Practical Muon Accelerates Projected Feature Learning in Scaling-Law Models

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

We ask whether optimizer geometry can improve feature-learning scaling laws in the two-layer linear student–teacher model of Bordelon et al. [3]. Starting from their projected feature-learning dynamics, we replace the feature update with practical Newton–Schulz Muon while preserving the finite-width projection bottleneck. As a proxy theory, we derive a projected partial-polar population flow and show that it removes the small projected-gradient norm factor inherited by projected SGD on hard source-condition tasks. Empirically, practical NS_5 Muon yields positive fixed-compute width-scaling exponents on hard tasks, whereas projected SGD has negative fixed-compute width exponents over the same sweep. Readout-Hessian diagnostics show a smaller quantile-gap decay exponent, consistent with stronger low-rank feature alignment. These results suggest that matrix orthogonalization can materially change finite-compute feature-learning scaling.

1. Introduction

Feature learning can improve neural scaling laws when the target is outside the RKHS of the initial kernel. In the projected two-layer linear model of Bordelon et al. [3], this improvement is driven by the growth of an effective dynamical kernel. We study a complementary question: if the model and finite-width projection bottleneck are fixed, can the *optimizer* further accelerate this feature-learning mechanism?

Our answer is that Muon-like matrix orthogonalization supplies a plausible and measurable acceleration mechanism [2, 6, 8]. The population projected-feature descent direction is rank one. Projected SGD therefore scales its update by the singular value of this rank-one direction, while a partial-polar Muon proxy removes that singular value. The resulting same-state vector-field comparison predicts a late-time advantage when the projected residual gradient becomes small. We test this prediction with practical Newton–Schulz Muon at finite compute.

Contributions. We formulate a projected partial-polar proxy that preserves the BAP finite-width bottleneck, derive its local acceleration ratio over projected SGD, and show that practical Muon produces positive fixed-compute width exponents on hard tasks together with stronger readout-Hessian spectral diagnostics.

2. Model and Baseline

We use the student–teacher setup of Bordelon et al. [3]. Let $\boldsymbol{\psi}_\infty(x) \in \mathbb{R}^M$ have diagonal covariance $\mathbf{\Lambda}$. A width-N student is

$$f(x,t) = \frac{1}{N}\mathbf{w}(t)^\top \mathbf{A}(t)\boldsymbol{\psi}_\infty(x), \qquad y(x) = \mathbf{w}^\star \cdot \boldsymbol{\psi}_\infty(x), \tag{1}$$

with residual

$$\mathbf{v}^0(t) = \mathbf{w}^\star - \frac{1}{N}\mathbf{A}(t)^\top \mathbf{w}(t). \tag{2}$$

The source-capacity conditions are [3, 4]

$$\lambda_k \sim k^{-\alpha}, \qquad \sum_{\ell>k} \lambda_\ell (w^\star_\ell)^2 \sim k^{-\alpha\beta}, \qquad \alpha > 1. \tag{3}$$

We call $\beta < 1$ hard because the target lies outside the initial RKHS in the BAP scaling convention, and $\beta > 1$ easy. Projected SGD updates the readout and feature matrix by

$$\mathbf{w}(t+1) - \mathbf{w}(t) = \eta\, \mathbf{A}(t)\frac{\mathbf{\Psi}_\infty(t)^\top \mathbf{\Psi}_\infty(t)}{B}\mathbf{v}^0(t), \tag{4}$$

$$\mathbf{A}(t+1) - \mathbf{A}(t) = \eta\gamma_{\text{fl}}\, \mathbf{w}(t)\mathbf{v}^0(t)^\top \frac{\mathbf{\Psi}_\infty(t)^\top \mathbf{\Psi}_\infty(t)}{B} P_0, \qquad P_0 = \frac{1}{N}\mathbf{A}(0)^\top \mathbf{A}(0). \tag{5}$$

The fixed random Gram matrix P_0 has rank at most N and is the finite-width feature bottleneck. In the population limit, define

$$\mathbf{q}(t) = P_0 \mathbf{\Lambda}\mathbf{v}^0(t). \tag{6}$$

The projected feature gradient becomes

$$G_A(t) = \gamma_{\text{fl}}\, \mathbf{w}(t)\mathbf{q}(t)^\top. \tag{7}$$

For hard tasks, BAP's projected feature-learning exponent is

$$\chi_{\text{BAP}}(\beta) = \frac{2\beta}{1+\beta}, \qquad K_{\text{eff}}(t) \sim t^{a_{\text{BAP}}}, \qquad a_{\text{BAP}} = \frac{1-\beta}{1+\beta}. \tag{8}$$

Thus the relevant comparison is Muon versus an already feature-learning projected-SGD baseline, not Muon versus lazy training [5].

