OPTIMIZING LARGE LANGUAGE MODELS WITH CHAINED LMOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon has motivated a growing family of optimizers that compose multiple matrix normalizations, but these methods remain fragmented and lack a unified perspective. We introduce *chained linear minimization oracles (chained LMOs)*, which cast these methods as compositions of LMOs. Despite their empirical success, many chains fall outside the standard LMO framework and can diverge on smooth convex objectives. To explain why composition can nevertheless help, we turn to linear associative memory and show that chaining can improve over Muon under anisotropic embeddings. Empirically, we propose *TensorChain*, a novel optimizer within the framework that stacks compatible weight matrices across different layers and normalizes the 3d tensor across its axes. In Qwen3 0.6B and 1.7B pretraining, TensorChain outperforms all chained baselines in average token efficiency, with average token savings of 9.6% over Muon at matched validation loss.

1 INTRODUCTION

After the success of the Muon optimizer (Jordan et al., 2024) in pretraining large language models, numerous theoretical and empirical follow-up works have been proposed. Consider minimizing a differentiable objective $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ over a weight matrix $W \in \mathbb{R}^{m\times n}$. Muon maintains a momentum matrix $M_t \in \mathbb{R}^{m\times n}$, the exponential moving average (EMA) of the gradient, and updates the weights along its matrix sign:

$$M_t = \beta M_{t-1} + (1-\beta)\nabla f(W_t), \quad W_{t+1} = W_t - \eta_t \,\mathrm{msign}(M_t). \tag{1}$$

Here, $\beta \in [0, 1)$ is the momentum coefficient and $\eta_t > 0$ is the learning rate. For a matrix M with singular value decomposition $M = U\Sigma V^\top$, the matrix sign is defined as $\mathrm{msign}(M) = UV^\top$; in practice, it is approximated by a few Newton–Schulz iterations rather than by an explicit SVD.

A principled way to understand this update is through the linear minimization oracle (LMO) (Pethick et al., 2025; Kovalev, 2025). For a norm ball $\mathcal{D} = \{X : \|X\| \le \rho\}$, define

$$\mathrm{lmo}_{\mathcal{D}}(G) \in \arg\min_{X\in\mathcal{D}} \langle G, X\rangle.$$

Over the spectral-norm ball, $\mathrm{lmo}_{\mathcal{D}}(G) = -\rho\ \mathrm{msign}(G)$, so Muon is exactly the iteration

$$M_t = \beta M_{t-1} + (1-\beta)\nabla f(W_t), \quad W_{t+1} = W_t + \eta_t \,\mathrm{lmo}_{\mathcal{D}}(M_t), \tag{2}$$

with $\rho = 1$. This iteration updates in the steepest descent direction under the geometry of $\mathcal{D}$, and inherits the standard convergence guarantees under the corresponding norm-smoothness assumptions (Pethick et al., 2025). Other norms recover other familiar normalizations: up to sign and scale, the max-row- and max-column-ℓ_2 balls yield row normalization (RN) and column normalization (CN), and the entrywise-ℓ_∞ ball yields the elementwise sign (Sgn).

Recent Muon variants augment the matrix-sign step with exactly these normalizations, applied before the matrix sign (Chang et al., 2026; Liu et al., 2026), after it (Zhang et al., 2026b; Li et al., 2025), on both sides (Si et al., 2025), or in repeated alternation (Dewulf et al., 2026). Although these methods were proposed separately, each with its own motivation and method-specific analysis, they share a common structure: each is a composition of LMOs.

We introduce **chained LMOs** as a unifying framework for these variants and evaluate them in a unified experimental protocol. Given constraint sets $\mathcal{D}_1, \ldots, \mathcal{D}_k$, we define the chained LMO

$$T(G) := (-1)^{k+1}\big(\mathrm{lmo}_{\mathcal{D}_k} \circ \cdots \circ \mathrm{lmo}_{\mathcal{D}_1}\big)(G), \tag{3}$$

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.

and the corresponding iteration

$$M_t = \beta M_{t-1} + (1-\beta)\nabla f(W_t), \quad W_{t+1} = W_t + \eta_t T(M_t). \tag{4}$$

The factor $(-1)^{k+1}$ accounts for the negation introduced by each LMO. Muon is the single-LMO case $k = 1$ with the spectral-norm ball, while many recent variants become instances of eq. (4), either directly or, for methods with Adam-like state, after turning off the EMA terms following Bernstein & Newhouse (2024) (Table 1).

Having identified chained LMOs as a common representation of Muon variants, a natural question is whether the favorable theory of individual LMOs survives composition. We show that for proposed Muon variants in the literature (Zhang et al., 2026b; Liu et al., 2026; Chang et al., 2026; Si et al., 2025; Li et al., 2025; Dewulf et al., 2026), it does not. This exposes the limitations of the existing LMO perspective and motivates a complementary view through linear associative memory, where we show that chaining can itself provide a learning advantage. Empirically, we compare these variants under a common protocol. Since each was originally evaluated in a different setup, their reported results are not directly comparable; a controlled comparison is therefore valuable in that it isolates the effect of the chain itself from differences in architecture, scale, and tuning. We further introduce a novel normalization axis that outperforms existing chained baselines under the same protocol, suggesting that the chained-LMO design space remains far from exhausted. Specifically,

- **A unified theory of chained LMOs.** We introduce chained LMOs as a common representation of recent Muon variants and characterize the limits of the standard LMO perspective. Many practical chains are not LMOs of any fixed constraint set (Section 2.1). For positively aligned chains such as $\mathrm{RN} \circ \mathrm{msign}$, a ν-approximate LMO framework recovers convergence guarantees, but with unavoidable dimension dependence (Proposition 4). Other chains admit negatively aligned directions and can diverge even on smooth convex objectives (Proposition 1).
- **A provable learning advantage from chaining.** Building on recent analyses of Muon in linear associative memory (Wang et al., 2026; Kim et al., 2026), we show that different normalizations can address distinct sources of anisotropy. Under anisotropic answer embeddings and power-law pair frequencies, $\mathrm{msign} \circ \mathrm{RN}$ can achieve asymptotically perfect retrieval while both Muon and gradient descent fail.
- **A new normalization axis and unified empirical comparison.** We introduce **TensorChain**, which treats the layer dimension of stacked parameter matrices as a novel axis of normalization. Under a common training setup (Section 5.1), TensorChain outperforms existing chained Muon baselines in the 1M-token batch regime, and remains competitive in the 4M-token batch regime.

The rest of the paper is organized as follows. In Section 2, we show that many chains in Table 1 are not LMOs of any fixed constraint set, and analyze their convergence and divergence on smooth objectives. Section 3 offers a complementary perspective through linear associative memory, where chaining provably improves learning. Section 4 shows our exploration of different chains and introduces TensorChain, which is compared with existing chained-LMO methods under a matched training protocol Section 5. Section 6 concludes the paper. Related optimization frameworks and their relation to chained LMOs are discussed in Appendix A.1.

2 Optimizing with chained LMOs

We work with the stateless chains in Table 1 (Appendix A.2) and write $N = N_k \circ \cdots \circ N_1$ for the corresponding chain of normalization maps; since each N_i is odd and invariant to positive rescaling, the chained LMO (equation 3) is $T = -N$ up to scale. As $-N(G) \in \arg\min_{X \in \mathcal{K}} \langle G, X \rangle$ exactly when $N(G) \in \arg\max_{X \in -\mathcal{K}} \langle G, X \rangle$, chained LMO is an LMO over some set if and only if N is a *linear **maximization** oracle* over some set $\mathcal{C}$, that is, $N(G) \in \arg\max_{X \in \mathcal{C}} \langle G, X \rangle$ for every G. For each chain, we therefore ask whether N is a linear maximization oracle over any set.

2.1 Chained LMOs are not necessarily LMOs

Appendix A.2 shows that many Muon variants optimize with chained LMOs. We now make precise the claim from the introduction that these chains generally leave the LMO class, so that the standard LMO theory does not apply to them. The claim has to be established chain by chain, since some

Table 1: **Existing and proposed Muon variants as chained LMOs.** Methods with Adam-like updates reduce to chains once the EMA is turned off. RN/CN: row/column normalization; msign: matrix sign; Sgn: elementwise sign; *Chain conv.*: whether the stateless chain drives the gradient norm to zero. †Also considers $\mathrm{RN} \circ \mathrm{CN} \circ \mathrm{msign}$ and $\mathrm{CN} \circ \mathrm{RN} \circ \mathrm{msign}$.

Method	LMO Chain	Chain conv.	Notes
Muon+ (Zhang et al., 2026b)	$\mathrm{RN}/\mathrm{CN} \circ \mathrm{msign}^{\dagger}$	✔	–
NorMuon (Li et al., 2025)	$\mathrm{RN} \circ \mathrm{msign}$	✔	EMA
MuonEq (Chang et al., 2026)	$\mathrm{msign} \circ \mathrm{RN}/\mathrm{CN}$	✔	–
Muon^2 (Liu et al., 2026)	$\mathrm{msign} \circ \mathrm{Sgn}$	✘	EMA
AdaMuon (Si et al., 2025)	$\mathrm{Sgn} \circ \mathrm{msign} \circ \mathrm{Sgn}$	✘	EMA
Aurora (Dewulf et al., 2026)	$(\mathrm{msign} \circ \mathrm{RN})^{\circ 2}$	✘	tall only; accumulated row scaling

compositions do stay inside the class: for instance, positive row scaling preserves entrywise signs, so $\mathrm{Sgn} \circ \mathrm{RN} = \mathrm{Sgn}$ and the chain $-(-\mathrm{Sgn}) \circ (-\mathrm{RN}) = -\mathrm{Sgn}$, which is indeed the LMO of the entrywise-ℓ_∞ ball. We consider three groups: the chains used in practice (Table 1), finite alternations of RN and CN, and their limit $\mathcal{E}$.

Theorem 1. *Let $m, n \geq 5$, and let normalizations map zero rows and columns to zero. There is no fixed set $\mathcal{K} \subseteq \mathbb{R}^{m \times n}$ that satisfies: for every G in its domain*

$$T(G) \in \arg\max_{X \in \mathcal{K}} \langle G, X \rangle$$

for each map T listed below:

- *(i)* $\mathrm{msign} \circ \mathrm{RN}$*,* $\mathrm{msign} \circ \mathrm{RN} \circ \mathrm{CN}$*,* $\mathrm{RN} \circ \mathrm{msign}$*,* $\mathrm{CN} \circ \mathrm{RN} \circ \mathrm{msign}$*, and* $\mathrm{RN} \circ \mathrm{CN} \circ \mathrm{msign}$*, for $m > n$.*
- *(ii) Alternation with the matrix sign:* $(\mathrm{msign} \circ \mathrm{RN})^{\circ k}$ *and* $(\mathrm{RN} \circ \mathrm{msign})^{\circ k}$*, for every $k \geq 1$ and $m > n$.*
- *(iii) Every finite composition of* RN *and* CN *that contains both, in particular* $(\mathrm{RN} \circ \mathrm{CN})^{\circ k}$*; its limit $\mathcal{E}$; and* $\mathrm{msign} \circ \mathcal{E}$*, the last two on matrices with no zero entries.*
- *(iv) Every finite composition of* msign *and* Sgn *that contains both.*

Hence the map $-T$ fails to be a LMO of any fixed set.

The statement concerns the exact matrix sign and the stateless chains, and it covers every chain in Table 1 up to transpose. Normalizing after msign can be vacuous: on tall inputs of full column rank, the columns of $\mathrm{msign}(G)$ are already orthonormal, so $\mathrm{CN} \circ \mathrm{msign} = \mathrm{msign}$, and symmetrically, on wide inputs of full row rank, $\mathrm{RN} \circ \mathrm{msign} = \mathrm{msign}$. The latter is why the restriction $m > n$ exists.

The proof relies on properties that a linear maximization oracle over a fixed set $\mathcal{C}$ must satisfy.

Lemma 1. *Let $N(G) \in \arg\max_{X \in \mathcal{C}} \langle G, X \rangle$ for every G, for a fixed set $\mathcal{C}$, and let $Y = N(G)$ for some G. If $\|N(Y)\|_F = \|Y\|_F$, then $N(Y) = Y$.*

Lemma 2. *Let $N(G) \in \arg\max_{X \in \mathcal{C}} \langle G, X \rangle$ for every G, for a fixed set $\mathcal{C}$. If $N(-G) = -N(G)$, then $\langle G, N(G) \rangle \geq 0$.*

For (i)–(ii) and the finite compositions in (iii), the normalizations keep undoing each other: on a suitable input, applying the chain a second time changes the output but not its norm, contradicting Lemma 1. For $\mathrm{msign} \circ \mathcal{E}$ and (iv), the first map is invariant to a family of transformations of its input: diagonal rescalings for $\mathcal{E}$, and changes of singular values or entry magnitudes for msign and Sgn. Transforming G within this family leaves the output unchanged but can make it negatively aligned with G; since these chains are odd, this contradicts Lemma 2. The limit $\mathcal{E}$ is idempotent, odd, and positively aligned, so both lemmas hold for it; we instead violate the LMO condition directly, by finding G and H with $\langle G, \mathcal{E}(G) \rangle < \langle G, \mathcal{E}(H) \rangle$.

2.2 Convergence behavior of chained LMOs

An LMO update always has a negative alignment with the gradient, which makes it a descent direction. When there exists a positive alignment, which is possible for chained LMOs, we can construct a counterexample where an optimization method diverges even for smooth functions (Proposition 1).

Proposition 1. *Let T be positively scale invariant with $\langle G, T(G)\rangle > 0$ for some G. For the convex, 1-smooth quadratic $f(X) = \langle G, X\rangle^2/(2\|G\|_F^2)$ with $X_0 = G/\|G\|_F^2$, the iterates $X_{t+1} = X_t + \eta_t T(\nabla f(X_t))$ satisfy $f(X_t) \to \infty$ for every step size sequence with $\sum_t \eta_t = \infty$.*

The counterexamples of Theorem 1 give such inputs for every composition of msign and Sgn that contains both, including the stateless chains of Muon[2] and AdaMuon, and for $\mathrm{msign} \circ \mathcal{E}$ in all sufficiently large shapes. The length-4 chain of Aurora, $(\mathrm{msign} \circ \mathrm{RN})^{\circ 2}$, also admits such input for $n = 2$ and $m \geq 258$ (Lemma B.11). None of these methods can therefore have a convergence guarantee even over smooth functions.

Approximate LMOs. For chains with negative alignment, we measure how far they are from an exact LMO. We formalize this notion using approximate LMOs.

Definition 1 (Approximate LMO)**.** *Let $\mathcal{K}$ be a compact set symmetric with respect to the origin, and define $m_{\mathcal{K}}(G) = \min_{X \in \mathcal{K}} \langle G, X\rangle$. For $\nu \in (0, 1]$, a map U is a* ν-approximate LMO *of $\mathcal{K}$ if $U(G) \in \mathcal{K}$ and $\langle G, U(G)\rangle \leq \nu\, m_{\mathcal{K}}(G)$ for every G.*

The feasibility condition $U(G) \in \mathcal{K}$ prevents the approximation factor from being increased trivially by rescaling U. It also implies $\nu \leq 1$, where $\nu = 1$ implies exact LMO. Using the same machinery from the LMO framework, we can prove convergence of approximate-LMOs.

Proposition 2. *Let f have an L-Lipschitz gradient, let $\Delta = f(X_0) - \inf f$, and let U be a ν-approximate LMO of $\mathcal{K}$. Define $R := \max_{X \in \mathcal{K}} \|X\|_F$.*

(i) Gradient descent. *With $X_{t+1} = X_t + \eta\, U(\nabla f(X_t))$ and $\eta = \sqrt{2\Delta/(LR^2T)}$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} h_{\mathcal{K}}(\nabla f(X_t)) \leq \frac{R}{\nu}\sqrt{\frac{2L\Delta}{T}}.$$

(ii) SGD with momentum. *With stochastic gradients of variance at most σ^2, $X_{t+1} = X_t + \eta\, U(M_t)$ with the momentum of Equation* (4)*, $1 - \beta = T^{-1/2}$ and $\eta = cT^{-3/4}$, the same average in expectation is $O\big(\nu^{-1}[\Delta/c + R(LRc + \sigma)]T^{-1/4}\big)$.*

Compared with an exact LMO, both rates lose exactly a factor $1/\nu$. Two chains from Table 1, $\mathrm{msign} \circ \mathrm{RN}$ and $\mathrm{RN} \circ \mathrm{msign}$, fall into the class of approximate-LMOs (Proposition 3).

Proposition 3. *On all $m \times n$ matrices, $-\mathrm{RN} \circ \mathrm{msign}$ is a $1/\sqrt{m}$-approximate LMO of $\mathcal{K}_{\mathrm{row}} = \{X : \max_i \|X_{i:}\|_2 \leq 1\}$, and $-\mathrm{msign} \circ \mathrm{RN}$ is a $1/\sqrt{m}$-approximate LMO of the spectral-norm ball. By transposition, the same holds for CN with $1/\sqrt{n}$.*

This gives convergence guarantees for the stateless versions of NorMuon, Muon+ with a single normalization and MuonEq, with additional $\sqrt{m}$ appearing in the upper bound. The additional dimension factor cannot be removed within this analysis: any set for which a map is feasible must contain the convex hull of its image, and a larger set only makes the approximation factor smaller.

Proposition 4. *Let $n = 2$. For every closed convex set $\mathcal{K}$ containing the image, the best approximation factor satisfies $\nu \leq 3(m-1)^{-1/4}$ for $-\mathrm{RN} \circ \mathrm{msign}$ (for odd m) and $\nu \leq 3/\sqrt{m-1}$ for $-\mathrm{msign} \circ \mathrm{RN}$.*

Both bounds tend to zero, so any guarantee derived from Proposition 2 degrades with dimension. A tighter analysis may need to go beyond uniform approximation, e.g. by bounding the approximation factor along the trajectory, which is out of scope for this work.

3 Provable Benefit of Chaining in Linear Associative Memory Learning

Section 2 shows that the standard worst-case guarantees of LMO-based optimization do not generally survive chaining: some chains can diverge even on smooth convex objectives, while others admit only dimension-dependent approximate-LMO guarantees. Any positive theory for chaining must therefore exploit additional problem structure beyond generic smoothness.

We instead demonstrate a benefit on the linear associative memory (Appendix A.3), following recent theoretical work on Muon (Wang et al., 2026; Kim et al., 2026; Li et al., 2026). Muon helps there

by whitening the anisotropy of the label frequencies, which suffices when the frequencies are the only source of anisotropy, as for the isotropic Gaussian embeddings of Kim et al. (2026). With a second source, such as anisotropic embeddings, Muon alone may not suffice, and chains such as $\mathrm{msign} \circ \mathrm{RN}$ can provably help. Anisotropic embeddings also make the model a closer abstraction of trained transformers, whose hidden representations are dominated by a few outlier coordinates (Timkey & Van Schijndel, 2021; Dettmers et al., 2022). We make this argument precise in the following sections.

We store N facts with queries $y_i \in \mathbb{R}^n$ and answers $v_i = \Sigma x_i \in \mathbb{R}^d$, where $y_i \sim \mathcal{N}(0, I_n/n)$ and $x_i \sim \mathcal{N}(0, I_d/d)$ are independent, and $\Sigma = \mathrm{diag}(\sigma) \in \mathbb{R}^{d\times d}$ has positive entries. Fact i occurs with probability $p_i = i^{-\alpha}/\sum_{r=1}^{N} r^{-\alpha}$ for a fixed $\alpha > 1$. A memory $W \in \mathbb{R}^{d\times n}$ is trained with the cross-entropy loss and evaluated by its retrieval accuracy,

$$\mathcal{L}(W) = \sum_{i=1}^{N} p_i \left(-v_i^\top W y_i + \log \sum_{j=1}^{N} e^{v_j^\top W y_i} \right), \quad \mathrm{Acc}(W) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left\{ v_i^\top W y_i > \max_{j\neq i} v_j^\top W y_i \right\}.$$

Following prior theoretical analysis of linear associative memory (Kim et al., 2026), we study the retrieval accuracy after a single update from $W_0 = 0$. Let $G = -\nabla\mathcal{L}(0)$ denote the negative gradient at initialization. We compare three methods, gradient descent, Muon, and row normalization followed by Muon:

$$W_{\mathrm{GD}} = G, \quad W_{\mathrm{Muon}} = \mathrm{msign}(G), \quad W_{\mathrm{chain}} = \mathrm{msign}(\mathrm{RN}(G)).$$

Step sizes are irrelevant because retrieval accuracy is invariant to positive rescaling of W. This chain is tailored to the anisotropy in our setting, and other forms of anisotropy may give provable advantages to other chains.

Asymptotic regime. Fix α, an infinite positive sequence $\sigma = (\sigma_j)_{j\geq 1}$, and dimension sequences N_n, d_n. For each n, the instance uses $N = N_n$ facts, query dimension n, answer dimension $d = d_n$, and the scales $\sigma_{1:d_n}$. For a method $\mathrm{M} \in \{\mathrm{GD}, \mathrm{Muon}, \mathrm{chain}\}$, write

$$\mathrm{Acc}_n^{\mathrm{M}} = \mathrm{Acc}\big(W_{\mathrm{M}};\, N_n, n, d_n, \sigma_{1:d_n}\big)$$

for the random accuracy of the instance. The quantity of interest is its limit in probability as $n \to \infty$.

Assumption 1 (Scaling)**.** *Fix* $0 < \varepsilon < (\beta - 1/2)/2$ *and* $\gamma > 1/(1-\beta)$. *Eventually,* $n^{1+(1-\beta+\varepsilon)/\alpha} \leq N_n \leq n^{1+(1/2-\varepsilon)/\alpha}$ *and* $n^{1/(1-\beta)} \leq d_n \leq n^{\gamma}$.

Assumption 2 (Anisotropy)**.** *For some fixed* $1/2 < \beta < 1$ *and constants* $0 < c_\sigma \leq C_\sigma < \infty$*, we have* $c_\sigma j^{-\beta} \leq \sigma_j \leq C_\sigma j^{-\beta}$ *for every* $j \geq 1$.

The scaling regime (Assumption 1) is roughly $n \ll N \ll d$: there are more facts than query dimensions, so retrieval requires extracting a weak fact-specific signal, while the large answer dimension provides many coordinates over which that signal can accumulate.

Assumption 2 ensures that $\sum_j \sigma_j^2$ remains bounded while $\sum_{j\leq d} \sigma_j$ diverges. Thus row normalization, which removes the answer-coordinate scales before the matrix sign, allows the chain to accumulate signal across coordinates in a regime where GD and Muon fail.

The core result is that the asymptotic one-step accuracy of chained Muon tends to 1, whereas those of Muon and GD go to 0.

Theorem 2 (One-step separation)**.** *Under Assumptions 1 and 2, as* $n \to \infty$*,*

$$\left(\mathrm{Acc}_n^{\mathrm{chain}}, \mathrm{Acc}_n^{\mathrm{Muon}}, \mathrm{Acc}_n^{\mathrm{GD}}\right) \xrightarrow{\mathbb{P}} (1, 0, 0).$$

The separation is driven by how the answer-coordinate scales enter the three updates. For GD and Muon, the relevant normalized score quantities depend on quadratic and quartic powers of σ_a; because $\sum_a \sigma_a^2 < \infty$ and $\sum_a \sigma_a^4 < \infty$, their normalized score vectors have only $O(n)$ total energy, while the correct-answer signal remains too weak to overcome the maximum of the $N \gg n$ incorrect scores. By contrast, row normalization removes the answer-coordinate scales before the matrix-sign step, since $\mathrm{RN}(\Sigma X K) = \mathrm{RN}(X K)$. The useful contributions can then add coherently in proportion to $\sum_a \sigma_a$, while the residual fluctuations are controlled by $\sqrt{\sum_a \sigma_a^2}$. Since $\sum_{a\leq d} \sigma_a \to \infty$ while the quadratic and quartic sums remain bounded, the chained update can accumulate signal across answer coordinates in a regime where GD and Muon cannot, yielding the separation above. The full proof is deferred to Appendix C.

