Rank-One Spectral Normalization Accelerates Projected Feature Learning

Chryseis Xinyi Liu CHRYSEISLIU@GMAIL.COM
Independent Researcher

Manuel Paez MANUEL.PAEZ@COLUMBIA.EDU
Columbia University

Abstract

We ask whether optimizer-induced matrix normalization can improve feature-learning scaling laws in the projected two-layer linear student–teacher model of Bordelon et al. [3]. In this model the population feature-gradient direction is rank one after applying the fixed finite-width projection bottleneck. Projected SGD therefore scales its feature update by the singular value of this rank-one matrix, while polar normalization removes that singular-value factor. We show that zero-momentum practical NS_5 Muon reduces exactly to this polar direction up to a fixed scalar in the rank-one population setting. Empirically, practical NS_5 Muon yields positive fixed-compute width-scaling exponents where projected SGD has negative exponents. Mechanism controls inspired by recent Freon and Kaon critiques of Muon-specific geometry [11] show that exact polar updates, Frobenius-normalized gradients, norm-matched random spectra, and Freon with $c = 1/2$ reproduce Muon's projected-population scaling once this scalar is matched. Thus the evidence supports a broad rank-one spectral-normalization mechanism, with practical Muon as an efficient implementation, rather than a Muon-unique geometric effect.

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?

This question is motivated by the recent success of Muon, a hidden-layer optimizer that applies Newton–Schulz matrix zeroth-power iterations to matrix-shaped updates [1, 7, 9]. A common interpretation is that Muon helps because it approximates a particular non-Euclidean or polar geometry. At the same time, Shumaylov et al. [11] challenge this uniqueness story by introducing Freon, a family of Schatten/quasi-norm spectral updates, and Kaon, a control that preserves gradient singular vectors while replacing singular values by random spectra. Their results suggest that the exact target geometry may be less important than broader spectral reshaping and step-size effects.

Our answer in the projected feature-learning model is that rank-one spectral normalization supplies a plausible and measurable acceleration mechanism. 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 polar-normalized update removes that singular value. Practical zero-momentum NS_5 Muon implements the same direction up to a fixed scalar in this setting. The resulting same-state

© C.X. Liu & M. Paez.

vector-field comparison predicts a late-time advantage when the projected residual gradient becomes small. We test this prediction with practical Newton–Schulz Muon and with spectral controls at finite compute.

Contributions. We formulate a projected partial-polar update that preserves the BAP finite-width bottleneck, derive its local acceleration ratio over projected SGD, and show that zero-momentum NS_5 Muon reduces to this update up to a scalar in the rank-one population setting. We then use polar, Frobenius-normalized, random-spectrum, Freon, and truncation controls to identify which parts of the Muon-like update are actually responsible for the observed finite-compute width-scaling improvement. The resulting claim is deliberately narrower and stronger: projected feature learning is accelerated by rank-one spectral normalization, not by a Muon-unique geometry.

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 $\boldsymbol{\Lambda}$. 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

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

We explicitly define the loss as $\mathcal{L}(t) = \frac{1}{2}\mathbf{v}^0(t)^\top \boldsymbol{\Lambda}\mathbf{v}^0(t)$ to ensure gradient consistency throughout the continuous-time dynamics. The source-capacity conditions are [3, 4]

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

Equivalently, $\lambda_k (w_k^\star)^2 \sim k^{-(\alpha\beta+1)}$, so $(w_k^\star)^2 \sim k^{-\alpha(\beta-1)-1}$. Thus the initial-kernel RKHS norm $\sum_k (w_k^\star)^2$ is finite for $\beta > 1$ and infinite for $\beta < 1$. 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

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

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

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

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

The projected feature gradient becomes

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

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

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

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

3. Rank-One Spectral Normalization Mechanism

Practical Muon applies a Newton–Schulz zeroth-power map to a matrix momentum buffer [2, 6–8]. 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 $\gamma_{\text{fl}}\, \|\mathbf{w}(t)\|\, \left\|P_0\mathbf{\Lambda}\mathbf{v}^0(t)\right\|$.

Remark (Feature-Rate Normalization). Polar normalization removes the internal singular-value factor $\gamma_{\text{fl}}\, \|\mathbf{w}(t)\|\, \left\|P_0\mathbf{\Lambda}\mathbf{v}^0(t)\right\|$ from the projected feature update. It does not remove all update-scale dependence: the external learning rate, aspect-ratio scaling, finite NS_5 scalar, and momentum still determine the realized discrete dynamics.

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

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

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.

