), justifying the single-forward-pass LQA estimate used in our spectral analysis.

D Spectral Probing of an AdamW Trajectory

The spectral profile in Figure 1 is measured on checkpoints produced by a Muon run, which raises a natural question: is the measured structure a property of the Transformer loss landscape, or an artefact of the particular trajectory that the update rule of Muon carves through it? To answer this, we repeat the complete out-of-sample probing analysis of Section 4 on a standard AdamW trajectory of the same 64M-parameter model under the identical protocol. The gradient of a gradient batch $\mathcal{B}_\mathrm{g}$ is integrated into the first-moment (momentum) buffer stored at each AdamW checkpoint, and the same Scion-scaled per-rank probes of Equation (4) are measured against a held-out batch with the validated LQA estimate. Figure 4 shows the same overall structure as the Muon trajectory. The loss-optimal step rises by more than an order of magnitude from the sensitive head into a mostly flat, far more tolerant bulk, and the deep tail again contributes only a modest residual share of the cumulative loss reduction.

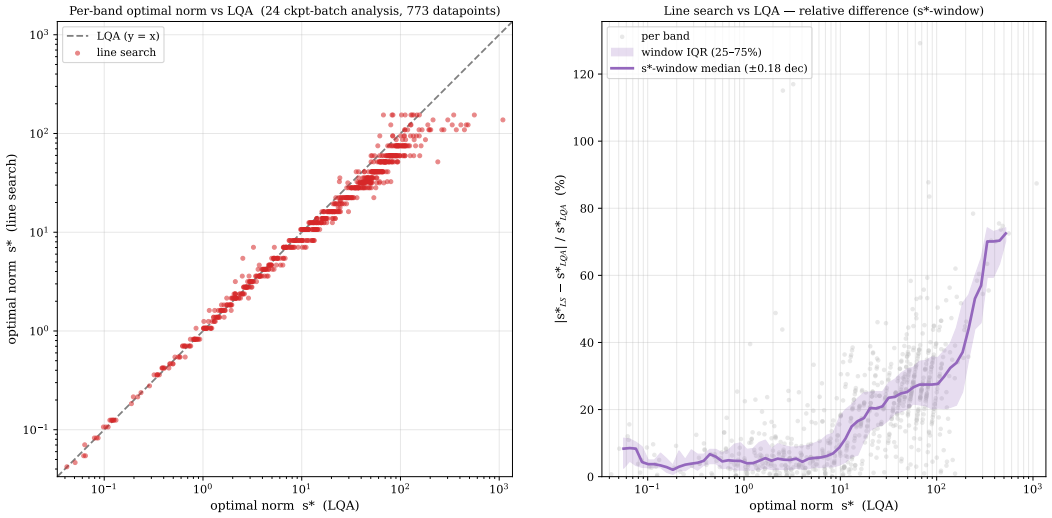

Figure 3: **Validation of the local-quadratic estimate against the exact line search. Left:** model-wide loss-optimal update norm s_d^*, estimated by the LQA (x-axis) and by the golden-section line-search reference (y-axis), for the 773 probe directions sampled from 24 checkpoint–batch analyses spanning all training stages and both Muon and AdamW trajectories; the dashed line marks perfect agreement ($y = x$). **Right:** relative difference between the two as a function of the LQA estimate, with the windowed median and interquartile range overlaid. The median deviation stays near 5% over the operationally relevant range, rising only for extremely flat directions where the loss minimum is shallow.

The one systematic difference appears at the head and is expected. The whitened update of Muon drives every singular direction at the same shared global scale. The head's measured loss-optimal step remains close to this applied scale (the middle panel of Figure 1), suggesting operation near the Edge-of-Stability. The AdamW update, in contrast, is not normalised. The exact update norm realised along the trajectory varies, and the head's loss-optimal step size inherits this variation, so it varies relatively more freely across checkpoints. As a consequence, the optimal step size profile is more variable across checkpoints than along the Muon trajectory (at early checkpoints it even rises across the bulk before the sharp near-head decline), although the overall trend of a sensitive head and a more tolerant bulk still holds. This close structural agreement shows that the measured anisotropic profile persists across the evaluated Muon and AdamW trajectories, rather than being specific to the optimiser that generated the trajectory.

E Spectral Profiles Across Model Widths

The probing analyses of Section 4 were run on a width-512 (64M) model. To check that the measured structure is not specific to that scale, the complete out-of-sample probing analysis is repeated on a Muon trajectory of the 124M (width-768) model, matching the smallest trained scale of Section 6 (probing-run setup in Appendix J). Figure 5 gives the same three-panel analysis as Figure 1; Figure 6 compares the near-head transition at both widths using a linear step-size axis against log rank.

The qualitative observations in Section 4 persist at width 768. The profile remains highly anisotropic and stable across checkpoints, with a sensitive head near the applied Muon scale, a mostly flat tolerant bulk, and an approximately log-rank-linear near-head transition.

The most relevant visible scale effect is a modest upward shift in the approximate plateau-onset rank, from about 32 at width 512 to about 40 at width 768 (the right panel of Figure 6). Because the transition is gradual, these values should be interpreted as approximate rather than sharply identified thresholds. This shift is consistent with the square-root width heuristic used in Section 5, which predicts $k \approx 39$ at width 768. With only two directly probed widths, however, we do not treat this agreement as evidence establishing a general scaling law.

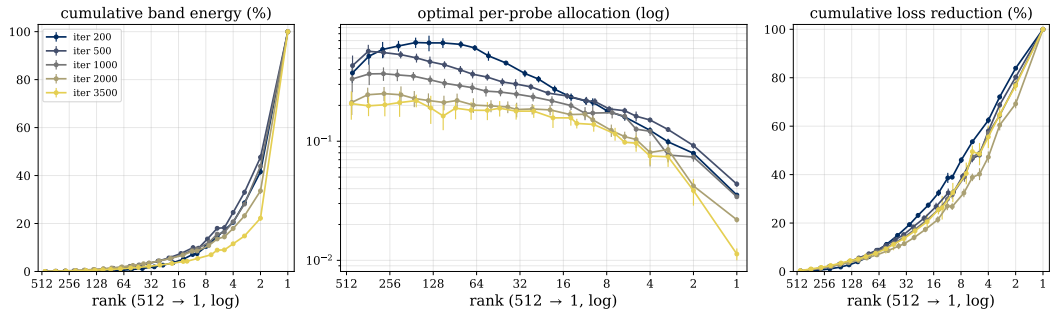

Figure 4: **Out-of-sample loss-optimal step size profile along an AdamW trajectory** (same 64M-parameter "modded-nanogpt" model as Figure 1, checkpoints from iterations 200–3500). Out-of-sample spectral probing (Section 4) of the updated first-moment (momentum) buffer, with each per-rank probe's optimal step size measured on a disjoint out-of-sample batch; the horizontal axis runs from the deep tail (rank 512) to the head (rank 1, log scale). **Left:** cumulative spectral energy, again concentrated in the leading ranks. **Middle:** per-rank optimal spectral allocation η_d^* (the loss-optimal step along each probe; LQA estimate; log axis). As with Muon, the bulk is mostly flat and the optimal allocation drops sharply near the head, which tolerates a far smaller step. Because the AdamW update is unnormalised, the realised update norm varies, and the head's loss-optimal step therefore varies more freely across checkpoints; the slower convergence of AdamW also prolongs the burn-in, giving a more unstable profile in early checkpoints (iterations 200–500) that settles to the same head–bulk shape later (iterations 1000–3500). **Right:** cumulative loss reduction (LQA estimate): as for Muon, the leading ranks account for most of the achievable reduction and the deep tail contributes a modest residual share. The close structural agreement with Figure 1 shows that the sensitive-head and tolerant-bulk structure persists across the evaluated Muon and AdamW trajectories.

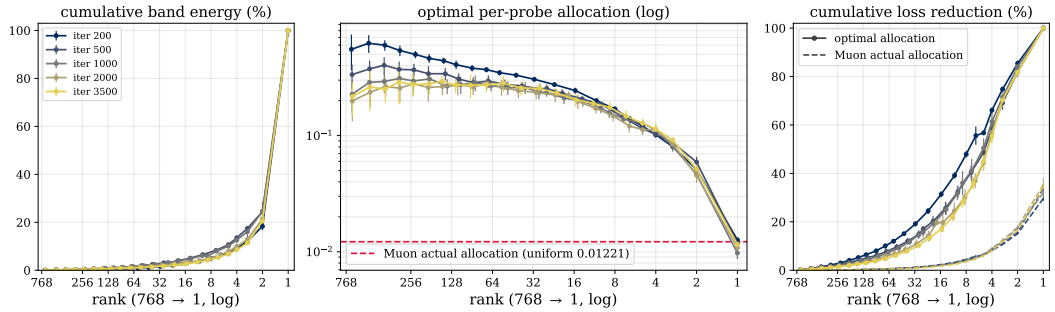

Figure 5: **Out-of-sample loss-optimal step size profile at width 768** (124M model, Muon trajectory). Panels, axes, and protocol as in Figure 1. The qualitative structure of the width-512 analysis is reproduced.

F In-Sample Spectral Probing

The out-of-sample probing of Section 4 measures the step size that reduces loss the most on *unseen* data. For completeness, we also report the *in-sample* variant, in which each probe's optimal step size is measured on the *same* batch whose gradient formed the buffer ($\mathcal{B}_o = \mathcal{B}_g$). Related curvature analyses in prior works also evaluate updates in-sample, although they measure different quantities. We report it in the same rank-organised, three-panel layout as the main-text out-of-sample figure (Figure 1), so the two regimes can be compared panel for panel.

