On How Muon Reshapes Skill Learning Dynamics

author names withheld

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

Abstract

Spectrum-aware optimizers, particularly Muon, exhibit more favorable empirical scaling than optimizers like SGD and Adam. Prior work explains this through more balanced learning in settings with imbalanced labels or input-output associations. In this work, we show that Muon enables more balanced skill acquisition in settings with task-level imbalance. Using a multi-task sparse parity-based setup, we show that with Muon, sub-tasks are learned more in parallel than with other optimizers. We introduce a Skill Acquisition Lag (SAL) metric to quantify parallel versus sequential learning. In a simple in-context linear regression setting, we show that Muon's convergence is independent of both task- and input-level imbalance, in contrast to (normalized) GD.

1. Introduction

Neural scaling laws describe the empirical observation that mean test loss $\mathcal{L}$ decreases predictably with model size N, data size D, and compute C: $\mathcal{L} \propto X^{-\alpha}$ for each scaling variable X [3, 5]. Gaining insight on the relationship between how these variables, and various choices made for training, and the resulting scaling exponent α is important for resource allocation [3]. One crucial aspect is the role of the optimization algorithm in the compute-scaling law. While most modern networks are trained with Adam [7], recent work has shown that spectrum-aware optimizers such as Muon [4], which applies Newton-Schultz orthogonalization to the gradient update, exhibit a more favorable empirical scaling law [9]. In this work, we seek to explain this observation.

To probe this question, we adopt the framework of Michaud et al. [12]. They put forth the *quantization hypothesis*, which posits that the knowledge to be acquired by a model decomposes into discrete skills or *quanta* whose usage frequencies follow a power law $p_k \propto k^{-\alpha}$. As scale increases, the model acquires skills in decreasing order of frequency, and averaging the loss over many such acquisitions yields a smooth power law in compute, model size, and data. Within this framework, since the scaling exponent α is governed by the rate of skill acquisition, a natural mechanism for improving α is to make it more parallel rather than sequential.

A separate line of work studying the effectiveness of spectrum-aware optimizers suggests that they do enable more balanced learning, albeit at the level of data rather than tasks. These results concern imbalance at the level of labels or input-output associations, where certain classes [13], tokens [14], or fact pairs [6, 8] appear less frequently than others. We ask whether the same benefits extend to imbalance at the *task* level, where different sub-tasks are acquired at different rates. Under the quantization hypothesis, an optimizer that learns low-frequency skills at the same rate as high-frequency ones leads to an improvement in the exponent α. We hypothesize that Muon does exactly this, providing a concrete mechanism for its improved scaling behavior over Adam. We make this connection explicit through our contributions:

- On the multi-task sparse parity benchmark of Michaud et al. [12] (see Section 2 for details), we show that Muon enables more parallel learning of sub-tasks as compared to NMD (the normalized counterpart of GD with momentum) and Adam.
- To quantify this, we introduce the Skill Acquisition Lag (Δ_{SAL}) metric (Section 4), which measures the degree to which sub-tasks are learned in parallel rather than sequentially. We find that Muon achieves substantially lower Δ_{SAL} than NMD and Adam.
- We complement these findings with a theoretical analysis of in-context linear regression with a simplified linear attention model. Under both input- and task-level imbalance, we prove that Muon's updates decouple across the eigenmodes of the covariance structure, yielding a uniform convergence rate across tasks, in contrast to (N)GD.

2. Background

2.1. The Quantization Model of Neural Scaling

To study neural scaling laws, Michaud et al. [12] proposed modeling the knowledge or skills learned by neural networks as a combination of one or more discrete, indivisible units or *quanta*. The use frequencies of these quanta follow a power law $p_k \propto k^{-(\alpha+1)}$ (for $k \in [n_{\mathrm{tasks}}]$), with exponent α. Under this model, a network that has learned τ quanta achieves loss $\mathcal{L}_\tau - \mathcal{L}_\infty \propto \tau^{-\alpha}$. This relation is translated into power laws in terms of size of data, parameters, and compute.