Heuristic 1 (Polar Growth Advantage). For hard tasks $0 < \beta < 1$, the same-state ratio (12) suggests a polar advantage on intervals where $P_0\mathbf{\Lambda}\mathbf{v}^0(t)$ remains aligned with the projected teacher direction and $\gamma_{\text{fl}}\, \|\mathbf{w}(t)\|\, \left\|P_0\mathbf{\Lambda}\mathbf{v}^0(t)\right\|$ is small. We therefore expect faster effective-kernel growth in this polar-dominant regime, but we do not claim a proved global asymptotic inequality over the full coupled trajectory. The supplement gives the 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, source exponents $\beta \in \{0.2, 0.5, 0.8, 1.2, 1.5\}$, and multiple seeds per width. The comparison includes projected SGD, AdamW [10], practical NS_5 Muon, exact polar updates, Frobenius-normalized gradients, norm-matched random-spectrum updates, Freon with $c = 1/2$, and top-singular-value truncation. Freon and the random-spectrum control are included to connect this controlled feature-learning setting with the Freon/Kaon evidence of Shumaylov et al. [11]: they test whether Muon's exact singular-value target is necessary, while preserving the singular-vector directions supplied by the gradient. Frobenius-normalized gradients divide by the Frobenius norm, random-spectrum controls preserve the singular vectors but replace singular values by

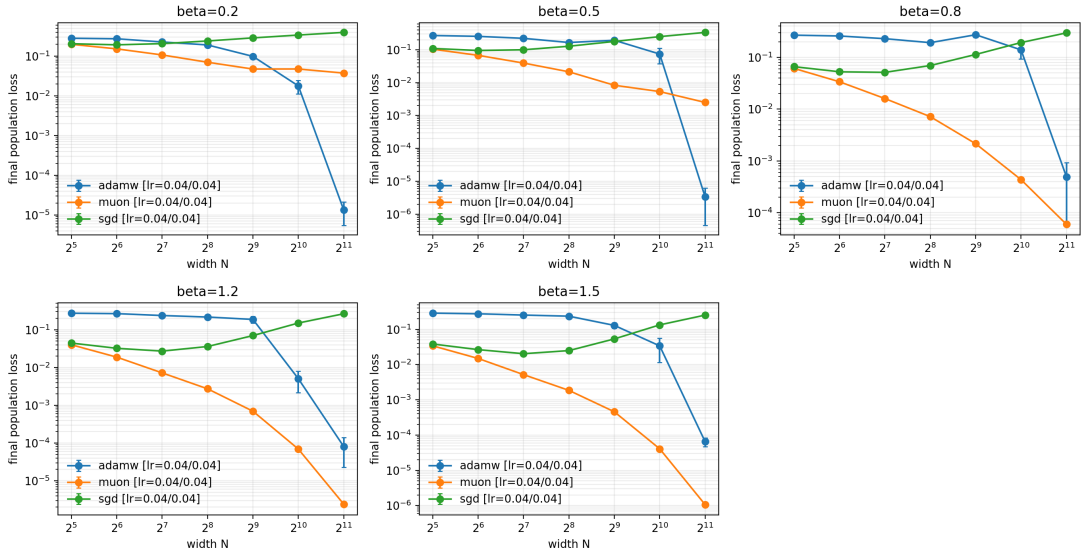

Figure 1: Fixed-compute final loss versus width for projected population dynamics. Practical NS_5 Muon improves with width on hard tasks, whereas projected SGD has negative fixed-compute width exponents over the same sweep.

norm-matched random values, Freon applies the $c = 1/2$ spectral transform, and truncation zeros the largest singular values. The linear projected model also exposes a special rank-one limit in which these distinctions can collapse. Muon uses NS_5 on the feature matrix with zero momentum, square-root aspect-ratio update scaling, and SGD on the readout. We fit $\mathcal{L}(N, t_0) \sim N^{-\chi_{\text{width}}}$ 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 approximately $0.41, 0.91, 1.62$ in the projected population sweep. We treat this as finite-compute evidence rather than an asymptotic theorem for practical Muon; bootstrap plots, the AdamW baseline, leave-one-width-out checks, and intervals are in the supplement.

The control grid gives the key identification result. In the rank-one projected population setting, zero-momentum NS_5 Muon has cosine 1 with exact polar and a fixed scalar 0.696436 under our implementation's normalization. After matching this scalar in the feature learning rate, exact polar, Frobenius-normalized gradients, random spectra with preserved singular vectors, and Freon $c = 1/2$ all produce the same fitted exponents as Muon. For $\beta = 0.2, 0.5, 0.8, 1.2, 1.5$, Muon's exponents are $0.410, 0.906, 1.616, 2.181, 2.307$, while the matched normalized controls give $0.411, 0.905, 1.598, 2.217, 2.354$. Projected SGD and top-singular-value truncation remain negative across the same sweep.