Observation. Figure 7 reports the in-sample profile along the same Muon trajectory, in the same rank-organised, three-panel layout as the out-of-sample figures. The qualitative structure matches the out-of-sample analysis (a sensitive head with a loss-optimal step close to Muon's applied scale and a far more tolerant bulk), but two quantitative differences appear. First, the in-sample optimal

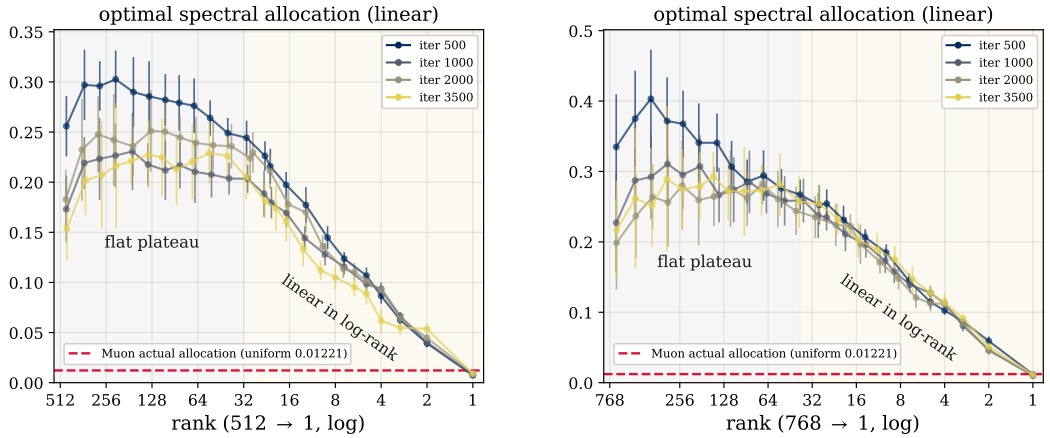

Figure 6: **The near-head decline of the loss-optimal step size profile is close to linear in log rank.** We replot the probed spectral profile with a *linear* y-axis against log rank. **Left:** the width-512 (64M) model (the middle panel of Figure 1); **Right:** the width-768 (124M) model (see Appendix E). At both scales, the loss-optimal step decreases approximately linearly with log rank from the bulk plateau to the head. The visually estimated plateau onset shifts from rank ~ 32 at width 512 to rank ~ 40 at width 768, consistent with the square-root width heuristic used for SAMuon in Section 5.

allocation declines *monotonically* from the deep tail to the head rather than holding a mostly flat bulk, and the anisotropy spans two to three orders of magnitude, wider than the out-of-sample gap. Second, the cumulative loss reduction under the optimal allocation accrues steadily from the tail, so the bulk, including its deep-tail directions, appears far more productive than out-of-sample, where the reduction concentrates at the head. Both effects are consistent with in-sample probing partly measuring the optimiser's ability to refit the gradient batch that formed the update; the out-of-sample analysis in the main text removes this confounder and yields the more conservative, generalisation-relevant profile that motivates SAMuon. The per-module sign alignment of Appendix L supports the refitting explanation of the in-sample behaviour: tail directions in the attention projections align poorly with the gradients of unseen batches, so their large in-sample tolerance mainly reflects refitting of the sampled batch. The same in-sample inflation across the bulk, including the deep tail, also occurs along an AdamW trajectory (Figure 9), confirming that this batch-refitting confounder is a property of the probe, not of the optimiser that generated the trajectory.

Figure 8 overlays the two regimes directly along the same Muon trajectory. The out-of-sample optimal allocation (solid) is uniformly more conservative than the in-sample one (dashed) across the entire bulk, with the gap widest in the deep tail, where the in-sample allocation runs several times larger. The per-probe loss-reduction panel tells the same story from the capacity side. For the in-sample case, every probe across the spectrum yields a large reduction; out-of-sample, the tail's contribution collapses and the reduction climbs only towards the head. The in-sample tail gain that does not survive out-of-sample is refitting of the gradient batch, while the residual that does survive is the generalisation-relevant headroom that SAMuon targets. Note that the same in-sample-versus-out-of-sample gap holds along an AdamW trajectory (Figure 10), so this separation is generalisable rather than specific to Muon.

G Construction of the Spectral Allocation Comparison

This section details the construction of the curves in Figure 2, which compares the spectral scaling each optimiser applies against the measured loss-optimal step size profile. All curves are computed at checkpoint 1000 of the out-of-sample probing analyses (Section 4), averaged across out-of-sample batches, and head-normalised so that the top singular direction (rank 1, largest σ) sits at 1. A single checkpoint is shown for visual clarity. The left panel plots the spectral allocation against spectral rank (log axes). The right panel plots the same allocation against singular value σ (linear axes), making explicit the transfer function each optimiser applies to the spectrum.

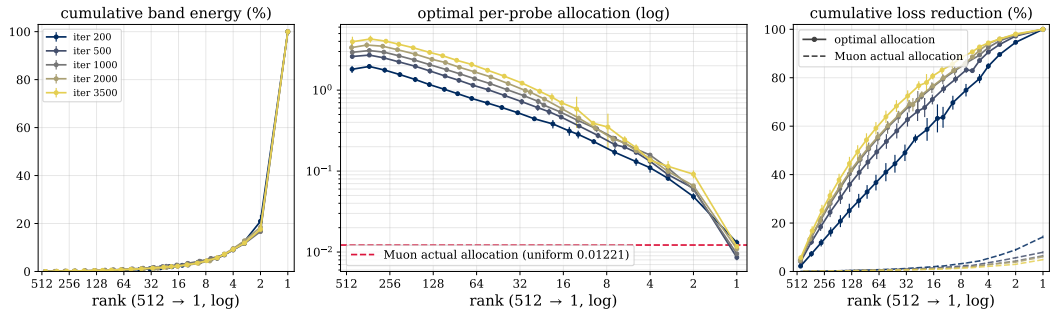

Figure 7: **In-sample loss-optimal step size profile along a Muon trajectory**, in the same panel layout as the out-of-sample Figure 1 (same 64M-parameter "modded-nanogpt" model and checkpoints), but with each probe's optimal step size measured on the *same* batch whose gradient formed the buffer ($\mathcal{B}_o = \mathcal{B}_g$). Two differences from the out-of-sample profile stand out. First, the entire bulk is raised well above its out-of-sample level (middle panel), widening the anisotropy. Second, the deep tail contributes far more to the idealised cumulative loss reduction (right panel). Both indicate that, in-sample, the bulk mainly yields refitting of the batch that formed the update, a gain that does not transfer to unseen data, which is exactly the confounder the main-text out-of-sample protocol removes.

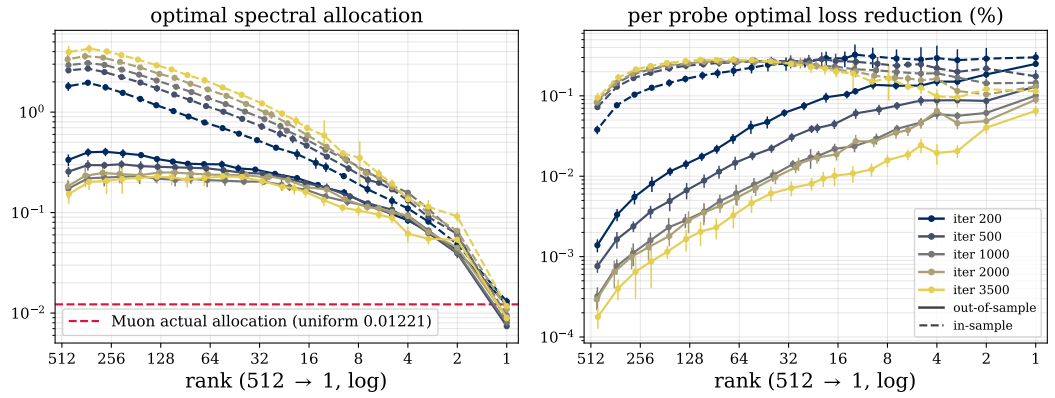

Figure 8: **In-sample versus out-of-sample spectral profiles along a Muon trajectory** (64M-parameter "modded-nanogpt", checkpoints from iterations 200–3500). Solid curves are the out-of-sample (main-text) measurement; dashed curves are the in-sample variant. **Left:** per-rank optimal spectral allocation. The in-sample allocation is systematically higher across the entire bulk, several times the out-of-sample level. Both decline sharply towards the head, where the loss-optimal step is close to Muon's applied scale (red dashed, 0.01221). **Right:** per-probe optimal loss reduction (log axis). In-sample, every probe across the spectrum, particularly the deep tail, yields a large reduction; out-of-sample, the per-probe reduction in the tail is far smaller and only climbs towards the head. The gap between the two reflects that the tail spectrum of the gradient batch mainly contributes to overfitting that batch, with limited transferable capacity to out-of-sample batches.