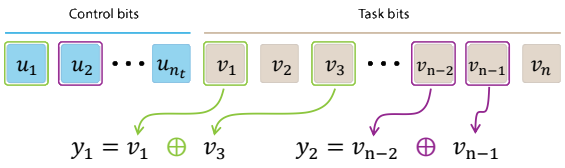

Figure 1: Illustration of the multi-task sparse parity setup with n_t control bits, n task bits, $m = 2$ tasks. Sub-task 1 (green) label is $y_1 = v_1 \oplus v_3$ from $S_1 = \{1, 3\}$; Sub-task 2 (purple) label is $y_2 = v_{n-1} \oplus v_{n-1}$ from $S_2 = \{n-2, n-1\}$.

In the realm of data-scaling: assuming multi-epoch training, D samples and sufficient network capacity, a quanta k is learned only if the dataset contains at least τ tokens involving it, meaning only quanta with $p_k \geq \tau/D$ are acquired. Since $p_k \propto k^{-(\alpha+1)}$, the number of learned quanta scales as $q \propto D^{1/(\alpha+1)}$, yielding the data-scaling law $\mathcal{L}(D) \approx \mathcal{L}_\infty + C_D D^{-\alpha/(\alpha+1)}$. An analogous argument gives the respective step-scaling exponent under GD: quantum k grows as $t_k \propto 1/p_k \propto k^{\alpha+1}$, yielding $\alpha_S^{\mathrm{GD}} = \alpha/(\alpha+1)$.

In Michaud et al. [12], authors propose the *multi-task sparse parity* benchmark as a concrete testbed to empirically validate the quantization model. This benchmark involves a binary classification problem on bit strings $\boldsymbol{x} = [\boldsymbol{u}, \boldsymbol{v}] \in \{0,1\}^{n_{\mathrm{tasks}}+n}$, where $\boldsymbol{u} \in \{0,1\}^{n_{\mathrm{tasks}}}$ denotes a one-hot task selector, and $\boldsymbol{v} \in \{0,1\}^n$ denote the task bits. A sub-task k corresponds to selecting a fixed subset $S_k \subset \{1, \ldots, n\}$ of $|S_k| = m$ task bits. To generate an input-label pair, we first sample a sub-task k with probability $p_k \propto k^{-(\alpha+1)}$, set $u_k = 1$, and draw $\boldsymbol{v}$ uniformly. Then, the label is given by an m-sparse parity: $\boldsymbol{y} = \bigoplus_{i \in S_k} \boldsymbol{v}_i$.

2.2. Optimizers

We consider three optimizers in this work. **NMD** (Normalized Momentum Descent) is momentum-based GD with the update normalized by its Frobenius norm. **Adam** [7] computes per-parameter updates by scaling the accumulated gradients by the inverse square root of their second moments.

Muon [4] is a spectrum-aware, matrix-valued optimizer that orthogonalizes an accumulated gradients via Newton-Schulz iterations. See Appendix A for the update rules and further details.

3. Muon Enables Faster Skill Acquisition

As discussed in the previous section, under the quantization model [12], skills are learned in decreasing order of use frequencies, *i.e.*, the time to learn skill k scales as $t_k \propto 1/p_k \propto k^{\alpha+1}$. A direct consequence of this sequential learning is that the compute-scaling exponent $\alpha_S = \alpha/(\alpha+1)$, which is strictly less than α and reflects a bottleneck imposed by the rarest skills which are acquired at the slowest rate. Within this model, note that an optimizer that enables more parallel rather than sequential skill acquisition would lead to a better exponent α_S. For instance, considering the other extreme where all tasks are learned in parallel, *i.e.*, $t_k \propto 1$ leads to an improved exponent $\alpha_S = \alpha$.