To leave the exact rank-one population limit, we also test mini-batch training with a momentum buffer, the first setting in this linear model where the matrix update source can become genuinely full rank. In a 675-run grid with momentum 0.95, batch size 512, widths 64 to 1024, and tuned SGD/AdamW/RMSProp baselines, exact polar, Frobenius-normalized, and Freon controls remain broadly competitive with practical NS_5 Muon. Muon is best or near-best for some harder source

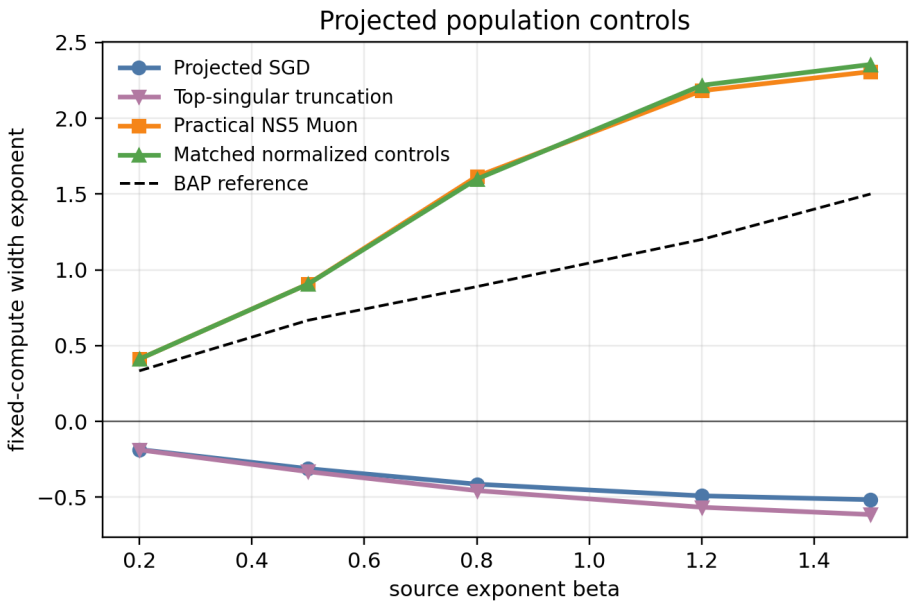

Figure 2: Projected population controls at fixed compute. Practical NS_5 Muon overlaps with norm-matched exact polar, Frobenius-normalized, random-spectrum, and Freon $c = 1/2$ controls. Projected SGD has negative width exponents, and top-singular-value truncation is a negative control: in the rank-one population setting it removes the only feature-update direction, showing that low rank alone is not sufficient. The result supports broad rank-one spectral normalization rather than a Muon-unique geometry.

exponents, but it is not uniquely separated from the broader matrix-normalization family. AdamW is a strong adaptive baseline for easy source conditions and reaches numerical-floor losses for $\beta = 0.2, 0.5$, but it is much weaker than the matrix spectral-normalization methods for $\beta \geq 0.8$. Random-spectrum controls are intermediate: preserving singular-vector subspaces plus broad normalization is powerful, but outside the rank-one limit the singular-value transform is not irrelevant.

For spectral diagnostics we use the conditional readout Hessian block

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

Because the instantaneous population feature update is rank one, it is natural to track whether a learned low-rank component separates from the weighted random background in H_{ww}. We use this only as a spiked-covariance diagnostic: the cumulative trajectory need not be a single rank-one perturbation, and the power-law covariance means that a Marchenko–Pastur threshold is not exact. Appendix Figure 4 shows larger readout-Hessian gaps for practical Muon than for projected SGD at fixed width, consistent with 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.