- *Loss-optimal step size profile*: the per-rank loss-optimal step size measured by out-of-sample spectral probing (the measured profile of the Muon trajectory at checkpoint 1000 in Figure 1).
- *SGD*: a virtual update that assigns each rank a scale proportional to its average singular value, that is, the allocation produced when SGD applies the momentum buffer directly.
- *Adam*: the true Adam update $\mathbf{\Delta}^{\mathrm{A}}$ projected onto each rank of the singular basis of its own buffer, $a_i = \boldsymbol{u}_i^\top \mathbf{\Delta}^{\mathrm{A}} \boldsymbol{v}_i$, that is, the diagonal of $\mathbf{U}^\top \mathbf{\Delta}^{\mathrm{A}} \mathbf{V}$. Only these diagonal terms are tracked. The off-diagonal terms $\boldsymbol{u}_i^\top \mathbf{\Delta}^{\mathrm{A}} \boldsymbol{v}_j$ with $i \neq j$, which describe how Adam couples distinct singular directions, are ignored. This update is sampled from checkpoint 1000 of the AdamW trajectory.
- *Muon*: exact whitening, a uniform scale across the spectrum.

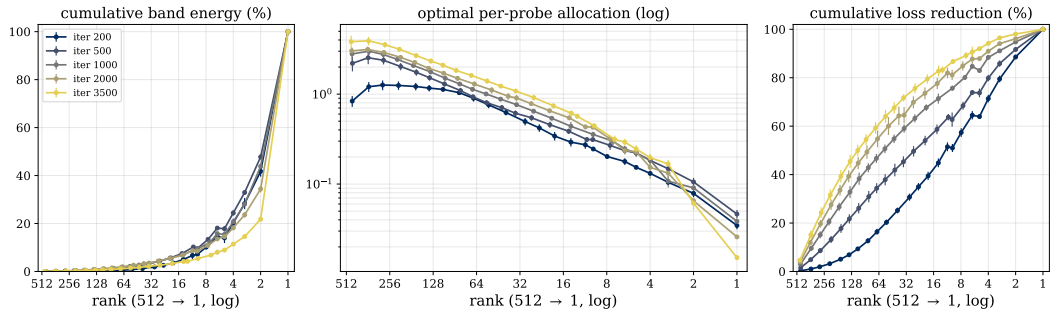

Figure 9: **In-sample loss-optimal step size profile along an AdamW trajectory**, the AdamW analogue of the Muon Figure 7 and the in-sample counterpart of the out-of-sample Figure 4 (same 64M-parameter modded-nanogpt model and checkpoints; three-panel layout as Figure 1), with each probe's optimal step size measured on the *same* batch whose gradient formed the buffer ($\mathcal{B}_o = \mathcal{B}_g$). As for Muon (Figure 7), the bulk is lifted above its out-of-sample level (middle panel) and the deep tail contributes far more of the idealised cumulative loss reduction (right panel), the same batch-refitting signatures that the main-text out-of-sample protocol removes.

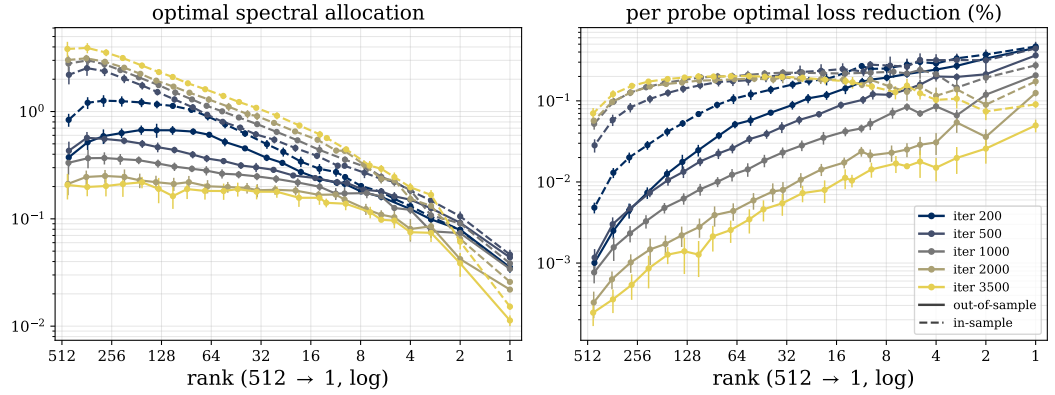

Figure 10: **In-sample versus out-of-sample spectral profiles along an AdamW trajectory** (64M-parameter "modded-nanogpt", checkpoints from iterations 200–3500), the AdamW analogue of the Muon Figure 8; panels and axes as there, with solid curves out-of-sample and dashed in-sample. As for Muon, the in-sample allocation is systematically more permissive across the bulk (widest in the deep tail) and every in-sample probe yields a large loss reduction, whereas out-of-sample the tail's contribution collapses; the gap is the batch-refitting capacity that does not transfer to unseen data.

- *SAMuon-lite*: the two-level allocation of Section 5 with $k = 1$, holding the head at the same scale as Muon and amplifying the bulk uniformly to $\gamma = 5$, the measured head gap.
- *SAMuon*: the log-rank-linear allocation of Equation (9) at $\gamma = 7.07$, its tuned optimum at all batch sizes (Table 3), ramping from the anchored head to the flat bulk boost over the leading k ranks.

Both the SAMuon and SAMuon-lite allocations are drawn at their post-warmup target profiles. The comparison summarises the argument of Section 4 that the SGD allocation is anti-correlated with the measured profile; Adam dampens but does not overcome the head's dominance; Muon reallocates several times the scale towards the bulk relative to SGD; SAMuon-lite holds the head at the same scale as Muon and lifts the bulk uniformly towards the measured profile; and SAMuon follows the measured near-head decline itself, thereby providing the closest alignment among the five methods.

H SAMUON AND SAMUON-LITE ALGORITHM

Algorithm 1 SAMuon / SAMuon-lite, shown for a single 2D weight matrix

Input: initial weights $\mathbf{W}^{(1)}$, momentum buffer $\mathbf{M}^{(0)} = \mathbf{0}$, Scion matrix scaling $\kappa = \sqrt{d_{\text{out}}/d_{\text{in}}}$
Hyperparameters: η (effective LR), μ (momentum), γ (bulk boost), T (total steps); profile rank k (width rule, Appendix J) and target scales $\{s_i\}_{i=1}^{k}$ (Equation (9)) for SAMuon only
for $t = 1, \ldots, T$ **do**
  Compute gradient $\mathbf{G}^{(t)}$ on batch $\mathcal{B}_t$
  $\mathbf{M}^{(t)} \leftarrow \mu\,\mathbf{M}^{(t-1)} + (1-\mu)\,\mathbf{G}^{(t)}$
  $\gamma_t \leftarrow 1 + (\gamma - 1)\,w_t$ ▷ cosine-scheduled spectral warmup
  Variant #1: SAMuon (log-rank-linear profile):
    $(\boldsymbol{u}_i, \boldsymbol{v}_i)_{i \le k} \leftarrow \mathrm{LowRankSVD}(\mathbf{M}^{(t)}, k)$ ▷ `torch.svd_lowrank()`
    $s_i(t) \leftarrow 1 + (s_i - 1)\,w_t$, for $i = 1, \ldots, k$ ▷ leading-rank spectral warmup
    $\mathbf{O}_{\text{SA}}^{(t)} \leftarrow \gamma_t\,\mathrm{NS}(\mathbf{M}^{(t)}) - \sum_{i=1}^{k}(\gamma_t - s_i(t))\,\boldsymbol{u}_i\boldsymbol{v}_i^{\top}$ ▷ Equation (8)
  Variant #2: SAMuon-lite ($k = 1$, simple two-level profile):
    $(\boldsymbol{u}_1, \boldsymbol{v}_1) \leftarrow \mathrm{PowerIter}(\mathbf{M}^{(t)})$
    $\mathbf{O}_{\text{SA}}^{(t)} \leftarrow \gamma_t\,\mathrm{NS}(\mathbf{M}^{(t)}) - (\gamma_t - 1)\,\boldsymbol{u}_1\boldsymbol{v}_1^{\top}$ ▷ Equation (8) with $k = 1$
  $\mathbf{W}^{(t+1)} \leftarrow \mathbf{W}^{(t)} - \eta\,\kappa\,\mathbf{O}_{\text{SA}}^{(t)}$
end for

I CONVERGENCE PROOF FOR THE IDEALISED SAMUON UPDATE

In this section, we establish convergence guarantees for the idealised exact-whitening form of the SAMuon / SAMuon-lite variants of Algorithm 1. By extending the convergence analysis framework of standard Muon (see Theorem 4.1 in Shen et al. (2026)) to accommodate the head-anchored spectral allocation of Section 5.1, we bound the iteration-averaged expected gradient nuclear norm. Although the spectral shaping differs substantially from uniform whitening, the analysis depends on it only through its bounds, so a single argument covers SAMuon, SAMuon-lite, and other related profiles while mathematically formalising the trade-off introduced by the bulk boost γ.

I.1 SETUP AND NOTATION

Following the standard matrix-structured analysis of Muon, we state the result for one two-dimensional parameter matrix, matching the blockwise form of the optimiser update. Accordingly, the theorem below is a single-matrix guarantee; SAMuon applies the same update independently to every eligible matrix in the model. Consider the task of minimising a non-convex objective function $l : \mathbb{R}^{m\times n} \to \mathbb{R}$. Let $\mathbf{W}_t \in \mathbb{R}^{m\times n}$ denote the parameter matrix at iteration t, and let $r = \min(m, n)$ denote the maximum possible rank of the matrix.

The optimiser maintains an exponential moving average (EMA) of the stochastic gradients, initialised at zero as in Algorithm 1, and reshapes its spectrum with the head-anchored allocation of Section 5.1.