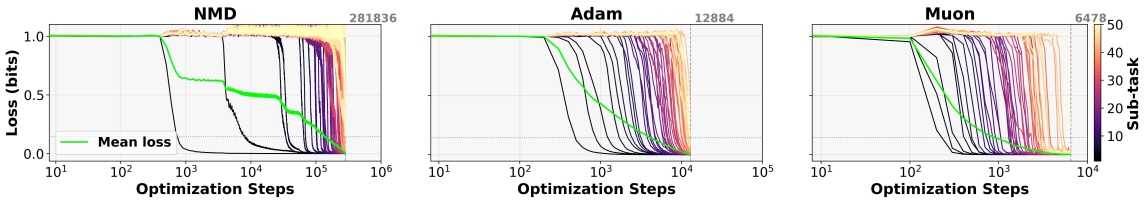

Figure 2: Comparison of per-task loss convergence for NMD, Adam and Muon. With NMD, skill acquisition is sequential, where per-task loss converges in decreasing order of task frequency. Adam and Muon enable more parallel skill acquisition, where gaps between per-task loss convergence are significantly smaller, leading to faster mean loss convergence. Experimental setup and hyper-parameters explained in Appendix B.1.

Motivated from prior work showing that Muon enables more balanced learning under imbalanced settings [11, 13], we hypothesize that these benefits are also observed at the task level. To test this hypothesis, we compare the per-task convergence behaviour for the three optimizers: NMD, Adam, and Muon on the multi-task sparse parity benchmark. The results in Fig. 2 show that for NMD, skill acquisition is sequential: per-task losses for rarer sub-tasks start reducing only after the network learns more common sub-tasks. On the other hand, for both Adam and Muon skill acquisition is more parallel, with the effect being more prominent for Muon. This is accompanied by faster mean loss convergence rate. In the next section, we make this finding more concrete by introducing a metric to measure the relative sub-task convergence gaps.

4. Quantifying Parallel vs. Sequential Skill Acquisition

In this section, we introduce *skill acquisition lag* (Δ_{SAL}), a summary statistic that captures the degree of parallel versus sequential skill acquisition. While per-task convergence curves provide a detailed account of learning dynamics, SAL enables a quantitative comparison across optimizers.

Definition 1 (Skill Acquisition Lag (Δ_{SAL})) *Fix a loss threshold $\tau > 0$. For each sub-task $k \in [n_{\mathrm{tasks}}]$, let $T_k(\tau) := \min\{\, t : \mathcal{L}_k(t) \le \tau \,\}$ be the first step at which sub-task k reaches loss $\le \tau$, and let $T_{\mathrm{first}}(\tau) := \min_k T_k(\tau)$ denote the convergence time of the fastest learned sub-task. Define*

the p-coverage time

$$T(p,\tau) := \min\left\{t : \left|\{k : T_k(\tau) \le t\}\right| \ \ge\ \lceil p n_{\text{tasks}} \rceil\right\}, \tag{1}$$

the first step at which at least a fraction p *of sub-tasks have each reached loss* $\le \tau$. *The* Skill Acquisition Lag *for task fraction* p *and loss threshold* τ *is*

$$\Delta_{\text{SAL}}(p,\tau) = \frac{T(p,\tau) - T_{\text{first}}(\tau)}{T(p,\tau)} \in [0,1). \tag{2}$$

Consider $p = 1$. Then, for parallel acquisition, *i.e.*, all sub-tasks converge simultaneously with the first one, $\Delta_{\text{SAL}} = 0$, whereas for sequential acquisition, $\Delta_{\text{SAL}} = \alpha^{n_{\text{tasks}}}$.

Results. Fig. 3 shows the p-coverage time (left) and SAL (right). We find that for NMD, both p-coverage time and SAL values are much larger, indicating sequential skill acquisition, as predicted by the quantization model. On the other hand, Muon (and Adam) exhibit much smaller p-coverage time and SAL values, indicating more parallel skill acquisition.