3. Projected Polar Mechanism

Practical Muon applies a Newton–Schulz zeroth-power map to a matrix momentum buffer [6**?** , 7]. To isolate its deterministic geometric effect, we study the zero-momentum exact-polar population proxy applied only to $\mathbf{A}$:

$$\dot{\mathbf{w}}(t) = \eta\, \mathbf{A}(t)\mathbf{\Lambda}\mathbf{v}^0(t), \tag{9}$$

$$\dot{\mathbf{A}}_{\text{pMuon}}(t) = \eta \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}\frac{(P_0\mathbf{\Lambda}\mathbf{v}^0(t))^\top}{\|P_0\mathbf{\Lambda}\mathbf{v}^0(t)\|}. \tag{10}$$

Equation (10) preserves the projected right direction but removes the singular value of (7), namely $\gamma_{\mathrm{fl}} \|\mathbf{w}(t)\| \left\|P_0 \mathbf{\Lambda} \mathbf{v}^0(t)\right\|$. For the spike diagnostic

$$m(t) = \frac{1}{N} \frac{\mathbf{w}(t)^\top \mathbf{A}(t) \mathbf{w}^\star}{\|\mathbf{w}^\star\|_{\mathbf{\Lambda}}}, \qquad \|\mathbf{w}^\star\|_{\mathbf{\Lambda}}^2 = (\mathbf{w}^\star)^\top \mathbf{\Lambda} \mathbf{w}^\star, \tag{11}$$

the readout contribution is the same for both methods. The feature contributions at a fixed state therefore satisfy

$$R_{II}(t) := \frac{\dot{m}_A^{\mathrm{pMuon}}(t)}{\dot{m}_A^{\mathrm{BAP}}(t)} = \frac{1}{\gamma_{\mathrm{fl}} \|\mathbf{w}(t)\| \, \|P_0 \mathbf{\Lambda} \mathbf{v}^0(t)\|}. \tag{12}$$

This is a local vector-field ratio, not a trajectory domination theorem. It predicts a late-time polar advantage when easy modes have been learned and the projected residual-gradient norm is small.

Conjecture 1 (Polar growth advantage) *For hard tasks $0 < \beta < 1$, if $P_0 \mathbf{\Lambda} \mathbf{v}^0(t)$ remains nontrivially aligned with the projected teacher direction, then the projected partial-polar flow has effective kernel growth $K_{\mathrm{eff,pMuon}}(t) \sim t^{a_{\mathrm{pMuon}}(\beta)}$ with $a_{\mathrm{pMuon}}(\beta) > a_{\mathrm{BAP}}(\beta)$ in the polar-dominant regime of (12). Consequently $\chi_{\mathrm{pMuon}} = \beta(1 + a_{\mathrm{pMuon}}) > \chi_{\mathrm{BAP}}$.*

The supplement gives the full spike calculation, the relation to practical momentum Muon, and the extreme-hard-task blindspot of this rank-one proxy.

4. Experiments and Diagnostics

We run projected population-gradient experiments for 10,000 steps with widths from 32 to 2048, five seeds per width, and source exponents $\beta \in \{0.2, 0.5, 0.8, 1.2, 1.5\}$. The completed grid contains 525 runs: SGD, AdamW [9], and Muon across all widths, seeds, and source exponents. All optimizers preserve the BAP feature projection. Muon uses NS_5 on the feature matrix with zero momentum, square-root aspect-ratio update scaling, and SGD on the readout. We fit $\mathcal{L}(N, t_0) \sim N^{-\chi_{\mathrm{width}}}$ at fixed compute; positive χ_{width} means loss improves with width. For the hard tasks $\beta = 0.2, 0.5, 0.8$, the BAP asymptotic reference exponents $2\beta/(1+\beta)$ are $0.33, 0.67, 0.89$, while Muon's fitted fixed-compute exponents are $0.42, 0.92, 1.62$. We treat this as finite-compute evidence rather than an asymptotic theorem for practical Muon; the bootstrap plot, full numerical table, AdamW baseline, leave-one-width-out checks, and intervals are in the supplement. For spectral diagnostics we use the conditional readout Hessian block

$$H_{ww}(t) = \frac{1}{N^2} \mathbf{A}(t) \mathbf{\Lambda} \mathbf{A}(t)^\top. \tag{13}$$