For $t = 1, \ldots, T$, the update rule is

$$\mathbf{M}_0 = \mathbf{0}, \tag{10}$$
$$\mathbf{M}_t = \mu\mathbf{M}_{t-1} + (1-\mu)\mathbf{G}_t \tag{11}$$
$$\mathbf{Q}_t = \mathrm{SAMuon}(\mathbf{M}_t) \tag{12}$$
$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta\mathbf{Q}_t \tag{13}$$

where η is a constant effective step size, $\mu \in [0, 1)$ is the momentum parameter, and $\mathbf{G}_t$ is the stochastic mini-batch gradient. Because this is a single-matrix result, the fixed positive Scion scale κ is absorbed into η. As in Section 3, each momentum buffer is assumed to have rank r. Here $\mathrm{SAMuon}(\cdot)$ denotes the anisotropic spectral scaling of Equation (8). Following Shen et al. (2026), we idealise the Newton–Schulz iteration as exact whitening. Then each singular direction receives exactly its prescribed scale, with the head remaining at the same unit scale as in Muon, while $\mathbf{Q}_t$ shares the singular vectors of $\mathbf{M}_t$. The guarantee therefore covers this idealised exact-whitening

update, as is standard for analyses of this type; the finite-step Newton–Schulz, randomised-SVD, and power-iteration approximations run by the deployed algorithms lie outside the bound.

Consequently, if the Singular Value Decomposition (SVD) of the momentum buffer is $\mathbf{M}_t = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, the preconditioned update evaluates to $\mathbf{Q}_t = \mathbf{U}\hat{\mathbf{\Sigma}}\mathbf{V}^\top$ with modified singular values

$$\hat{\sigma}_i = \begin{cases} s_i(t) & i \le k \quad \text{(the leading-rank transition, with the head anchored at } \hat{\sigma}_1 = s_1(t) = 1) \\ \gamma_t & i > k \quad \text{(uniformly boosted ranks),} \end{cases} \tag{14}$$

following the head-anchored allocation of Equation (8); the two-level SAMuon-lite is the case $k = 1$, and SAMuon takes the log-rank-linear scales of Equation (9). The analysis below depends on this scaling only through its bounds, with the amplification fixed at $\gamma \ge 1$. As the warmup raises the whole profile monotonically from the all-ones Muon scaling towards its target over the first T_0 steps (so $1 \le s_i(t) \le \gamma_t \le \gamma$ throughout), all modified singular values satisfy

$$1 \;\le\; \hat{\sigma}_i \;\le\; \gamma \qquad \text{for all } i \text{ and } t. \tag{15}$$

All bounds below are therefore stated in terms of the post-warmup amplification γ, and hold uniformly over training. More generally, the same analysis applies to any diagonal scaling in the exact singular-vector coordinates of $\mathbf{M}_t$ with coefficients in $[1, \gamma]$; it does not require head anchoring or a particular rank ordering of the scales.

I.2 Assumptions

We rely on the following standard assumptions regarding the loss landscape and the stochastic oracle, following Shen et al. (2026). Both landscape assumptions are idealisations relative to Transformer practice: a single global smoothness constant is assumed and the variance bound is assumed uniform along the trajectory.

Assumption I.1 (L-Smoothness)**.** The objective function l has L-Lipschitz continuous gradients with respect to the Frobenius norm. For any $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{m\times n}$:

$$l(\mathbf{W}) \le l(\mathbf{W}') + \langle \nabla l(\mathbf{W}'), \mathbf{W} - \mathbf{W}' \rangle + \frac{L}{2}\|\mathbf{W} - \mathbf{W}'\|_F^2 \tag{16}$$

Assumption I.2 (Bounded Variance)**.** Let $\mathcal{F}_t$ denote the optimisation history before $\mathbf{G}_t$ is sampled. At each step t, the stochastic oracle provides a conditionally unbiased gradient estimator with bounded conditional variance:

$$\mathbb{E}[\mathbf{G}_t \mid \mathcal{F}_t] = \nabla l(\mathbf{W}_t), \tag{17}$$

$$\mathbb{E}[\|\mathbf{G}_t - \nabla l(\mathbf{W}_t)\|_F^2 \mid \mathcal{F}_t] \le \frac{\sigma^2}{B}, \tag{18}$$

where B is the mini-batch size.

Assumption I.3 (Lower-Bounded Objective)**.** The objective function $l(\mathbf{W})$ is bounded from below by a global minimum l^*.

I.3 Main convergence theorems

Theorem I.4 (Ergodic Convergence of the Idealised SAMuon Update)**.** *Under assumptions I.1 to I.3, using the idealised exact-whitening SAMuon update with constant learning rate η for T iterations yields the following bound on the expected nuclear norm of the true gradient:*

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla l(\mathbf{W}_t)\|_*] \le & \frac{l(\mathbf{W}_1) - l^*}{\eta T} + \frac{Lr\gamma^2\eta}{2} \\ & + \sqrt{r}(1+\gamma)\left(\frac{\sigma\sqrt{1-\mu}}{\sqrt{B(1+\mu)}} + \frac{\mu\|\nabla l(\mathbf{W}_1)\|_F}{(1-\mu)T} + \frac{\sqrt{r}\gamma\eta\mu L}{1-\mu}\right). \end{aligned} \tag{19}$$

The following choices recover the stochastic and deterministic convergence orders of Muon while retaining the same zero-buffer initialisation as Algorithm 1.

Corollary I.5 (Stochastic and Deterministic Convergence Rates). *Let $\mathcal{D} = l(\mathbf{W}_1) - l^*$ and regard γ as fixed.*

(i) In the stochastic setting, set $B = 1$, $\eta = \sqrt{(1-\mu)\mathcal{D}/(rTL)}$, and $1 - \mu = \min\{\sqrt{L\mathcal{D}}/(\sigma\sqrt{T}), 1\}$. Then

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla l(\mathbf{W}_t)\|_*] \le \mathcal{O}_\gamma\left(\sqrt[4]{\frac{r^2 L\mathcal{D}\sigma^2}{T}} + \sqrt{\frac{rL\mathcal{D}}{T}} + \frac{\sqrt{r}\,\sigma}{\sqrt{T}}\right). \tag{20}$$

(ii) In the deterministic setting ($\sigma = 0$), set $\eta = \sqrt{\mathcal{D}/(rTL)}$ and hold $\mu < 1$ fixed independently of T. Then

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla l(\mathbf{W}_t)\|_* \le \mathcal{O}_\gamma\left(\sqrt{\frac{rL\mathcal{D}}{T}}\right). \tag{21}$$

The bounds follow by substituting these choices into Theorem I.4 and using the standard consequence $\|\nabla l(\mathbf{W}_1)\|_F \le \sqrt{2L\mathcal{D}}$ of L-smoothness and lower boundedness; the zero-initialisation term is lower order in both cases. Here $\mathcal{O}_\gamma$ suppresses constants depending on the fixed amplification bound γ (and, in the deterministic case, on the fixed momentum coefficient).

These are bounds on the iteration-averaged expected gradient nuclear norm, equivalently on the expected gradient nuclear norm at an iterate sampled uniformly from $\{\mathbf{W}_1, \dots, \mathbf{W}_T\}$, rather than on the final iterate. The stochastic rate uses learning-rate and momentum choices that depend on the horizon T, unlike the fixed momentum used in our experiments. For fixed $\mu < 1$ and batch size B with $\sigma > 0$, the theorem retains a nonvanishing stochastic-noise term and does not establish convergence of this stationarity measure to zero.

Remark I.6. Setting $\gamma = 1$ (and hence $\gamma_t \equiv 1$, $\hat{\sigma}_i \equiv 1$) specialises Theorem I.4 to Muon with the zero-buffer initialisation of Algorithm 1. This initialisation produces the vanishing initial-gradient term in the theorem; the first-gradient initialisation analysed by Shen et al. (2026) instead produces a vanishing initial-noise term. Both conventions give the stochastic and deterministic convergence orders stated in corollary I.5.

The bound makes the cost of amplification explicit: increasing γ increases the stochastic noise term through $1 + \gamma$ and the deterministic penalties through γ^2 and $\gamma(1 + \gamma)$. The head anchor ($\hat{\sigma}_1 = 1$) is motivated by the measured head–bulk difference in loss-optimal step sizes (Section 4), rather than by the bound, which does not distinguish which rank receives which scale. Both proposed profiles satisfy the same guarantee at a given γ, since their scales remain within $[1, \gamma]$.