An important direction for future work is uncovering the mechanism behind how the spectral design of Muon enables more parallel skill acquisition in neural networks. In the next section, we analyze a simple setup with task-level imbalance to get some insight.

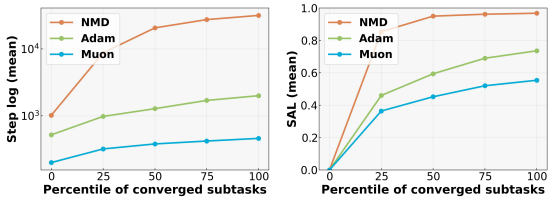

Figure 3: Convergence dynamics and SAL values across optimizers on the multi-task sparse parity benchmark. **Left:** *p-coverage* time averaged across different model widths demonstrates that Muon reaches each percentile of converged subtasks in fewest steps. **Right:** $\Delta_{\text{SAL}}(p)$ as a function of coverage p. Smaller SAL values indicate more parallel skill acquisition. Details of hyperparameters are in Appendix B.2.

5. Theoretical Analysis on In-Context Linear Regression

Data Generation. We consider a synthetic in-context learning (ICL) linear regression problem, where each task is specified by a weight vector $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_w) \in \mathbb{R}^d$. For $t \in [T+1]$, inputs are sampled as $\boldsymbol{x}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_x) \in \mathbb{R}^d$ with labels $y_t = \boldsymbol{x}_t^\top \boldsymbol{w}$, where $\boldsymbol{x}_q = \boldsymbol{x}_{T+1}$. Let $\boldsymbol{\Sigma}_w = \boldsymbol{U}_w \boldsymbol{\Lambda}_w \boldsymbol{U}_w^\top$ and $\boldsymbol{\Sigma}_x = \boldsymbol{U}_x \boldsymbol{\Lambda}_x \boldsymbol{U}_x^\top$ denote the eigenvalue decompositions, with $\lambda_i^{(w)} \propto i^{-\gamma}$ and $\lambda_i^{(x)} \propto i^{-\beta}$, corresponding to task and input covariance imbalance, respectively. Let $\kappa_x := \lambda_{\max}^{(x)}/\lambda_{\min}^{(x)}$ and $\kappa_w := \lambda_{\max}^{(w)}/\lambda_{\min}^{(w)}$ denote the condition numbers of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_w$, respectively.

Model. We consider a simple reparameterized linear self-attention model, with trainable paramater $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ (see Lu et al. [10], Ma et al. [11], Zhang et al. [15, 16] for details). The query prediction is through the bilinear form $\hat{y}_q = (\hat{\boldsymbol{\Sigma}}_x \boldsymbol{w})^\top \boldsymbol{Q} \boldsymbol{x}_q$, where $\hat{\boldsymbol{\Sigma}}_x = \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t \boldsymbol{x}_t^\top$.

The training objective to be minimized is the mean squared error over the population, *i.e.*,

$$\mathcal{L}_{\gamma,\beta}(\boldsymbol{Q}) := \tfrac{1}{2}\mathbb{E}_{\boldsymbol{w},\boldsymbol{x}_{1:T},\boldsymbol{x}_q}\left[(\hat{y}_q - \boldsymbol{w}^\top \boldsymbol{x}_q)^2\right] = \tfrac{1}{2}\operatorname{tr}\left((\boldsymbol{\Sigma}_x \boldsymbol{Q} - \boldsymbol{I})\,\boldsymbol{\Sigma}_x\,(\boldsymbol{\Sigma}_x \boldsymbol{Q} - \boldsymbol{I})^\top\,\boldsymbol{\Sigma}_w\right).$$

In Ma et al. [11], the authors consider a similar setting but with $\gamma = 0$ (no task-level imbalance) and compare convergence of GD and SpecGD (spectral GD), a canonical form of Muon without accumulation and with exact matrix operations. They show that GD takes $\Omega(\sqrt{\kappa}\log(1/\epsilon))$ steps, where $\kappa = \kappa_x^3$, to converge to loss $\le \epsilon$, whereas SpecGD converges in $O(\log(1/\epsilon))$ steps.