4 EXPLORING THE SPACE OF CHAINS

Section 3 suggests that equalizing the update across coordinates is what makes chaining useful. We push this further by exploring three directions in the space of chained LMOs. First, we introduce **TensorChain**, which adds a new normalization axis across layers. Second, we explore **Aurora-LN**, a simple long-side generalization of Aurora that applies the same chain to both tall and wide matrices. Third, motivated by recent work replacing Newton–Schulz with simpler axis-wise normalization (Glentis et al., 2025; Deng et al., 2026), we study whether the spectral step itself can be replaced by iterated row–column normalization, which we refer to as **SinkhornChain**.

Algorithm 1 TensorChain

Require: Momenta $M_{1:g} \in (\mathbb{R}^{m\times n})^g$
$T \leftarrow \mathrm{Stack}(M_1, \dots, M_g) \in \mathbb{R}^{g\times m\times n}$
$A \leftarrow \mathrm{AN}_{\mathrm{layer}} \circ \mathrm{AN}_{\mathrm{col}} \circ \mathrm{AN}_{\mathrm{row}}$
$T \leftarrow A^{\circ k}(T)$
$T \leftarrow \mathrm{NS}_{\mathrm{slice}}(T)$
$T \leftarrow A^{\circ k}(T)$
return $\mathrm{Unstack}(T)$

Algorithm 2 Aurora-LN

Require: momentum $M \in \mathbb{R}^{m\times n}$
if $m \geq n$ **then**
  $U \leftarrow (\mathrm{NS} \circ \mathrm{RN} \circ \mathrm{NS} \circ \mathrm{RN})(M)$
else
  $U \leftarrow \left[(\mathrm{NS} \circ \mathrm{RN} \circ \mathrm{NS} \circ \mathrm{RN})(M^\top)\right]^\top$
end if
return U

4.1 TENSORCHAIN: NORMALIZING ACROSS LAYERS

Existing chains normalize only within individual weight matrices, through row, column, or spectral normalization, and treat each layer independently. TensorChain instead jointly normalizes parameters of the same shape across layers (Algorithm 1).

Parameters with the same shape and role, for example, the query projections W_Q of transformer blocks, are partitioned into groups. For a group of size g, we stack their momenta into a tensor $M \in \mathbb{R}^{g\times m\times n}$, whose first axis indexes layers. Normalizing fibers along this axis prevents individual layers from dominating the corresponding coordinates of the group's update.

Let $\mathrm{AN}_{\mathrm{layer}}, \mathrm{AN}_{\mathrm{row}}, \mathrm{AN}_{\mathrm{col}}$ denote ℓ_2-normalization along the corresponding tensor axes, define

$$\mathcal{A} := \mathrm{AN}_{\mathrm{layer}} \circ \mathrm{AN}_{\mathrm{col}} \circ \mathrm{AN}_{\mathrm{row}},$$

and let $\mathrm{NS}_{\mathrm{slice}}$ apply Newton–Schulz independently to each matrix slice. TensorChain maps

$$\mathrm{TensorChain}(M) = \mathcal{A}^{\circ k} \circ \mathrm{NS}_{\mathrm{slice}} \circ \mathcal{A}^{\circ k}(M),$$

whose slices give the per-layer updates.

TensorChain fits directly into the chained-LMO framework of Section 2 when we assume exact msign instead of Newton–Schulz. We will denote this as idealized TensorChain. Each fiber normalization is an exact LMO over a product of ℓ_2 balls, while slice-wise matrix sign is an exact LMO over the corresponding product of spectral-norm balls.

Proposition 5 (Idealized TensorChain as chained LMO)**.** *For $k \geq 1$, Idealized TensorChain is a composition of exact LMOs over product norm balls, but the resulting map is not itself an LMO of any fixed constraint set when $m, n \geq 5$. Moreover, Idealized TensorChain map admits a negative-alignment input and can diverge on a convex smooth objective under non-summable step sizes.*

A closely related approach is Teon (Zhang et al., 2026a), which also stacks matrices across layers but applies Newton–Schulz to a 2d matricization of the tensor. TensorChain instead applies Newton–Schulz independently to each matrix slice, with axis-wise normalization before and after. Empirical comparison with the Teon optimizer can be found in Appendix F.7.

We can consider the overhead of TensorChain in two axes: additional flops due to additional normalization, and additional communication cost to construct the tensor. Fiber normalizations cost $O(gmn)$ per group, negligible compared with the $O(gmn\min(m, n))$ Newton–Schulz steps, so TensorChain matches Muon's optimizer FLOPs up to lower-order terms. Under FSDP, layer-axis normalization couples matrices within a group and requires additional communication. Empirical performance of TensorChain can be found in section 5.2.

4.2 AURORA-LN: NORMALIZING ALONG THE LONG SIDE

Aurora (Dewulf et al., 2026) includes several design choices beyond the normalization chain itself: the Aurora update is applied only to tall matrices, while wide matrices fall back to Muon, and a β parameter accumulates the diagonal row preconditioner across steps. From the perspective of chaining, we consider a simplified variant that removes this accumulation and applies the same chain along the long side of every matrix. For tall and square matrices we apply the chain directly, while for wide matrices we apply it to the transpose and transpose the resulting update back. We refer to this variant as *Aurora-LN* (Algorithm 2, LN stands for Longside Normalization).

Long-side or wide-matrix variants of Aurora-style normalization have previously appeared in modded-nanoGPT experiments (Jordan et al., 2024), so we do not present Aurora-LN as a novel optimizer or algorithmic primitive. Rather, we revisit this design choice as a controlled simplification of Aurora. Our experiments in Section 5.2 show that this simplified variant can be more effective than the default Aurora implementation.

4.3 REPLACING NEWTON–SCHULZ WITH SINKHORN NORMALIZATION

A complementary direction is to ask whether chaining can eliminate the spectral step altogether. Motivated by recent work on inexpensive normalizations replacing Newton–Schulz (Deng et al., 2026; Glentis et al., 2025), we consider *SinkhornChain*, which alternates row and column rescaling as in Sinkhorn balancing, but uses ℓ_2 rather than ℓ_1 normalization. For an $m \times n$ matrix with $m > n$, each row–column normalization sweep costs $O(mn)$ FLOPs, whereas each Newton–Schulz step costs $O(mn^2)$. Thus, for a fixed number of iterations, SinkhornChain has substantially lower arithmetic complexity than Newton–Schulz.

However, none of our tested normalization variants or layerwise replacement configurations outperformed Muon; the strongest replaces NS only on Q, K, reaching validation loss 2.5945 versus Muon's 2.5905. Nevertheless, we still report these results, as in large-scale training the computational cost of Muon may be problematic and Sinkhorn shows a possibility of replacing it. Full details and ablations are given in Appendix E.

5 EXPERIMENTS

5.1 SETUP

Models and data. We train Qwen3 models (Yang et al., 2025) of sizes 0.6B and 1.7B on FineWeb (Penedo et al., 2024) with sequence length 4096, using pre-norm RMSNorm (Zhang & Sennrich, 2019), QK-norm, untied embeddings, and the default `torchtitan` initialization (Liang et al., 2025). We report loss on a held-out FineWeb split of 26M tokens.

Training. We evaluate token batch sizes of 1M and 4M. The 0.6B model uses 8-way FSDP, while the 1.7B model uses 4-way data parallelism and 8-way FSDP. Learning rate schedules of both cosine decay with 10% warmup and WSD results (10% warmup and 10% decay) were evaluated. Matrix parameters use the optimizer under comparison with Nesterov momentum, with each update U rescaled by $0.2\sqrt{mn}/\|U\|_F$ (Liu et al., 2025). Embeddings, output, and normalization parameters use AdamW with learning rate 8×10^{-3}, and all parameters use decoupled weight decay 0.1 (Loshchilov & Hutter, 2017). We use a fixed random seed.

Baselines and tuning. We compare the methods in Table 1 together with the two explorations in Section 4, TensorChain and Aurora-LN. We use the official implementation where available and our own otherwise, with the same five-step `bfloat16` Newton–Schulz for all methods. For TensorChain we grouped all the weights that have the same role as a same group, for simplicity. On 0.6B at $1\times$ Chinchilla (20 tokens per non-embedding parameter; Hoffmann et al. 2022), we sweep learning rates over $\{1, 2, 3, 5, 8\} \times 10^{-3}$ for WSD and $\{5, 8, 12, 16, 20\} \times 10^{-3}$ for cosine decay.

Across these sweeps, we observe that most baselines attain their best validation loss at the same learning rate within each schedule: 3×10^{-3} under WSD and 12×10^{-3} under cosine decay. This motivates a matched-learning-rate comparison in which the selected baselines are evaluated at a shared learning rate. AdaMuon is the only clear exception. Under WSD, it shares the same optimum and shows no clear advantage. However, under cosine decay, its optimum shifts to a

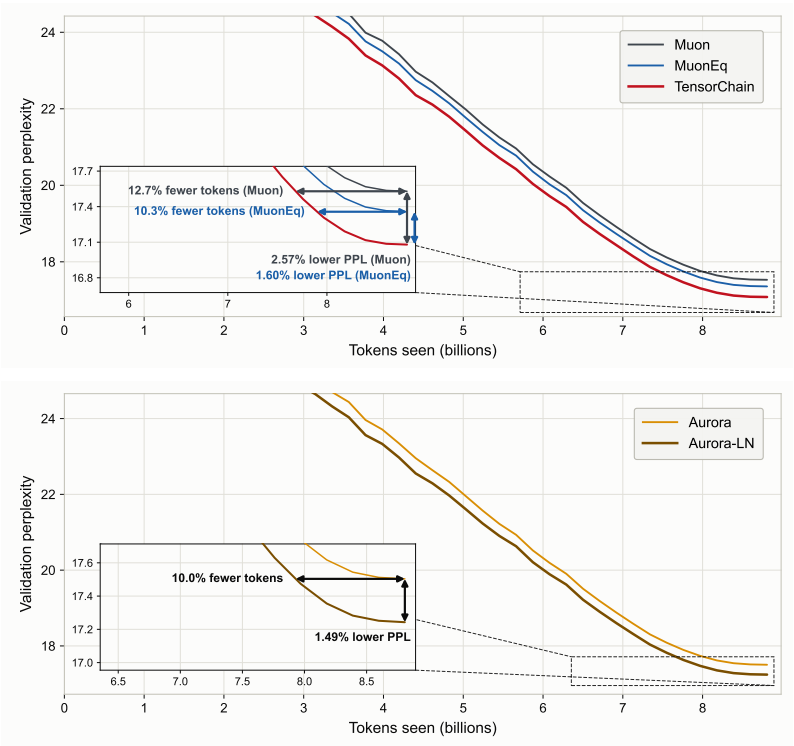


Figure 1: **Top:** Comparison of Muon, MuonEq and TensorChain; **Bottom:** Comparison of Aurora and Aurora-LN, both setups are training on Qwen3-0.6B with a 1M-token batch at $1\times$ Chinchilla under cosine LR. For other setups see section F.6.

substantially larger learning rate and its performance improves markedly, though it remains worse than TensorChain. We therefore exclude AdaMuon from the matched-learning-rate comparison, report its full sweep separately in Appendix F, and leave a systematic study to future work.

Among the remaining baselines, we select MuonEq, Aurora-LN, and Muon^2 as the best-performing representatives using one Newton–Schulz, two Newton–Schulz, and additional momentum-dependent state, respectively, while retaining Muon and Aurora as references. For $4\times$ Chinchilla, we rescale it by $4^{-0.32} \approx 0.64$ from the learning rate tuned at 1x Chinchilla (Bjorck et al., 2025).

5.2 Results

Table 2: Final validation loss under WSD and cosine schedules. Bold marks the best method in each row. The two methods on the right (Aurora, Aurora-LN) use Newton–Schulz twice, whereas other methods apply Newton–Schulz once. Muon^2 has an additional momentum state. (*) were introduced in Section 4.

Schedule	Model	Batch	Tokens	Muon	MuonEq	TensorChain*	Muon^2	Aurora	Aurora-LN*
WSD	0.6B	1M	8.8B	2.9201	2.9151	**2.9000**	2.9104	2.9184	2.9116
WSD	1.7B	4M	112B	2.5310	2.5279	**2.5236**	2.5264	2.5301	2.5242
Cosine	0.6B	1M	8.8B	2.8639	2.8540	**2.8379**	2.8497	2.8624	2.8474
Cosine	0.6B	4M	35B	2.7527	2.7434	**2.7344**	2.7443	2.7487	2.7391
Cosine	1.7B	1M	28.1B	2.5974	2.5924	**2.5857**	2.5885	2.5964	2.5882
Cosine	1.7B	4M	112B	2.4823	2.4784	2.4747	2.4744	2.4810	**2.4734**

Main comparison. TensorChain shows a consistent advantage over the more directly comparable chained baselines. Under WSD, it achieves the best final validation loss in both reported settings; under cosine decay, it is best in three of the four settings and outperforms Muon and MuonEq through-

out. The only cases in which TensorChain is surpassed are by methods that use either an additional Newton–Schulz application or additional momentum-dependent state. Thus, among variants with comparable computational and state complexity, TensorChain provides a consistent improvement across model scales, batch sizes, and learning-rate schedules.

The original Aurora performs only slightly better than Muon in our experiments, whereas the simplified Aurora-LN variant consistently improves over Aurora and becomes the best-performing method in the largest cosine setting. This suggests that seemingly minor choices in how a chain is applied across matrix shapes can have a substantial effect on performance. Figure 1 visualizes these two comparisons: TensorChain against Muon and MuonEq, and Aurora-LN against the original Aurora.

Table 3: Token-efficiency gain of TensorChain: percentage$(\%)$ reduction in tokens needed to reach a common validation-loss target, estimated by linear interpolation between adjacent checkpoints. Positive values favor TensorChain.

Model	Batch	Tokens	Aurora-LN	Muon^2	MuonEq	Aurora	Muon
0.6B	1M	8.8B	8.2	9.0	10.3	12.4	12.8
0.6B	4M	35B	5.4	7.6	7.4	8.9	10.0
1.7B	1M	28.1B	3.9	4.0	6.2	7.5	7.8
1.7B	4M	112B	-1.9	-0.5	6.7	7.4	7.8

The token-efficiency comparison in Table 3 makes the same trend more explicit. We report the percentage reduction in training tokens required to reach the same validation loss, estimated by linear interpolation between adjacent checkpoints; within each fixed-batch setting, this is equivalently the reduction in training steps. Across the four settings, TensorChain requires on average 9.6% fewer tokens than Muon and 7.7% fewer than MuonEq.

TensorChain grouping. We compare consecutive, stride, greedy load-balanced, and all-compatible-layer grouping (Table 4); implementation details are deferred to Section F.5. Consecutive grouping is consistently worse with validation loss gap of ≈ 0.007 for WSD and ≈ 0.01 for cosine, while stride, greedy, and all performed similarly compared to consecutive grouping. This suggests that mixing matrices across distant layers matters.

Table 4: TensorChain grouping ablation on Qwen3-0.6B, 1M-token batch, $1\times$ Chinchilla. Each entry reports the best validation loss from the corresponding full learning-rate sweep.

Schedule	Consecutive	Stride	Greedy	All
WSD	2.9070	**2.8990**	2.8994	2.9000
Cosine	2.8474	**2.8357**	2.8370	2.8379

6 Conclusion

We introduced the chained-LMO framework as a unified view of recent Muon variants. Although chaining generally falls outside the standard LMO framework and therefore may lose its usual smooth optimization guarantee, it can still provide substantial practical benefits. To close this gap, we analyzed linear associative memory with multiple sources of anisotropy and proved that chaining can yield a strict learning advantage. Empirically, we introduced TensorChain, which adds normalization across the layer axis, and showed that it can outperform existing chained variants in language-model pretraining.

An interesting direction for future work is to better understand how TensorChain's gains vary across model scales and training settings, including the particularly strong improvements observed at 0.6B. On the theory side, extending the associative-memory analysis to more general chains and forms of anisotropy would clarify when chaining should help. Finally, TensorChain naturally extends to additional tensor axes arising in multi-head attention and mixture-of-experts models, where understanding the potential gains is an interesting direction for future work.

AI USE STATEMENT

In this paper, AI tools were used in the following capacities:

- Assistance in proof verification and organization
- Assistance in novel proofs in Section 3; specifically on probabilistic rigor
- Coding assistants were used to help implement the proposed methods, run experimental pipelines, and retrieve and organize experimental results.

REFERENCES

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

Johan Bjorck, Alon Benhaim, Vishrav Chaudhary, Furu Wei, and Xia Song. Scaling optimal lr across token horizons. In *International Conference on Learning Representations*, volume 2025, pp. 83640–83657, 2025.

Da Chang, Qiankun Shi, Lvgang Zhang, Yu Li, Ruijie Zhang, Yao Lu, Yongxiang Liu, and Ganzhao Yuan. Muoneq: Balancing before orthogonalization with lightweight equilibration. *arXiv preprint arXiv:2603.28254*, 2026.

Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves a constrained optimization: As lyapunov predicts. In *International Conference on Learning Representations*, volume 2024, pp. 35404–35439, 2024.

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

Michael Crawshaw, Chirag Modi, Mingrui Liu, and Robert M Gower. An exploration of non-euclidean gradient descent: Muon and its many variants. *arXiv preprint arXiv:2510.09827*, 2025.

Shenyang Deng, Zhuoli Ouyang, Tianyu Pang, Zihang Liu, Ruochen Jin, Shuhua Yu, and Yaoqing Yang. Rmnp: Row-momentum normalized preconditioning for scalable matrix-based optimization. *arXiv preprint arXiv:2603.20527*, 2026.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35: 30318–30332, 2022.

Alec Dewulf, Dhruv Pai, Li Yang, Ashley Zhang, and Ben Keigwin. Aurora: A leverage-aware spectral optimizer. *arXiv preprint arXiv:2606.27715*, 2026.

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

Roger Fletcher. A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322, 1970.

Athanasios Glentis, Jiaxiang Li, Andi Han, and Mingyi Hong. A minimalist optimizer design for llm pretraining. In *ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models*, 2025.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

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, 2024. URL `https://github.com/KellerJordan/modded-nanogpt`.

Juno Kim, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D Lee. Sharp capacity scaling of spectral optimizers in learning associative memory. *arXiv preprint arXiv:2603.26554*, 2026.

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

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

Tim Tsz-Kit Lau and Weijie Su. Symmetry-compatible principle for optimizer design: Embeddings, lm heads, swiglu mlps, and moe routers. *arXiv preprint arXiv:2605.18106*, 2026.

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

Binghui Li, Kaifei Wang, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws. *arXiv preprint arXiv:2602.05725*, 2026.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025.

Wanchao Liang, Tianyu Liu, Less Wright, Will Constable, Andrew Gu, Chien-Chin Huang, Iris Zhang, Wei Feng, Howard Huang, Junjie Wang, et al. Torchtitan: One-stop pytorch native solution for production ready llm pretraining. In *International Conference on Learning Representations*, volume 2025, pp. 92161–92181, 2025.

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

Ziyue Liu, Ruijie Zhang, Zhengyang Wang, Yequan Zhao, Yupeng Su, Zi Yang, and Zheng Zhang. Muon2: Boosting muon via adaptive second-moment preconditioning. *arXiv preprint arXiv:2604.09967*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Bohao Ma, Junyu Zhang, and Chuan He. Schattor: Schatten-family methods for deep learning optimization. *arXiv preprint arXiv:2606.15702*, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022a.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022b.

Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*, 2023.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Thomas Pethick. When to use what schatten-p norm in deep learning? *arXiv preprint arXiv:2606.15268*, 2026.

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

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026.

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

William Timkey and Marten Van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, 2021.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Andrey Veprikov, Arman Bolatov, Aleksandr Bogdanov, Samuel Horváth, Aleksandr Beznosikov, Martin Takáč, and Slavomir Hanzely. Preconditioned norms: A unified framework for steepest descent, quasi-newton and adaptive methods. *arXiv preprint arXiv:2510.10777*, 2025.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Tan. Muon outperforms adam in tail-end associative memory learning. In *International Conference on Learning Representations*, volume 2026, pp. 37382–37419, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Ruijie Zhang, Yequan Zhao, Ziyue Liu, Zhengyang Wang, Dongyang Li, Yupeng Su, Sijia Liu, and Zheng Zhang. Teon: Tensorized orthonormalization beyond layer-wise muon for large language model pre-training. *arXiv preprint arXiv:2601.23261*, 2026a.

Ruijie Zhang, Yequan Zhao, Ziyue Liu, Zhengyang Wang, Yupeng Su, Liyan Tan, and Zheng Zhang. Muon+: Towards more effective muon via one additional normalization step for llm pre-training. *arXiv preprint arXiv:2602.21545*, 2026b.

Shu Zhong, Mingyu Xu, Tenglong Ao, and Guang Shi. Understanding transformer from the perspective of associative memory. *arXiv preprint arXiv:2505.19488*, 2025.

A Appendix: Related work

A.1 Frameworks around Muon and variants

Bernstein & Newhouse (2024) take a descent step that minimizes the local quadratic model

$$\operatorname{argmin}_{\Delta W} \langle \nabla f(W), \Delta W \rangle + \tfrac{\lambda}{2} \|\Delta W\|^2$$

with the norm $\|\cdot\|$ fixed in advance. The solution splits into a step size and a direction, $-(\|\nabla f(W)\|^\dagger/\lambda) \cdot \arg\max_{\|X\|=1} \langle \nabla f(W), X \rangle$. Taking $\|\cdot\|$ to be the induced operator norm $\|W\|_{\alpha\to\beta} = \max_x \|Wx\|_\beta / \|x\|_\alpha$ gives a family of optimizers including popular adaptive algorithms. Switching off their EMAs, Adam (Kingma & Ba, 2014), Shampoo (Gupta et al., 2018) and Prodigy (Mishchenko & Defazio, 2023) are each steepest descent under a particular member: $(\alpha, \beta) = (\ell_1, \ell_\infty)$ for Adam and Prodigy, which differ only in their step-size rule, and (ℓ_2, ℓ_2) for Shampoo, which is a positive scaling of the Muon update when the EMA is switched off. Pethick et al. (2025) write the same direction as a *linear minimization oracle*

$$\operatorname{lmo}_{\mathcal{B}}(\nabla f(W)) = \operatorname{argmin}_{X \in \mathcal{B}} \langle \nabla f(W), X \rangle$$

over a norm ball, and interpret existing normalized gradient methods as optimizing with LMOs. By understanding the induced norm as how the output is amplified in a linear layer under bounded input, Pethick et al. (2025) propose the Scion optimizer, whose design principle is to bound the RMS norm of intermediate layers in neural network training. Kovalev (2025) interprets the constraint set $\mathcal{B}$ as a trust region, and both works prove convergence guarantees for their respective classes of algorithms.

These frameworks cast optimizer choice as an appropriate choice of norm or constraint set, and subsequent work varies that choice along several axes. Lau et al. (2025) write vector preconditioning and matrix-aware updates as the same subproblem under different norms: the time-dependent quadratic metric gives the variable-metric methods that date back to Fletcher (1970) and are made data-driven in AdaGrad (Duchi et al., 2011), while the spectral norm gives $\|G\|_{\mathrm{nuc}}\ \mathrm{msign}(G)$, the solution of Bernstein & Newhouse (2024). Crawshaw et al. (2025) aggregate per-layer norms into a product norm, whose oracle decomposes into scaled per-layer oracles; this expresses the standard practice of running Muon and Adam on separate parameter classes. Others vary the geometry itself: the Schatten-p family interpolates between the Frobenius and spectral cases (Ma et al., 2026; Pethick, 2026), and preconditioned matrix norms widen the family further (Veprikov et al., 2025).

Certain frameworks deviate from the norm-choice template. Lion-$\mathcal{K}$ (Chen et al., 2024) replaces the sign map by the subgradient of a convex $\mathcal{K}$, and shows that the iterates converge to KKT points of $\min_X F(X)$ subject to $\mathcal{K}^*(X) \le 1/\lambda$ for decoupled weight decay λ. Muon is the special case $\mathcal{K} = \|\cdot\|_{\mathrm{tr}}$ (Chen et al., 2025). Shumaylov et al. (2026) instead exhibit updates of the form $(GG^\top)^{-c}G$ whose best-performing exponents fall in the quasi-norm regime $c > 1/2$, where the update is provably not the LMO of any unitarily invariant norm ball; they establish convergence directly, through alignment and a descent potential rather than through the oracle.