I.4 Detailed proofs

Proof of Theorem I.4. **Step 1: The Descent Lemma and the Anisotropic Penalty**
We begin by substituting the SAMuon parameter update $\mathbf{W}_{t+1} - \mathbf{W}_t = -\eta\mathbf{Q}_t$ into the L-smoothness inequality (assumption I.1):

$$l(\mathbf{W}_{t+1}) \le l(\mathbf{W}_t) - \eta\langle\nabla l(\mathbf{W}_t), \mathbf{Q}_t\rangle + \frac{L\eta^2}{2}\|\mathbf{Q}_t\|_F^2 \tag{22}$$

Because $\mathbf{Q}_t$ is constrained by the spectral bounds of Equation (15), its singular values are bounded above by γ. The squared Frobenius norm is geometrically bounded independently of the raw gradient scale:

$$\|\mathbf{Q}_t\|_F^2 = \sum_{i=1}^{r}\hat{\sigma}_i^2 \le \sum_{i=1}^{r}\gamma^2 = r\gamma^2 \tag{23}$$

Step 2: Decomposing the Descent Direction
We evaluate the inner product between the true gradient and the anisotropic update by adding and subtracting the momentum buffer $\mathbf{M}_t$:

$$\langle\nabla l(\mathbf{W}_t), \mathbf{Q}_t\rangle = \langle\mathbf{M}_t, \mathbf{Q}_t\rangle + \langle\nabla l(\mathbf{W}_t) - \mathbf{M}_t, \mathbf{Q}_t\rangle \tag{24}$$

Step 3: The Nuclear Norm Alignment
Since $\mathbf{M}_t$ and $\mathbf{Q}_t$ share identical singular vectors by design, their trace inner product collapses into a weighted sum of the singular values of $\mathbf{M}_t$. Because Equation (15) guarantees $\hat{\sigma}_i \geq 1$ for all dimensions, we obtain:

$$\langle \mathbf{M}_t, \mathbf{Q}_t \rangle = \sum_{i=1}^{r} \sigma_i \hat{\sigma}_i \geq \sum_{i=1}^{r} \sigma_i = \|\mathbf{M}_t\|_* \tag{25}$$

Step 4: Bounding the Amplified Tracking Error
Let $\mathbf{\Delta}_t = \nabla l(\mathbf{W}_t) - \mathbf{M}_t$ represent the tracking error. Applying the Cauchy–Schwarz inequality, followed by the upper bound $\|\mathbf{Q}_t\|_F \leq \sqrt{r}\gamma$ derived from Equation (23) gives:

$$\langle \mathbf{\Delta}_t, \mathbf{Q}_t \rangle \geq -\|\mathbf{\Delta}_t\|_F \|\mathbf{Q}_t\|_F \geq -\sqrt{r}\gamma \|\mathbf{\Delta}_t\|_F \tag{26}$$

Substituting Equations (25) and (26) into Equation (24) yields:

$$\langle \nabla l(\mathbf{W}_t), \mathbf{Q}_t \rangle \geq \|\mathbf{M}_t\|_* - \sqrt{r}\gamma \|\mathbf{\Delta}_t\|_F \tag{27}$$

Relating $\|\mathbf{M}_t\|_*$ to the true gradient via the reverse triangle inequality ($\|\mathbf{A}\|_* \geq \|\mathbf{B}\|_* - \|\mathbf{A} - \mathbf{B}\|_*$) and the matrix norm bound $\|\mathbf{\Delta}_t\|_* \leq \sqrt{r}\|\mathbf{\Delta}_t\|_F$ gives:

$$\|\mathbf{M}_t\|_* \geq \|\nabla l(\mathbf{W}_t)\|_* - \sqrt{r}\|\mathbf{\Delta}_t\|_F \tag{28}$$

Combining these establishes a lower bound for the descent direction, exposing the γ tracking multiplier:

$$\langle \nabla l(\mathbf{W}_t), \mathbf{Q}_t \rangle \geq \|\nabla l(\mathbf{W}_t)\|_* - \sqrt{r}(1+\gamma)\|\mathbf{\Delta}_t\|_F \tag{29}$$

Step 5: Bounding the EMA Tracking Error
Define the moving average of the true gradients by $\mathbf{C}_0 = \mathbf{0}$ and

$$\mathbf{C}_t = \mu \mathbf{C}_{t-1} + (1-\mu)\nabla l(\mathbf{W}_t). \tag{30}$$

The tracking error decomposes as

$$\|\mathbf{\Delta}_t\|_F \leq \|\mathbf{M}_t - \mathbf{C}_t\|_F + \|\nabla l(\mathbf{W}_t) - \mathbf{C}_t\|_F. \tag{31}$$

For the first term, unrolling the two moving averages gives

$$\mathbf{M}_t - \mathbf{C}_t = (1-\mu)\sum_{i=1}^{t} \mu^{t-i}\big(\mathbf{G}_i - \nabla l(\mathbf{W}_i)\big). \tag{32}$$

The conditional unbiasedness in assumption I.2 makes the summands a martingale-difference sequence. Hence the cross terms vanish after taking expectations, and Cauchy–Schwarz gives

$$\mathbb{E}[\|\mathbf{M}_t - \mathbf{C}_t\|_F] \leq \sqrt{\mathbb{E}[\|\mathbf{M}_t - \mathbf{C}_t\|_F^2]} \tag{33}$$

$$\leq \frac{\sigma}{\sqrt{B}}(1-\mu)\sqrt{\sum_{i=1}^{t}\mu^{2(t-i)}} \leq \frac{\sigma\sqrt{1-\mu}}{\sqrt{B(1+\mu)}}. \tag{34}$$

For the second term, let $\mathbf{D}_t = \nabla l(\mathbf{W}_t) - \mathbf{C}_t$. Since $\mathbf{C}_0 = \mathbf{0}$, $\mathbf{D}_1 = \mu\nabla l(\mathbf{W}_1)$. For $t \geq 2$, L-smoothness and Equation (23) give

$$\|\mathbf{D}_t\|_F \leq \mu\|\mathbf{D}_{t-1}\|_F + \mu\|\nabla l(\mathbf{W}_t) - \nabla l(\mathbf{W}_{t-1})\|_F \tag{35}$$

$$\leq \mu\|\mathbf{D}_{t-1}\|_F + \mu L\eta\sqrt{r}\gamma. \tag{36}$$

Unrolling this recursion yields

$$\|\mathbf{D}_t\|_F \leq \mu^t\|\nabla l(\mathbf{W}_1)\|_F + \frac{\sqrt{r}\gamma\eta\mu L}{1-\mu}. \tag{37}$$

Combining the two components therefore gives

$$\mathbb{E}[\|\mathbf{\Delta}_t\|_F] \leq \frac{\sigma\sqrt{1-\mu}}{\sqrt{B(1+\mu)}} + \mu^t\|\nabla l(\mathbf{W}_1)\|_F + \frac{\sqrt{r}\gamma\eta\mu L}{1-\mu}. \tag{38}$$

Step 6: Telescoping the Descent Bound
Taking the expectation of the descent lemma (Equation (22)) and substituting Equations (23) and (29) gives:

$$\mathbb{E}[l(\mathbf{W}_{t+1})] \leq \mathbb{E}[l(\mathbf{W}_t)] - \eta\mathbb{E}[\|\nabla l(\mathbf{W}_t)\|_*] + \eta\sqrt{r}(1+\gamma)\mathbb{E}[\|\mathbf{\Delta}_t\|_F] + \frac{L\eta^2 r\gamma^2}{2} \tag{39}$$

Rearranging gives the following bound on the expected nuclear norm of the true gradient:

$$\eta\mathbb{E}[\|\nabla l(\mathbf{W}_t)\|_*] \leq \mathbb{E}[l(\mathbf{W}_t)] - \mathbb{E}[l(\mathbf{W}_{t+1})] + \eta\sqrt{r}(1+\gamma)\mathbb{E}[\|\mathbf{\Delta}_t\|_F] + \frac{L\eta^2 r\gamma^2}{2} \tag{40}$$

Summing this inequality from $t = 1$ to T, noting that $l(\mathbf{W}_{T+1}) \geq l^*$ (assumption I.3), and dividing by ηT yields:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla l(\mathbf{W}_t)\|_*] \leq \frac{l(\mathbf{W}_1) - l^*}{\eta T} + \frac{Lr\gamma^2\eta}{2} + \frac{\sqrt{r}(1+\gamma)}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{\Delta}_t\|_F] \tag{41}$$

Substituting Equation (38) and using $T^{-1}\sum_{t=1}^{T}\mu^t \leq \mu/((1-\mu)T)$ proves Theorem I.4. □

J Detailed experimental setup

In this section, we provide the exact hyperparameters and architectural details to ensure full reproducibility of our results. Our codebase and training pipeline are built directly upon the open-source implementation provided by Pethick et al. (2025).

J.1 Motivation for using Scion as the "Muon" baseline

In our experiments, Muon serves as a crucial strong baseline. Instead of the vanilla Muon implementation, we are in fact using the improved Muon version, Scion (Pethick et al., 2025). We select Scion over the vanilla Muon implementation for three strategic reasons: (i) *Efficiency:* Scion replaces the auxiliary Adam optimiser for embeddings with a lightweight signed update rule, reducing memory overhead; (ii) *Stability:* It employs a theoretically grounded normalisation scheme that decouples the learning rate from model scale, facilitating consistent hyperparameter tuning; and (iii) *Performance:* Scion consistently demonstrates improved empirical convergence and large-batch scalability over vanilla Muon.

J.2 Model architecture

All models are decoder-only Transformers based on the "modded-nanogpt" architecture (Jordan et al., 2024a; Pethick et al., 2025), which modifies GPT-2 by replacing LayerNorm with RMSNorm, replacing absolute positional embeddings with Rotary Positional Embeddings (RoPE) (Su et al., 2024), and removing bias terms to improve arithmetic efficiency. The 64M and 124M models used for the spectral probing analyses, and the 124M, 300M, and 1B models used for the training experiments, all share the same depth (12 layers) and head dimension, so width is the only scale axis among them. The vocabulary size is padded to 50,304 to improve kernel efficiency. The complete architectural specifications are summarised in Table 2.