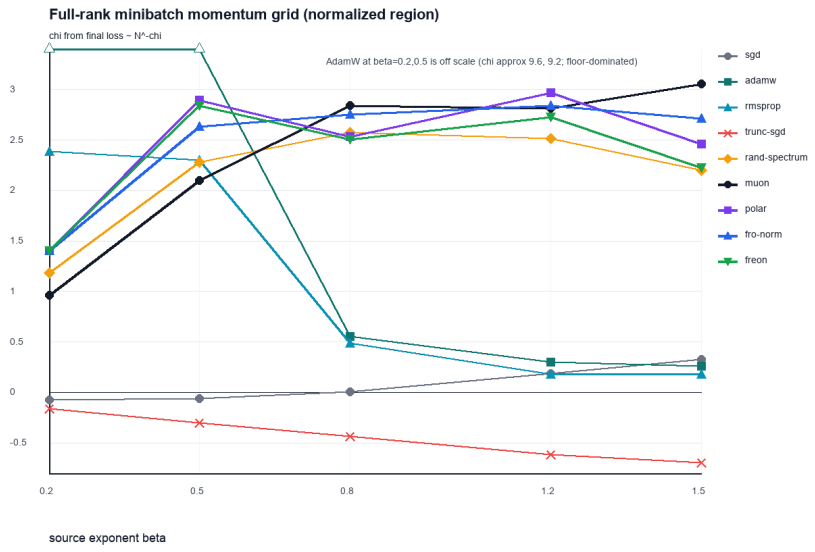

Figure 3: Full-rank mini-batch momentum controls at fixed compute. The plotted region clips AdamW's floor-dominated $\beta = 0.2, 0.5$ exponents, which are approximately 9.6 and 9.2. Across harder source exponents, exact polar, Frobenius-normalized, and Freon updates remain competitive with practical NS_5 Muon, while random-spectrum updates are intermediate and truncation remains poor.

5. Discussion

The theoretical and empirical claims have different scope. The theory is a rank-one population argument showing how polar normalization removes the small singular value of the projected feature gradient. In the zero-momentum projected population setting, practical NS_5 Muon does not introduce a distinct direction: it equals the same polar-normalized update up to a fixed scalar. The defensible conclusion is therefore mechanism plus finite-compute validation: rank-one spectral normalization accelerates projected feature learning relative to projected SGD, and practical Muon is one efficient implementation of this normalized update.

The extreme-hard regime remains a useful stress test. The deterministic ratio argument is inconclusive for $\beta < 1/3$, yet the finite-compute population experiments still show positive width scaling at $\beta = 0.2$. At present we treat this as an empirical finite-time fact rather than evidence for a second stochastic mechanism. In this linear two-layer model, mini-batch noise alone does not make the instantaneous feature-gradient matrix full rank: it still has the form $\mathrm{outer}(\mathbf{w}, \widehat{\mathbb{E}}[(f-y)x])/N$, followed by the fixed BAP right projection. Momentum is therefore the first tested setting where the update source can accumulate multiple rank-one directions into a full-rank buffer. The completed mini-batch-plus-momentum grid strengthens the main caution: practical Muon is strong, but exact polar/Freon/Frobenius controls are broadly competitive and often better, so the evidence supports broad spectral normalization rather than a Muon-unique mechanism.

References

[1] 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.

[2] Åke Björck and Clazett Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971.

[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] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008.

[7] 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/.

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

[9] 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.

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

[11] 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, 2026. URL https://arxiv.org/abs/2605.11181.

Appendix A. Author Contributions

Chryseis Xinyi Liu led the project, identified central questions, designed and implemented the optimizer controls, ran and analyzed the empirical studies, and wrote parts of the paper. Manuel Paez contributed to mathematical derivations, experimental design discussions, and part of the paper writing.

Appendix B. Supplementary Theory

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

B.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

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

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

where $\mathbf{q}(t) = P_0\mathbf{\Lambda}\mathbf{v}^0(t)$. Differentiating $m(t) = \frac{1}{N}\frac{\mathbf{w}(t)^\top \mathbf{A}(t)\mathbf{w}^\star}{\|\mathbf{w}^\star\|_{\mathbf{\Lambda}}}$ gives

$$\dot{m}(t) = \frac{1}{N\,\|\mathbf{w}^\star\|_{\mathbf{\Lambda}}}\left[\dot{\mathbf{w}}(t)^\top \mathbf{A}(t)\mathbf{w}^\star + \mathbf{w}(t)^\top \dot{\mathbf{A}}(t)\mathbf{w}^\star\right]. \tag{16}$$

Because $\mathbf{A}(0)$ is initialized with i.i.d. Gaussian entries and $\mathbf{w}^\star$ is a fixed deterministic vector independent of $\mathbf{A}$, concentration of measure yields $\mathbb{E}[\mathbf{A}(0)^\top \mathbf{A}(0)] = NI$. Thus, its action on the target vector yields $\mathbf{A}(0)^\top \mathbf{A}(0)\mathbf{w}^\star = N\mathbf{w}^\star + \mathcal{O}(\sqrt{N})$. The readout term is:

$$\dot{m}_w(t) \approx \frac{\eta}{\|\mathbf{w}^\star\|_{\mathbf{\Lambda}}}\mathbf{v}^0(t)^\top \mathbf{\Lambda}\mathbf{w}^\star. \tag{17}$$