These frameworks share one structural feature: a single geometry determines the update map, and the same geometry supplies the analysis. An LMO over a norm ball is aligned with the negative gradient, $\langle G, \operatorname{lmo}_{\mathcal{B}}(G) \rangle = -\rho\|G\|_*$, which yields a descent lemma under the matching norm-smoothness assumption; Lion-$\mathcal{K}$ replaces this with a Lyapunov function built from $\mathcal{K}$. Chained LMOs do not in general inherit either structure, because the LMO class is not closed under composition. Even when every constituent map is an LMO, many practical chains are not the LMO of any fixed constraint set (Section 2), nor in general the subgradient of a convex $\mathcal{K}$, so neither the oracle-based nor the Lyapunov-based argument transfers. We therefore analyze chains directly, in the spirit of Shumaylov et al. (2026). For positively aligned chains, an α-approximate LMO framework recovers convergence, at the cost of an unavoidable dimension dependence. Chains that admit negative-alignment directions, in contrast, can diverge even on smooth convex objectives (Section 2.2).

Closest to our setting is Lau & Su (2026), whose contribution is a design principle: update maps derived from the symmetry group of each weight block, with NorMuon, Muon+, MuonEq and Aurora (Li et al., 2025; Zhang et al., 2026b; Chang et al., 2026; Dewulf et al., 2026) placed within that perspective. We instead study the effect of chaining itself: how composed updates relate to existing frameworks, whether they converge, and how they can improve linear associative memory learning. Their construction composes row-norm and one-sided spectral maps only, so the elementwise chains of Si et al. (2025) and Liu et al. (2026) fall outside it; our framework places no such restrictions.

A.2 Existing Muon variants as chained LMOs

Recent Muon variants add normalization at different stages of the orthogonalization pipeline. Post-orthogonalization methods correct the row- or column-wise imbalance left after the matrix-sign step. NorMuon (Li et al., 2025) does so with a row-wise second-moment state, while Muon+ (Zhang et al., 2026b) applies stateless row, column, column–row, or row–column normalization directly after msign. Pre-orthogonalization methods instead precondition the momentum before Newton–Schulz to improve the spectrum entering the orthogonalization step: MuonEq (Chang et al., 2026) uses row/column equilibration, and Muon^2 (Liu et al., 2026) uses Adam-style second-moment preconditioning. AdaMuon (Si et al., 2025) acts on both sides, applying an elementwise sign before orthogonalization and a second-moment normalization afterward, which combines coordinate-wise adaptivity with orthogonal updates. Aurora (Dewulf et al., 2026) alternates row normalization and the matrix sign on tall matrices through the chain $(\mathrm{msign}\circ\mathrm{RN})^{\circ 2}$, so that the update is approximately balanced in both its spectrum and its row norms.

Some of these optimizers are stateful because of exponential moving averages. Bernstein & Newhouse (2024) recover norm-based steepest-descent methods from Adam (Kingma & Ba, 2014), Shampoo (Gupta et al., 2018), and Prodigy (Mishchenko & Defazio, 2023) by switching off their EMAs, and we use the same reduction to expose the underlying normalization maps. For example, Adam with $\beta_1 = \beta_2 = 0$ reduces to Sgn; similarly, NorMuon reduces to $\mathrm{RN}\circ\mathrm{msign}$, Muon^2 to $\mathrm{msign}\circ\mathrm{Sgn}$, and AdaMuon to $\mathrm{Sgn}\circ\mathrm{msign}\circ\mathrm{Sgn}$. We use this reduction only to isolate the normalization chain, not to identify the full stateful optimizer with its stateless counterpart.

A.3 Linear Associative Memory and Muon

An associative memory stores key–value pairs $\{(k_i, v_i)\}_{i=1}^N$, with $k_i \in \mathbb{R}^{d_k}$ and $v_i \in \mathbb{R}^{d_v}$, in a matrix $W \in \mathbb{R}^{d_v\times d_k}$ so that presenting k_i recalls v_i; for orthonormal keys, $W = \sum_i v_i k_i^\top$ does so exactly. Attention can be read this way (Zhong et al., 2025): without the softmax normalization it computes $o_t = \sum_{i\le t} \exp(q_t^\top k_i/\sqrt{d_k})\, v_i = S_t\phi(q_t)$ with $S_t = \sum_{i\le t} v_i\phi(k_i)^\top$, where ϕ is the infinite-dimensional feature map of the exponential kernel, and FFN layers are similar through ReLU kernels. From an interpretability perspective, factual associations in a transformer can be edited by a rank-one update to an MLP projection (Meng et al., 2022a;b).

A growing line of work relates the success of Muon to associative memory. Wang et al. (2026) find empirically that Muon's advantage over Adam comes mainly from the value/output and FFN weights, precisely the parameters that act as associative memories. To study this theoretically, Wang et al. (2026); Kim et al. (2026); Li et al. (2026) turn to the linear associative memory, where W retrieves pair i correctly when $v_i^\top W k_i > v_j^\top W k_i$ for all $j \neq i$, and is trained with the cross-entropy loss $\mathcal{L}$ on these logits, with pair i occurring with frequency p_i. When the frequencies are imbalanced, Muon provably retrieves better than gradient descent, because it *whitens* the frequencies. At initialization, $-\nabla\mathcal{L}(0) \approx \sum_i p_i\, v_i k_i^\top$, so gradient descent learns each pair at a rate proportional to p_i. With orthonormal embeddings, this sum is a singular value decomposition with singular values p_i, and $\mathrm{msign}\left(\sum_i p_i\, v_i k_i^\top\right) = \sum_i v_i k_i^\top$ gives every pair the same weight. Wang et al. (2026); Li et al. (2026) analyze orthonormal embeddings, and Kim et al. (2026) extend the result to Gaussian embeddings.

B APPENDIX: PROOFS IN SECTION 2

Sign convention. Throughout the paper, LMO means *linear minimization oracle*:

$$\mathrm{lmo}_K(G) \in \arg\min_{X\in K}\langle G, X\rangle.$$

We write $N = N_k \circ \cdots \circ N_1$ for a chain of normalization maps and $T = -N$ for the signed update used in minimization. For the norm balls and odd oracle selections considered here, $\mathrm{lmo}_{\mathcal{D}_j} = -N_j$, so

$$T = (-1)^{k+1}\big(\mathrm{lmo}_{\mathcal{D}_k} \circ \cdots \circ \mathrm{lmo}_{\mathcal{D}_1}\big) = -N, \qquad X_{t+1} = X_t + \eta_t T(M_t) = X_t - \eta_t N(M_t).$$

For any fixed set C, with $-C := \{-X : X \in C\}$,

$$N(G) \in \arg\max_{Z\in C}\langle G, Z\rangle \quad \Longleftrightarrow \quad -N(G) \in \arg\min_{X\in -C}\langle G, X\rangle.$$

Thus the non-LMO proofs below test whether N is a *linear maximization oracle*; their conclusions concern the signed update $T = -N$ as a linear minimization oracle. The reflected set is the same set only when $C = -C$, as for the norm balls used here.

We write $\mathrm{msign}(G) = UV^\top$ for the canonical polar factor of $G = U\Sigma V^\top$, with zero singular directions mapped to zero, and Sgn for the entrywise sign with $\mathrm{Sgn}(0) = 0$. Row and column normalization are denoted by RN and CN, respectively. The proofs use two necessary conditions for a linear maximization oracle.

B.1 NON-LMO RESULTS

Lemma B.1 (Lemma 1 of the paper)**.** *Let $N(G) \in \arg\max_{X\in K}\langle G, X\rangle$ for every G, for a fixed set K, and let $Y = N(G)$. If $\|N(Y)\|_F = \|Y\|_F$, then $N(Y) = Y$.*

Proof. Since every oracle output belongs to K, $Y \in K$. Optimality at the input Y gives

$$\langle Y, N(Y)\rangle \geq \langle Y, Y\rangle = \|Y\|_F^2.$$

By Cauchy–Schwarz and the assumed equality of norms,

$$\langle Y, N(Y)\rangle \leq \|Y\|_F\|N(Y)\|_F = \|Y\|_F^2.$$

Hence equality holds in Cauchy–Schwarz. The equal norms imply $N(Y) = Y$, including the case $Y = 0$. □

Lemma B.2 (Lemma 2 of the paper)**.** *If $N(G) \in \arg\max_{X\in K}\langle G, X\rangle$ for every G, for a fixed set K, and $N(-G) = -N(G)$, then*

$$\langle G, N(G)\rangle \geq 0.$$

Proof. Because $N(-G) = -N(G)$ is itself an output of the oracle, it belongs to K. Therefore optimality at G implies $\langle G, N(G)\rangle \geq \langle G, -N(G)\rangle = -\langle G, N(G)\rangle$. Thus $2\langle G, N(G)\rangle \geq 0$. □

Consequently, to rule out a linear maximization-oracle representation for N, it suffices to find G such that $N(N(G)) \neq N(G)$ but $\|N(N(G))\|_F = \|N(G)\|_F$, or, for an odd map, to find G such that $\langle G, N(G)\rangle < 0$. By the sign convention above, either argument rules out an LMO representation for the signed update $-N$.

Lemma B.3 (Tall matrix-sign and row-normalization witnesses)**.** *Assume $m > n$. Each normalization map N listed in parts (i) and (ii) of Theorem 1 has an input G for which $N(N(G)) \neq N(G)$ and $\|N(N(G))\|_F = \|N(G)\|_F$.*

Proof. Assume throughout this subsection that $m > n$, and put $r = m - n \geq 1$. For $0 < t < 1$, define $A(t) \in \mathbb{R}^{m\times n}$ by the rows

$$A(t)_{1,:} = (\sqrt{1-t}, \sqrt{t}, 0, \ldots, 0),$$
$$A(t)_{2,:} = (\sqrt{1-t}, -\sqrt{t}, 0, \ldots, 0),$$

followed by r copies of $e_2^\top$, and then $e_3^\top, \dots, e_n^\top$. Thus $A(t)$ has exactly $2 + r + (n-2) = m$ rows. Every row has Euclidean norm one, so $\mathrm{RN}(A(t)) = A(t)$. Moreover, $A(t)^\top A(t) = \mathrm{diag}\big(2(1-t),\, r+2t,\, 1, \dots, 1\big)$.

Define $Q(t) := \mathrm{msign}(A(t)) = A(t)\big(A(t)^\top A(t)\big)^{-1/2}$. Since the columns of $A(t)$ are orthogonal, column normalization produces exactly the same matrix: $\mathrm{CN}(A(t)) = Q(t)$. Since $Q(t)$ has orthonormal columns, $\mathrm{msign}(Q(t)) = Q(t), \mathrm{CN}(Q(t)) = Q(t)$. The only nontrivial transition is row normalization of $Q(t)$. For either of its first two rows, the squared second coordinate after normalization is $\frac{t/(r+2t)}{\frac{1}{2}+t/(r+2t)} = \frac{2t}{r+4t}$. Hence, with $\phi(t) := \frac{2t}{r+4t}$, we have the exact identities

$$\begin{aligned} &\mathrm{RN}(A(t)) = A(t), \\ &\mathrm{CN}(A(t)) = \mathrm{msign}(A(t)) = Q(t), \\ &\mathrm{CN}(Q(t)) = \mathrm{msign}(Q(t)) = Q(t), \\ &\mathrm{RN}(Q(t)) = A(\phi(t)). \end{aligned} \tag{5}$$

The map ϕ is strictly increasing: $\phi'(t) = \frac{2r}{(r+4t)^2} > 0$. At $t_0 = \frac{1}{2}$, we have $\phi(t_0) = \frac{1}{r+2} < \frac{1}{2} = t_0$. Strict monotonicity therefore gives $t_0 > \phi(t_0) > \phi^2(t_0) > \cdots$ for every finite number of iterations.

We now apply this one-dimensional dynamics to the maps in parts (i)–(ii) of Theorem 1.

For part (i), equation 5 gives the following transitions. The map $N = \mathrm{msign} \circ \mathrm{RN}$ sends $A(t)$ to $Q(t)$, and a second application sends it to $Q(\phi(t))$. The map $\mathrm{RN} \circ \mathrm{msign}$ sends $A(t)$ to $A(\phi(t))$, then to $A(\phi^2(t))$. The maps $\mathrm{msign} \circ \mathrm{RN} \circ \mathrm{CN}$ and $\mathrm{CN} \circ \mathrm{RN} \circ \mathrm{msign}$ both send $A(t)$ to $Q(\phi(t))$, then to $Q(\phi^2(t))$. Finally, $\mathrm{RN} \circ \mathrm{CN} \circ \mathrm{msign}$ has the same transitions as $\mathrm{RN} \circ \mathrm{msign}$ on this family, because $\mathrm{CN}(Q(t)) = Q(t)$. All five maps are therefore nonidempotent at $t = t_0$. Along these witness orbits, outputs have Frobenius norm $\sqrt{n}$ when the final operation is msign or CN, and $\sqrt{m}$ when it is RN. Lemma B.1 excludes a fixed-set linear maximization-oracle representation for each normalization map, and hence an LMO representation for its negative.

For part (ii), let $F = \mathrm{msign} \circ \mathrm{RN}$. Then $F(A(t)) = Q(t)$ and $F(Q(t)) = Q(\phi(t))$, so for every $k \geq 1$, $F^k(A(t)) = Q(\phi^{k-1}(t))$. Applying the same map F^k once more replaces $\phi^{k-1}(t)$ by $\phi^{2k-1}(t)$, which is distinct at $t = t_0$.

For the reverse composition $H = \mathrm{RN} \circ \mathrm{msign}$, the same argument gives the distinct iterates $A(\phi^k(t_0))$ and $A(\phi^{2k}(t_0))$. Again the outputs along these witness orbits have constant Frobenius norm, so Lemma B.1 proves part (ii). □

Lemma B.4 (Finite mixed row/column-normalization words)**.** *Every finite composition of* RN *and* CN *that contains both is not a linear maximization oracle of a fixed set* $\mathcal{K}$*.*

Proof. Repeated adjacent copies of the same normalization may be removed because $\mathrm{RN}^2 = \mathrm{RN}, \mathrm{CN}^2 = \mathrm{CN}$. Thus it suffices to consider an alternating sequence of RN and CN.

Put $r = m-1, s = n-1$, and fix any constant $a > 1$. For $t > 0$, define the positive row-normalized matrix $R(t)$ by its squared entries:

$$R(t)_{ij}^2 = \begin{cases} \dfrac{at}{at+s}, & i = 1,\ j = 1, \\ \dfrac{1}{at+s}, & i = 1,\ j \geq 2, \\ \dfrac{t}{t+s}, & i \geq 2,\ j = 1, \\ \dfrac{1}{t+s}, & i \geq 2,\ j \geq 2. \end{cases}$$

Let $C_1(t)$ denote the squared norm of the first column and $C_2(t)$ the squared norm of any other column: $C_1(t) = \frac{at}{at+s} + r\frac{t}{t+s}$, and $C_2(t) = \frac{1}{at+s} + r\frac{1}{t+s}$. Column normalization multiplies the squared first-to-other coordinate ratio in every row by $C_2(t)/C_1(t)$. Row normalization preserves these within-row ratios. Therefore $\mathrm{RN}(\mathrm{CN}(R(t))) = R(F(t))$, where $F(t) = t\frac{C_2(t)}{C_1(t)} = \frac{(1+ra)t+ms}{am\,t+s(a+r)}$.

This scalar map is strictly increasing. Indeed, $F'(t) = \frac{rs(a-1)^2}{\big(am\,t+s(a+r)\big)^2} > 0$. The fixed-point equation $F(t) = t$ is a quadratic with positive leading coefficient and negative constant term: $am\,t^2 + \big[s(a+r) - (1+ra)\big]t - ms = 0$. It therefore has exactly one positive root, denoted $t_\star$.

Moreover, $F(1)-1 = \frac{n(1-a)}{am+s(a+r)} < 0$. Thus $1 > t_\star$. Since F is increasing, $t_\star < F(t) < t\,(t > t_\star)$. Consequently $1 > F(1) > F^2(1) > \cdots > t_\star$, and no two finite iterates coincide.

Suppose first that a reduced chain of RN and CN ends in RN. Starting from $R(t)$, every effective CN-then-RN pair applies one copy of F; any initial RN is inert. Hence for some integer $k \geq 1$, $N(R(t)) = R(F^k(t))$. Applying the same word again gives $N(N(R(t))) = R(F^{2k}(t))$. At $t = 1$ these are distinct.

If instead the word ends in CN, transpose the entire construction. Transposition exchanges RN and CN, and $\langle G, X\rangle = \langle G^\top, X^\top\rangle$. Thus a linear maximization-oracle representation for the original word would induce one for the transposed RN-ending word, which we just ruled out.

Finally, on the positive witness family, an RN-ending word has output norm $\sqrt{m}$, while a CN-ending word has output norm $\sqrt{n}$. Lemma B.1 therefore proves that every finite composition containing both RN and CN is not a linear maximization oracle of any fixed set. □

The limiting row/column equilibration map $\mathcal{E}$ uses the compatible normalization $\|\mathcal{E}(G)_{i,:}\|_2^2 = n/m$ and $\|\mathcal{E}(G)_{:,j}\|_2^2 = 1$, so every equilibrated matrix has squared Frobenius norm n. The following lemma identifies the balanced representative inside a positive diagonal orbit.

Lemma B.5 (Uniqueness in a positive diagonal orbit)**.** *Let B, C be strictly positive matrices in the same positive diagonal orbit: $C = D_1 B D_2, D_1, D_2 \succ 0$. If B and C have the same squared row sums and the same squared column sums, then $B = C$.*

Proof. Write $S_{ij} = B_{ij}^2, U_{ij} = C_{ij}^2$. For suitable scalars α_i, β_j, $\log \frac{U_{ij}}{S_{ij}} = \alpha_i + \beta_j$. Therefore

$$\sum_{ij}(U_{ij} - S_{ij}) \log \frac{U_{ij}}{S_{ij}} = \sum_i \alpha_i \sum_j (U_{ij} - S_{ij}) + \sum_j \beta_j \sum_i (U_{ij} - S_{ij}) = 0,$$

because the row and column marginals agree. But each summand satisfies $(u - s)\log(u/s) \geq 0$ for $u, s > 0$, with equality only when $u = s$. Hence $U = S$, and positivity gives $B = C$. □

Lemma B.6 (The limiting equilibration map is not a linear maximization oracle)**.** *The limiting equilibration map E in Theorem 1 - (iii) is not a linear maximization oracle of any fixed set.*

Proof. If B is already balanced and $L, R \succ 0$ are diagonal, then

$$\mathcal{E}(LBR) = B. \tag{6}$$

We now construct a positive balanced matrix explicitly. First suppose $m \geq n$. Set $b = \frac{1}{\sqrt{2m}}, a = \sqrt{\frac{n+1}{2m}}, c = \frac{1}{\sqrt{m}}$. Let the first n rows form the $n \times n$ matrix $(a-b)I_n + b\mathbf{1}\mathbf{1}^\top$, and let each of the remaining $m - n$ rows equal $c\mathbf{1}^\top$. Call the resulting matrix B.

For a top row, $a^2 + (n-1)b^2 = \frac{n+1}{2m} + \frac{n-1}{2m} = \frac{n}{m}$. For a lower row, $nc^2 = \frac{n}{m}$. Each column therefore has squared norm $\frac{n}{m} + (m-n)c^2 = \frac{n}{m} + \frac{m-n}{m} = 1$. Thus B is balanced.

When $m < n$, use $B_{m,n} = \sqrt{\frac{n}{m}}\, B_{n,m}^\top$, which has squared row norms n/m and squared column norms one.

We prove that E is not a linear maximization oracle by a local perturbation inside the balanced manifold. Let $L = \mathrm{diag}(2, 1, \ldots, 1), R = \mathrm{diag}(2, 1, \ldots, 1)$, and define $G = LBR$. By equation 6, $E(G) = B$. Let Z be supported on the leading 2×2 block with $Z_{11} = Z_{22} = 1$, $Z_{12} = Z_{21} = -1$. Its row sums and column sums are zero. For sufficiently small t, define $B(t)$ entrywise by $B(t)_{ij} =$

$\sqrt{B_{ij}^2 + tZ_{ij}}$. All entries remain positive. Since Z has zero row and column sums, $B(t)$ has exactly the same squared row and column norms as B, so $B(t)$ is balanced.

Hence $E(B(t)) = B(t)$, so $B(t)$ is itself an output of E. On the other hand, $\frac{d}{dt}\langle G, B(t)\rangle\big|_{t=0} = \frac{1}{2}\sum_{ij} L_{ii}R_{jj}Z_{ij}$. Only the first 2×2 block contributes, giving $\frac{d}{dt}\langle G, B(t)\rangle\big|_{t=0} = \frac{1}{2}(2-1)(2-1) = \frac{1}{2} > 0$. Thus for all sufficiently small $t > 0$, $\langle G, B(t)\rangle > \langle G, B\rangle = \langle G, E(G)\rangle$. If E were a linear maximization oracle of a fixed set K, then every output $B(t)$ would belong to K, contradicting optimality of $E(G) = B$ at the input G.

Therefore the limiting equilibration E is not a linear maximization oracle of any fixed set. □

Now we find explicit counterexamples of Lemma 2 for $\mathrm{msign} \circ \mathcal{E}$ and the iteration of sign and msign, proving these chains are not linear maximization oracles.

Lemma B.7 (Negative alignment after equilibration)**.** *For every shape covered by Theorem 1, the odd map* $\mathrm{msign} \circ \mathcal{E}$ *admits an input* G *such that*

$$\langle G, (\mathrm{msign} \circ \mathcal{E})(G)\rangle < 0.$$

Hence $\mathrm{msign} \circ \mathcal{E}$ *is not a linear maximization oracle of any fixed set.*

Proof. Write $N = \mathrm{msign} \circ \mathcal{E}$. Both constituents are odd, so N is odd. By Lemma B.2, it is therefore enough to find G with $\langle G, \mathrm{msign}(\mathcal{E}(G))\rangle < 0$.

First suppose $m > n$, and use the balanced positive matrix B from the previous subsection. Put $r = m - n$, $s = a + (n-1)b$, $\lambda = s^2 + nrc^2$. The leading $n \times n$ block of B is $(a-b)I_n + bJ_n$. A direct calculation gives

$$B^\top B = (a-b)^2\left(I - \frac{1}{n}J_n\right) + \lambda\,\frac{1}{n}J_n.$$

Hence, writing $Q = \mathrm{msign}(B)$, the leading $n \times n$ block of Q is

$$Q_{\mathrm{top}} = I - \frac{1}{n}J_n + \frac{s}{\sqrt{\lambda}}\frac{1}{n}J_n. \tag{7}$$

Since $r > 0$, $\lambda > s^2$, and therefore every off-diagonal entry of the block equation 7 is strictly negative: $(Q_{\mathrm{top}})_{ij} = \frac{1}{n}\left(\frac{s}{\sqrt{\lambda}} - 1\right) < 0$, $i \neq j$. Choose, say, the entry $(1,2)$, and set $\delta = -B_{12}Q_{12} > 0$. For a scalar $M \geq 1$, let $L_M = \mathrm{diag}(M, 1, \ldots, 1)$, $R_M = \mathrm{diag}(1, M, 1, \ldots, 1)$, and define $G_M = L_M B R_M$. By diagonal invariance, $E(G_M) = B$, so $N(G_M) = Q$. The selected $(1,2)$ contribution to $\langle G_M, Q\rangle$ is $-\delta M^2$. Every other entry receives weight at most M. Moreover $\sum_{ij}|B_{ij}Q_{ij}| \leq \|B\|_F\|Q\|_F = n$, because B has squared Frobenius norm n and Q has n orthonormal columns. Therefore $\langle G_M, Q\rangle \leq -\delta M^2 + nM$. Taking $M > \frac{n}{\delta}$ gives $\langle G_M, N(G_M)\rangle < 0$. For $m < n$, transpose the corresponding $n \times m$ witness. Compatible equilibration obeys the transpose relation up to the positive scalar $\sqrt{n/m}$, and the polar factor removes this scalar. Thus the same negative inner product is obtained.