In the following, we extend this result to task-level imbalance and compare NGD vs. SpecGD.

Convergence Analysis. Similar to Ma et al. [11], we make the following assumption.

Assumption 1 *The matrices $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_w$ share a common eigenbasis $\boldsymbol{V}^* \in \mathbb{R}^{d\times d}$ so that $\mathbf{\Sigma}_x = \boldsymbol{V}^* \mathbf{\Lambda}_x \boldsymbol{V}^{*\top}, \mathbf{\Sigma}_w = \boldsymbol{V}^* \mathbf{\Lambda}_w \boldsymbol{V}^{*\top}$.*

In this setting, the effective condition number $\kappa = \kappa_x^3 \kappa_w$. For NGD, the update direction is the same as GD (up to a scaling factor). Since the objective $\mathcal{L}_{\gamma,\beta}$ is strongly convex, the classical lower bound for first-order methods [2, Theorem 2.3] applies, yielding the same iteration complexity as GD [11], *i.e.*, it depends on $\sqrt{\kappa}$. The following result shows that in contrast to NGD, the iteration complexity for SpecGD is free from any dependence on the condition numbers κ_x and κ_w.

Theorem 2 *Initialize $\boldsymbol{Q}_0 = \mathbf{0}$ and adopt the learning rate schedule $\eta_t = \frac{C_\eta}{\sigma_{\min}(\mathbf{\Sigma}_x)} \rho^t$ with $C_\eta \geq 1$ and $\rho \in [1/2, 1)$. Then, under Ass. 1, for any target loss $\varepsilon > 0$, SpecGD iterates achieve $\|\boldsymbol{Q}_T - \mathbf{\Sigma}_x^{-1}\| \leq \varepsilon$ whenever $T \geq \frac{1}{1-\rho} \log\Big(\frac{C_\eta}{\sigma_{\min}(\mathbf{\Sigma}_x)\,\varepsilon}\Big)$.*

Here, we note that SpecGD's spectral orthogonalization decouples the dynamics into independent scalar sequences $\theta_{i,t+1} = \theta_{i,t} - \eta_t \cdot \mathtt{sign}(\lambda_i^{(x)} \theta_{i,t} - 1)$, in which the per-mode contraction is identical across i, yielding a κ-independent rate (see Appendix C.1 for the proof). In contrast, (N)GD updates scale proportionally to $\lambda_i^{(x)2} \lambda_i^{(w)}$ along eigenmode i, leading to a κ-dependent rate.

Experimental Results. In Appendix C.2, we present experimental results comparing Muon and GD under varying input (κ_x) or task (κ_w) imbalance. While convergence for GD slows substantially, Muon remains largely insensitive to the imbalance.

6. Conclusion

We studied how Muon's spectral design reshapes task learning dynamics under frequency imbalance. First, we noticed that on the multi-task sparse parity benchmark, Muon results in substantially smaller gaps between convergence times across sub-tasks. Then, to quantify this effect, we introduced the Δ_{SAL} metric, showing that Muon maintains lower values across all coverage percentiles, and further verifying a more parallel skill acquisition schedule. Moreover, in a simple in-context linear regression setting, we proved that Muon's per-mode convergence rate is independent of both task- and input-level imbalance, unlike (N)GD. A natural next step is to study how more parallel acquisition that we observe translates into a measurably improved step-scaling exponent α_S. More broadly, our findings suggest that designing optimizers that further encourage parallel skill acquisition is a promising path toward better scaling behavior.

References

[1] Jeremy Bernstein. Deriving muon, 2025. URL https://jeremybernste.in/writing/deriving-muon.

[2] Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021.

[3] 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.