If training forms an aligned low-rank component in $\mathbf{A}(t)$, the top eigenvalue of this weighted covariance should separate from the background. Because $\mathbf{\Lambda}$ has a power-law spectrum, this is only BBP-like in spirit [1]; we use the top spectral gap as an empirical outlier diagnostic rather than an exact threshold theorem. The supplementary Hessian-gap plot shows that Muon has a consistently larger gap than projected SGD at fixed width and a smaller fitted gap-decay exponent, matching the prediction of stronger low-rank feature alignment. The normalized H_{ww} trace scales close to the expected N^{-2} law for both methods, so we use it mainly as a normalization check rather than as the differentiating alignment diagnostic.

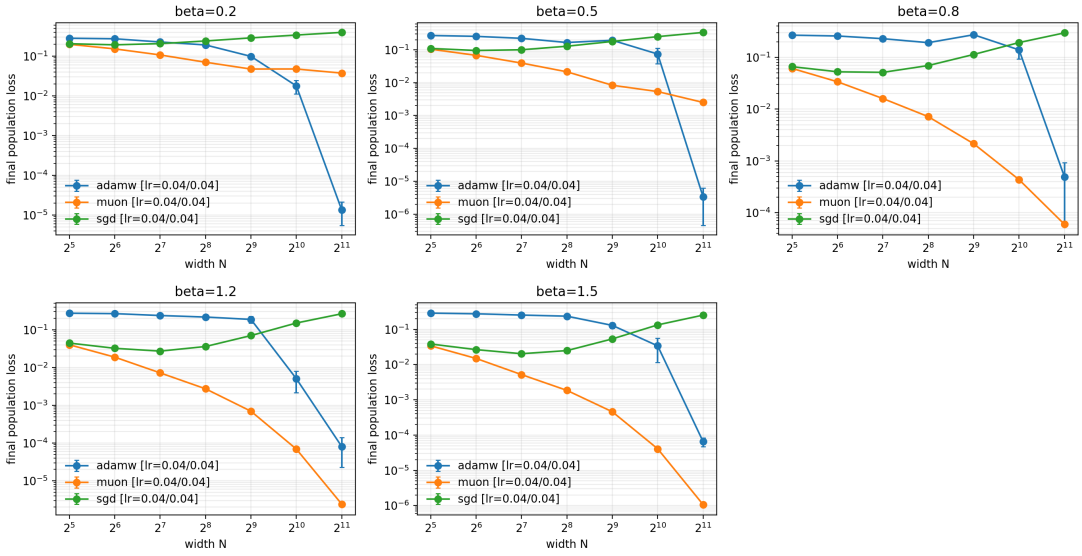

Figure 1: Fixed-compute final loss versus width for projected population dynamics. Practical NS_5 Muon improves with width on hard tasks, whereas projected SGD has negative fixed-compute width exponents over the same sweep.

5. Discussion

The theoretical and empirical claims have different scope. The theory is an idealized rank-one population proxy that explains how polar normalization can remove the small singular value of the projected feature gradient. The experiments use practical Newton–Schulz Muon at finite compute, where the optimizer differs from the proxy through finite-step Newton–Schulz orthogonalization, update scaling, and, in broader settings, momentum or stochastic gradients. The defensible conclusion is therefore mechanism plus finite-compute validation: the proxy explains why a Muon-like update should help, and the practical NS_5 experiments show robust positive fixed-compute width scaling under the BAP projection.

References

[1] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697, 2005.

[2] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *NeurIPS 2024 Workshop on Optimization for Machine Learning*, 2024. URL https://openreview.net/forum?id=ux18f5nOpD.

[3] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):084002, 2025.

[4] Andrea Caponnetto and Ernesto De Vito. Fast rates for regularized least-squares algorithm. *MIT CSAIL Technical Report*, 2005.

[5] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[7] Zdislav Kovarik. Some iterative methods for improving orthonormality. *SIAM Journal on Numerical Analysis*, 7(3):386–389, 1970.

[8] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025. doi: 10.48550/arXiv.2502.16982.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Appendix A. Supplementary Theory

This appendix records the derivations omitted from the five-page main text. All quantities use the same notation as the main paper. The fixed matrix $P_0 = N^{-1}\mathbf{A}(0)^\top\mathbf{A}(0)$ is never dropped; it is the finite-width Gram bottleneck throughout.