Now suppose $m = n = q \geq 3$. Consider first the 3×3 positive circulant matrix

$$B_3 = \frac{1}{\sqrt{14}}\begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{pmatrix}.$$

Every row and every column has norm one. Furthermore, $B_3^\top B_3 = \frac{3}{14}I_3 + \frac{11}{14}J_3$, and direct evaluation of the polar factor gives

$$Q_3 := \mathrm{msign}(B_3) = \frac{1}{3}\begin{pmatrix} 1-\sqrt{3} & 1 & 1+\sqrt{3} \\ 1+\sqrt{3} & 1-\sqrt{3} & 1 \\ 1 & 1+\sqrt{3} & 1-\sqrt{3} \end{pmatrix}.$$

In particular, $(Q_3)_{11} = \frac{1-\sqrt{3}}{3} < 0$. For $q > 3$, let $B_0 = B_3 \oplus I_{q-3}$. This matrix is balanced and invertible, but has zero entries. To obtain a strictly positive witness, define the squared-entry matrix $S_\varepsilon = (1-\varepsilon)(B_0 \circ B_0) + \varepsilon \frac{J_q}{q}$, where $\circ$ denotes the Hadamard product, and set $(B_\varepsilon)_{ij} = \sqrt{(S_\varepsilon)_{ij}}$. Every row and column of S_ε sums to one, so B_ε is balanced. It is strictly positive. Since $B_\varepsilon \to B_0$ and B_0 is invertible, B_ε is invertible for all sufficiently small $\varepsilon > 0$. Continuity of the polar factor at invertible matrices gives $\mathrm{msign}(B_\varepsilon) \longrightarrow Q_3 \oplus I_{q-3}$. Hence, for sufficiently small fixed ε, $(B_\varepsilon)_{11} > 0$, $[\mathrm{msign}(B_\varepsilon)]_{11} < 0$. We can therefore repeat the preceding diagonal-amplification argument, now scaling the first row and first column. For sufficiently large M, $G_M = L_M B_\varepsilon L_M$ satisfies $E(G_M) = B_\varepsilon$ and $\langle G_M, \mathrm{msign}(B_\varepsilon)\rangle < 0$. Thus $\mathrm{msign} \circ E$ has a negative-alignment witness in every shape $m, n \geq 5$. Since it is odd, Lemma B.2 proves that it is not a linear maximization oracle of any fixed set. □

Lemma B.8 (Mixed matrix-sign and entrywise-sign words)**.** *Every finite composition of* msign *and* Sgn *that contains both admits an input G with negative alignment,*

$$\langle G, N(G)\rangle < 0.$$

Hence no such composition is a linear maximization oracle of any fixed set.

Proof. Write $P = \mathrm{msign}$ and $S = \mathrm{Sgn}$. Since both maps are idempotent, every finite chain containing both reduces to an alternating chain of length at least two. Both maps are odd, so by Lemma B.2 it suffices to construct a negative-alignment witness for every alternating chain.

We first consider square matrices of size $q \geq 5$. Suppose the first applied operation is S. Put $k = q-1$, $J = \mathbf{1}\mathbf{1}^\top$, $B = J - 2I$, and $A = B + 2e_1e_1^\top$. Thus A has $+1$ in its $(1,1)$ entry, -1 on the remaining diagonal, and $+1$ off the diagonal.

Let $Q = P(A)$. To determine its signs, let $u = k^{-1/2}(0, 1, \ldots, 1)^\top$ and decompose $\mathbb{R}^q = \mathrm{span}\{e_1, u\} \oplus \{e_1, u\}^\perp$. On the orthogonal complement, $A = -2I$, while on $\mathrm{span}\{e_1, u\}$ it is represented by

$$H_A = \begin{pmatrix} 1 & \sqrt{k} \\ \sqrt{k} & k-2 \end{pmatrix}.$$

Since $\det H_A = -2$, its eigenvalues have opposite signs. Setting $D = \sqrt{k^2 - 2k + 9}$, its matrix sign is

$$\mathrm{msign}(H_A) = \frac{1}{D}\begin{pmatrix} 3-k & 2\sqrt{k} \\ 2\sqrt{k} & k-3 \end{pmatrix}.$$

It follows that Q has negative diagonal entries and positive off-diagonal entries for $k \geq 4$. Hence $S(Q) = B$.

The eigenvalues of B are $q-2$ along $\mathbf{1}$ and -2 on $\mathbf{1}^\perp$, so $P(B) = 2J/q - I =: R$. For $q \geq 5$, R again has negative diagonal and positive off-diagonal entries, hence $S(R) = B$. Thus every alternating chain beginning with S outputs Q, B, or R.

Now take $G_t = B + (t+1)e_1e_1^\top$. Since $S(G_t) = A$, the possible chain outputs remain Q, B, R, independently of t. Their $(1,1)$ entries are all strictly negative. Therefore $\langle G_t, X\rangle = \langle B, X\rangle + (t+1)X_{11} \to -\infty$ for every $X \in \{Q, B, R\}$. Taking t sufficiently large gives a negative-alignment witness for every chain beginning with S.

Now suppose the first applied operation is P. Let $w = (8q, 8q, 1, \ldots, 1)^\top$, let $d_w = \|w\|_2^2$, and define the Householder reflection $Q = I - 2ww^\top/d_w$. Its diagonal entries are positive and its off-diagonal entries are negative, so $S(Q) = 2I - J =: B$. Since B has eigenvalue $2-q$ along $\mathbf{1}$ and eigenvalue 2 on $\mathbf{1}^\perp$, we have $P(B) = I - 2J/q =: R$, and again $S(R) = B$. Thus every such alternating chain eventually alternates between B and R.

It remains to choose an input with polar factor Q and negative alignment with both B and R. Define $z = (I - ww^\top/d_w)\mathbf{1}$. Then $z \perp w$, so $Qz = z$. Moreover, if $r = \|z\|_2^2$, then

$$r = q - \frac{(17q-2)^2}{128q^2+q-2} > \frac{q}{2}$$

for $q \geq 5$. Consequently $z^\top Bz = 2r - r^2 < 0$ and $z^\top Rz = r - 2r^2/q < 0$.

For $\gamma > 0$, set $G_\gamma = Q(I + \gamma zz^\top)$. The second factor is positive definite, so $P(G_\gamma) = Q$. Since $Qz = z$, we also have $G_\gamma = Q + \gamma zz^\top$. Therefore both $\langle G_\gamma, B\rangle$ and $\langle G_\gamma, R\rangle$ tend to $-\infty$ as $\gamma \to \infty$. Taking γ sufficiently large gives a negative-alignment witness for every chain beginning with P.

Finally, for rectangular matrices, take $q = \min\{m, n\}$ and zero-pad the corresponding $q \times q$ witness. Both P and S commute with zero padding, and Frobenius inner products are unchanged. Lemma B.2 therefore rules out a linear maximization-oracle representation for every finite composition containing both msign and Sgn. □

Theorem B.9 (Theorem 1 of the paper)**.** *Let $m, n \geq 5$, and let normalizations map zero rows and columns to zero. For each map $N : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ listed below, there is no fixed set $K \subseteq \mathbb{R}^{m\times n}$ such that*

$$N(G) \in \arg\max_{X\in K} \langle G, X\rangle$$

for every G:

- *(i)* $\mathrm{msign}\circ\mathrm{RN}$*,* $\mathrm{msign}\circ\mathrm{RN}\circ\mathrm{CN}$*,* $\mathrm{RN}\circ\mathrm{msign}$*,* $\mathrm{CN}\circ\mathrm{RN}\circ\mathrm{msign}$*, and* $\mathrm{RN}\circ\mathrm{CN}\circ\mathrm{msign}$ *for $m > n$.*
- *(ii) Alternation with matrix sign:* $(\mathrm{msign}\circ\mathrm{RN})^{\circ k}$ *and* $(\mathrm{RN}\circ\mathrm{msign})^{\circ k}$ *for every $k \geq 1$ and $m > n$.*
- *(iii) Every finite composition of* RN *and* CN *that contains both, in particular* $(\mathrm{RN}\circ\mathrm{CN})^{\circ k}$*; its limit $\mathcal{E}$; and* $\mathrm{msign}\circ\mathcal{E}$*, with the last two considered on matrices with no zero entries.*
- *(iv) Every finite composition of* msign *and* Sgn *that contains both.*

Equivalently, for every listed normalization map N, the signed update $-N$ is not an LMO of any fixed set.

Proof. Parts (i) and (ii) follow from Lemma B.3 and Lemma B.1. The finite-word claim in part (iii) follows from Lemma B.4 and Lemma B.1; the limit E is excluded by Lemma B.6, and $\mathrm{msign}\circ E$ is excluded by Lemmas B.7 and B.2. Part (iv) follows from Lemmas B.8 and B.2. □

B.2 Convergence and non-convergence guarantees

Proposition B.10 (Proposition 1 of the paper)**.** *Let T be positively scale invariant with $\langle G, T(G)\rangle > 0$ for some G. For the convex, 1-smooth quadratic*

$$f(X) = \frac{\langle G, X\rangle^2}{2\|G\|_F^2}$$

with $X_0 = G/\|G\|_F^2$, the iterates $X_{t+1} = X_t + \eta_t T(\nabla f(X_t))$ satisfy $f(X_t) \to \infty$ for every nonnegative step-size sequence with $\sum_t \eta_t = \infty$.

Proof. Let $a := \langle G, T(G)\rangle > 0$. The Hessian of f is the rank-one orthogonal projector $H[Z] = G\langle G, Z\rangle/\|G\|_F^2$, so f is convex, nonnegative, and 1-smooth. Put $c_t = \langle G, X_t\rangle$. Since $c_0 = 1$, whenever $c_t > 0$, $\nabla f(X_t) = (c_t/\|G\|_F^2)G$ is a positive multiple of G. Positive scale invariance gives $T(\nabla f(X_t)) = T(G)$. Thus $c_{t+1} = c_t + \eta_t a$, and induction yields

$$c_t = 1 + a\sum_{j<t}\eta_j > 0, \qquad X_t = X_0 + \left(\sum_{j<t}\eta_j\right) T(G).$$

Consequently,

$$f(X_t) = \frac{\left(1 + a\sum_{j<t}\eta_j\right)^2}{2\|G\|_F^2} \longrightarrow \infty.$$

□

The witnesses in Lemmas B.8 and B.7 have $\langle G, N(G)\rangle < 0$ for the *normalization map* N. Their signed updates $T = -N$ therefore satisfy $\langle G, T(G)\rangle > 0$, as required by Proposition B.10. Equivalently, the stateless iteration $X_{t+1} = X_t - \eta_t N(\nabla f(X_t))$ diverges. The Aurora claim follows by the same sign conversion from the next lemma.

Lemma B.11 (Negative alignment for two Aurora cycles)**.** *If $n = 2$ and $m \geq 258$, then $(\mathrm{msign} \circ \mathrm{RN})^{\circ 2}$ has a full-column-rank input G with nonzero entries such that*

$$\left\langle G, (\mathrm{msign} \circ \mathrm{RN})^{\circ 2}(G)\right\rangle < 0.$$

Proof. Let $m = r + 2$ with $r \geq 256$, put $s = \sqrt{5 + 1/r}$, and define

$$q = \frac{1}{s}\left(\underbrace{r^{-1}, \ldots, r^{-1}}_{r \text{ entries}}, -2, 1\right)^{\top}, \quad Q = \begin{bmatrix} m^{-1/2}\mathbf{1} & q \end{bmatrix}.$$

Then $\mathbf{1}^T q = 0$ and $\|q\| = 1$, so $Q^T Q = I_2$. Write $c = \sqrt{m}/s$ and $a = c/r$. The rows of $\mathrm{RN}(Q)$ have slopes $a, -2c, c$, with the first slope repeated r times. Let $C = \mathrm{RN}(Q)^T \mathrm{RN}(Q)$. Direct summation gives $C_{22} < 3$, $C_{11} \leq m$, and

$$cC_{12} \geq \frac{c^2}{1 + a^2} - \frac{1}{2} = \frac{(r+2)r^2}{5r^2 + 2r + 2} - \frac{1}{2} \geq \frac{r}{6} - \frac{1}{2}.$$

For $h = \sqrt{\det C}$, the inverse square root of a positive definite 2×2 matrix is a positive scalar multiple of $\left(\begin{smallmatrix} C_{22}+h & -C_{12} \\ -C_{12} & C_{11}+h \end{smallmatrix}\right)$. Hence the first coordinate of the last row of $V := \mathrm{msign}(\mathrm{RN}(Q))$ has the sign of $C_{22} + h - cC_{12}$. Since $h \leq \sqrt{3m}$ and, for $r \geq 256$, $r/6 - 1/2 > 3 + \sqrt{3(r+2)}$, this coordinate is strictly negative.

Now let $A_\varepsilon = \mathrm{RN}([\mathbf{1}, \varepsilon q])$. A QR expansion gives $\mathrm{msign}(A_\varepsilon) \to Q$ as $\varepsilon \downarrow 0$: the second QR diagonal is $\varepsilon + O(\varepsilon^3)$, the off-diagonal term is $O(\varepsilon^3)$, and the polar factor of the resulting upper-triangular 2×2 matrix tends to the identity. Therefore

$$(\mathrm{msign} \circ \mathrm{RN})^2([\mathbf{1}, \varepsilon q]) \longrightarrow V.$$

For sufficiently small fixed $\varepsilon > 0$, the inner product between the last row of A_ε and the last row of this output is negative. Multiplying only the last input row by a sufficiently large positive scalar leaves the first row-normalization, and hence the chained output, unchanged, while making the total alignment negative. The input remains full column rank with nonzero entries. This proves the claimed negative-alignment example for $(\mathrm{msign} \circ \mathrm{RN})^{\circ 2}$ when $n = 2$ and $m \geq 258$. □

Definition B.12 (Definition 1 of the paper)**.** *Let K be a compact set containing the origin and define*

$$m_K(G) := \min_{X \in K} \langle G, X\rangle.$$

*For $\nu \in (0, 1]$, a map U is a ν-*approximate LMO *of K if, for every G,*

$$U(G) \in K, \qquad \langle G, U(G)\rangle \leq \nu m_K(G).$$

Write $h_K(G) := \max_{X \in K}\langle G, X\rangle$. The nonnegative minimization gap is $-m_K(G) = h_K(-G) = h_{-K}(G)$. For an origin-symmetric set $K = -K$, this equals $h_K(G)$, the notation used for the norm-ball bounds in the main text.

Proposition B.13 (Proposition 2 of the paper)**.** *Let f have an L-Lipschitz gradient, let $\Delta = f(X_0) - \inf f < \infty$, and let U be a ν-approximate LMO of a compact set K containing the origin. Write $R := \max_{X \in K} \|X\|_F$.*

(i) Gradient descent. With $X_{t+1} = X_t + \eta U(\nabla f(X_t))$ and

$$\eta = \sqrt{\frac{2\Delta}{LR^2T}}, \qquad \frac{1}{T}\sum_{t=0}^{T-1}[-m_K(\nabla f(X_t))] \leq \frac{R}{\nu}\sqrt{\frac{2L\Delta}{T}}.$$

(ii) SGD with momentum. With unbiased stochastic gradients of conditional variance at most σ^2,

$$M_t = \beta M_{t-1} + (1-\beta)\widehat{g}_t, \qquad X_{t+1} = X_t + \eta U(M_t),$$

$1-\beta = T^{-1/2}$, $\eta = cT^{-3/4}$, and fixed initialization with finite initial tracking error, the same average in expectation is

$$O\Big(\nu^{-1}\big[\Delta/c + R(LRc+\sigma)\big]T^{-1/4}\Big).$$

For $K = -K$, both statements can equivalently be written with $h_K(\nabla f(X_t))$ in place of $-m_K(\nabla f(X_t))$.

Proof. Feasibility implies $\|U(v)\|_F \le R$. The minimum-value function is R-Lipschitz: $|m_K(g) - m_K(v)| \le R\|g-v\|_F$. The robust descent-alignment inequality is

$$\langle g, U(v)\rangle \le \nu m_K(g) + (1+\nu)R\|g-v\|_F. \tag{8}$$

Indeed,

$$\langle g, U(v)\rangle = \langle v, U(v)\rangle + \langle g-v, U(v)\rangle \le \nu m_K(v) + R\|g-v\|_F,$$

and the Lipschitz bound for m_K gives equation 8.

For part (i), let $g_t = \nabla f(X_t)$. Smoothness and approximate optimality give

$$f(X_{t+1}) \le f(X_t) + \nu\eta m_K(g_t) + \frac{LR^2\eta^2}{2}.$$

Summing and using $f(X_0) - f(X_T) \le \Delta$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1}[-m_K(g_t)] \le \frac{\Delta}{\nu\eta T} + \frac{LR^2\eta}{2\nu}.$$

The stated choice of η balances the two terms and proves part (i). The degenerate cases $R = 0$ or $\Delta = 0$ have zero minimization gap along the corresponding constant iterates and can be treated separately.

For part (ii), write $\widehat{g}_t = g_t + \zeta_t$, where $\mathbb{E}[\zeta_t \mid \mathcal{F}_t] = 0$ and $\mathbb{E}[\|\zeta_t\|_F^2 \mid \mathcal{F}_t] \le \sigma^2$. Let $e_t = M_t - g_t$. Applying equation 8 with $v = M_t$ and then smoothness gives

$$\nu\eta[-m_K(g_t)] \le f(X_t) - f(X_{t+1}) + (1+\nu)R\eta\|e_t\|_F + \frac{LR^2\eta^2}{2}.$$

Thus it remains to bound $T^{-1}\sum_t \mathbb{E}\|e_t\|_F$. For $t \ge 1$, the tracking recursion is

$$e_t = \beta e_{t-1} + \beta(g_{t-1} - g_t) + (1-\beta)\zeta_t.$$

Since $\|X_t - X_{t-1}\|_F \le R\eta$, smoothness gives $\|g_t - g_{t-1}\|_F \le LR\eta$. Unrolling the recursion, bounding the drift by the triangle inequality, and using martingale orthogonality for the noise yields

$$\mathbb{E}\|e_t\|_F \le \beta^{t+1}E_0 + \frac{\beta}{1-\beta}LR\eta + \sigma\sqrt{\frac{1-\beta}{1+\beta}}, \tag{9}$$

where $E_0 := \mathbb{E}\|M_{-1} - g_0\|_F < \infty$. Averaging and summing the descent inequality gives

$$\frac{1}{T}\sum_{t<T}\mathbb{E}[-m_K(g_t)] \le \frac{1}{\nu}\left[\frac{\Delta}{\eta T} + CR\left(\frac{LR\eta}{1-\beta} + \sigma\sqrt{1-\beta} + \frac{E_0}{(1-\beta)T}\right) + \frac{LR^2\eta}{2}\right]$$

for a universal constant C. Setting $1-\beta = T^{-1/2}$ and $\eta = cT^{-3/4}$ gives

$$\frac{1}{T}\sum_{t<T}\mathbb{E}[-m_K(g_t)] = O\Big(\nu^{-1}[\Delta/c + R(LRc+\sigma)]T^{-1/4}\Big).$$

For fixed initialization, the initial-error and direct-smoothness terms have lower orders $T^{-1/2}$ and $T^{-3/4}$, respectively. If $r\mathbb{B}_F \subseteq K$ for some $r > 0$, then $-m_K(g) \ge r\|g\|_F$, so these bounds also control the gradient norm. Here $\mathbb{B}_F$ is the Frobenius unit ball. The proof uses momentum before applying U; an EMA of a statistic internal to U is not covered. □

Proposition B.14 (proposition 3 of the paper). *Let* $K_{\mathrm{row}} := \{X : \max_i \|X_{i,:}\|_2 \le 1\}$. *Then* $-(\mathrm{RN}\circ \mathrm{msign})$ *is a* $1/\sqrt{m}$*-approximate LMO of* K_{row}*, while* $-(\mathrm{msign}\circ\mathrm{RN})$ *is a* $1/\sqrt{m}$*-approximate LMO of the spectral-norm unit ball. By transposition, the same holds for* CN *with* $1/\sqrt{n}$.

Proof. We prove the two row-normalized statements; the column-normalized versions follow by transposition. The zero input is immediate, so assume $G \neq 0$. First let $Q = \mathrm{msign}(G)$ and write the polar factorization $G = QH$, where $H = (G^TG)^{1/2} \succeq 0$. Let D row-normalize the nonzero rows of Q, and set $D_{ii} = 1$ on zero rows, so that $\mathrm{RN}(Q) = DQ$. Since QQ^T is an orthogonal projection, every row of Q has norm at most one and hence $D \succeq I$. Also $HQ^TQ = H$. Therefore

$$\langle G, \mathrm{RN}(Q)\rangle = \mathrm{tr}(HQ^TDQ) = \mathrm{tr}\, H + \mathrm{tr}\big(HQ^T(D-I)Q\big) \ge \mathrm{tr}\, H = \|G\|_*.$$

The support function of $K_{\mathrm{row}} = \{X : \max_i \|X_{i,:}\|_2 \le 1\}$ is $h_{K_{\mathrm{row}}}(G) = \sum_i \|G_{i,:}\|_2$. Since $\sum_i \|G_{i,:}\|_2 \le \sqrt{m}\|G\|_F \le \sqrt{m}\|G\|_*$, we obtain

$$\langle G, (\mathrm{RN}\circ\mathrm{msign})(G)\rangle \ge \frac{1}{\sqrt{m}} h_{K_{\mathrm{row}}}(G).$$

Because $K_{\mathrm{row}} = -K_{\mathrm{row}}$, the signed output $-\mathrm{RN}(Q)$ is feasible and

$$\langle G, -\mathrm{RN}(Q)\rangle \le -\frac{1}{\sqrt{m}} h_{K_{\mathrm{row}}}(G) = \frac{1}{\sqrt{m}} m_{K_{\mathrm{row}}}(G).$$

This is the required approximate *minimization* inequality.

For the reverse order, let $A = \mathrm{RN}(G)$ and write $G = \mathrm{diag}(d_1, \dots, d_m)A$, where $d_i = \|G_{i,:}\|_2$. For each nonzero row, $\|A_{i,:}\|_2 = 1$. Put $M = A^TA$. On $\mathrm{range}(M)$, $(M^\dagger)^{1/2} \succeq \|A\|_{\mathrm{op}}^{-1} I$, and each $A_{i,:}^T$ belongs to this range. Hence, with $P = \mathrm{msign}(A) = A(M^\dagger)^{1/2}$,

$$\langle A_{i,:}, P_{i,:}\rangle \ge \frac{1}{\|A\|_{\mathrm{op}}} \ge \frac{1}{\sqrt{m}}$$

for every nonzero row. Zero rows have $d_i = 0$ and contribute nothing. Thus

$$\langle G, P\rangle \ge \frac{1}{\sqrt{m}} \sum_i d_i \ge \frac{1}{\sqrt{m}} \|G\|_*.$$

The last inequality is the nuclear-norm triangle inequality applied to the single-row decomposition of G. Write $K_{\mathrm{sp}} := \{X : \|X\|_{\mathrm{op}} \le 1\}$. Since $\|P\|_{\mathrm{op}} \le 1$, the signed output $-P$ is feasible and

$$\langle G, -P\rangle \le -\frac{1}{\sqrt{m}} \|G\|_* = \frac{1}{\sqrt{m}} m_{K_{\mathrm{sp}}}(G).$$

Thus $-(\mathrm{msign}\circ\mathrm{RN})$ is a $1/\sqrt{m}$-approximate LMO. □

Proposition B.15 (Proposition 4 of the paper). *Let* $n = 2$. *For every closed convex set* K *containing the image of the* signed *map in question, the best uniform approximation factor satisfies*

$$\nu \le 3(m-1)^{-1/4}$$

for $-(\mathrm{RN}\circ\mathrm{msign})$ *when* m *is odd, and*

$$\nu \le \frac{3}{\sqrt{m-1}}$$

for $-(\mathrm{msign}\circ\mathrm{RN})$.