J.3 Training hyperparameters

The optimiser-comparison experiments train on the FineWeb dataset (Penedo et al., 2024) (100B-token sample; ODC-By 1.0 licence) at global batch sizes of 1M, 2M, and 4M tokens (1024, 2048, and 4096 sequences). The 124M and 300M models use a $\sim$10B-token budget, with iteration counts of 9600, 4800, and 2400 at the three batch sizes, and the 1B model a $\sim$20B-token budget. For the 124M model this budget is $\sim 4\times$ the Chinchilla-optimal allocation (Hoffmann et al., 2022), falling to $\sim 1.7\times$ at 300M and $\sim 1\times$ at 1B.

Table 2: **Model architectures.** The 64M, 124M, 300M and 1B models used for probing and/or training experiments. All utilise a head dimension of 128 and a depth of 12 layers.

Specification	**64M**	**124M**	**300M**	**1B**
Layers (L)	12	12	12	12
Model Dimension (d_{model})	512	768	1280	2560
Attention Heads (H)	4	6	10	20
Head Dimension (d_h)	128	128	128	128
Vocabulary Size	50,304	50,304	50,304	50,304
Context Length	1024	1024	1024	1024

Each optimiser configuration is evaluated with a single seed. Within each model-scale and batch-size setting, all optimisers use the same model initialisation and identical data ordering. All training runs use NVIDIA A100 (80GB) GPUs with bfloat16 mixed precision.

Learning-rate schedule. Following Pethick et al. (2025), we use a trapezoidal learning-rate schedule with a constant phase followed by linear decay over the final 28.5% of training tokens. AdamW additionally uses a linear learning-rate warmup over the first 5% of training. The separate spectral warmup for the SAMuon variants is specified below.

Optimiser Settings. Muon and both SAMuon variants use a momentum coefficient of $\mu = 0.9$ in all experiments.

We tune AdamW's learning rate, Muon's spectral radius, and each SAMuon variant's bulk boost separately for each batch size on the 124M model. The remaining settings are fixed. These hyperparameters are transferred across model scales, with the single exception of the AdamW learning rate, which is halved at each scale step ($2^{-8.5} \to 2^{-9.5} \to 2^{-10.5}$ at batch size 1024) following the optimal-learning-rate trend for AdamW reported in Pethick et al. (2025). For AdamW we sweep the learning rate on a grid around the Pethick et al. (2025) value, with successive values differing by a factor of 2. For Muon we sweep the spectral radius `radius_2d` (the learning-rate factor on the 2D, Muon-effective modules) over the $\sqrt{2}$-spaced grid $\{35.36, 50, 70.71, 100, 141.42\}$. Since this sweep is flat around its optimum, both SAMuon instantiations instead run at the fixed radius 50, the Scion default, and sweep only the bulk boost γ over $\{3.54, 5, 7.07, 10, 14.14\}$, a grid chosen around the measured head gap. The complete sweeps are reported in Table 5, and the per-batch optima are collected in Table 3.

Probing-run training setup. The main-text spectral probing analyses are run on the 64M and 124M models, following the Pethick et al. (2025) protocol: 5100 training steps at batch size 512, using the optimiser hyperparameters specified by Pethick et al. (2025). For the smallest 64M model, the probed Muon trajectory reaches a final validation loss of 3.4292, and the AdamW trajectory probed in Appendix D reaches 3.4913 under the same protocol. The width-768 (124M) Muon trajectory probed in Appendix E follows exactly the 124M setup of the same protocol, with the learning rate transferred from the 64M model following Pethick et al. (2025), and reaches a final validation loss of 3.2812.

Profile warmup schedule. The spectral shaping is itself warmed up. In the full-budget training runs, the *cosine* warmup weight $w_t = \frac{1}{2}\big(1-\cos(\pi \min(t/T_0,\, 1))\big)$ spans the first 30% of training iterations ($T_0 = 0.3\,T$), interpolating the whole profile from the Muon scaling through $\gamma_t = 1 + (\gamma - 1)\, w_t$ and $s_i(t) = 1 + (s_i - 1)\, w_t$, after which it is held at its target. The warmup span mirrors the trapezoidal schedule's final linear-decay phase (28.5% of training) with a matching $\sim 30\%$ ramp at the start. This lets the model first settle onto a stable Muon trajectory before the shaping fully engages, and applies identically to both instantiations (for SAMuon-lite, it reduces to the single-scale ramp of γ_t). The shortened runs used only for the token-efficiency measurement instead retain the fixed warmup horizon $T_0 = 0.3\,N$ of the corresponding Muon baseline, as detailed in Appendix K.

This warmup is not arbitrary since the spectral anisotropy itself *emerges* slowly over early training. Figure 11 tracks the out-of-sample profile along the Muon trajectory and shows that the head gap rises from a smaller gap (≈ 2 at iteration 50) towards its final level (≈ 5) over the first few hundred iterations, while the head's concentration of spectral energy and directional curvature also climbs at

the same time. Applying the full shaping from the start would therefore overdrive a spectrum that is not yet highly anisotropic. Warming the profile instead lets the optimiser track the anisotropy as it emerges and supplies a measurement-grounded justification for delaying the full shaping (the span itself mirrors the schedule's decay phase, as above). The final head gap also anchors the bulk-boost *search*. The γ grid of our experiments is chosen around the measured gap ($\approx$5), and the optima sit at or moderately above it, namely $\gamma^* = 7.07$ for SAMuon at all batch sizes and between 7.07 and 10 for SAMuon-lite (Table 5).

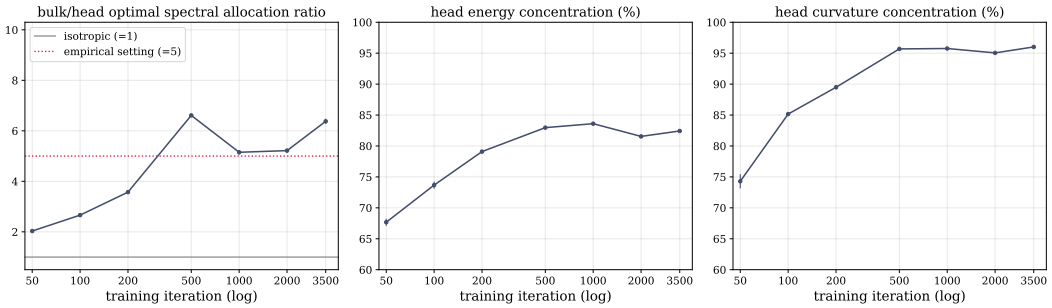

Figure 11: **The spectral anisotropy emerges over early training, motivating the slow warmup.** Out-of-sample probing along the Muon trajectory (64M-parameter "modded-nanogpt"), against training iteration (log axis). **Left:** the head gap (rank-2 over rank-1 optimal spectral allocation ratio, the largest single gap in the measured profile) rises from a smaller gap ($\approx$ 2) at iteration 50 towards its final level of $\approx$ 5 (red dotted), on which the γ search grid of our experiments is centred, within a few hundred iterations, and stays there; loss-optimal step sizes are nonuniform throughout. **Middle:** the head's concentration of spectral energy, $\sigma_1^2 / \sum_i \sigma_i^2$; **Right:** the head's concentration of directional curvature, $\big(\Delta\boldsymbol{\theta}^{(1)\top}\mathbf{H}\,\Delta\boldsymbol{\theta}^{(1)}\big) \big/ \sum_d \Delta\boldsymbol{\theta}^{(d)\top}\mathbf{H}\,\Delta\boldsymbol{\theta}^{(d)}$, using the rank-$d$ directional curvature of Equation (7); both concentrate over the same window, from $\sim$70% to $\sim$85–95%. Early in training the loss-optimal step size profile is less anisotropic, so SAMuon and SAMuon-lite delay the full anisotropic shaping and warm the profile from the uniform scaling of Muon to its target over the first 30% of training (much slower than standard warmup schedules).

J.4 Newton–Schulz Parameters

We adopt the Newton–Schulz iteration and its polynomial coefficients from the LaCT implementation (Zhang et al., 2026a). We use five Newton–Schulz iterations, each with three polynomial coefficients (5×3). Our implementation evaluates each iteration using four matrix multiplications. For a normalised input $\mathbf{X}_1$, iteration ℓ applies

$$\mathbf{X}_{\ell+1} = a_\ell \mathbf{X}_\ell + b_\ell \big(\mathbf{X}_\ell \mathbf{X}_\ell^\top\big)\mathbf{X}_\ell + c_\ell \big(\mathbf{X}_\ell \mathbf{X}_\ell^\top\big)^2 \mathbf{X}_\ell, \qquad \ell = 1, \ldots, 5. \tag{42}$$

The iteration-specific coefficients in Table 4 are adopted unchanged, without learning or tuning them in this work. Figure 12 shows the resulting response after five iterations. Numerically evaluating the scalar recurrence with these coefficients gives a maximum absolute deviation from one of approximately 0.0318 over normalised input singular values $\sigma \in [0.02, 1]$.

Table 3: **Optimiser hyperparameters.** Fixed settings and selected hyperparameters are reported below. Tuned optima are listed for batch sizes 1024/2048/4096, with the full sweeps in Table 5. Both SAMuon instantiations run at the fixed spectral radius 50. The profile rank k of SAMuon follows the fixed width rule described below and is not tuned. All values transfer across model scales, except the AdamW learning rate, which is halved at each scale step.