A.1. Population Projected-SGD Spike ODE

In the population limit, $B^{-1}\mathbf{\Psi}_\infty(t)^\top\mathbf{\Psi}_\infty(t) \to \mathbf{\Lambda}$, so BAP's projected continuous-time feature dynamics have rank-one form

$$\dot{\mathbf{w}}(t) = \eta\,\mathbf{A}(t)\mathbf{\Lambda}\mathbf{v}^0(t), \tag{14}$$

$$\dot{\mathbf{A}}_{\mathrm{BAP}}(t) = \eta\gamma_{\mathrm{fl}}\,\mathbf{w}(t)\mathbf{v}^0(t)^\top\mathbf{\Lambda}P_0 = \eta\gamma_{\mathrm{fl}}\,\mathbf{w}(t)\mathbf{q}(t)^\top, \tag{15}$$

where $\mathbf{q}(t) = P_0\mathbf{\Lambda}\mathbf{v}^0(t)$. Differentiating

$$m(t) = \frac{1}{N}\frac{\mathbf{w}(t)^\top\mathbf{A}(t)\mathbf{w}^\star}{\|\mathbf{w}^\star\|_{\mathbf{\Lambda}}} \tag{16}$$

gives

$$\dot{m}(t) = \frac{1}{N\,\|\mathbf{w}^\star\|_{\mathbf{\Lambda}}}\left[\dot{\mathbf{w}}(t)^\top\mathbf{A}(t)\mathbf{w}^\star + \mathbf{w}(t)^\top\dot{\mathbf{A}}(t)\mathbf{w}^\star\right]. \tag{17}$$

Using the standard directional approximation $N^{-1}\mathbf{A}(t)^\top \mathbf{A}(t)\mathbf{w}^\star \approx \mathbf{w}^\star$ along the teacher direction, the readout term is

$$\dot{m}_w(t) \approx \frac{\eta}{\|\mathbf{w}^\star\|_{\mathbf{\Lambda}}} \mathbf{v}^0(t)^\top \mathbf{\Lambda} \mathbf{w}^\star. \tag{18}$$

The projected-SGD feature contribution is

$$\dot{m}_A^{\mathrm{BAP}}(t) = \frac{\eta\gamma_{\mathrm{fl}} \|\mathbf{w}(t)\|^2}{N \|\mathbf{w}^\star\|_{\mathbf{\Lambda}}} \mathbf{q}(t)^\top \mathbf{w}^\star \tag{19}$$

$$= \frac{\eta\gamma_{\mathrm{fl}} \|\mathbf{w}(t)\|^2}{N \|\mathbf{w}^\star\|_{\mathbf{\Lambda}}} \mathbf{v}^0(t)^\top \mathbf{\Lambda} P_0 \mathbf{w}^\star. \tag{20}$$

Thus

$$\dot{m}_{\mathrm{BAP}}(t) \approx \frac{\eta}{\|\mathbf{w}^\star\|_{\mathbf{\Lambda}}} \mathbf{v}^0(t)^\top \mathbf{\Lambda} \mathbf{w}^\star + \frac{\eta\gamma_{\mathrm{fl}} \|\mathbf{w}(t)\|^2}{N \|\mathbf{w}^\star\|_{\mathbf{\Lambda}}} \mathbf{v}^0(t)^\top \mathbf{\Lambda} P_0 \mathbf{w}^\star. \tag{21}$$

The approximation is directional, not an operator-norm claim that $N^{-1}\mathbf{A}^\top \mathbf{A}$ is close to the identity when $N < M$.

A.2. Projected Partial-Polar Spike ODE

The rank-one feature gradient has singular value $\gamma_{\mathrm{fl}} \|\mathbf{w}(t)\| \|\mathbf{q}(t)\|$. The exact partial-polar proxy replaces it by

$$\mathrm{ppolar}(G_A(t)) = \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} \frac{\mathbf{q}(t)^\top}{\|\mathbf{q}(t)\|}. \tag{22}$$

Substitution into the product rule gives the same readout term and the feature term

$$\dot{m}_A^{\mathrm{pMuon}}(t) = \frac{\eta}{N \|\mathbf{w}^\star\|_{\mathbf{\Lambda}}} \mathbf{w}(t)^\top \left(\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} \frac{\mathbf{q}(t)^\top}{\|\mathbf{q}(t)\|} \right) \mathbf{w}^\star \tag{23}$$

$$= \frac{\eta \|\mathbf{w}(t)\|}{N \|\mathbf{w}^\star\|_{\mathbf{\Lambda}}} \frac{\mathbf{v}^0(t)^\top \mathbf{\Lambda} P_0 \mathbf{w}^\star}{\|P_0 \mathbf{\Lambda} \mathbf{v}^0(t)\|}. \tag{24}$$