Proof. For an unbounded K, interpret m_K as an infimum; if it is $-\infty$ in a tested direction, no positive approximation factor is possible. Set $n = 2$. First let $N = \mathrm{RN}\circ\mathrm{msign}$ and $U = -N$. Put $r = m-1$, and assume r is even. Define $p, q \in \mathbb{R}^m$ by

$$p_1 = \sqrt{1 - r^{-1/2}}, \qquad p_i = r^{-3/4} \quad (i \ge 2),$$

$$q_1 = 0, \qquad q_i = \begin{cases} r^{-1/2}, & 2 \le i \le r/2+1, \\ -r^{-1/2}, & r/2+2 \le i \le m. \end{cases}$$

Then $\|p\| = \|q\| = 1$ and $p^\top q = 0$, so $Q = [p, q]$ has orthonormal columns. For $G = Q\,\mathrm{diag}(1, \varepsilon)$, $\varepsilon > 0$, we have $\mathrm{msign}(G) = Q$. With $a = r^{-3/4}$ and $b = r^{-1/2}$, direct rowwise normalization gives

$$\langle G, N(G)\rangle = \sqrt{1 - r^{-1/2}} + r\frac{a^2 + \varepsilon b^2}{\sqrt{a^2 + b^2}} \le 3$$

for sufficiently small ε, for example $0 < \varepsilon \le r^{-1}$. Define Q' by $Q'_{1,:} = e_2^\top$ and $Q'_{i,:} = r^{-1/2} e_1^\top$ for $i \ge 2$. Then $(Q')^\top Q' = I_2$, so $N(Q') = \mathrm{RN}(Q')$, whose first row is $e_2^\top$ and remaining rows are $e_1^\top$. Since $-N(Q') \in K$,

$$m_K(G) \le \langle G, -N(Q')\rangle = -ra = -r^{1/4}.$$

If U is a ν-approximate LMO of K, then

$$-3 \le \langle G, U(G)\rangle \le \nu m_K(G) \le -\nu r^{1/4},$$

which proves $\nu \le 3(m-1)^{-1/4}$.

For the reverse chain, let $N = \mathrm{msign} \circ \mathrm{RN}$ and $U = -N$. Set $\varepsilon = m^{-2}$ and define $G \in \mathbb{R}^{m\times 2}$ by

$$G_{1,:} = e_1^\top, \qquad G_{i,:} = \varepsilon e_1^\top \ (2 \le i \le m-1), \qquad G_{m,:} = \varepsilon e_2^\top.$$

Then $A = \mathrm{RN}(G)$ has $m-1$ copies of $e_1^\top$ and one copy of $e_2^\top$, so $A^\top A = \mathrm{diag}(m-1, 1)$. Hence

$$\langle G, N(G)\rangle = \frac{1 + (m-2)\varepsilon}{\sqrt{m-1}} + \varepsilon \le \frac{3}{\sqrt{m-1}}.$$

Take Q' with first row $e_1^\top$ and all remaining rows $e_2^\top/\sqrt{m-1}$. It has orthonormal columns and satisfies $N(Q') = Q'$. Thus $-Q' \in K$, and

$$m_K(G) \le \langle G, -Q'\rangle = -1 - \frac{\varepsilon}{\sqrt{m-1}} \le -1.$$

Approximate minimization gives

$$-\frac{3}{\sqrt{m-1}} \le \langle G, U(G)\rangle \le \nu m_K(G) \le -\nu,$$

so $\nu \le 3/\sqrt{m-1}$. Both bounds vanish as $m \to \infty$, proving Proposition 4. □

C Appendix: Proofs in section 3

We use the model and scaling assumptions of Section 3. For clarity, we first specify the notation used throughout this appendix. Let

$$X = [x_1, \dots, x_N] \in \mathbb{R}^{d\times N}, \qquad Y = [y_1, \dots, y_N] \in \mathbb{R}^{n\times N},$$

where the entries of X and Y are independent centered Gaussians with variances $1/d$ and $1/n$, respectively, and X is independent of Y. Write $p_i = i^{-\alpha}/\sum_{j=1}^N j^{-\alpha}$, and define

$$P_N = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top, \qquad s = Yp, \qquad K = P_N\,\mathrm{diag}(p) Y^\top \in \mathbb{R}^{N\times n}.$$

Differentiating the cross-entropy loss at zero gives the exact identity

$$G = -\nabla\mathcal{L}(0) = \sum_{i=1}^N p_i \left(\Sigma x_i - \frac{1}{N}\sum_{j=1}^N \Sigma x_j\right) y_i^\top = \Sigma X K.$$

In particular,

$$K^\top e_i = p_i y_i - \frac{s}{N}, \qquad K^\top K = \sum_{i=1}^N p_i^2 y_i y_i^\top - \frac{1}{N} s s^\top.$$

We abbreviate the three updates by

$$W_G = G, \qquad W_M = \mathrm{pol}(G), \qquad W_C = \mathrm{pol}(\mathrm{RN}(G)).$$

Here pol is the exact canonical polar factor, equivalently msign in the main text. All inverse matrix powers below are taken on the full-column-rank event, which has probability one for all sufficiently large n. The statements concern these exact updates, not a fixed number of Newton–Schulz iterations.

Set

$$h = (n/N)^{\alpha}, \qquad L = \log(enNd), \qquad s_1 = \sum_{a=1}^{d} \sigma_a, \qquad s_2 = \sum_{a=1}^{d} \sigma_a^2.$$

The assumptions imply $L \asymp \log n$, $N/n \to \infty$, $d/n \to \infty$, and

$$n^{-1/2+\epsilon} \le h \le n^{-(1-\beta+\epsilon)}, \qquad s_1 \asymp d^{1-\beta} \ge cn, \qquad c \le s_2 \le C,$$

$$h \to 0, \qquad hn^{1-\beta} \le n^{-\epsilon}, \qquad \frac{\sqrt{L/n}}{h} \to 0, \qquad \frac{\sqrt{n}}{Nh} \to 0,$$

$$L^{\alpha}\sqrt{\frac{n}{d}} \to 0, \qquad \frac{\sqrt{ns_2}\, L^{\alpha+2}}{s_1 h} \le Cn^{-\epsilon} L^{\alpha+2} \to 0.$$

Constants c, C may change between displays and depend only on the fixed parameters in the assumptions. A subscript b also permits dependence on a fixed $b \in (0,1)$. The eigenvalues of a positive semidefinite matrix are indexed in nonincreasing order. Conditional $O_{\mathbb{P}}$ bounds below are uniform over the specified query event; bounds for an index $i \ge bN$ are uniform over deterministic indices in this range.

The proof is organized by conditioning: first on the query realization Y, then on the observed gradient G, and finally on the residual score randomness left after conditioning on (Y, G).

C.1 Query-side event and population geometry

Lemma C.1 (High-probability regularity of the queries)**.** *There are constants $c, C > 0$ such that the event $\mathcal{E}_Y$ defined by*

$$\max_{i \le N} \|y_i\|_2 \le 2, \qquad \left\| \frac{1}{N} YY^{\top} \right\|_{\mathrm{op}} \le \frac{C}{n}, \tag{10}$$

$$cn^{-2\alpha} I \preceq K^{\top}K \preceq CI, \qquad c \le \mathrm{tr}(K^{\top}K) \le C, \tag{11}$$

$$\mathrm{tr}\big((K^{\top}K)^{-1/2}\big) \ge cn^{\alpha+1}, \qquad \|(K^{\top}K)^{-1/2} s\|_2 \le C\sqrt{n}, \tag{12}$$

$$\lambda_{\ell}(K^{\top}K) \ge c\ell^{-2\alpha} \qquad (1 \le \ell \le \lfloor A_* L \rfloor), \tag{13}$$

and

$$\sup_{B \ne 0} \frac{1}{N} \# \left\{ i : \|By_i\|_2 \le \eta \frac{\|B\|_F}{\sqrt{n}} \right\} \le C\left(\eta + \sqrt{\frac{n}{N}}\right) \qquad (0 < \eta < 1) \tag{14}$$

satisfies $\mathbb{P}_Y(\mathcal{E}_Y^c) \to 0$. Here $A_* > 0$ is a fixed sufficiently large constant, and the supremum in equation 14 is over all nonzero real matrices with n columns and any finite number of rows. The inequalities hold simultaneously for all B and η.

For a deterministic index i, define

$$s_{-i} = \sum_{j \ne i} p_j y_j, \qquad K_{-i} = P_N\big(\mathrm{diag}(p) - p_i e_i e_i^{\top}\big) Y^{\top}.$$

Let $\mathcal{E}_{Y,-i}$ consist of equation 11–equation 13 with K, s replaced by K_{-i}, s_{-i}. Thus $\mathcal{E}_{Y,-i}$ is measurable with respect to $Y_{-i} = (y_j)_{j \ne i}$; it does not include conditions involving y_i. For every fixed $b \in (0,1)$,

$$\sup_{i \ge bN} \mathbb{P}_Y((\mathcal{E}_Y \cap \mathcal{E}_{Y,-i})^c) \to 0.$$

Proof. Gaussian norm and singular-value bounds give equation 10 with probability tending to one, since $\log N = o(n)$ and $N/n \to \infty$. Write

$$S = \sum_{j=1}^{N} p_j^2 y_j y_j^{\top}, \qquad K^{\top}K = S - \frac{1}{N} ss^{\top}.$$

To control the centering term, represent $y_j = \varepsilon_j \widetilde{y}_j$, where the ε_j are independent Rademacher variables, independent of the Gaussian vectors $\widetilde{y}_j$. Conditional on $\widetilde{Y}$, let $A = [p_j \widetilde{y}_j]_j$. Then

$$v = s^\top S^{-1} s = \varepsilon^\top A^\top (AA^\top)^{-1} A\varepsilon.$$

The matrix between the signs is a rank-n orthogonal projection. Consequently

$$\mathbb{E}_\varepsilon v = n, \qquad \mathrm{Var}_\varepsilon(v) = 2\sum_{j\neq k} \big[A^\top (AA^\top)^{-1} A\big]_{jk}^2 \leq 2n.$$

Thus $v \leq 2n$ with probability at least $1 - 2/n$. On this event, for all sufficiently large n,

$$(1 - v/N)S \preceq K^\top K \preceq S, \qquad s^\top (K^\top K)^{-1} s = \frac{v}{1 - v/N} \leq Cn.$$

The equality is the rank-one inverse formula. This proves the second bound in equation 12 and allows spectral bounds for S to be transferred to $K^\top K$ with fixed constants.

The first $4n$ queries satisfy $\sum_{j\leq 4n} y_j y_j^\top \succeq cI$ except with probability e^{-cn}. Since $p_j \geq cn^{-\alpha}$ on this block, $S \succeq cn^{-2\alpha} I$. The upper trace bound follows from $\max_j \|y_j\| \leq 2$ and $\sum_j p_j^2 \leq C$. The first term $p_1^2 \|y_1\|^2$ gives a positive lower trace bound with probability tending to one. For $q = \lfloor A_* L \rfloor = o(n)$, the $n \times q$ matrix $[y_1, \ldots, y_q]$ has all singular values in $[c, C]$ except with probability e^{-cn}. The nonzero eigenvalues of its weighted Gram matrix therefore give $\lambda_\ell(S) \geq c p_\ell^2 \geq c\ell^{-2\alpha}$ for $\ell \leq q$. This proves equation 11 and equation 13.

For the inverse trace, put $r = \lfloor n/4 \rfloor$ and $S_{>r} = \sum_{j>r} p_j^2 y_j y_j^\top$. For a fixed unit vector u, $nu^\top S_{>r} u$ is a weighted sum of independent squared standard Gaussians. Its weights obey

$$\sum_{j>r} p_j^2 \leq Cn^{1-2\alpha}, \qquad \sum_{j>r} p_j^4 \leq Cn^{1-4\alpha}, \qquad \max_{j>r} p_j^2 \leq Cn^{-2\alpha}.$$

The Gaussian moment-generating-function bound, at a tail parameter of order n, followed by a $1/4$-net of the sphere, yields $\|S_{>r}\|_{\mathrm{op}} \leq Cn^{-2\alpha}$ except with probability e^{-cn}. The first r terms of S have rank at most r. Hence at least $n - r$ eigenvalues of $K^\top K \preceq S$ are at most $Cn^{-2\alpha}$, and

$$\mathrm{tr}\big((K^\top K)^{-1/2}\big) \geq cn^{\alpha+1}.$$

We next prove equation 14, including its uniformity. Write $z_i = \sqrt{n}\, y_i$ and $a_n = \sqrt{n/N}$. On equation 10, $\|N^{-1} \sum_i z_i z_i^\top\|_{\mathrm{op}} \leq C_0$ for a fixed C_0. For a fixed unit vector v and $0 < t \leq 1$,

$$\mathbb{P}(|v^\top z_i| \leq 2t) \leq C_1 t.$$

A Bernoulli Chernoff bound therefore gives

$$\mathbb{P}\left\{\frac{1}{N}\#\{i : |v^\top z_i| \leq 2t\} > C_2 t\right\} \leq e^{-cNt}$$

for a sufficiently large fixed C_2; when $C_2 t \geq 1$ the event is empty. For each dyadic $t \in [a_n, 1]$, also including the smallest dyadic t not below a_n, take a sphere net of mesh $\delta = c_0 t^{3/2}$. Its cardinality is at most $(3/\delta)^n$. A union bound over these nets has failure probability at most

$$C \log(e/a_n) \exp\{Cn \log(e/a_n) - cNa_n\} = o(1),$$

because $N/n \to \infty$ and $Na_n/n = \sqrt{N/n} \gg \log(eN/n)$. For an arbitrary unit vector u, choose a net point v within δ. If $|u^\top z_i| \leq t$ but $|v^\top z_i| > 2t$, then $|(u - v)^\top z_i| > t$. Thus

$$\frac{1}{N}\#\{i : |(u-v)^\top z_i| > t\} \leq \frac{1}{Nt^2} \sum_i ((u-v)^\top z_i)^2 \leq \frac{C_0 \delta^2}{t^2} \leq Ct.$$

Combining these bounds and using monotonicity between dyadic thresholds and below a_n proves, simultaneously for all unit u and all $t > 0$,

$$\frac{1}{N}\#\{i : |u^\top z_i| \leq t\} \leq C(t + a_n).$$

For any nonzero B, diagonalize

$$\frac{B^\top B}{\|B\|_F^2} = \sum_k w_k u_k u_k^\top, \qquad w_k \geq 0, \quad \sum_k w_k = 1.$$

If $\|Bz_i\| \leq \eta \|B\|_F$, at least half of the weight is on indices k for which $|u_k^\top z_i| \leq \sqrt{2}\eta$. Therefore

$$\begin{aligned}\frac{1}{N}\#\{i : \|Bz_i\| \leq \eta\|B\|_F\} &\leq 2\sum_k w_k \frac{1}{N}\#\{i : |u_k^\top z_i| \leq \sqrt{2}\eta\} \\ &\leq C(\eta + a_n).\end{aligned}$$

This is equation 14; it is pathwise and allows B to depend on the data.

Finally, for a fixed $b > 0$, a deterministic index $i \geq bN$ lies outside the first $4n$ and first $\lfloor A_* L \rfloor$ queries once n is large. Deleting its weighted term leaves these blocks intact and only decreases the positive semidefinite tail matrix. The same centering argument applies with its weight set to zero; the projection still has rank n. All failure bounds are independent of the particular deterministic $i \geq bN$. This proves the final assertion. □

Lemma C.2 (Bulk leverage on $\mathcal{E}_Y$)**.** *On $\mathcal{E}_Y$, $t_j := (K^\top K)^{-1/2} K^\top e_j$ satisfies $\|t_j\|_2 \leq 1$ for every j. For every fixed $b > 0$,*

$$\max_{i \geq bN} \|t_i\|_2 \leq C_b h.$$

Proof. The matrix $K(K^\top K)^{-1/2}$ has orthonormal columns, so every row has norm at most one. For $i \geq bN$,

$$\|t_i\|_2 \leq p_i \|(K^\top K)^{-1/2}\|_{\mathrm{op}} \|y_i\|_2 + \frac{1}{N}\|(K^\top K)^{-1/2} s\|_2.$$

On $\mathcal{E}_Y$, the first term is at most $C_b N^{-\alpha} n^\alpha = C_b h$, while the second is $C\sqrt{n}/N = o(h)$. Absorbing the latter term into the constant proves the result. □

On $\mathcal{E}_Y$, define the response matrices for a fresh $\xi \sim N(0, I_n)$ by

$$S_\xi = \xi^\top K^\top K \, \xi, \qquad C_Y = \mathbb{E}\frac{\xi\xi^\top}{S_\xi}, \qquad H_Y = \mathbb{E}\frac{\xi\xi^\top}{\sqrt{S_\xi}}, \qquad J_Y = H_Y C_Y^{-1/2}.$$

Once Y is fixed, these are deterministic matrices.

Lemma C.3 (Response moments on $\mathcal{E}_Y$)**.** *On $\mathcal{E}_Y$,*

$$cI \preceq C_Y \preceq CI, \qquad cI \preceq J_Y \preceq I.$$

The matrices C_Y, H_Y, J_Y commute with $K^\top K$. For every fixed positive integer r, for all sufficiently large n,

$$\sup_{Y \in \mathcal{E}_Y} \mathbb{E}\big[(\xi^\top K^\top K \xi)^{-r}\big] \leq C_r < \infty.$$

If

$$(1-e)K^\top K \preceq \widetilde{K}^\top \widetilde{K} \preceq (1+e)K^\top K, \qquad 0 \leq e \leq \frac{1}{2},$$

and $\widetilde{J}$ is defined from $\widetilde{K}^\top \widetilde{K}$ in the same way, then

$$\|\widetilde{J} - J_Y\|_{\mathrm{op}} \leq Ce.$$

The same conclusions apply to K_{-i} on $\mathcal{E}_{Y,-i}$.

Proof. For a fixed r, choose a fixed integer $q > 2r$. In an eigenbasis of $K^\top K$, equation 13 gives

$$S_\xi \geq c q^{-2\alpha} \sum_{\ell=1}^{q} \xi_\ell^2.$$

The negative r-th moment of χ_q^2 is finite; more explicitly it is $2^{-r}\Gamma(q/2 - r)/\Gamma(q/2)$. This proves the inverse-moment bound, uniformly on $\mathcal{E}_Y$.

Sign symmetry in this eigenbasis makes C_Y and H_Y diagonal. Put

$$c_j = \mathbb{E}\frac{\xi_j^2}{S_\xi}, \qquad h_j = \mathbb{E}\frac{\xi_j^2}{\sqrt{S_\xi}}.$$

Cauchy–Schwarz gives $h_j^2 \le c_j$. Also

$$1 \le h_j\,\mathbb{E}[\xi_j^2\sqrt{S_\xi}], \qquad \mathbb{E}[\xi_j^2 S_\xi] = \operatorname{tr}(K^\top K) + 2\lambda_j(K^\top K) \le C.$$

Since $\mathbb{E}[\xi_j^2\sqrt{S_\xi}] \le \sqrt{\mathbb{E}[\xi_j^2 S_\xi]}$, it follows that $h_j \ge c$. A further application of Cauchy–Schwarz and the inverse second-moment bound gives $c_j \le (\mathbb{E}\xi_j^4)^{1/2}(\mathbb{E}S_\xi^{-2})^{1/2} \le C$. Thus $c \le c_j \le C$ and $c \le h_j/\sqrt{c_j} \le 1$, proving the spectral bounds and commutation statements.

The relative Loewner comparison gives $|\widetilde{S}_\xi^{-1} - S_\xi^{-1}| \le CeS_\xi^{-1}$ and $|\widetilde{S}_\xi^{-1/2} - S_\xi^{-1/2}| \le CeS_\xi^{-1/2}$ pointwise. Testing the resulting matrix expectations against arbitrary vectors shows $\|\widetilde{C} - C_Y\|_{\mathrm{op}} + \|\widetilde{H} - H_Y\|_{\mathrm{op}} \le Ce$. Both covariance matrices are bounded below by a fixed positive multiple of the identity. The integral formula

$$D^{-1/2} = \frac{1}{\pi}\int_0^\infty t^{-1/2}(D + tI)^{-1}\,dt$$

and the resolvent identity therefore give $\|\widetilde{C}^{-1/2} - C_Y^{-1/2}\|_{\mathrm{op}} \le Ce$. Taking the difference of the products $\widetilde{H}\widetilde{C}^{-1/2}$ and $H_Y C_Y^{-1/2}$ proves the last bound. Only the Gram-matrix conditions were used, so the same proof applies on $\mathcal{E}_{Y,-i}$. □

Lemma C.4 (Population retrieval margin on $\mathcal{E}_Y$)**.** *Fix $b \in (0,1)$. There is $c_b > 0$ such that for every fixed $\varepsilon > 0$,*

$$\sup_{i \ge bN} \mathbb{P}_Y\left(t_i^\top J_Y y_i < c_b h \;\text{ or }\; \max_{j \ne i}|t_j^\top J_Y y_i| > \varepsilon h \,\middle|\, \mathcal{E}_Y\right) \to 0.$$

Proof. We first record the relative square-root estimate needed below. For positive definite A_0, A_1, suppose

$$\delta = \|A_0^{-1/2}(A_1 - A_0)A_0^{-1/2}\|_F \le \frac{1}{2}.$$

Then

$$\|A_1^{-1/2}A_0^{1/2} - I\|_{\mathrm{op}} \le C\delta.$$

Indeed, let $A(t) = A_0 + t(A_1 - A_0)$ and $H(t) = A(t)^{1/2}$. Differentiating $H(t)^2 = A(t)$ gives $H\dot{H} + \dot{H}H = A_1 - A_0$. If $F(t) = H(t)^{-1}(A_1 - A_0)H(t)^{-1}$, then in an eigenbasis of $H(t)$ with eigenvalues $\lambda_j > 0$,

$$[H^{-1}\dot{H}]_{jk} = \frac{\lambda_k}{\lambda_j + \lambda_k}F(t)_{jk}, \qquad \|H^{-1}\dot{H}\|_F \le \|F(t)\|_F.$$

Since $A(t) \succeq (1 - t\delta)A_0$,

$$\|F(t)\|_F \le \frac{\delta}{1 - t\delta}, \qquad \|H(t)^{-1}A_0^{1/2}\|_{\mathrm{op}} \le (1 - t\delta)^{-1/2}.$$

It follows that

$$\left\|\frac{d}{dt}\big(H(t)^{-1}A_0^{1/2}\big)\right\|_F \le \frac{\delta}{(1 - t\delta)^{3/2}}.$$

Integration from zero to one proves the claim. In particular, this estimate uses the Frobenius norm of the relative perturbation, not a condition-number-dependent additive perturbation bound.