The projected-SGD feature contribution is

$$\begin{aligned}\dot{m}_A^{\text{BAP}}(t) &= \frac{\eta\gamma_{\text{fl}}\,\|\mathbf{w}(t)\|^2}{N\,\|\mathbf{w}^\star\|_{\mathbf{\Lambda}}}\mathbf{q}(t)^\top \mathbf{w}^\star \\ &= \frac{\eta\gamma_{\text{fl}}\,\|\mathbf{w}(t)\|^2}{N\,\|\mathbf{w}^\star\|_{\mathbf{\Lambda}}}\mathbf{v}^0(t)^\top \mathbf{\Lambda} P_0\mathbf{w}^\star.\end{aligned} \tag{18}$$

Thus

$$\dot{m}_{\text{BAP}}(t) \approx \frac{\eta}{\|\mathbf{w}^\star\|_{\mathbf{\Lambda}}}\mathbf{v}^0(t)^\top \mathbf{\Lambda}\mathbf{w}^\star + \frac{\eta\gamma_{\text{fl}}\,\|\mathbf{w}(t)\|^2}{N\,\|\mathbf{w}^\star\|_{\mathbf{\Lambda}}}\mathbf{v}^0(t)^\top \mathbf{\Lambda} P_0\mathbf{w}^\star. \tag{19}$$

B.2. Projected Partial-Polar Spike ODE

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

$$\text{ppolar}(G_A(t)) = \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} \frac{\mathbf{q}(t)^\top}{\|\mathbf{q}(t)\|}. \tag{20}$$

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

$$\begin{aligned} \dot{m}_A^{\text{pMuon}}(t) &= \frac{\eta}{N \|\mathbf{w}^\star\|_{\mathbf{\Lambda}}} \mathbf{w}(t)^\top \left(\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} \frac{\mathbf{q}(t)^\top}{\|\mathbf{q}(t)\|} \right) \mathbf{w}^\star \\ &= \frac{\eta \|\mathbf{w}(t)\|}{N \|\mathbf{w}^\star\|_{\mathbf{\Lambda}}} \frac{\mathbf{v}^0(t)^\top \mathbf{\Lambda} P_0 \mathbf{w}^\star}{\|P_0 \mathbf{\Lambda} \mathbf{v}^0(t)\|}. \end{aligned} \tag{21}$$

The ratio of feature contributions is therefore

$$\frac{\dot{m}_A^{\text{pMuon}}(t)}{\dot{m}_A^{\text{BAP}}(t)} = \frac{1}{\gamma_{\text{fl}} \|\mathbf{w}(t)\| \|P_0 \mathbf{\Lambda} \mathbf{v}^0(t)\|}. \tag{22}$$

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

$$\frac{\mathbf{q}(t)^\top \mathbf{w}^\star}{\|\mathbf{q}(t)\|} = \|\mathbf{w}^\star\| \cos \theta_{\text{proj}}(t) \tag{23}$$

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

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

Practical Muon forms a buffer

$$B_A(t) = \mu B_A(t-1) + G_A(t), \tag{24}$$

applies a fixed Newton–Schulz map $O_A(t) = \text{NS}_5(B_A(t))$ and updates $\mathbf{A}(t+1) = \mathbf{A}(t) + \eta O_A(t)$ up to sign convention. In the zero-momentum population experiments, $B_A(t) = G_A(t)$ is rank one. If $G = suv^\top$ with $s > 0$ and $\|u\| = \|v\| = 1$, the implementation first normalizes by the Frobenius norm, so $X_0 = uv^\top$. The NS_5 iteration is an odd polynomial map of the form

$$X_{k+1} = aX_k + (bX_kX_k^\top + c(X_kX_k^\top)^2)X_k, \qquad (a, b, c) = (3.4445, -4.7750, 2.0315). \tag{25}$$

Therefore every iterate remains $X_k = z_k uv^\top$, with scalar recurrence

$$z_{k+1} = az_k + bz_k^3 + cz_k^5, \qquad z_0 = 1. \tag{26}$$

After five iterations this gives $z_5 \simeq 0.696436$. Thus, in the rank-one population setting used for the main Muon experiments,

$$\text{NS}_5(G_A(t)) = 0.696436 \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} \frac{(P_0 \mathbf{\Lambda} \mathbf{v}^0(t))^\top}{\|P_0 \mathbf{\Lambda} \mathbf{v}^0(t)\|}. \tag{27}$$

This is exact up to floating-point arithmetic and the implementation's aspect-ratio scaling. Hence practical NS_5 Muon and exact polar have the same trajectory after matching the feature learning rate by this scalar.