The ratio of feature contributions is therefore

$$\frac{\dot{m}_A^{\mathrm{pMuon}}(t)}{\dot{m}_A^{\mathrm{BAP}}(t)} = \frac{1}{\gamma_{\mathrm{fl}} \|\mathbf{w}(t)\| \|P_0 \mathbf{\Lambda} \mathbf{v}^0(t)\|}. \tag{25}$$

The numerator is not assumed to be perfectly aligned. Equivalently,

$$\frac{\mathbf{q}(t)^\top \mathbf{w}^\star}{\|\mathbf{q}(t)\|} = \|\mathbf{w}^\star\| \cos\theta_{\mathrm{proj}}(t), \tag{26}$$

where $\theta_{\mathrm{proj}}(t)$ is the angle between the projected residual gradient and the teacher direction. The proxy advantage requires this angular factor not to vanish too quickly.

A.3. Practical Muon versus the Rank-One Proxy

Practical Muon does not polarize the instantaneous population gradient. For a matrix parameter it forms a buffer

$$B_A(t) = \mu B_A(t-1) + G_A(t), \tag{27}$$

applies a fixed Newton–Schulz map $O_A(t) = \mathrm{NS}_5(B_A(t))$, and updates $\mathbf{A}(t+1) = \mathbf{A}(t) + \eta O_A(t)$ up to sign convention. Even if each population gradient $G_A(t)$ is rank one, the buffer $B_A(t) = \sum_{s\le t} \mu^{t-s} G_A(s)$ need not be rank one. Practical Muon can therefore be richer than the deterministic proxy, especially in stochastic mini-batch training where

$$\mathbb{E}[\mathrm{polar}(G_{\mathrm{batch}})] \neq \mathrm{polar}(\mathbb{E}[G_{\mathrm{batch}}]). \tag{28}$$

A.4. Extreme-Hard-Task Blindspot

At initialization, typically $\|\mathbf{w}(0)\| \sim \sqrt{N}$ and $\left\|P_0 \mathbf{\Lambda} \mathbf{v}^0(0)\right\| = \mathcal{O}(1)$ on the projected teacher component, so the polar feature contribution need not dominate immediately. Using BAP's hard-task asymptotics as a state proxy near crossover gives

$$\|\mathbf{w}(t)\| \left\|P_0 \mathbf{\Lambda} \mathbf{v}^0(t)\right\| \approx t^{\frac{1-3\beta}{2(1+\beta)}} \tag{29}$$

up to projection-dependent constants. This predicts late-time pMuon dominance when $\beta > 1/3$. For $\beta < 1/3$, the rank-one deterministic argument is inconclusive; empirical gains there should be interpreted as evidence for behavior beyond this proxy, not as a contradiction.

Appendix B. Supplementary Hessian Diagnostics

The parameter vector is $\theta = (\mathbf{w}, \mathrm{vec}(\mathbf{A}))$. The loss is not jointly quadratic because

$$\mathbf{v}^0 = \mathbf{w}^\star - \frac{1}{N}\mathbf{A}^\top \mathbf{w} \tag{30}$$

is bilinear. Conditional Hessian blocks are nevertheless useful diagnostics. Ignoring a harmless factor of two from the convention for the squared loss,

$$H_{ww}(t) = \frac{1}{N^2}\mathbf{A}(t)\mathbf{\Lambda}\mathbf{A}(t)^\top, \qquad H_{AA}(t) = \frac{1}{N^2}\mathbf{w}(t)\mathbf{w}(t)^\top \otimes \mathbf{\Lambda}. \tag{31}$$

The feature block has eigenvalues

$$\left\{ \frac{\|\mathbf{w}(t)\|^2}{N^2}\lambda_j \right\}_{j=1}^{M} \cup \{0\}^{N(M-1)}. \tag{32}$$

Thus H_{AA} is highly degenerate and shaped by $\mathbf{\Lambda}$; it should not be described as having a Marchenko–Pastur random bulk. The more natural BBP-like diagnostic is the readout block H_{ww}. As a diagnostic ansatz, write

$$\mathbf{A}(t) = \mathbf{A}_{\mathrm{noise}}(t) + s(t)u(t)\tilde{v}(t)^\top. \tag{33}$$

Then

$$H_{ww}(t) \approx \frac{1}{N^2}\mathbf{A}_{\mathrm{noise}}(t)\mathbf{\Lambda}\mathbf{A}_{\mathrm{noise}}(t)^\top + \frac{s(t)^2}{N^2}\big(\tilde{v}(t)^\top \mathbf{\Lambda}\tilde{v}(t)\big)u(t)u(t)^\top + \text{cross terms}. \tag{34}$$

When the aligned component is sufficiently large relative to the weighted background, the top eigenvalue should separate. Because $\mathbf{\Lambda}$ is power-law, we use BBP language only as a spiked-covariance heuristic.