Parameter	**AdamW**	**Muon (Scion)**	**SAMuon-lite**	**SAMuon**
Global Settings				
Momentum	0.9	0.9	0.9	0.9
Weight Decay	0.0	0.0	0.0	0.0
Gradient Clip	1.0	None	None	None
LR Schedule	Trapezoidal	Trapezoidal	Trapezoidal	Trapezoidal
LR Warmup	5%	0%	0%	0%
Hyperparameter sweeps at 124M (optima at bs 1024/2048/4096)				
LR / radius grid	2^x, unit steps in x	$\{35.36, \ldots, 141.42\}$	fixed at 50	fixed at 50
Optimal LR / radius	$2^{-8.5}/2^{-9}/2^{-8.5}$	$70.71/50/70.71$	50	50
Bulk boost γ grid	–	–	$\{3.54, \ldots, 14.14\}$	$\{3.54, \ldots, 14.14\}$
Optimal γ	–	–	$10/7.07/10$	7.07 (all)
Profile rank k	–	–	1	$39/50/71$
Profile warmup	–	–	30% (cosine)	30% (cosine)
Transfer to 300M and 1B				
Across model scales	LR halved per step	unchanged	unchanged	k by width rule

Table 4: Newton–Schulz coefficients adopted from LaCT (Zhang et al., 2026a).

Iteration ℓ	a_ℓ	b_ℓ	c_ℓ
1	4.0848	−6.8946	2.9270
2	3.9505	−6.3029	2.6377
3	3.7418	−5.5913	2.3037
4	2.8769	−3.1427	1.2046
5	2.8366	−3.0525	1.2012

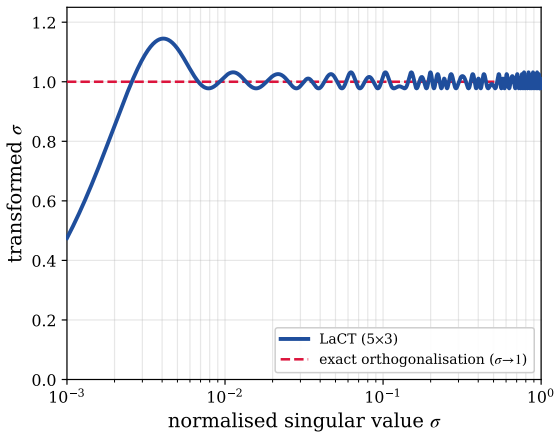

Figure 12: Response of the 5×3 Newton–Schulz scheme after five iterations with the LaCT coefficients in Table 4. The dashed line denotes exact orthogonalisation, which maps every non-zero singular value to one. Over normalised input singular values $\sigma \in [0.02, 1]$, the maximum absolute deviation from one is approximately 0.0318.

Table 5: **Hyperparameter searches at 124M** (width 768, FineWeb-100B, ∼10B tokens), for each batch size. Entries are final validation loss, and the per-batch optimum of each sweep is in bold. AdamW is swept over learning rate (2^x), Muon over its spectral radius, and both SAMuon instantiations over the bulk boost γ at the re-anchored radius 50, the Scion default. The Muon optimum shifts with batch size ($70.71/50/70.71$), whereas the SAMuon optimum is $\gamma = 7.07$ at all batch sizes. Searches at the superseded radius 70.71 are omitted for the SAMuon instantiations.

Hyperparameter	bs 1024	bs 2048	bs 4096
AdamW: learning rate ($\log_2$)			
-10.5	3.2363	–	–
-10	–	3.2620	–
-9.5	3.2113	–	3.3266
-9	–	**3.2410**	–
-8.5	**3.2105**	–	**3.3050**
-8	–	3.2462	–
-7.5	3.2360	–	3.3392
-7	–	3.2963	–
-6.5	–	–	3.4059
Muon: spectral radius			
35.36	3.1625	3.1768	3.2046
50	3.1641	**3.1726**	3.1979
70.71	**3.1622**	3.1743	**3.1953**
100	3.1641	3.1757	3.1965
141.42	3.1649	3.1761	3.1986
SAMuon: γ at radius 50			
3.54	3.1473	3.1535	3.1734
5	3.1445	3.1502	3.1686
7.07	**3.1435**	**3.1484**	**3.1658**
10	3.1442	3.1486	3.1662
14.14	–	3.1504	–
SAMuon-lite: γ at radius 50			
3.54	3.1462	3.1553	3.1724
5	3.1445	3.1523	3.1699
7.07	3.1449	**3.1511**	3.1680
10	**3.1438**	3.1516	**3.1678**
14.14	–	3.1550	–

J.5 Full sweep results

Table 5 reports all sweep results underlying the tuned hyperparameters in Table 3, covering the AdamW learning rate, the Muon spectral radius, and the bulk boost of both SAMuon instantiations. Note that SAMuon attains a best-of-sweep loss at least as low as SAMuon-lite at all three batch sizes, with the two effectively tied at batch size 1024, and that its optimum is at $\gamma = 7.07$ in all cases whereas the SAMuon-lite optimum varies.

J.6 SAMuon / SAMuon-lite configurations

Both instantiations run at the fixed spectral radius 50 and add the bulk boost γ, swept per batch size on the 124M model, with the selected optima summarised in Table 3 and the complete sweeps reported in Table 5. The bulk boost is ramped from the Muon value over the cosine profile warmup spanning the first 30% of training. Setting $\gamma = 1$ recovers Muon exactly. SAMuon-lite captures the head with a rank-1 power iteration (Section 5). SAMuon instead requires the leading k singular pairs of the momentum buffer, computed per step with `torch.svd_lowrank`. Its profile rank follows the width rule $k(d_{\text{model}}) = \left\lfloor 32\sqrt{d_{\text{model}}/512} \right\rfloor$ below, *i.e.* $k = 39/50/71$ at the trained widths $768/1280/2560$, held fixed and not tuned.

Low-rank estimator hyperparameters. For the 124M, 300M, and 1B models, respectively, SAMuon uses $k = 39/50/71$, with `torch.svd_lowrank` parameters $q = 44/55/76$ (five oversampling dimensions) and $n_{\text{iter}} = 4/5/6$. At the same model scales, SAMuon-lite uses a rank-one power iteration with $10/12/14$ iterations.

Choice of the profile rank k. The plateau onset measured on the width-512 probing model is $k \approx 32$ (the left panel of Figure 6); the width rule $k(d_{\text{model}}) = \left\lfloor 32\sqrt{d_{\text{model}}/512} \right\rceil$ transfers this calibration across model widths, giving $k = 39/50/71$ at the trained widths $768/1280/2560$. We present the square-root rule as a calibrated heuristic rather than a derived law. However, we believe this heuristic should be reliable because the profile is mostly flat beyond the log-rank-linear band, so a moderately mis-set k mostly re-labels plateau ranks whose target scale is close to γ regardless. Under this rule SAMuon attains the best or joint-best final loss at every trained width (Section 6.2). A power-law dependence of the spectral structure of the Muon momentum matrices on model size has been independently documented by Magakyan et al. (2026) and LiMuon (Huang et al., 2026) provides a precedent for running a randomised SVD of the momentum buffer inside a Muon-style optimiser. Directly measuring the plateau onset at the two largest trained widths, rather than transferring it, is left for future work; the width-768 probing of Appendix E shows a modest shift of the approximate plateau onset in the direction predicted by the rule, although additional widths are needed to test the proposed scaling.

J.7 Wall-clock cost of the optimiser update step

To quantify the per-iteration overhead discussed in Section 5.2, we time one full training iteration of the 1B model under its real batch-size-1024 configuration on a single NVIDIA RTX 6000 Ada GPU; this benchmark serves the wall-clock evaluation only, while the training runs of Section 6.1 execute on A100 GPUs. The timing run also uses the 5×3 Newton–Schulz scheme. Per iteration, the GPU performs 64 microbatch forward+backward passes. Table 6 reports the GPU wall-clock breakdown. The forward+backward computation dominates at 91.2% of the Muon iteration, with the Newton–Schulz whitening accounting for the remaining 8.8%. On top of this, the rank-1 power iteration of SAMuon-lite adds 41.2 ms (0.5% of the Muon iteration), confirming its low wall-clock overhead relative to Muon. The rank-k `torch.svd_lowrank` call of SAMuon adds 657.8 ms (7.4%), nearly as much as the Newton–Schulz whitening itself despite its far lower FLOP count ($O(mnk)$ against $O(mnr)$, Section 5.2), which reflects the unoptimised kernel rather than the arithmetic cost of the method.

Table 6: **Per-GPU wall-clock cost of one training iteration** (1B model, batch-size-1024 configuration), measured on a single NVIDIA RTX 6000 Ada GPU. The full Muon iteration is the forward+backward computation plus the Newton–Schulz whitening (8916 ms in total), and percentages are relative to this total. The SAMuon and SAMuon-lite rows give the *additional* cost of each variant's head estimate on top of the Muon iteration.

Component	ms/iter	% of Muon iter
Forward + backward ($\times 64$ microbatches, compiled, bf16)	8133.8	91.2%
Muon: Newton–Schulz (5×3)	782.1	8.8%
SAMuon: + `torch.svd_lowrank`	657.8	7.4%
SAMuon-lite: + power iteration (bf16)	41.2	0.5%

K Measuring the token-efficiency improvement

For each model and batch-size setting, we train the tuned-Muon baseline for N iterations and use its final validation loss as the target. We then ask how short a SAMuon or SAMuon-lite run can be while still reaching that target. Every candidate run of length T uses a complete learning-rate schedule, including the final warmdown, compressed into T iterations. We therefore compare final losses from fully annealed runs, rather than comparing the loss curves at an intermediate iteration.