Fix a deterministic $i \ge bN$, and work on $\mathcal{E}_Y \cap \mathcal{E}_{Y,-i}$. Let J_{-i} denote the response operator obtained from $K_{-i}^\top K_{-i}$. Direct expansion gives

$$K^\top K - K_{-i}^\top K_{-i} = p_i^2(1 - 1/N)y_i y_i^\top - \frac{p_i}{N}(y_i s_{-i}^\top + s_{-i}y_i^\top).$$

Set

$$e_i := \left\| (K_{-i}^\top K_{-i})^{-1/2} \big(K^\top K - K_{-i}^\top K_{-i}\big)(K_{-i}^\top K_{-i})^{-1/2} \right\|_F .$$

Since $p_i n^\alpha \asymp_b h$,

$$e_i \le C_b \left(h^2 + \frac{h\sqrt{n}}{N} \right) = O_b(h^2).$$

Hence

$$(1 - e_i) K_{-i}^\top K_{-i} \preceq K^\top K \preceq (1 + e_i) K_{-i}^\top K_{-i}.$$

Lemma C.3 and the relative square-root estimate just proved (with $A_0 = K_{-i}^\top K_{-i}$ and $A_1 = K^\top K$) give

$$\|J_Y - J_{-i}\|_{\mathrm{op}} \le C_b h^2, \tag{15}$$

$$\left\| (K^\top K)^{-1/2} (K_{-i}^\top K_{-i})^{1/2} - I \right\|_{\mathrm{op}} \le C_b h^2, \tag{16}$$

$$\left\| (K^\top K)^{-1/2} J_Y - (K_{-i}^\top K_{-i})^{-1/2} J_{-i} \right\|_{\mathrm{op}} \le C_b n^\alpha h^2. \tag{17}$$

For $j \neq i$, define

$$t_j^{-i} := (K_{-i}^\top K_{-i})^{-1/2} \left(p_j y_j - \frac{s_{-i}}{N} \right),$$

and $t_i^{-i} := -(K_{-i}^\top K_{-i})^{-1/2} s_{-i}/N$. These are rows of $K_{-i}(K_{-i}^\top K_{-i})^{-1/2}$, so $\|t_j^{-i}\|_2 \le 1$. Also,

$$t_j = (K^\top K)^{-1/2} (K_{-i}^\top K_{-i})^{1/2} t_j^{-i} - \frac{p_i}{N} (K^\top K)^{-1/2} y_i,$$

hence

$$\max_{j \neq i} \|t_j - t_j^{-i}\|_2 \le C_b (h^2 + h/N). \tag{18}$$

Conditional on a realization of Y_{-i} satisfying $\mathcal{E}_{Y,-i}$, the deleted objects are fixed and $y_i \sim N(0, I_n/n)$ is independent. At this step we do not condition on $\mathcal{E}_Y$, which also involves y_i. Each wrong deleted response is centered Gaussian with variance at most $1/n$. A Gaussian tail bound and a union bound give

$$\max_{j \neq i} |(t_j^{-i})^\top J_{-i} y_i| = O_{\mathbb{P}}\left(\sqrt{\frac{L}{n}} \right) = o_{\mathbb{P}}(h).$$

The replacement error from (t_j^{-i}, J_{-i}) to (t_j, J_Y) is $o(h)$ by equation 15, equation 18, and $\|y_i\| \le 2$.

For the correct coordinate, put

$$A_i := (K_{-i}^\top K_{-i})^{-1/2} J_{-i}.$$

This matrix is symmetric positive definite and independent of y_i, with

$$\operatorname{tr} A_i \ge c n^{\alpha+1},\ \|A_i\|_{\mathrm{op}} \le C n^\alpha,\ \|A_i\|_F \le C n^{\alpha+1/2}.$$

Therefore

$$\mathbb{E}[y_i^\top A_i y_i \mid Y_{-i}] = \frac{1}{n} \operatorname{tr} A_i \ge c n^\alpha,$$

$$\begin{aligned} \operatorname{Var}(y_i^\top A_i y_i \mid Y_{-i}) &= \frac{2}{n^2} \|A_i\|_F^2 \\ &\le C n^{2\alpha - 1}. \end{aligned}$$

Chebyshev gives $y_i^\top A_i y_i \ge c n^\alpha/2$ with failure $O(1/n)$. Finally,

$$t_i^\top J_Y y_i = p_i y_i^\top (K^\top K)^{-1/2} J_Y y_i - \frac{1}{N} s^\top (K^\top K)^{-1/2} J_Y y_i.$$

The matrix replacement error is $O_b(h^3)$ after multiplying by p_i, while

$$\frac{1}{N} \left| s^\top (K^\top K)^{-1/2} J_Y y_i \right| \le C \frac{\sqrt{n}}{N} = o(h).$$

Since $p_i n^\alpha \asymp_b h$, the correct coordinate is at least $c_b h - o(h)$. The preceding bounds show that the margin failure event, intersected with $\mathcal{E}_Y \cap \mathcal{E}_{Y,-i}$, has probability tending to zero, with a bound depending on i only through $i \ge bN$. Since $\sup_{i \ge bN} \mathbb{P}_Y((\mathcal{E}_Y \cap \mathcal{E}_{Y,-i})^c) \to 0$ and $\mathbb{P}_Y(\mathcal{E}_Y) \to 1$, conditioning on $\mathcal{E}_Y$ yields the stated supremum. □

C.2 Condition on Y: randomness revealed through G

Fix $Y \in \mathcal{E}_Y$. For the a-th row of X, define

$$\xi_a = \sqrt{d}\,(K^\top K)^{-1/2} K^\top X_{a,:}^\top.$$

Conditional on Y, the vectors $\xi_1, \dots, \xi_d$ are independent $N(0, I_n)$, and

$$G_{a,:} = \frac{\sigma_a}{\sqrt{d}} \xi_a^\top (K^\top K)^{1/2}.$$

The conditional law of the ξ_a does not depend on Y, so the family $\{\xi_a\}_{a=1}^d$ is independent of Y. Conversely, (Y, G) determines all ξ_a, because every σ_a is positive.

Define the observed-gradient matrices

$$\mathcal{H} = d(K^\top K)^{-1/2} G^\top G (K^\top K)^{-1/2},$$
$$\mathcal{J} = d(K^\top K)^{-1/2} G^\top \Sigma^2 G (K^\top K)^{-1/2}.$$

Conditional on Y,

$$\mathcal{H} = \sum_{a=1}^d \sigma_a^2 \xi_a \xi_a^\top, \qquad \mathcal{J} = \sum_{a=1}^d \sigma_a^4 \xi_a \xi_a^\top.$$

Lemma C.5 (Observed-gradient estimates)**.** *There is $C < \infty$ such that*

$$\sup_{Y \in \mathcal{E}_Y} \mathbb{P}_\xi\big(\|\mathcal{H}\mathcal{J}^{-1}\mathcal{H}\|_{\mathrm{op}} > Cn \mid Y\big) \to 0,$$

and

$$\sup_{Y \in \mathcal{E}_Y} \mathbb{E}_\xi\Big[\sqrt{d}\,\|G(K^\top K)^{-1/2}\|_* \mid Y\Big] \le Cn^{3/2-\beta}.$$

For every fixed $\delta > 0$,

$$\sup_{Y \in \mathcal{E}_Y} \mathbb{P}_\xi\left(\left|\frac{1}{n}\operatorname{tr}\mathcal{H} - \sum_{a\ge 1}\sigma_a^2\right| > \delta \text{ or } \left|\frac{1}{n}\operatorname{tr}\mathcal{J} - \sum_{a\ge 1}\sigma_a^4\right| > \delta \;\middle|\; Y\right) \to 0.$$

Proof. Split the first $2n$ rows from the rest: $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_1$ and $\mathcal{J} = \mathcal{J}_0 + \mathcal{J}_1$. Let Ξ_0 have rows $\xi_a^\top$, $a \le 2n$, and let B have rows $\sigma_a^2 \xi_a^\top$ on the same block. Then $\mathcal{J}_0 = B^\top B$ and $\mathcal{H}_0 = \Xi_0^\top B = B^\top \Xi_0$. Since $B(B^\top B)^{-1} B^\top$ is an orthogonal projection,

$$\mathcal{H}_0 \mathcal{J}^{-1} \mathcal{H}_0 \preceq \mathcal{H}_0 \mathcal{J}_0^{-1} \mathcal{H}_0 \preceq \Xi_0^\top \Xi_0 \preceq CnI$$

except with probability e^{-cn}. The same block gives $\mathcal{J} \succeq cn^{-4\beta} \Xi_0^\top \Xi_0 \succeq cn^{1-4\beta} I$. On each dyadic block of rows with indices of order $2^k n$, the maximum weight σ_a^2 is at most $C(2^k n)^{-2\beta}$ and the unweighted Gaussian Gram matrix has operator norm at most $C 2^k n$, with failure $e^{-c 2^k n}$. Summing over these blocks, using $2\beta > 1$, gives

$$\|\mathcal{H}_1\|_{\mathrm{op}} \le C \sum_{k \ge 1} (2^k n)^{1-2\beta} \le C n^{1-2\beta}$$

except with probability Ce^{-cn}. Therefore $\mathcal{H}_1 \mathcal{J}^{-1} \mathcal{H}_1 \preceq CnI$. Finally,

$$(\mathcal{H}_0 + \mathcal{H}_1)\mathcal{J}^{-1}(\mathcal{H}_0 + \mathcal{H}_1) \preceq 2\mathcal{H}_0 \mathcal{J}^{-1} \mathcal{H}_0 + 2\mathcal{H}_1 \mathcal{J}^{-1} \mathcal{H}_1,$$

which proves the first assertion.

The matrix $\sqrt{d}\, G(K^\top K)^{-1/2}$ has rows $\sigma_a \xi_a^\top$. Bound the nuclear norm of the first n rows by the sum of their row norms, and bound that of the remaining rows by $\sqrt{n}$ times their Frobenius norm. Jensen's inequality gives

$$\begin{aligned}\mathbb{E}_\xi\Big[\sqrt{d}\,\|G(K^\top K)^{-1/2}\|_* \mid Y\Big] &\le \sqrt{n} \sum_{a \le n} \sigma_a + \sqrt{n}\left(n \sum_{a > n} \sigma_a^2\right)^{1/2} \\ &\le Cn^{3/2-\beta}.\end{aligned}$$

For the trace statements,

$$\mathbb{E}_\xi \frac{\operatorname{tr}\mathcal{H}}{n} = s_2, \qquad \operatorname{Var}_\xi\left(\frac{\operatorname{tr}\mathcal{H}}{n}\right) = \frac{2}{n}\sum_{a\le d}\sigma_a^4 \le \frac{C}{n}.$$

The analogous variance for $\operatorname{tr}\mathcal{J}/n$ is $2\sum_{a\le d}\sigma_a^8/n \le C/n$. The corresponding finite sums converge to the stated infinite sums, so Chebyshev's inequality completes the proof. All these laws and bounds are uniform in Y. □

For any update $W \in \mathbb{R}^{d\times n}$, write

$$m_i(W) = K(K^\top K)^{-1}G^\top W y_i, \qquad m_{ji}(W) = e_j^\top m_i(W).$$

When $\|\Sigma W y_i\|_2 > 0$, define

$$a^{(i)}(W) = \frac{\sqrt{d}\, m_i(W)}{\|\Sigma W y_i\|_2}.$$

If $Wy_i = 0$, set $a^{(i)}(W) = 0$; all scores then vanish and strict retrieval fails. This convention covers zero denominators in the statements below. For the two baselines, nonzero queries have nonzero denominators almost surely.

Corollary C.6 (Conditional mean-energy bound)**.** *For any update W with $\operatorname{col}(W) \subseteq \operatorname{col}(G)$, on the event $\|\mathcal{H}\mathcal{J}^{-1}\mathcal{H}\|_{\mathrm{op}} \le Cn$,*

$$\|a^{(i)}(W)\|_2^2 \le Cn \qquad \textit{for every } i.$$

Proof. The case $Wy_i = 0$ follows from the convention above. Otherwise, write $Wy_i = Gq$ and $\widetilde{q} = (K^\top K)^{1/2}q$. Direct substitution gives

$$a^{(i)}(W) = K(K^\top K)^{-1/2}\frac{\mathcal{H}\widetilde{q}}{\sqrt{\widetilde{q}^\top \mathcal{J}\widetilde{q}}}.$$

The matrix $K(K^\top K)^{-1/2}$ has orthonormal columns. Hence

$$\|a^{(i)}(W)\|_2^2 \le \|\mathcal{H}\mathcal{J}^{-1/2}\|_{\mathrm{op}}^2 = \|\mathcal{H}\mathcal{J}^{-1}\mathcal{H}\|_{\mathrm{op}} \le Cn.$$

□

Lemma C.7 (Correct conditional mean for gradient descent)**.** *Fix $b \in (0,1)$ and $u > 0$. Then*

$$\sup_{Y\in\mathcal{E}_Y}\sup_{i\ge bN}\mathbb{P}_\xi\Big(|a_i^{(i)}(W_G)| > u \mid Y\Big) \to 0.$$

Proof. For $W_G = G$, set $q = (K^\top K)^{1/2}y_i$ and $t = t_i$. The case $q = 0$ is immediate. Otherwise,

$$a_i^{(i)}(W_G) = \frac{t^\top \mathcal{H} q}{\sqrt{q^\top \mathcal{J} q}}.$$

Write $e = q/\|q\|$ and $t = \langle t, e\rangle e + t_\perp$. The Gaussian coordinates along e and $t_\perp$ are independent. Conditioning on those along e gives the distributional identity

$$a_i^{(i)}(W_G) \overset{d}{=} \langle t, e\rangle \frac{\sum_{a\le d}\sigma_a^2 g_a^2}{\sqrt{\sum_{a\le d}\sigma_a^4 g_a^2}} + \|t_\perp\|_2 g_0,$$

where $g_0, g_1, g_2, \ldots$ are independent standard Gaussians. Under a common coupling across d, the numerator is bounded by $\sum_{a\ge 1}\sigma_a^2 g_a^2 < \infty$ almost surely, and the denominator is at least $\sigma_1^2|g_1| > 0$ almost surely. Thus the displayed ratio is uniformly tight, independently of Y and i. Lemma C.2 gives $\|t_i\| \le C_b h$ for $i \ge bN$, and $h \to 0$. This proves the assertion. □

Lemma C.8 (Correct conditional mean for Muon)**.** *For every fixed $u > 0$ and $\delta > 0$,*

$$\sup_{Y\in\mathcal{E}_Y}\mathbb{P}_\xi\left[\frac{1}{N}\#\{i : |a_i^{(i)}(W_M)| > u\} > \delta \;\middle|\; Y\right] \to 0.$$

Proof. Put $W = W_M$, and define

$$\nu_i = t_i^\top (K^\top K)^{-1/2} G^\top W y_i, \qquad v_i = \|\Sigma W y_i\|_2.$$

For $v_i > 0$, $a_i^{(i)}(W) = \sqrt{d}\,\nu_i / v_i$. The correct expansion is obtained with

$$A = (K^\top K)^{-1} G^\top W, \qquad z = A^\top s = W^\top G (K^\top K)^{-1} s : \qquad \nu_i = p_i y_i^\top A y_i - \frac{1}{N} z^\top y_i.$$

On $\mathcal{E}_Y$, pathwise even when A depends on the data,

$$\frac{1}{N}\sum_i |y_i^\top A y_i| \le \operatorname{tr}\left(|\operatorname{sym} A| \frac{YY^\top}{N} \right) \le \frac{C}{n} \|A\|_*,$$

where $\operatorname{sym} A = (A + A^\top)/2$ and $\|\operatorname{sym} A\|_* \le \|A\|_*$. Since $\|W\|_{\mathrm{op}} \le 1$, Lemma C.5 gives

$$\begin{aligned} \mathbb{E}_\xi[\sqrt{d}\,\|A\|_* \mid Y] &\le \|(K^\top K)^{-1/2}\|_{\mathrm{op}}\, \mathbb{E}_\xi[\sqrt{d}\,\|G(K^\top K)^{-1/2}\|_* \mid Y] \\ &\le C n^{\alpha+3/2-\beta}. \end{aligned}$$

Also, although W depends on G, its operator norm bound implies

$$\begin{aligned} \mathbb{E}_\xi[d\|z\|_2^2 \mid Y] &\le \mathbb{E}_\xi[d\|G(K^\top K)^{-1} s\|_2^2 \mid Y] \\ &= s_2 \|(K^\top K)^{-1/2} s\|_2^2 \le Cn. \end{aligned}$$

Consequently $\sqrt{d}\,\|A\|_* = O_{\mathbb{P}}(n^{\alpha+3/2-\beta})$ and $\sqrt{d}\,\|z\|_2 = O_{\mathbb{P}}(\sqrt{n})$, uniformly in Y. For fixed $b > 0$, use $p_i \le C_b N^{-\alpha}$ on $i \ge bN$ and

$$\frac{1}{N}\sum_i |z^\top y_i| \le \left(z^\top \frac{YY^\top}{N} z \right)^{1/2} \le \frac{C}{\sqrt{n}} \|z\|_2$$

to obtain

$$\begin{aligned} \frac{1}{N}\sum_{i \ge bN} |\sqrt{d}\,\nu_i| &\le C_b \frac{N^{-\alpha}}{n} \sqrt{d}\,\|A\|_* + \frac{C}{N\sqrt{n}} \sqrt{d}\,\|z\|_2 \\ &= O_{\mathbb{P},b}\left(h n^{1/2-\beta} + N^{-1} \right). \end{aligned} \tag{19}$$

Since W_M is the polar factor of the full-column-rank matrix G,

$$\begin{aligned} \|\Sigma W_M\|_F^2 &= \operatorname{tr}\left[(G^\top G)^{-1} G^\top \Sigma^2 G \right] = \operatorname{tr}(\mathcal{H}^{-1}\mathcal{J}) \\ &\ge \frac{\operatorname{tr}\mathcal{J}}{\|\mathcal{H}\|_{\mathrm{op}}} \ge \frac{\operatorname{tr}\mathcal{J}}{\operatorname{tr}\mathcal{H}}. \end{aligned}$$

The trace limits in Lemma C.5 have positive finite values. Hence $\|\Sigma W_M\|_F \ge c_0 > 0$ with conditional probability tending to one, uniformly in Y. On this event, apply equation 14 pathwise with $B = \Sigma W_M$. For each fixed $\eta \in (0,1)$, all but at most $CN(\eta + \sqrt{n/N})$ indices satisfy

$$v_i \ge c_0 \eta / \sqrt{n}.$$

No independence of B and Y is required. Counting the exceptional indices and then using Markov's inequality for the remaining deterministic list of numerators gives

$$\frac{1}{N} \#\{ i : |a_i^{(i)}(W_M)| > u \} \le b + \frac{1}{N} + C\eta + C\sqrt{\frac{n}{N}} + \frac{\sqrt{n}}{u c_0 \eta N} \sum_{i \ge bN} |\sqrt{d}\,\nu_i|.$$

By equation 19, the last term is

$$O_{\mathbb{P},b}\left(\frac{h n^{1-\beta} + \sqrt{n}/N}{u\eta} \right) = o_{\mathbb{P}}(1).$$

For a given $\delta > 0$, choose b, η small and fixed so that $b + C\eta < \delta/2$, and then send $n \to \infty$. The exceptional observed-gradient event also has probability tending to zero uniformly in Y, which proves the assertion. □

For the chained update, the answer-coordinate scales cancel exactly: $\mathrm{RN}(G) = \mathrm{RN}(XK)$, since every σ_a is positive. Define

$$\widetilde{W}_{a,:} = \frac{\xi_a^\top C_Y^{-1/2}}{\sqrt{d\,\xi_a^\top K^\top K \xi_a}}.$$

Lemma C.9 (Fixed-Y approximation for the chain)**.** *There is $C < \infty$ such that*

$$\sup_{Y \in \mathcal{E}_Y} \mathbb{P}_\xi\left(\|W_C - \widetilde{W}\|_{\mathrm{op}} > C L^{\alpha+1} \sqrt{\frac{n}{d}} \;\middle|\; Y \right) \to 0,$$

and

$$\sup_{Y \in \mathcal{E}_Y} \mathbb{P}_\xi\left(\max_{a \le d} \|(W_C)_{a,:}\|_2 > C L^{\alpha+1} \sqrt{\frac{n}{d}} \;\middle|\; Y \right) \to 0.$$

Moreover,

$$\sup_{Y \in \mathcal{E}_Y} \sup_{i \le N} \mathbb{P}_\xi\left(\left\| m_i(W_C) - \frac{s_1}{d} K (K^\top K)^{-1/2} J_Y y_i \right\|_\infty > C \frac{\sqrt{n s_2}}{d} L^{\alpha+2} \;\middle|\; Y \right) \to 0.$$

Proof. We give the concentration, polar perturbation, and conditional-mean steps separately. Throughout the proof $Y \in \mathcal{E}_Y$ is fixed.

Concentration of the proxy Gram matrix. Write

$$Z_a = \frac{C_Y^{-1/2} \xi_a \xi_a^\top C_Y^{-1/2}}{\xi_a^\top K^\top K \xi_a}, \qquad \widetilde{W}^\top \widetilde{W} = \frac{1}{d} \sum_{a=1}^{d} Z_a, \qquad \mathbb{E}_\xi Z_a = I.$$

Choose $q = \lceil AL \rceil$ with A a sufficiently large fixed constant; in Lemma C.1, take $A_* > 2A$. In an eigenbasis of $K^\top K$,

$$\xi_a^\top K^\top K \xi_a \ge c q^{-2\alpha} \sum_{\ell=1}^{q} \xi_{a,\ell}^2.$$

Let $\mathcal{T}_a$ be the event $\|\xi_a\|^2 \le Cn$ and $\sum_{\ell \le q} \xi_{a,\ell}^2 \ge q/2$. By choosing the fixed constants sufficiently large, Gaussian and chi-square tails give

$$\sup_{Y \in \mathcal{E}_Y} \mathbb{P}_\xi(\mathcal{T}_a^c \mid Y) \le e^{-20L}$$

for all sufficiently large n. On $\mathcal{T}_a$,

$$\xi_a^\top K^\top K \xi_a \ge c L^{1-2\alpha}, \qquad 0 \preceq Z_a \preceq C n L^{2\alpha-1} I.$$

For $\widehat{Z}_a = Z_a \mathbf{1}_{\mathcal{T}_a}$, positivity gives $\mathbb{E}\widehat{Z}_a \preceq I$ and

$$\mathbb{E}\widehat{Z}_a^2 \preceq C n L^{2\alpha-1} \mathbb{E}\widehat{Z}_a \preceq C n L^{2\alpha-1} I.$$

Thus the centered truncated sum has variance parameter at most $C d n L^{2\alpha-1}$ and summand norm at most $C n L^{2\alpha-1}$. The inverse moments in Lemma C.3 also give

$$\mathbb{E}\|Z_a\|_{\mathrm{op}}^2 \le C \mathbb{E}\left[\frac{\|\xi_a\|^4}{(\xi_a^\top K^\top K \xi_a)^2} \right] \le C n^2,$$

so Cauchy–Schwarz controls the truncation bias:

$$\|\mathbb{E}\widehat{Z}_a - I\|_{\mathrm{op}} \le C n e^{-10L}.$$

Apply matrix Bernstein to both signs of the centered truncated sum (Tropp, 2012, Theorem 6.1). Together with $\mathbb{P}(\bigcup_{a \le d} \mathcal{T}_a^c \mid Y) \le d e^{-20L}$, it gives

$$\|\widetilde{W}^\top \widetilde{W} - I\|_{\mathrm{op}} \le C \left(L^\alpha \sqrt{\frac{n}{d}} + L^{2\alpha} \frac{n}{d} \right) \le C L^\alpha \sqrt{\frac{n}{d}} =: e_n$$

with conditional probability tending to one uniformly in Y. The last inequality uses $L^\alpha \sqrt{n/d} \to 0$; the truncation bias is negligible on this scale.

A relative polar estimate. For positive definite matrices M, P with $\delta = \|P - I\|_{\mathrm{op}} \le 1/2$, we claim

$$\|\operatorname{pol}(PM) - I\|_{\mathrm{op}} \le C\big(1 + \log \operatorname{cond}(M)\big)\delta, \tag{20}$$

where $\operatorname{cond}(M) = \lambda_{\max}(M)/\lambda_{\min}(M)$. Here is a direct proof of the spectral-norm estimate. Set $E = P - I$ and write the polar decomposition

$$(I + tE)M = U(t)H(t), \qquad 0 \le t \le 1,$$

where $U(t)$ is orthogonal and $H(t)$ is positive definite. The factors are differentiable along this nonsingular path. Put $\Omega(t) = U(t)^\top \dot{U}(t)$ and

$$D(t) = U(t)^\top E(I + tE)^{-1}U(t).$$

The matrix $D(t)$ is symmetric, because E commutes with $I + tE$, and $\|D(t)\|_{\mathrm{op}} \le 2\delta$. Differentiating the polar decomposition and subtracting its transpose gives the Sylvester equation

$$\Omega H + H\Omega = DH - HD.$$

Its integral solution is

$$\Omega = \int_0^\infty e^{-sH}(DH - HD)e^{-sH}\,ds.$$

If $\mu = \lambda_{\min}(H)$ and $\Lambda = \lambda_{\max}(H)$, then

$$\|\Omega\|_{\mathrm{op}} \le 2\|D\|_{\mathrm{op}} \int_0^\infty \|e^{-sH}\|_{\mathrm{op}}\|He^{-sH}\|_{\mathrm{op}}\,ds \le C\|D\|_{\mathrm{op}}\big(1 + \log(\Lambda/\mu)\big).$$

For the last inequality, split the integral at $1/\Lambda$ and $1/\mu$: on the first interval use $\|He^{-sH}\|_{\mathrm{op}} \le \Lambda$; on the middle interval use $\|He^{-sH}\|_{\mathrm{op}} \le 1/(es)$; on the last interval the product is at most $\mu e^{-2s\mu}$. Moreover,

$$\frac{\Lambda}{\mu} = \operatorname{cond}((I + tE)M) \le \frac{1+\delta}{1-\delta}\operatorname{cond}(M) \le 3\operatorname{cond}(M).$$

Since $U(0) = I$ and $\|\dot{U}(t)\|_{\mathrm{op}} = \|\Omega(t)\|_{\mathrm{op}}$, integration over $t \in [0, 1]$ proves equation 20.