Appendix C. Supplementary Empirical Details

The main projected population sweep uses $N \in \{32, 64, 128, 256, 512, 1024, 2048\}$, five seeds per width, source exponents $\beta \in \{0.2, 0.5, 0.8, 1.2, 1.5\}$, population gradients, BAP feature projection, learning rates $\eta_w = \eta_A = 0.04$, zero Muon momentum, NS_5, and square-root aspect-ratio update scaling. The main text reports seed-averaged log–log fits for final loss and readout-Hessian spectral gap. The implementation logs the projected-gradient norm $\left\|P_0 \mathbf{\Lambda} \mathbf{v}^0\right\|$, the local-ratio proxy R_{II}, and readout-Hessian quantities hww_gap_q, hww_top_eig, and $hww\_trace\_over\_n$.

	Width exponent χ_{width}			Gap exponent δ_{gap}		
β	SGD	Muon	AdamW	SGD	Muon	AdamW
0.2	-0.18	0.42	1.86	0.91	0.88	2.63
0.5	-0.30	0.92	1.88	0.94	0.86	2.41
0.8	-0.41	1.62	1.03	0.96	0.86	1.81
1.2	-0.48	2.20	1.68	0.98	0.86	1.70
1.5	-0.51	2.34	1.55	0.98	0.85	1.71

Table 1: Seed-averaged power-law fits from the main projected population sweep (7 widths, 35 runs per optimizer/source exponent). AdamW is included as an optimizer baseline; the theory-matched comparison is projected SGD versus practical Muon.

β	Muon χ_{width}	leave-one-width-out range	bootstrap 95% CI
0.2	0.421	$[0.407, 0.457]$	$[0.408, 0.432]$
0.5	0.917	$[0.895, 0.964]$	$[0.900, 0.934]$
0.8	1.623	$[1.393, 1.801]$	$[1.598, 1.648]$
1.2	2.201	$[1.759, 2.479]$	$[2.172, 2.233]$
1.5	2.336	$[1.859, 2.624]$	$[2.315, 2.362]$

Table 2: Seed-bootstrap and leave-one-width-out checks for Muon's fixed-compute width exponents. The intervals remain positive across all measured source exponents.

For projected SGD, the hard-task seed-bootstrap intervals remain negative:

β	0.2	0.5	0.8
SGD bootstrap 95% CI	$[-0.189, -0.168]$	$[-0.320, -0.289]$	$[-0.426, -0.391]$

This gives a clean separation between projected SGD and practical Muon at fixed compute.

The normalized readout-Hessian trace exponent is close to 2 for both Muon and SGD: for $\beta = 0.2, 0.5, 0.8, 1.2, 1.5$, Muon gives $1.984, 1.981, 1.979, 1.974, 1.971$, while projected SGD gives $2.010, 2.015, 2.017, 2.019, 2.019$. This supports the stated normalization of H_{ww} but is not the primary alignment diagnostic. The learning-rate sweep over $\beta = 0.5, 0.8$, width 256, two seeds, and $\eta_w, \eta_A \in \{0.005, 0.01, 0.02, 0.04\}$ supports the main choice $\eta_w = \eta_A = 0.04$; it is used as tuning evidence rather than as the main scaling result.