With momentum, however, 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 rank-one proxy, especially in stochastic mini-batch training where

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

B.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 C. Supplementary Hessian Diagnostics

The parameter vector is $\theta = (\mathbf{w}, \text{vec}(\mathbf{A}))$. The loss is $\mathcal{L}(t) = \frac{1}{2}\mathbf{v}^0(t)^\top \mathbf{\Lambda} \mathbf{v}^0(t)$. It 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.

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

$$H_{AA}(t) = \frac{1}{N^2}\mathbf{w}(t)\mathbf{w}(t)^\top \otimes \mathbf{\Lambda}. \tag{32}$$

The feature block has eigenvalues

$$\left\{ \frac{\|\mathbf{w}(t)\|^2}{N^2}\lambda_j \right\}_{j=1}^{M} \cup \{0\}^{N(M-1)}. \tag{33}$$

Thus H_{AA} is highly degenerate and shaped by $\mathbf{\Lambda}$; it should not be described as having a Marchenko–Pastur random bulk. Its trace, however, captures the macroscopic scaling difference $\frac{\text{Tr}(H_{AA})}{N} \propto \frac{\|\mathbf{w}(t)\|^2 \text{Tr}(\mathbf{\Lambda})}{N^3}$. Since $\text{Tr}(\mathbf{\Lambda}) = \mathcal{O}(1)$ for $\alpha > 1$, observing the trace of H_{AA} gives direct empirical insight into the accelerated growth of the readout vector under Muon.

The more natural BBP-like diagnostic is the readout block H_{ww}. As a diagnostic ansatz, write

$$\mathbf{A}(t) = \mathbf{A}_{\text{noise}}(t) + s(t)u(t)\tilde{v}(t)^\top. \tag{34}$$

Then

$$H_{ww}(t) \approx \frac{1}{N^2}\mathbf{A}_{\text{noise}}(t)\mathbf{\Lambda}\mathbf{A}_{\text{noise}}(t)^\top + \frac{s(t)^2}{N^2}(\tilde{v}(t)^\top \mathbf{\Lambda}\tilde{v}(t))u(t)u(t)^\top + \text{cross terms}. \tag{35}$$

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

Appendix D. 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 also logs the projected-gradient norm $\left\|P_0\mathbf{\Lambda}\mathbf{v}^0\right\|$, the local-ratio proxy R_{II}, and readout-Hessian gap, top-eigenvalue, and trace diagnostics.

	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.

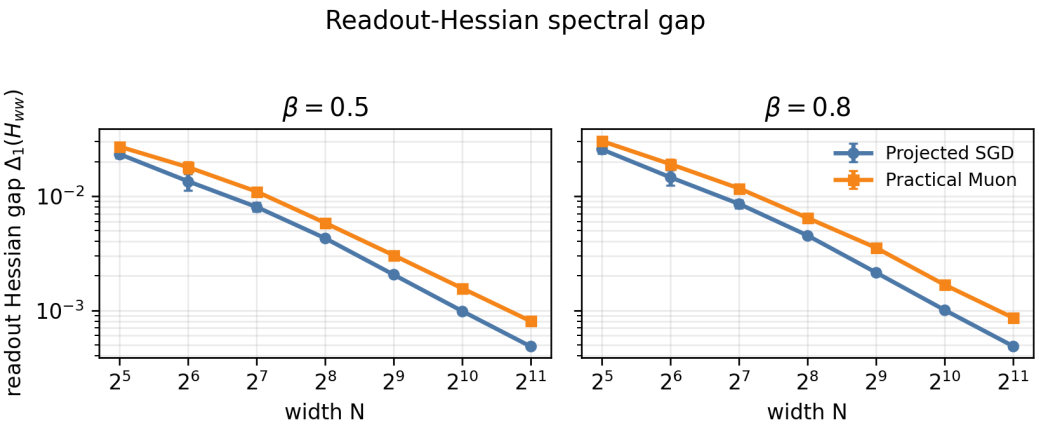

Figure 4: Readout-Hessian spectral gap diagnostic for the projected population sweep. Practical Muon has a consistently larger gap than projected SGD at fixed width for $\beta = 0.5$ and $\beta = 0.8$, supporting the interpretation that normalized matrix updates produce stronger low-rank feature alignment.