Application to W_C and the row bound. The factorization

$$\frac{\mathrm{RN}(G)}{\sqrt{d}} = \widetilde{W}\, C_Y^{1/2}(K^\top K)^{1/2}$$

is exact. On the proxy-Gram event, set

$$P = (\widetilde{W}^\top \widetilde{W})^{1/2}, \qquad V = \operatorname{pol}(\widetilde{W}), \qquad M = C_Y^{1/2}(K^\top K)^{1/2}.$$

Then $\widetilde{W} = VP$ and $V^\top V = I$. By Lemma C.3, the two factors in M commute, so M is symmetric positive definite. Its eigenvalues lie between $cn^{-\alpha}$ and C, and thus $1 + \log \operatorname{cond}(M) \le CL$. Left multiplication by an isometry preserves the polar factorization, so

$$W_C = V \operatorname{pol}(PM), \qquad \|P - I\|_{\mathrm{op}} \le Ce_n.$$

Applying equation 20 yields

$$\begin{aligned}\|W_C - \widetilde{W}\|_{\mathrm{op}} &= \|\operatorname{pol}(PM) - P\|_{\mathrm{op}} \\ &\le \|\operatorname{pol}(PM) - I\|_{\mathrm{op}} + \|P - I\|_{\mathrm{op}} \\ &\le CLe_n \le CL^{\alpha+1}\sqrt{\frac{n}{d}}.\end{aligned}$$

For the row bound, all $\mathcal{T}_a$ hold with conditional probability at least $1 - de^{-20L}$. On these events,

$$\max_{a \le d} \|\widetilde{W}_{a,:}\|_2 \le CL^{\alpha-1/2}\sqrt{\frac{n}{d}}.$$

The square matrix $\operatorname{pol}(PM)$ is orthogonal and $V = \widetilde{W}P^{-1}$, with $\|P^{-1}\|_{\mathrm{op}} \le 2$ for large n. Therefore

$$\|(W_C)_{a,:}\|_2 = \|V_{a,:}\|_2 \le 2\|\widetilde{W}_{a,:}\|_2 \le CL^{\alpha-1/2}\sqrt{\frac{n}{d}},$$

which is stronger than the stated row bound.

Conditional-mean concentration. Fix a deterministic i and write $b = C_Y^{-1/2} y_i$, so $\|b\| \le C$. The contribution from row a to coordinate j of $m_i(\widetilde{W})$ is

$$\frac{\sigma_a}{d} f_{aj}, \qquad f_{aj} = \frac{(t_j^\top \xi_a)(b^\top \xi_a)}{\sqrt{\xi_a^\top K^\top K \xi_a}}.$$

Its expectation is $\mathbb{E} f_{aj} = t_j^\top H_Y C_Y^{-1/2} y_i = t_j^\top J_Y y_i$. Gaussian moments, $\|t_j\| \le 1$, and the inverse-moment bound imply

$$\mathbb{E} f_{aj}^2 \le C$$

uniformly in a, j, i, Y. For each row, refine $\mathcal{T}_a$ by also requiring

$$\max_{j \le N} |t_j^\top \xi_a| \le C\sqrt{L}, \qquad |b^\top \xi_a| \le C\sqrt{L}.$$

Gaussian tails and a union bound over j allow the complement of this refined event to have probability at most e^{-20L}, after increasing fixed constants. On the refined event, $|f_{aj}| \le C L^{\alpha+1/2}$. The truncated variables remain independent across a; their total weighted variance is at most $C s_2/d^2$, and their centered summand bound is $C \max_a \sigma_a L^{\alpha+1/2}/d$. Their total bias is at most $C s_1 e^{-10L}/d$, which is negligible since $s_1 \le \sqrt{d s_2}$. Scalar Bernstein and a union bound over $j \le N$ now give

$$\begin{aligned}\max_{j \le N} \left| e_j^\top m_i(\widetilde{W}) - \frac{s_1}{d} t_j^\top J_Y y_i \right| &\le \frac{C}{d} \left(\sqrt{s_2 L} + \max_{a \le d} \sigma_a L^{\alpha+3/2} \right) \\ &\le C \frac{\sqrt{s_2}}{d} L^{\alpha+3/2}\end{aligned}$$

with conditional failure probability tending to zero. The constants depend on i only through $\|y_i\| \le 2$, so the same failure bound holds uniformly over deterministic $i \le N$.

Finally, the replacement error obeys

$$|e_j^\top (m_i(W_C) - m_i(\widetilde{W}))| \le \|G(K^\top K)^{-1/2} t_j\|_2 \, \|W_C - \widetilde{W}\|_{\mathrm{op}} \, \|y_i\|_2.$$

Here

$$\|G(K^\top K)^{-1/2} t_j\|_2^2 = \frac{1}{d} \sum_{a \le d} \sigma_a^2 (\xi_a^\top t_j)^2.$$

The weighted chi-square moment-generating-function bound, together with $\|t_j\| \le 1$, $\sum_a \sigma_a^4 \le s_2^2$, and a union bound over j, gives

$$\max_{j \le N} \|G(K^\top K)^{-1/2} t_j\|_2 \le C \sqrt{\frac{s_2 L}{d}}$$

with conditional probability tending to one uniformly in Y. Combining this with the operator-norm approximation and the proxy-mean bound yields

$$\left\| m_i(W_C) - \frac{s_1}{d} K (K^\top K)^{-1/2} J_Y y_i \right\|_\infty \le C \frac{\sqrt{n s_2}}{d} L^{\alpha+3/2} \le C \frac{\sqrt{n s_2}}{d} L^{\alpha+2},$$

which proves the last assertion. □

C.3 Condition on (Y, G): residual score randomness

Once (Y, G) is fixed, any update W measurable with respect to (Y, G), its conditional mean, and all query-side quantities are fixed. The remaining randomness is the component of X not revealed by $G = \Sigma X K$. Define the retrieval score

$$S_{ji}(W) = x_j^\top \Sigma W y_i.$$

Lemma C.10 (Exact conditional score law)**.** *Let W be any measurable function of (Y, G). For a fixed query y_i, write $\mathbf{S}_i(W) = (S_{1i}(W), \dots, S_{Ni}(W))^\top$. Conditional on (Y, G),*

$$\mathbf{S}_i(W) \sim N\left(m_i(W), \frac{\|\Sigma W y_i\|_2^2}{d} \left[I_N - K(K^\top K)^{-1} K^\top \right] \right),$$

where $m_i(W) = K(K^\top K)^{-1} G^\top W y_i$.

Proof. Conditional on Y, put $U = K(K^\top K)^{-1/2}$. For each answer-coordinate row,

$$X_{a,:}^\top = \frac{1}{\sqrt{d}} U\xi_a + x_{a,\perp}, \qquad x_{a,\perp} = (I_N - UU^\top)X_{a,:}^\top.$$

These two Gaussian components are independent, and the residual rows are mutually independent. Given (Y, G), the first components are fixed because

$$\xi_a = \sqrt{d}\,\sigma_a^{-1}(K^\top K)^{-1/2} G_{a,:}^\top.$$

The residual rows retain mean zero and covariance $d^{-1}(I_N - UU^\top)$. Substituting their decomposition into $\mathbf{S}_i(W) = X^\top \Sigma W y_i$ gives the stated covariance. For the mean, the observed component satisfies

$$\frac{1}{\sqrt{d}} \sum_{a=1}^{d} \sigma_a U \xi_a (W y_i)_a = K(K^\top K)^{-1} G^\top W y_i,$$

which is the claimed formula. □

Lemma C.11 (Wrong-score maximum)**.** *Suppose $n \to \infty$ and $N/n \to \infty$. Let P be any rank-n orthogonal projection in $\mathbb{R}^N$, let $a \in \mathbb{R}^N$ satisfy $\|a\|_2 \le C\sqrt{n}$ for a fixed C, and let $z \sim N(0, I_N - P)$. Put $\tau_n = \frac{1}{8}\sqrt{\log(N/n)}$. There is a deterministic sequence $\varepsilon_n \to 0$, depending only on n, N, C, such that for every P, a, i,*

$$\mathbb{P}\Big(a_i + z_i > \max_{j \neq i}(a_j + z_j)\Big) \le \mathbf{1}\{|a_i| > \tau_n\} + \varepsilon_n.$$

Proof. Write $z = (I_N - P)g$, where $g \sim N(0, I_N)$, and set $w = a - Pg$. Then $a + z = g + w$ and, since $\|Pg\|^2 \sim \chi^2_n$, $\|w\|^2 \le C'n$ except with probability e^{-cn}. Let $r = \sqrt{\log(N/n)}$ and $B_N = \#\{j : g_j > r\}$. For all sufficiently large n, $r \ge 1$ and

$$\mathbb{E}B_N = N\mathbb{P}(g_1 > r) \ge c\frac{\sqrt{Nn}}{r}.$$

A Chernoff bound gives $B_N \ge \mathbb{E}B_N/2$ except with probability $\exp(-c\sqrt{Nn}/r)$. On this event, if every coordinate of $g + w$ were below $r/2$, then

$$\|w\|^2 \ge \frac{r^2}{4} B_N \ge cr\sqrt{Nn} > C'n$$

for all sufficiently large n, a contradiction. Hence $\max_j(a_j + z_j) \ge r/2$ except on the two events just bounded. If $|a_i| \le r/8$, then $a_i + z_i \le r/4$ except with probability at most $e^{-r^2/128}$, because $\mathrm{Var}(z_i) \le 1$. Thus coordinate i cannot be the maximum on the intersection of these events. One may take, for sufficiently large n,

$$\varepsilon_n = e^{-cn} + \exp(-c\sqrt{Nn}/r) + e^{-r^2/128} \to 0.$$

The argument does not assume independence of the coordinates of z, or independence between g and w. □

Theorem C.12 (Theorem 2 of the paper)**.** *Fix $\alpha > 1$, $1/2 < \beta < 1$, and a fixed deterministic positive sequence satisfying*

$$c_\sigma a^{-\beta} \le \sigma_a \le C_\sigma a^{-\beta}\ (a \ge 1)$$

for constants $0 < c_\sigma \le C_\sigma < \infty$. Fix $0 < \epsilon < (\beta - 1/2)/2$ and $\Gamma > 1/(1-\beta)$. Let $n \to \infty$, and let $N = N(n)$ and $d = d(n)$ be integer sequences satisfying, eventually,

$$n^{\,1+(1-\beta+\epsilon)/\alpha} \le N \le n^{\,1+(1/2-\epsilon)/\alpha},\ n^{\,1/(1-\beta)} \le d \le n^{\Gamma}.$$

Then, on the same realized dataset,

$$\big(\mathrm{Acc}(W_C), \mathrm{Acc}(W_M), \mathrm{Acc}(W_G)\big) \xrightarrow{\mathbb{P}} (1, 0, 0).$$

Proof. Let $\mathcal{E}_X$ denote $\{\|\mathcal{H}\mathcal{J}^{-1}\mathcal{H}\|_{\mathrm{op}} \le Cn\}$. Lemma C.1 and Lemma C.5 give

$$\mathbb{P}(\mathcal{E}_Y^c) \to 0, \qquad \sup_{Y \in \mathcal{E}_Y} \mathbb{P}_\xi(\mathcal{E}_X^c \mid Y) \to 0.$$

Both W_G and W_M have their columns in $\mathrm{col}(G)$. On $\mathcal{E}_Y \cap \mathcal{E}_X$, the exact score law and Corollary C.6 show that, for a nonzero query, the scores normalized by $\|\Sigma W y_i\|/\sqrt{d}$ have the law

$$a^{(i)}(W) + z, \qquad \|a^{(i)}(W)\|_2 \le C\sqrt{n}, \qquad z \sim N\big(0, I_N - K(K^\top K)^{-1}K^\top\big).$$

The query is zero with probability zero; under the zero-denominator convention, such a query is not recovered in any case. The projection $K(K^\top K)^{-1}K^\top$ has rank n.

For GD, Lemma C.11 bounds conditional recovery probability by $\mathbf{1}\{|a_i^{(i)}(W_G)| > \tau_n\} + \varepsilon_n$ on the good events. For any fixed $u > 0$, eventually $\tau_n \ge u$. Lemma C.7 therefore implies, for every fixed $b \in (0,1)$,

$$\sup_{i \ge bN} \mathbb{P}\{i \text{ is recovered by } W_G\} \to 0.$$

Consequently

$$\mathbb{E}[\mathrm{Acc}(W_G)] \le b + \frac{1}{N} + \sup_{i \ge bN} \mathbb{P}\{i \text{ is recovered by } W_G\}.$$

Send $n \to \infty$ and then $b \downarrow 0$. Markov's inequality gives $\mathrm{Acc}(W_G) \xrightarrow{\mathbb{P}} 0$.

For Muon, for any fixed $u > 0$ and eventually $\tau_n \ge u$,

$$\frac{1}{N}\#\{i : |a_i^{(i)}(W_M)| > \tau_n\} \le \frac{1}{N}\#\{i : |a_i^{(i)}(W_M)| > u\} \xrightarrow{\mathbb{P}} 0$$

by Lemma C.8 and $\mathbb{P}(\mathcal{E}_Y^c) \to 0$. This fraction lies in $[0,1]$, so its expectation also tends to zero. Averaging Lemma C.11 over i on $\mathcal{E}_Y \cap \mathcal{E}_X$, and adding the probabilities of the complementary events, yields

$$\mathbb{E}[\mathrm{Acc}(W_M)] \to 0, \qquad \mathrm{Acc}(W_M) \xrightarrow{\mathbb{P}} 0.$$

For the chain, fix $b \in (0,1)$ and a deterministic $i \ge bN$. Lemma C.4 implies that, conditional on $\mathcal{E}_Y$, with probability tending to one uniformly over such i,

$$t_i^\top J_Y y_i \ge c_b h, \qquad \max_{j \ne i} |t_j^\top J_Y y_i| \le \frac{c_b}{4} h.$$

Fix a query realization satisfying this margin. Lemma C.9 then implies

$$m_{ii}(W_C) - \max_{j \ne i} m_{ji}(W_C) \ge \frac{c_b}{2}\frac{s_1}{d} h$$

except on an observed-gradient event whose conditional probability tends to zero uniformly in Y and deterministic i. Indeed, the ratio of its mean-approximation error to $(s_1/d)h$ is

$$\frac{\sqrt{ns_2}\, L^{\alpha+2}}{s_1 h} \le Cn^{-\epsilon} L^{\alpha+2} \to 0.$$

Intersect this event with the row-norm event in the same lemma, whose conditional failure probability also tends to zero uniformly in Y. On that intersection,

$$\frac{\|\Sigma W_C y_i\|_2}{\sqrt{d}} \le \frac{\|y_i\|_2}{\sqrt{d}} \left(\sum_{a \le d} \sigma_a^2 \|(W_C)_{a,:}\|_2^2\right)^{1/2} \le C\frac{\sqrt{ns_2}}{d} L^{\alpha+1}.$$

Now condition on (Y, G). Lemma C.10 shows that every centered score is Gaussian with variance bounded by the square of the last display. A union bound over $j \le N$ gives

$$\mathbb{P}\left(\max_{j \le N} |S_{ji}(W_C) - m_{ji}(W_C)| > C\frac{\sqrt{ns_2}}{d} L^{\alpha+3/2} \,\middle|\, Y, G\right) \to 0$$

uniformly over the stated good events. The fluctuation threshold is $o((s_1/d)h)$ by the same scaling calculation. Therefore the correct score is strictly greater than all wrong scores with probability tending to one. Combining the query, observed-gradient, and residual-score bounds yields

$$\sup_{i \geq bN} \mathbb{P}\{i \text{ is not recovered by } W_C\} \to 0.$$

This is a bound uniform over deterministic i, not a claim that a single union bound recovers every query. It suffices because

$$\mathbb{E}[1 - \mathrm{Acc}(W_C)] \leq b + \frac{1}{N} + \sup_{i \geq bN} \mathbb{P}\{i \text{ is not recovered by } W_C\}.$$

Send $n \to \infty$ and then $b \downarrow 0$, and use Markov's inequality to obtain $\mathrm{Acc}(W_C) \xrightarrow{\mathbb{P}} 1$.

All three updates are defined on the same realized dataset. A union bound over their three accuracy deviation events proves

$$\big(\mathrm{Acc}(W_C), \mathrm{Acc}(W_M), \mathrm{Acc}(W_G)\big) \xrightarrow{\mathbb{P}} (1, 0, 0).$$

Equivalently, for every fixed $\delta \in (0, 1)$,

$$\mathbb{P}\{\mathrm{Acc}(W_C) \geq 1 - \delta, \quad \max\{\mathrm{Acc}(W_M), \mathrm{Acc}(W_G)\} \leq \delta\} \to 1.$$

□

D APPENDIX: PROOFS IN SECTION 4

Proof of Proposition 5 of the paper. Write

$$\mathcal{T} = \mathcal{A}^{\circ k} \circ \mathrm{NS}_{\mathrm{slice}} \circ \mathcal{A}^{\circ k}, \qquad \mathcal{A} = \mathrm{AN}_{\mathrm{layer}} \circ \mathrm{AN}_{\mathrm{col}} \circ \mathrm{AN}_{\mathrm{row}}.$$

We first verify the chained-LMO representation. For an axis $a \in \{\mathrm{layer}, \mathrm{row}, \mathrm{col}\}$, let $\mathcal{F}_a$ denote the fibers along that axis and define

$$\mathcal{K}_a = \{Z : \|Z_f\|_2 \le 1 \text{ for every } f \in \mathcal{F}_a\}.$$

Since the Frobenius inner product decomposes over fibers,

$$\max_{Z \in \mathcal{K}_a} \langle M, Z \rangle = \sum_{f \in \mathcal{F}_a} \|M_f\|_2,$$

and a maximizer is obtained by normalizing every nonzero fiber independently. Hence AN_a is, up to the conventional sign, an exact LMO over $\mathcal{K}_a$.

Similarly, with

$$\mathcal{K}_{\mathrm{slice}} = \{Z : \|Z_{\ell,:,:}\|_{\mathrm{op}} \le 1 \text{ for every slice } \ell\},$$

we have

$$\max_{Z \in \mathcal{K}_{\mathrm{slice}}} \langle M, Z \rangle = \sum_{\ell} \|M_{\ell,:,:}\|_*,$$

whose maximizer is obtained by applying the matrix sign independently to each slice. Thus every constituent of $\mathcal{T}$ is an exact LMO, and $\mathcal{T}$ is a chained LMO.

It remains to show that the composition is not itself an LMO. For $G \in \mathbb{R}^{m \times n}$, let

$$\iota(G)_{1,:,:} = G, \qquad \iota(G)_{\ell,:,:} = 0 \quad (\ell > 1).$$

Row- and column-axis normalization preserve the signs and zero pattern of the nonzero slice. The subsequent layer-axis normalization acts on fibers of the form $(c, 0, \ldots, 0)$, and therefore

$$\mathcal{A}(\iota(G)) = \iota(\mathrm{Sgn}(G)).$$

Consequently, for every $k \ge 1$,

$$\mathcal{A}^{\circ k}(\iota(G)) = \iota(\mathrm{Sgn}(G)),$$

and hence

$$\mathcal{T}(\iota(G)) = \iota(\mathrm{Sgn} \circ \mathrm{msign} \circ \mathrm{Sgn}(G)).$$

By Lemma B.8, for every $m, n \ge 5$ there exists $G \in \mathbb{R}^{m \times n}$ such that

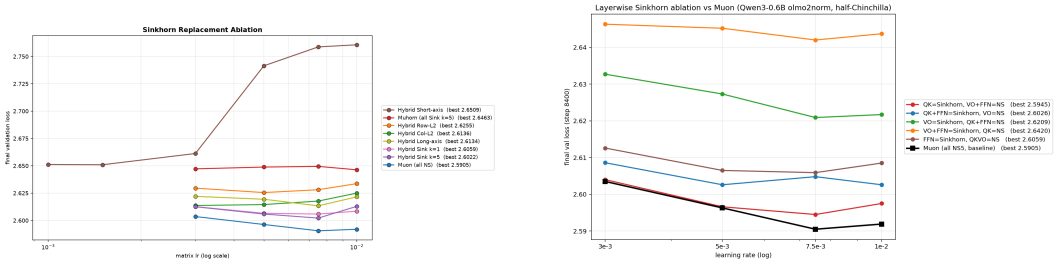
$$\langle G, \mathrm{Sgn} \circ \mathrm{msign} \circ \mathrm{Sgn}(G) \rangle < 0.$$

Hence

$$\langle \iota(G), \mathcal{T}(\iota(G)) \rangle < 0.$$

Thus TensorChain admits a negative-alignment input, and by Lemma 2, it is not a LMO of any fixed set. A smooth counterexample where TensorChain optimization diverges can be found by using Proposition 1. □

E APPENDIX: REPLACING NS WITH CHEAPER CHAINS



(a) FFN replacement ablation.

(b) Layerwise ablation.

Figure 2: Replacing Newton–Schulz with cheaper axis-wise normalizations on Qwen3-0.6B after 8.8B training tokens. (a) NS5 is retained on Q, K, V, O while the FFN uses the indicated replacement. (b) A single row–column equilibration sweep is applied to the indicated parameter classes, with NS5 retained elsewhere.

Experimental setup. We study whether inexpensive axis-wise normalizations can replace the Newton–Schulz (NS) step of Muon on the `olmo2norm` Qwen3-0.6B architecture. We use QK normalization, unfused QKV projections, Kaiming initialization, tied input/output embeddings, variable-length attention, selective activation checkpointing, and compilation. Training and validation use `fineweb_edu_350bt` and `fineweb_edu_350bt_eval`, respectively. We train for 8.8B tokens ($1\times$ Chinchilla) with sequence length 4096, global batch size 256 (1.05M tokens), and 8401 optimization steps. All runs use one node with 8 GPUs under replicated data parallelism and seed 42; validation is performed every 200 steps on one batch. Checkpoints are retained and training metrics are logged to TensorBoard.

We use 10% learning-rate warmup followed by linear decay to zero, matrix weight decay 0.1, momentum coefficient $\beta = 0.95$, and AdamW learning rate 8×10^{-3} for non-matrix parameters. All matrix updates use the same `s_mu_scale=max` scaling, $0.2\sqrt{\max(\text{fan_in}, \text{fan_out})}$. Muon uses five Newton–Schulz iterations (NS5). For each variant, we sweep the matrix learning rate over $\{3, 5, 7.5, 10\} \times 10^{-3}$ and report the best validation loss. For short-axis normalization, whose best point lies at the lower boundary of the sweep, we additionally evaluate 1×10^{-3} and 1.5×10^{-3}.

Replacing NS on the FFN. Our primary motivation is computational: the large FFN matrices account for the largest share of the NS computation, making them a natural target for a cheaper replacement. We therefore retain NS5 on the attention projections Q, K, V, O and replace NS on the FFN matrices W_1, W_2, W_3.

We compare row and column ℓ_2 normalization, normalization along the long or short matrix axis, and alternating row–column equilibration. For the latter, we test both a single sweep ($k = 1$) and five repeated sweeps ($k = 5$). We additionally evaluate row normalization and repeated row–column equilibration as replacements for NS on all matrix parameters. Among the FFN replacements, row–column equilibration performs best, while increasing from one to five sweeps yields only a small improvement. We therefore fix $k = 1$ for the subsequent weight-wise ablation.