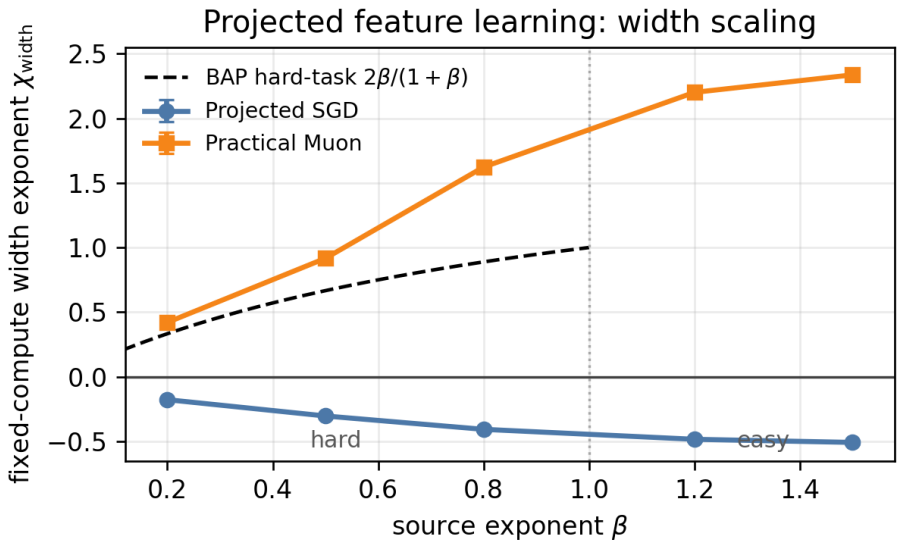

Figure 2: Fixed-compute width exponent with seed-bootstrap uncertainty. Practical Muon has positive fitted width exponents across the sweep, while projected SGD has negative exponents. The dashed curve is BAP's hard-task asymptotic reference $2\beta/(1+\beta)$; easy-task points are finite-compute empirical behavior rather than an asymptotic proxy-theory prediction.

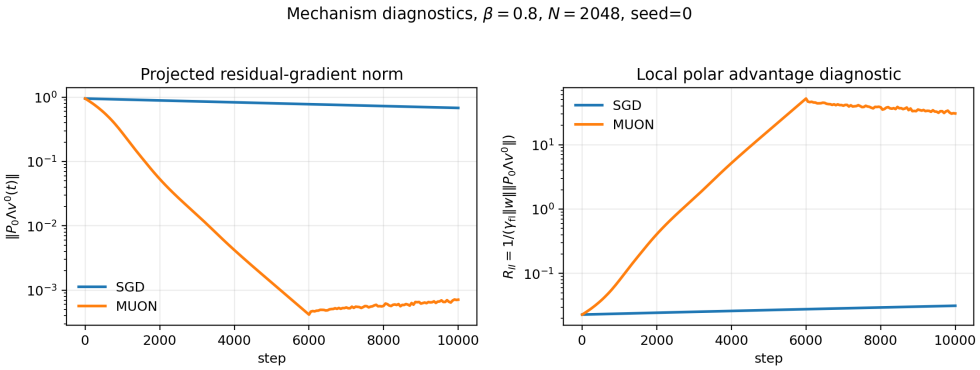

Figure 3: Mechanism diagnostic for $\beta = 0.8$, $N = 2048$, seed 0. Muon rapidly reduces the projected residual-gradient norm $\left\|P_0\mathbf{\Lambda}\mathbf{v}^0(t)\right\|$, making the same-state ratio $R_{II} = 1/(\gamma_{\text{fl}}\|\mathbf{w}\|\left\|P_0\mathbf{\Lambda}\mathbf{v}^0\right\|)$ large. Projected SGD does not enter this polar-dominant regime over the same horizon.

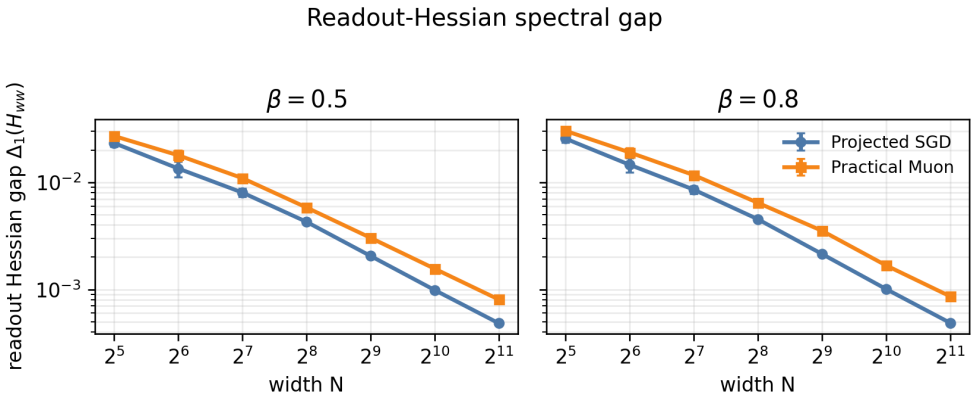

Figure 4: Readout-Hessian spectral gap for two hard tasks. Muon has a larger gap at fixed width and a smaller fitted decay exponent than projected SGD, consistent with stronger low-rank feature alignment in the readout block H_{ww}.