β	SGD	Trunc.	Muon	Polar	Fro-norm	Rand-spec.	Freon
0.2	-0.186	-0.188	0.410	0.411	0.411	0.411	0.411
0.5	-0.312	-0.332	0.906	0.905	0.905	0.905	0.905
0.8	-0.414	-0.457	1.616	1.598	1.598	1.598	1.598
1.2	-0.491	-0.567	2.181	2.217	2.217	2.217	2.217
1.5	-0.517	-0.615	2.307	2.354	2.354	2.354	2.354

Table 3: Projected population control sweep with three seeds per width. Polar, Frobenius-normalized, random-spectrum, and Freon $c = 1/2$ controls are norm-matched to the NS_5 scalar in practical Muon. Their agreement shows that the population mechanism is rank-one spectral normalization rather than Muon-unique geometry.

β	SGD	AdamW	RMSProp	Trunc.	Rand.	Muon	Polar	Fro-norm	Freon
0.2	-0.071	9.568	2.389	-0.164	1.185	0.962	1.401	1.407	1.401
0.5	-0.059	9.208	2.302	-0.301	2.278	2.101	2.895	2.637	2.841
0.8	0.007	0.554	0.486	-0.436	2.577	2.843	2.534	2.753	2.505
1.2	0.184	0.300	0.177	-0.614	2.517	2.814	2.969	2.839	2.730
1.5	0.330	0.257	0.181	-0.698	2.196	3.058	2.458	2.718	2.225

Table 4: Full-rank mini-batch-plus-momentum grid: fitted width exponents from 675 runs with widths 64 to 1024, three seeds per width, batch size 512, 3000 steps, and momentum 0.95. AdamW's low-β exponents are floor-dominated; at $N = 1024$ its mean losses are 4.5×10^{-15} and 3.2×10^{-15} for $\beta = 0.2, 0.5$.

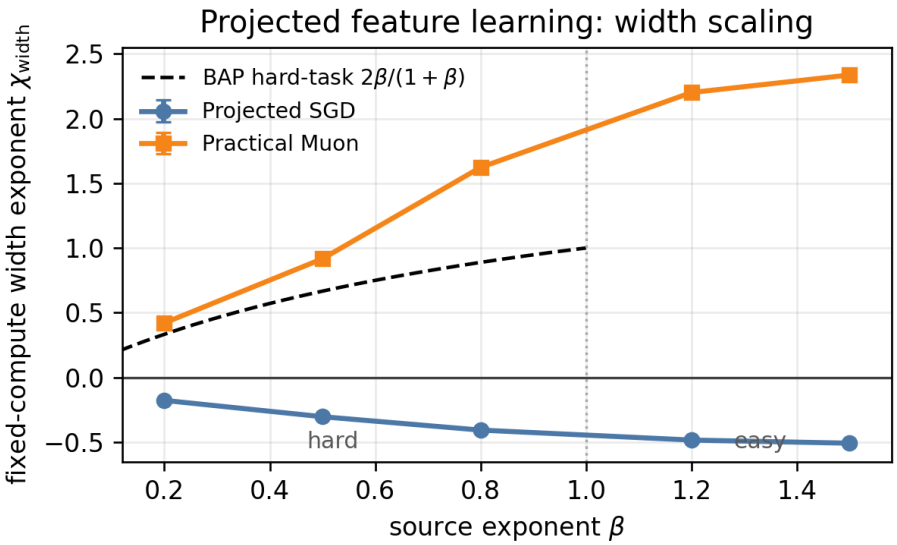

Figure 5: Projected population width exponents with uncertainty. Practical Muon has positive fixed-compute width exponents across the measured source exponents, while projected SGD remains negative. Bootstrap uncertainty is small relative to the optimizer separation, so the main conclusion is not driven by a single seed. The dashed curve is the BAP hard-task reference and is included as a guide, not as the fitted object.