Which weights can use the cheaper replacement? Having identified a single row–column sweep as the strongest inexpensive replacement, we next vary which parameter classes use it while retaining NS5 on all remaining matrices. We consider the FFN alone, Q, K alone, V, O alone, and combinations of these groups. The strongest configuration replaces NS only on Q, K, reaching validation loss 2.5945 compared with 2.5905 for Muon. Replacing additional parameter classes consistently degrades performance, and no configuration outperforms Muon.

Overall, cheap axis-wise normalization can approach NS when used selectively, but none of the tested replacements matches Muon. The small difference between one and five row–column sweeps also suggests that simply iterating equilibration further does not close the gap to spectral normalization.

Table 5: Selected Newton–Schulz replacement results on Qwen3-0.6B after 8.8B training tokens. Each entry reports the best result over the matrix learning-rate sweep. Δ denotes the difference from Muon.

Variant	Val. loss	Δ
Muon (NS5 on all matrices)	2.5905	–
NS5 on QKVO, replacement on FFN		
Row–column, $k = 5$	2.6022	+0.012
Row–column, $k = 1$	2.6059	+0.015
Long-axis ℓ_2	2.6134	+0.023
Column ℓ_2	2.6136	+0.023
Row ℓ_2	2.6255	+0.035
Short-axis ℓ_2	2.6509	+0.060
Row–column ($k = 1$) on listed weights, NS5 elsewhere		
Q, K	2.5945	+0.004
Q, K, FFN	2.6026	+0.012
FFN	2.6059	+0.015
V, O	2.6209	+0.030
V, O, FFN	2.6420	+0.052
All matrices (row–column, $k = 5$)	2.6463	+0.056

F Appendix: Raw Experiment Results

This section provides the complete experimental record underlying the comparisons in the main text. We report the full Qwen3-0.6B baseline sweeps under WSD and cosine schedules, the Muon learning-rate sweeps used for tuning, the full train/validation trajectories and endpoint losses, the TensorChain grouping ablation, and the cosine token-efficiency analysis. Unless otherwise stated, the experimental configuration follows Section 5.1. We retain the cosine results in addition to the main WSD experiments to document schedule-dependent differences in optimizer ranking and learning-rate sensitivity.

F.1 Full Qwen3-0.6B baseline comparison

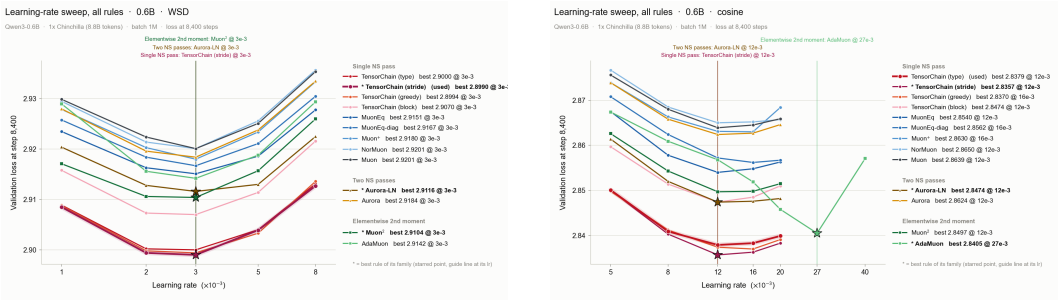
The two tables below report the complete common learning-rate grids for the Qwen3-0.6B, 1M-token-batch, $1\times$ Chinchilla baseline comparison. TensorChain grouping variants are reported separately in Section F.5. For AdaMuon, the learning rate grid was insufficient so we ran two additional experiments, $lr = 27e-3$ and $40e-3$ and could find that the optimal learning rate is $27e-3$ with val loss 2.8405, which is strictly worse than TensorChain's optimal val loss 2.8379.

Table 6: Full Qwen3-0.6B, 1M-token-batch, $1\times$ Chinchilla baseline sweep under WSD. Entries are final validation losses; methods are ordered by their best validation loss, and bold marks the best learning rate for each method.

Method	1×10^{-3}	2×10^{-3}	3×10^{-3}	5×10^{-3}	8×10^{-3}
TensorChain	2.9089	2.9002	**2.9000**	2.9040	2.9131
Muon[2]	2.9171	2.9106	**2.9104**	2.9157	2.9260
Aurora-LN	2.9204	2.9128	**2.9116**	2.9130	2.9225
AdaMuon	2.9290	2.9156	**2.9142**	2.9188	2.9294
MuonEq	2.9235	2.9163	**2.9151**	2.9186	2.9278
MuonEq-diag	2.9258	2.9184	**2.9167**	2.9211	2.9305
Muon+	2.9279	2.9203	**2.9180**	2.9234	2.9334
Aurora	2.9280	2.9196	**2.9184**	2.9238	2.9335
Muon	2.9299	2.9224	**2.9201**	2.9251	2.9354
NorMuon	2.9295	2.9214	**2.9201**	2.9256	2.9357

Table 7: Full Qwen3-0.6B, 1M-token-batch, $1\times$ Chinchilla baseline sweep under cosine decay. Entries are final validation losses; methods are ordered by their best validation loss, and bold marks the best learning rate for each method.

Method	5×10^{-3}	8×10^{-3}	12×10^{-3}	16×10^{-3}	20×10^{-3}
TensorChain	2.8501	2.8409	**2.8379**	2.8383	2.8399
AdaMuon	2.8674	2.8609	2.8568	2.8519	**2.8458**
Aurora-LN	2.8614	2.8520	**2.8474**	2.8476	2.8482
Muon[2]	2.8626	2.8543	**2.8497**	2.8498	2.8515
MuonEq	2.8675	2.8578	**2.8540**	2.8548	2.8563
MuonEq-diag	2.8708	2.8624	2.8572	**2.8562**	2.8567
Aurora	2.8739	2.8658	**2.8624**	2.8627	2.8646
Muon+	2.8739	2.8663	2.8631	**2.8630**	2.8684
Muon	2.8756	2.8680	**2.8639**	2.8645	2.8659
NorMuon	2.8767	2.8685	**2.8650**	2.8652	2.8657



(a) WSD learning rate tuning loss.

(b) COS learning rate tuning

Figure 3: Full Qwen3-0.6B baseline comparison.

F.2 MUON LEARNING-RATE SWEEPS

We report each tuning setup separately for Muon and use it for all baselines. The results are below.

Table 8: Muon learning-rate sweep for Qwen3-0.6B, 1M-token batch, 1× Chinchilla under WSD.

Matrix LR	1×10^{-3}	2×10^{-3}	3×10^{-3}	5×10^{-3}	8×10^{-3}
Final val. loss	2.9299	2.9224	**2.9201**	2.9251	2.9354

Table 9: Muon learning-rate sweep for Qwen3-1.7B, 4M-token batch, 1× Chinchilla under WSD.

Matrix LR	0.5×10^{-3}	1×10^{-3}	2×10^{-3}	4×10^{-3}	8×10^{-3}
Final val. loss	2.6878	2.6672	2.6523	**2.6493**	2.6589

Table 10: Muon learning-rate sweep for Qwen3-0.6B, 1M-token batch, 1× Chinchilla under cosine decay.

Matrix LR	1×10^{-3}	2×10^{-3}	3×10^{-3}	5×10^{-3}	8×10^{-3}	12×10^{-3}	16×10^{-3}	20×10^{-3}
Final val. loss	2.9201	2.8993	2.8857	2.8756	2.8680	**2.8639**	2.8645	2.8659

Table 11: Muon learning-rate sweep for Qwen3-0.6B, 4M-token batch, 1× Chinchilla under cosine decay.

Matrix LR	8×10^{-3}	12×10^{-3}	18×10^{-3}	27×10^{-3}	40×10^{-3}	60×10^{-3}
Final val. loss	2.9650	2.9568	2.9516	2.9462	**2.9459**	2.9468

Table 12: Muon learning-rate sweep for Qwen3-1.7B, 1M-token batch, 1× Chinchilla under cosine decay.

Matrix LR	4×10^{-3}	6×10^{-3}	9×10^{-3}	13.5×10^{-3}	20×10^{-3}
Final val. loss	2.6019	**2.5974**	2.5987	2.6042	2.6139

Table 13: Muon learning-rate sweep for Qwen3-1.7B, 4M-token batch, 1× Chinchilla under cosine decay.

Matrix LR	5×10^{-3}	8×10^{-3}	10×10^{-3}	15×10^{-3}	20×10^{-3}
Final val. loss	2.6228	2.6120	2.6091	**2.6047**	2.6049

F.3 WSD TRAINING AND VALIDATION TRAJECTORIES

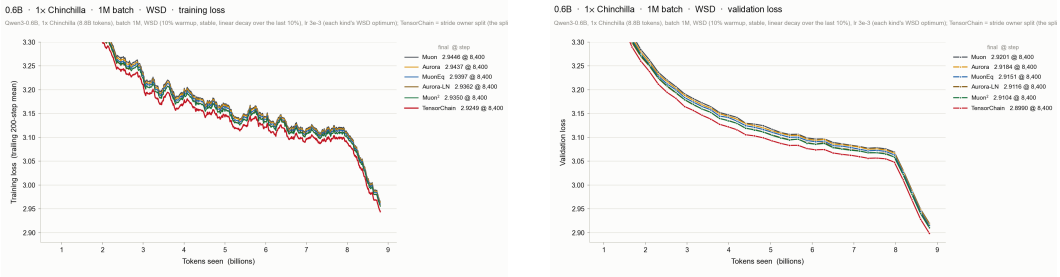
The following two tables aggregate the final losses across WSD setups. Since all methods within a setup use the same selected learning rate, the learning rate is omitted here.

Table 14: Final training loss for the WSD trajectory experiments. Bold marks the best method in each row.

Setup	Muon	Aurora	MuonEq	Muon^2	Aurora-LN	TensorChain
0.6B, 1M, 1×	2.9446	2.9437	2.9397	2.9350	2.9362	**2.9249**
1.7B, 4M, 4×	2.5254	2.5244	2.5226	2.5208	2.5188	**2.5181**

Table 15: Final validation loss for the WSD trajectory experiments. Bold marks the best method in each row.

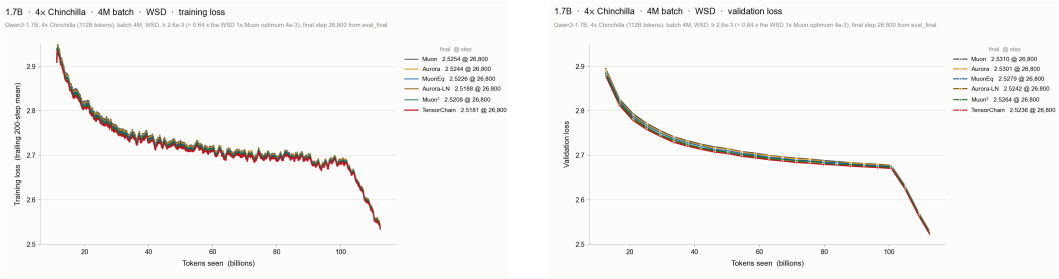
Setup	Muon	Aurora	MuonEq	Muon^2	Aurora-LN	TensorChain
0.6B, 1M, 1×	2.9201	2.9184	2.9151	2.9104	2.9116	**2.8990**
1.7B, 4M, 4×	2.5310	2.5301	2.5279	2.5264	2.5242	**2.5236**



(a) Training loss.

(b) Validation loss.

Figure 4: WSD trajectories for Qwen3-0.6B, 1M-token batch, 1× Chinchilla.



(a) Training loss. (b) Validation loss.

Figure 5: WSD trajectories for Qwen3-1.7B, 4M-token batch, $4\times$ Chinchilla.

F.4 Cosine training and validation trajectories

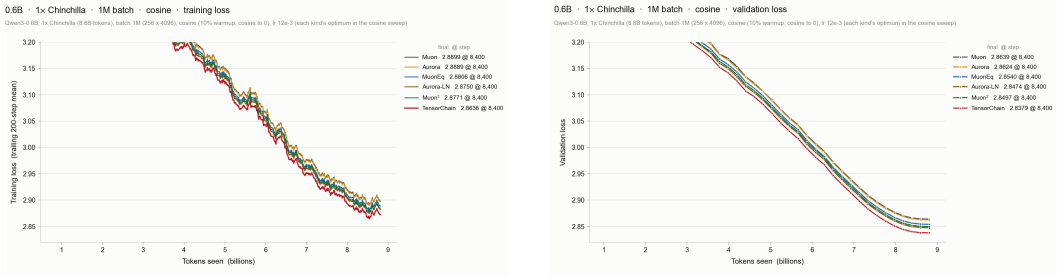
We analogously aggregate the four cosine settings. Again, learning rates are omitted because all methods within a given setup use the same selected learning rate.

Table 16: Final training loss for the cosine trajectory experiments. Bold marks the best method in each row.

Setup	Muon	Aurora	MuonEq	Muon^2	Aurora-LN	TensorChain
0.6B, 1M, $1\times$	2.8899	2.8889	2.8806	2.8771	2.8750	**2.8636**
0.6B, 4M, $4\times$	2.7934	2.7897	2.7845	2.7851	2.7804	**2.7748**
1.7B, 1M, $1\times$	2.6036	2.6017	2.5978	2.5941	2.5942	**2.5922**
1.7B, 4M, $4\times$	2.3859	2.3850	2.3814	2.3780	**2.3751**	2.3778

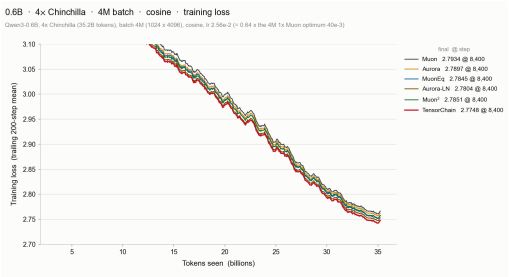
Table 17: Final validation loss for the cosine trajectory experiments. Bold marks the best method in each row.

Setup	Muon	Aurora	MuonEq	Muon^2	Aurora-LN	TensorChain
0.6B, 1M, $1\times$	2.8639	2.8624	2.8540	2.8497	2.8474	**2.8379**
0.6B, 4M, $4\times$	2.7527	2.7487	2.7434	2.7443	2.7391	**2.7344**
1.7B, 1M, $1\times$	2.5974	2.5964	2.5924	2.5885	2.5882	**2.5857**
1.7B, 4M, $4\times$	2.4823	2.4810	2.4784	2.4744	**2.4734**	2.4747

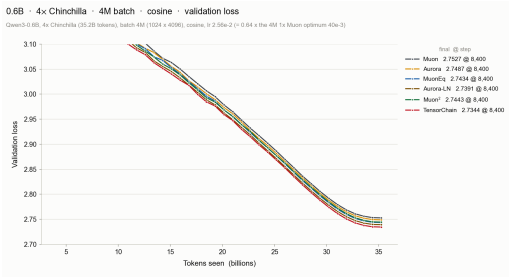


(a) Training loss. (b) Validation loss.

Figure 6: Cosine trajectories for Qwen3-0.6B, 1M-token batch, $1\times$ Chinchilla.

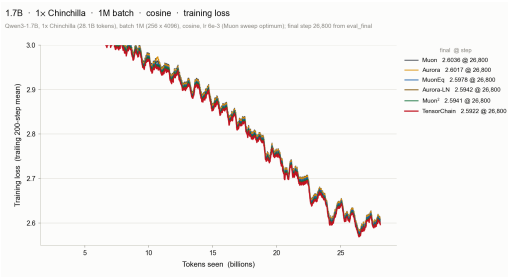


(a) Training loss.

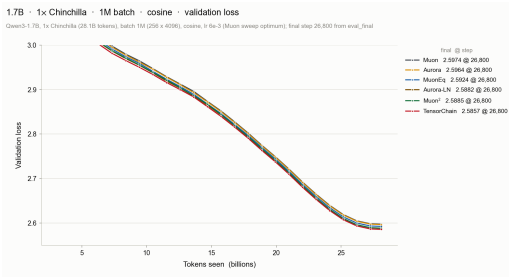


(b) Validation loss.

Figure 7: Cosine trajectories for Qwen3-0.6B, 4M-token batch, $4\times$ Chinchilla.

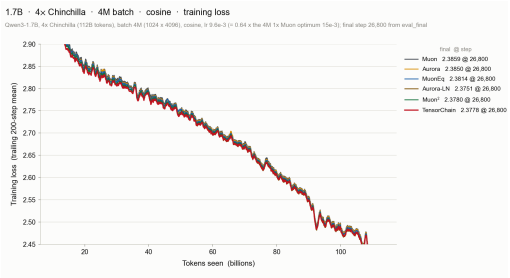


(a) Training loss.

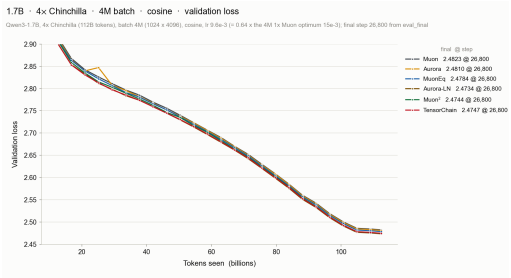


(b) Validation loss.

Figure 8: Cosine trajectories for Qwen3-1.7B, 1M-token batch, $1\times$ Chinchilla.



(a) Training loss.



(b) Validation loss.

Figure 9: Cosine trajectories for Qwen3-1.7B, 4M-token batch, $4\times$ Chinchilla.

F.5 TensorChain grouping ablation

TensorChain requires choosing which compatible matrices are grouped along the layer axis. For a 28-layer model, *consecutive* forms four contiguous groups, e.g., $\{1{:}7\}, \{8{:}14\}, \{15{:}21\}, \{22{:}28\}$, whereas *stride* interleaves depth, e.g., $\{1, 5, 9, \ldots, 25\}, \{2, 6, 10, \ldots, 26\}, \{3, 7, 11, \ldots, 27\}, \{4, 8, 12, \ldots, 28\}$. *Greedy* assigns compatible matrices to parameter-balanced groups, while *all* jointly groups all compatible matrices $\{1, 2, \cdots 28\}$. The results are extracted from the same Qwen3-0.6B full sweeps above.

Table 18: TensorChain grouping ablation on Qwen3-0.6B, 1M-token batch, $1\times$ Chinchilla. Each entry reports the best point from the corresponding full sweep.

	WSD		Cosine	
Grouping	Best LR	Val. loss	Best LR	Val. loss
Consecutive	3×10^{-3}	2.9070	12×10^{-3}	2.8474
Stride	3×10^{-3}	2.8990	12×10^{-3}	2.8357
Greedy	3×10^{-3}	2.8994	16×10^{-3}	2.8370
All	3×10^{-3}	2.9000	12×10^{-3}	2.8379

Consecutive grouping performs worse near its optimum than stride, greedy, and all-layer grouping. Since the latter three mix matrices across more distant depths despite using different assignment rules, this suggests that cross-depth mixing matters more than the precise partition.

F.6 Cosine token efficiency

The cosine schedule yields smooth validation trajectories throughout training, so we additionally compare methods by the number of training tokens required to reach the same validation loss. We do not use this metric for WSD because its terminal decay concentrates much of the loss reduction near the end of training, making within-run crossing times less informative.

For TensorChain and a baseline B, define the common target

$$\ell^\star(\mathrm{TC}, B) = \max\{\ell_{\mathrm{TC}}(D_{\max}), \ell_B(D_{\max})\},$$

the worse of the two final validation losses. Both methods therefore reach the target within the observed training horizon, so no extrapolation is required. For a method M, let

$$D_M(\ell) = \inf\Big\{D : \widehat{\ell}_M(D) \le \ell\Big\},$$

where $\widehat{\ell}_M(D)$ is obtained by linear interpolation between the two adjacent validation checkpoints bracketing the target. We define

$$\mathrm{TE}(\mathrm{TC}; B) = 100\left(1 - \frac{D_{\mathrm{TC}}(\ell^\star)}{D_B(\ell^\star)}\right)\%.$$

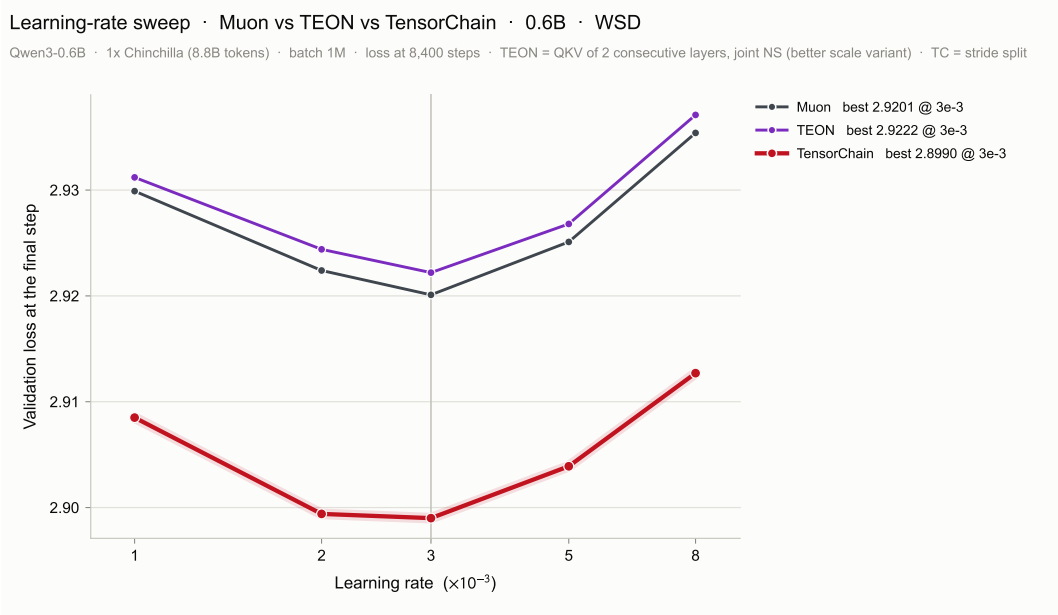
Positive values indicate that TensorChain reaches the same validation loss using fewer training tokens, while negative values favor the baseline. Since each pairwise comparison fixes the batch size, the same percentage is equivalently the reduction in optimization steps.

Table 19: Cosine-schedule token-efficiency gain of TensorChain. Each entry is the percentage reduction in training tokens required to reach the same validation-loss target. Positive values favor TensorChain.

Setting	Aurora-LN	Muon^2	MuonEq	Aurora	Muon
0.6B, 1M, $1\times$	8.2	9.0	10.3	12.4	12.8
0.6B, 4M, $4\times$	5.4	7.6	7.4	8.9	10.0
1.7B, 1M, $1\times$	3.9	4.0	6.2	7.5	7.8
1.7B, 4M, $4\times$	-2.0	-0.5	6.7	7.4	7.8

F.7 Comparison with the Teon optimizer

Since we could not find an open-sourced implementation of Teon (Zhang et al., 2026a), we implemented our own version of Teon. We concatenated the two consecutive Q, K, V weights and applied Newton–Schulz, following the suggestion of (Zhang et al., 2026a). We tuned the learning rate of Teon in Qwen3 0.6B pretraining setup, with 1M batch size and 1x chinchilla tokens, with WSD schedule with 10% warmup and decay. To our surprise, the Teon optimizer performed worse than Muon for our setup; The best validation loss for Teon was 2.9222, whereas Muon was 2.9201. Hence we decided to not include Teon in our baselines.


Learning-rate sweep · Muon vs TEON vs TensorChain · 0.6B · WSD
Qwen3-0.6B · 1x Chinchilla (8.8B tokens) · batch 1M · loss at 8,400 steps · TEON = QKV of 2 consecutive layers, joint NS (better scale variant) · TC = stride split
Muon best 2.9201 @ 3e-3
TEON best 2.9222 @ 3e-3
TensorChain best 2.8990 @ 3e-3
Validation loss at the final step
2.93
2.92
2.91
2.90
1
2
3
5
8
Learning rate (×10⁻³)