Table 7: **Token-efficiency improvement of SAMuon over tuned Muon at matched final validation loss.** N is the tuned-Muon schedule length, and $T^\star$ is the estimated SAMuon or SAMuon-lite schedule length that matches Muon's final loss, found using Algorithm 2 and checked with a confirmation run. SAMuon uses spectral radius 50; Muon uses its tuned radius for each batch size. The shaping warmup remains fixed at $0.30N$ absolute steps, making the reported improvements potentially conservative. The 1B runs use 20B tokens; the others use 10B.

				Token-effic. improv. $1 - T^\star/N$	
Model	Batch	Tokens	N	SAMuon	SAMuon-lite
124M (w768)	1024	10B	9600	20.3%	20.3%
	2048	10B	4800	23.8%	21.5%
	4096	10B	2400	24.0%	22.1%
300M (w1280)	1024	10B	9600	17.7%	15.9%
	2048	10B	4800	18.8%	17.1%
	4096	10B		*not trained*	
1B (w2560)	1024	20B	19200	13.3%	13.3%
	2048	20B		*not trained*	
	4096	20B	4800	15.6%	14.4%

If $T^\star$ is the estimated number of iterations needed to match the target, the token-efficiency improvement is

$$\text{improvement} = 1 - \frac{T^\star}{N}. \tag{43}$$

For example, matching the baseline after $0.8N$ iterations gives a 20% improvement.

K.1 Iteration matching

Longer fully annealed runs generally achieve lower final validation loss over the range considered. We therefore search for the run length at which the variant reaches the baseline target. We start with candidate lengths of $0.70N$ and $0.90N$, corresponding to improvements of 30% and 10%. The shorter run should finish above the target loss, whereas the longer run should match or improve upon it. If they do not bracket the target, we expand the search interval.

Within this interval, false-position interpolation uses the two endpoint losses to propose the next run length. We execute a complete, fully annealed run at that length and replace the shorter or longer endpoint according to whether the run misses or reaches the target. We repeat this process until the two endpoints differ by at most $0.005N$ iterations. Finally, we estimate $T^\star$ by linear interpolation between the runs nearest the target and execute a slightly longer confirmation run to check that the target is reached.

Final validation losses vary slightly between runs, partly because the optimiser uses a randomised low-rank SVD. The search therefore retains executed runs on both sides of the target rather than relying on a single interpolated estimate. The additional, slightly longer run checks that the target is reached near the interpolated matching length. The reported improvement remains based on the interpolated estimate.

The SAMuon shaping warmup is fixed at $0.30N$ absolute iterations and is not shortened with the candidate schedule. A matched run with $T < N$ therefore spends more than 30% of its iterations in warmup (between 34.6% and 39.5% for the settings in Table 7). This puts the shortened runs at a disadvantage, so the reported token-efficiency improvements are conservative.

L Per-Module Sign Alignment

The main-text probe of Section 4 applies each weight matrix's rank-d direction in the fixed orientation $-\boldsymbol{u}_{j,d}\boldsymbol{v}_{j,d}^\top$ that a real optimiser uses, and clusters by rank across weight matrices. The resulting profile therefore averages over an inter-block *sign* structure, *i.e.* how often a weight matrix's buffer

Algorithm 2 Iteration matching for the token-efficiency measurement.

Require: tuned-Muon schedule length N and final validation loss; SAMuon variant

Verify that a fully annealed N-iteration variant run reaches the baseline loss
Set the shorter and longer candidate lengths to $0.70N$ and $0.90N$
Train and evaluate fully annealed variant runs at both candidate lengths
Expand the interval if necessary until the shorter run misses the target and the longer run reaches it
while the candidate lengths differ by more than $0.005N$ **do**
    Choose the next length by false-position interpolation between the endpoint lengths and losses
    Train and evaluate a fully annealed variant run at the proposed length
    if the final loss is above the baseline target **then**
        Replace the shorter endpoint with this run
    else
        Replace the longer endpoint with this run
    end if
end while
Estimate $T^\star$ by linear interpolation between the runs nearest the target
Run a schedule half the final interval width longer than $T^\star$ and confirm that it reaches the target
return $1 - T^\star/N$

direction agrees with the out-of-sample gradient, which Figure 13 records. This agreement measures how trustworthy a direction is. A direction that agrees with the gradients of unseen batches makes transferable progress, whereas a direction that flips sign from batch to batch has a greater chance of disrupting training. Tail directions in the attention projections align poorly under this measure, which explains why the in-sample analysis of Appendix F is far more optimistic along the tail: in-sample, each direction is evaluated on the batch whose gradient formed the buffer, where its alignment is much better. The alignment also differs by weight-matrix type. The attention projections, and the query and key projections in particular, show the worst tail alignment, so their tail directions deserve the least trust; the QK-clip mechanism used in large-scale Muon training (Kimi Team, 2025) restrains exactly these matrices, so the probing framework provides results consistent with the training behaviour observed in practice. This cross-matrix agreement further determines whether contributions add or cancel within a rank, and is one source of the mostly flat bulk.

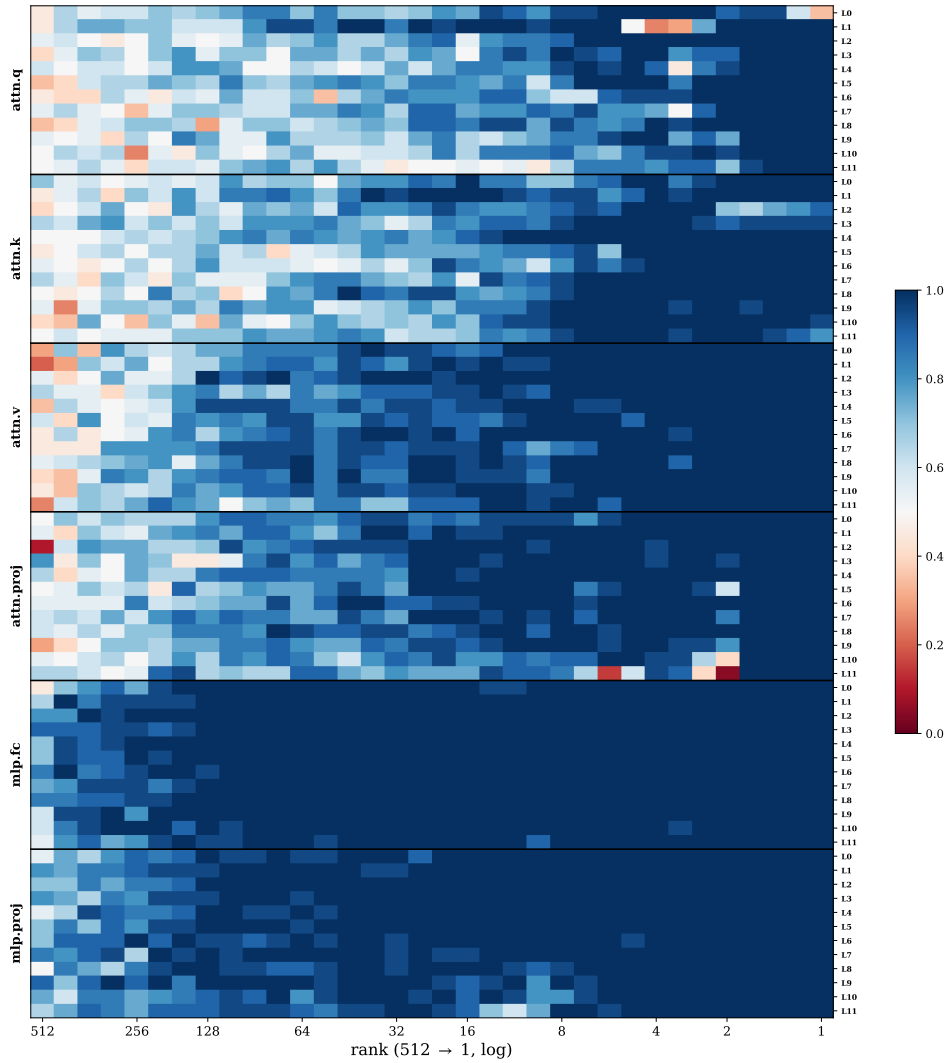

Figure 13: **Per-weight-matrix sign-alignment probability.** For the checkpoint at iteration 1000, the probability that a weight matrix's rank-d buffer direction $\boldsymbol{u}_{j,d}\boldsymbol{v}_{j,d}^{\top}$ is aligned (has positive projection) with the out-of-sample gradient, estimated over 20 held-out batches. Rows group the six weight types (attention query, key, value, and output projections; multilayer perceptron (MLP) up- and down-projections), each shown for all twelve layers L0–L11; columns run from the deep tail (rank 512) to the head (rank 1, log axis). Blue marks near-certain alignment, red near-certain anti-alignment, and white a coin-flip. Two patterns stand out. First, the top few ranks are almost always aligned (the deep-blue right edge across all weight matrices), so the head's orientation is quite deterministic. Second, in the tail the four attention projections drift towards 0.5 and below, *i.e.* their low-energy directions flip sign freely from batch to batch, whereas the MLP projections stay strongly aligned across almost the entire spectrum. The query and key projections show the worst tail alignment, so their tail directions are the least trustworthy; compare this observation with the QK-clip practice of Kimi Team (2025). This weight-matrix-dependent sign-flip structure is the inter-block variation that our per-rank clustering averages over (Appendix B); the cross-batch cancellation it produces in the tail is one source of the mostly flat bulk.