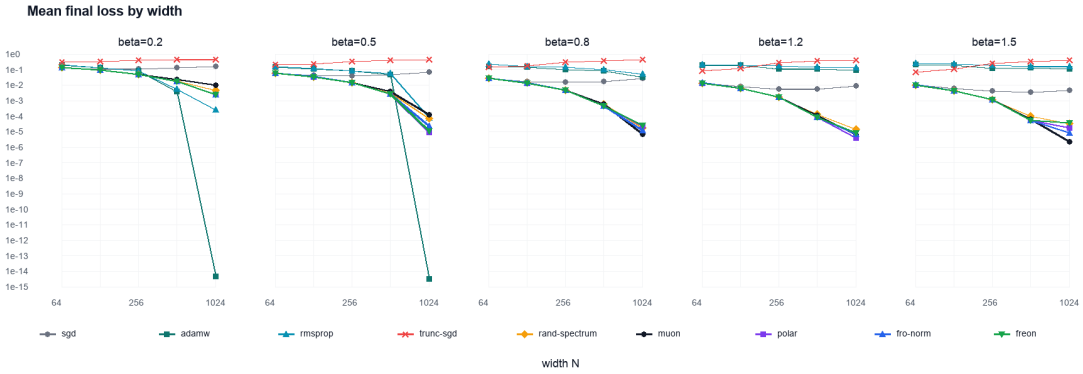

Figure 6: Mean final loss by width for the full-rank mini-batch momentum grid. The normalized matrix-update family improves with width on harder source exponents, while top-singular truncation remains poor. AdamW reaches numerical-floor loss for $\beta = 0.2, 0.5$, which explains its very large fitted exponents there, but it does not maintain that advantage for $\beta \geq 0.8$.

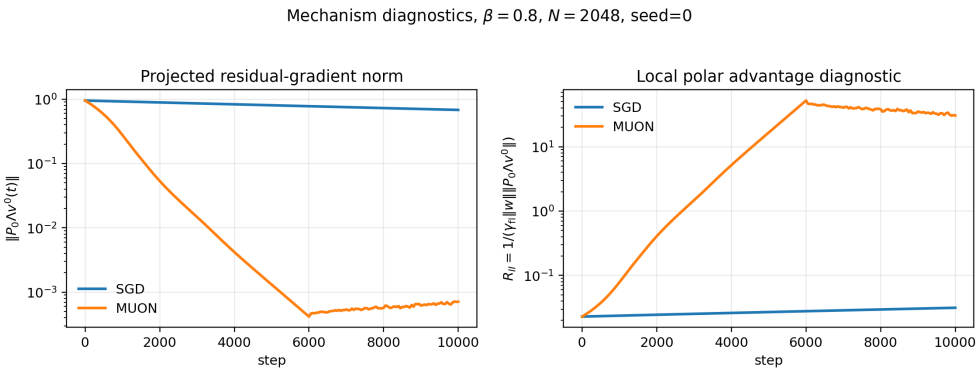

Figure 7: Mechanism diagnostics for $\beta = 0.8$, $N = 2048$. Practical Muon rapidly reduces the projected residual-gradient norm and produces a much larger local polar-advantage diagnostic R_{II} than projected SGD, supporting the local mechanism: normalization removes the small singular value that slows the projected feature update.

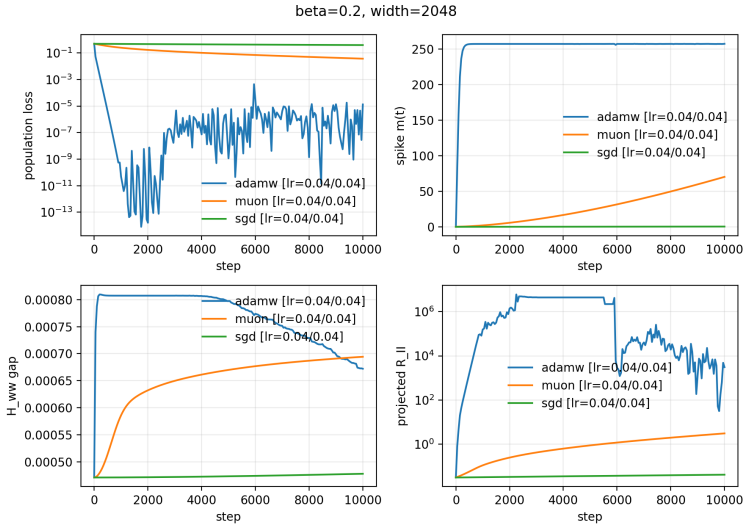

Figure 8: Representative time series for $\beta = 0.2$, $N = 2048$. AdamW can rapidly drive the population loss to a numerical floor in this low-β run but produces very large spike and ratio diagnostics; practical Muon improves more steadily, and projected SGD barely moves. This is a diagnostic caution: strong adaptive baselines must be reported, but their low-β behavior is not the same mechanism as the rank-one Muon/polar normalization result.

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

β	0.2	0.5	0.8
SGD 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 $\beta = 0.2, 0.5, 0.8, 1.2, 1.5$, the fitted exponents are Muon: $1.984, 1.981, 1.979, 1.974, 1.971$ and SGD: $2.010, 2.015, 2.017, 2.019, 2.019$. 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.