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

[5] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[6] Juno Kim, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D. Lee. Sharp capacity scaling of spectral optimizers in learning associative memory, 2026. URL https://arxiv.org/abs/2603.26554.

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

[8] Binghui Li, Kaifei Wang, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws, 2026. URL https://arxiv.org/abs/2602.05725.

[9] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Wenxuan Xu, Enfang Lu, Jian Yan, and Yutao Chen. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

[10] Yue M. Lu, Mary I. Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *arXiv preprint arXiv:2405.11751*, 2025.

[11] Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in Muon. *arXiv preprint*, 2026.

[12] Eric J Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[13] Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. *arXiv preprint arXiv:2510.22980*, 2025.

[14] Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Y F Tan. Muon outperforms Adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025.

[15] Ruiqi Zhang, Jingfeng Wu, and Peter Bartlett. In-context learning of a linear transformer block: Benefits of the mlp component and one-step gd initialization. *Advances in Neural Information Processing Systems*, 37:18310–18361, 2024.

[16] Yedi Zhang, Aaditya K. Singh, Peter E. Latham, and Andrew Saxe. Training dynamics of in-context learning in linear attention. *arXiv preprint arXiv:2501.16265*, 2025.

Appendix A. Optimizer Details for Section 2.2

We provide the update rules for the three optimizers considered in this paper. Let $\mathbf{W}_t$ denote the weight matrix at step t, $\nabla_t := \nabla\mathcal{L}(\mathbf{W}_t)$ the gradient, and $\eta > 0$ the learning rate.

A.1. NMD (Normalized Momentum Descent)

NMD consists of a momentum buffer in the form of a moving gradient average which is normalized before scaling by the learning rate to yield update

$$\begin{aligned} \mathbf{M}_t &= \beta\mathbf{M}_{t-1} + \nabla_t, \\ \mathbf{W}_{t+1} &= \mathbf{W}_t - \eta_t \frac{\mathbf{M}_t}{\|\mathbf{M}_t\|}. \end{aligned} \tag{3}$$

A.2. Adam

Adam [7] maintains per-parameter running estimates of the first and second moments of the gradient:

$$\begin{aligned} \mathbf{M}_t &= \beta_1\mathbf{M}_{t-1} + (1-\beta_1)\nabla_t, \\ \mathbf{V}_t &= \beta_2\mathbf{V}_{t-1} + (1-\beta_2)\nabla_t^2, \\ \mathbf{W}_{t+1} &= \mathbf{W}_t - \eta_t \frac{\hat{\mathbf{M}}_t}{\sqrt{\hat{\mathbf{V}}_t} + \epsilon}, \end{aligned} \tag{4}$$

where $\hat{\mathbf{M}}_t = \mathbf{M}_t/(1-\beta_1^t)$ and $\hat{\mathbf{V}}_t = \mathbf{V}_t/(1-\beta_2^t)$ are bias-corrected estimates, ∇_t^2 denotes the element-wise square, and $\epsilon > 0$ is a small constant for numerical stability. Adam adapts learning rates per parameter.

A.3. Muon

Muon [4] accumulates gradient momentum and then orthogonalizes the update via the matrix sign function (msign), computed efficiently through Newton-Schulz iterations:

$$\mathbf{M}_t = \beta\mathbf{M}_{t-1} + \nabla_t, \tag{5}$$

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t\sqrt{\tfrac{m}{n}}\,\mathrm{msign}(\mathbf{M}_t), \tag{6}$$

where $\mathbf{W}_t \in \mathbb{R}^{m\times n}$, $\mathrm{msign}(\mathbf{M}_t) = \mathbf{U}_t\mathbf{V}_t^\top$ given the SVD $\mathbf{M}_t = \mathbf{U}_t\mathbf{\Sigma}_t\mathbf{V}_t^\top$, and the factor $\sqrt{m/n}$ normalizes the update so that per-entry RMS is consistent across layers of different shape [1]. The msign function replaces all singular values with ones while preserving the left and right singular vectors, equivalent to projecting the momentum onto the nearest semi-orthogonal matrix. In our experiments, Muon is applied only to hidden-layer weight matrices; embedding and output parameters use Adam following standard practice [4, 9].