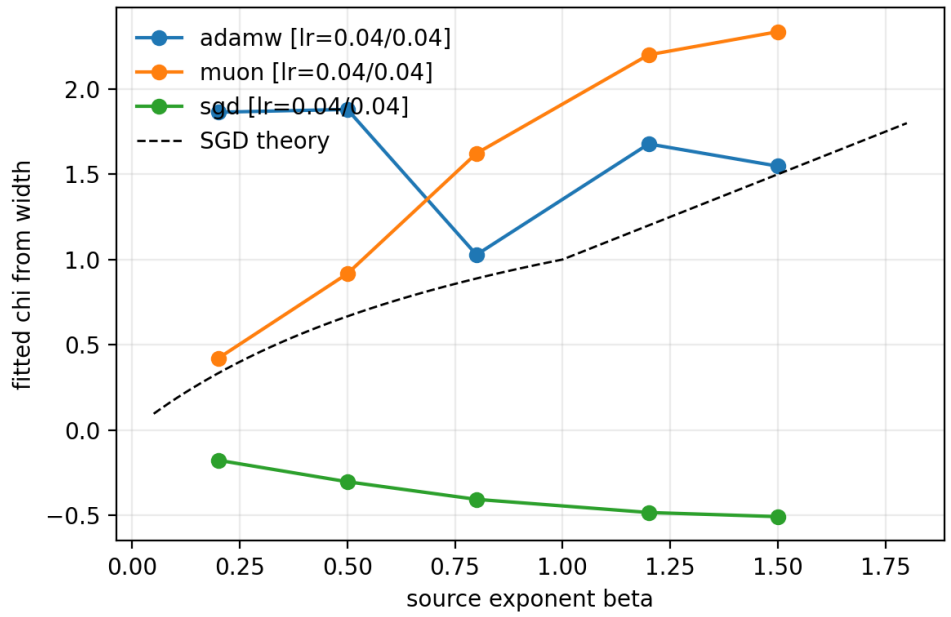

Figure 5: Supplementary view of fixed-compute width exponents as a function of the source exponent β.

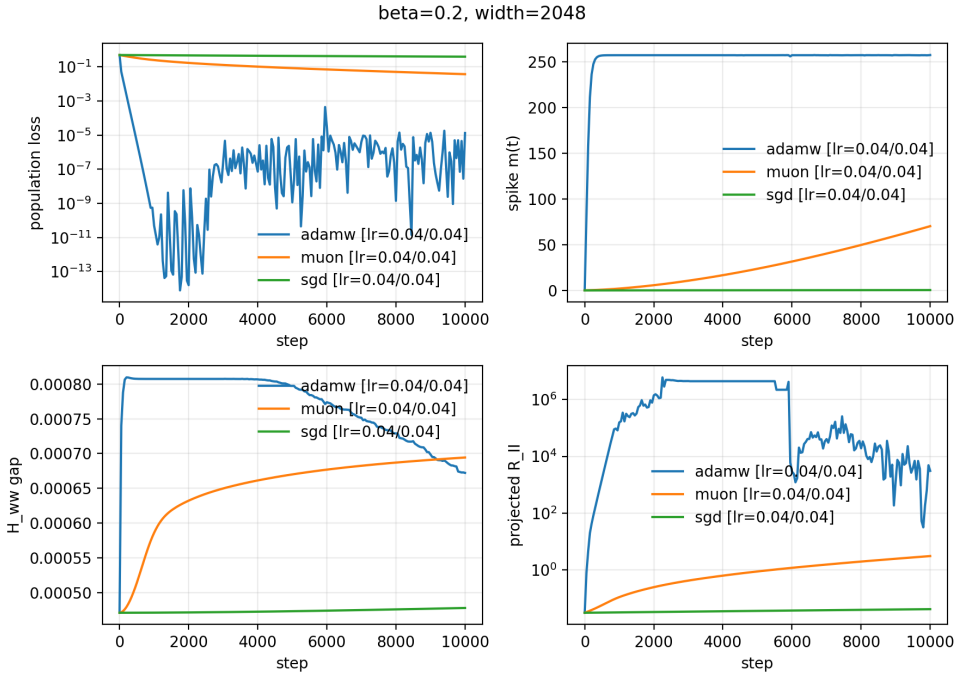

Figure 6: Representative time series for the hardest measured task ($\beta = 0.2$) at width $N = 2048$.