Appendix B. Omitted Details for Section 4

B.1. Experimental Setup

We utilize the multi-task sparse parity benchmark with $n_{\text{tasks}} = 50$ sub-tasks, $n = 100$ task bits, and $k = 3$ active bits per sub-task. Sub-task frequencies follow a Zipf distribution $p_i \propto i^{-\gamma}$ where

$\gamma = 1.4$. The model architecture is a single hidden layer ReLU multi-layer perceptron with 2000 hidden neurons, trained with batch size of $10,000$ for as many optimization steps as it takes to converge the last sub-task to target loss of $\tau = 0.1$ bits. For each optimizer, we search over learning rates $\{10^{-2}, 10^{-3}, 10^{-4}\}$ and report results at the best performing value, using configurations:

- NMD:$\beta = 0.9$, $\eta = 0.01$.
- Adam: $(\beta_1, \beta_2) \in \{(0.9, 0.999)\}$, $\eta = 0.001$.
- Muon: $\beta = 0.95$, $\eta = 0.01$ (hidden weights); Adam for embeddings [4].

B.2. Effect of Model Width

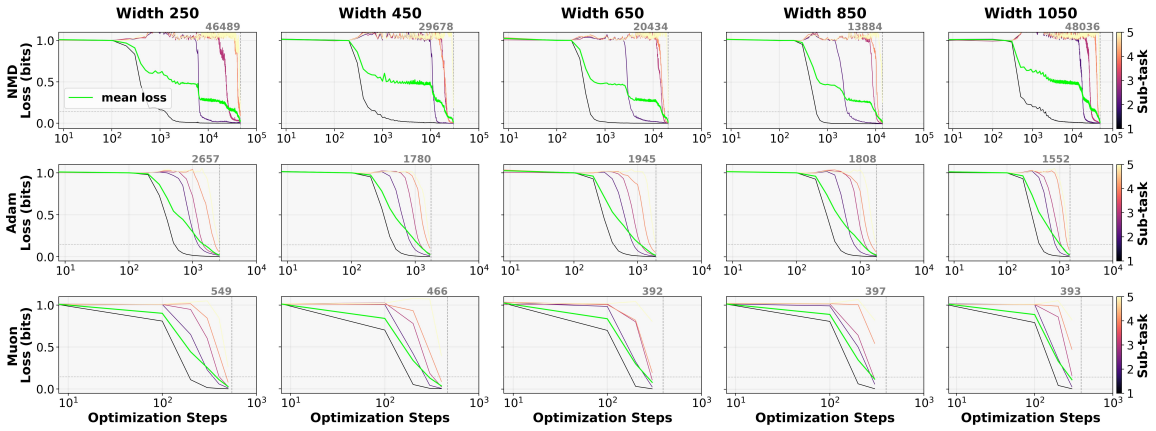

Figure 4: Loss curves on the multi-task sparse parity benchmark across model widths while varying optimizers demonstrates that Muon (bottom) consistently keeps a more parallel task acquisition schedule that Adam (middle) and NMD (top).

The figure above follows settings in Appendix B.1 but we set $n_{\text{tasks}} = 5$. We also vary model widths $\in \{250, 450, 650, 850, 1050\}$.

Appendix C. Omitted Details for Section 5

Training Objective. The objective is to minimize the expected squared prediction risk, which is written as

$$\begin{aligned}
\mathcal{L}_{\gamma,\nu}(\boldsymbol{Q}) &:= \frac{1}{2}\mathbb{E}_{\boldsymbol{w},\boldsymbol{x}_{1:T},\boldsymbol{x}_q}\left[\left(\hat{y}_q - \boldsymbol{w}^\top \boldsymbol{x}_q\right)^2\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{w},\boldsymbol{x}_{1:T},\boldsymbol{x}_q}\left[\left(\boldsymbol{w}^\top \boldsymbol{\Sigma}_x \boldsymbol{Q} \boldsymbol{x}_q - \boldsymbol{w}^\top \boldsymbol{x}_q\right)^2\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{w},\boldsymbol{x}_{1:T},\boldsymbol{x}_q}\left[\boldsymbol{w}^\top \left(\boldsymbol{\Sigma}_x \boldsymbol{Q} - \boldsymbol{I}\right) \boldsymbol{x}_q \boldsymbol{x}_q^\top \left(\boldsymbol{\Sigma}_x \boldsymbol{Q} - \boldsymbol{I}\right)^\top \boldsymbol{w}\right] \\
&= \frac{1}{2}\operatorname{tr}\Big(\left(\boldsymbol{\Sigma}_x \boldsymbol{Q} - \boldsymbol{I}\right) \mathbb{E}_{\boldsymbol{x}_q}\left[\boldsymbol{x}_q \boldsymbol{x}_q^\top\right] \left(\boldsymbol{\Sigma}_x \boldsymbol{Q} - \boldsymbol{I}\right)^\top \mathbb{E}_{\boldsymbol{w}}\left[\boldsymbol{w}\boldsymbol{w}^\top\right]\Big) \\
&= \frac{1}{2}\operatorname{tr}\Big(\left(\boldsymbol{\Sigma}_x \boldsymbol{Q} - \boldsymbol{I}\right) \boldsymbol{\Sigma}_x \left(\boldsymbol{\Sigma}_x \boldsymbol{Q} - \boldsymbol{I}\right)^\top \boldsymbol{\Sigma}_w\Big),
\end{aligned} \tag{7}$$

where the final equality uses $\mathbb{E}_{\boldsymbol{x}_q}[\boldsymbol{x}_q \boldsymbol{x}_q^\top] = \boldsymbol{\Sigma}_x$ and $\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}\boldsymbol{w}^\top] = \boldsymbol{\Sigma}_w$. The resulting objective is

$$\min_{\boldsymbol{Q}\in\mathbb{R}^{d\times d}} \mathcal{L}_{\gamma,\nu}(\boldsymbol{Q}) \;:=\; \frac{1}{2}\operatorname{tr}\Big(\left(\boldsymbol{\Sigma}_x \boldsymbol{Q} - \boldsymbol{I}\right) \boldsymbol{\Sigma}_x \left(\boldsymbol{\Sigma}_x \boldsymbol{Q} - \boldsymbol{I}\right)^\top \boldsymbol{\Sigma}_w\Big). \tag{8}$$

Note that the gradient of the objective $\mathcal{L}_{\gamma,\nu}$ is given by

$$\nabla\mathcal{L}_{\gamma,\nu}(\boldsymbol{Q}) = \boldsymbol{\Sigma}_x \boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_x \boldsymbol{Q} \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_x \boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_x. \tag{9}$$

C.1. Proof of Theorem 2

We first show that SpecGD decouples across eigenmodes.

Lemma 3 *For each $t \in \mathbb{Z}_{\geq 0}$, the SpecGD iterates satisfy $\boldsymbol{Q}_t = \boldsymbol{V}^* \boldsymbol{\Theta}_t \boldsymbol{V}^{*\top}$, where $\boldsymbol{\Theta}_t = \operatorname{diag}(\theta_{1,t}, \ldots, \theta_{d,t})$ evolves according to*

$$\boldsymbol{\Theta}_{t+1} = \boldsymbol{\Theta}_t - \eta_t \operatorname{diag\text{-}sign}(\boldsymbol{\Lambda}_x \boldsymbol{\Theta}_t - \boldsymbol{I}). \tag{10}$$

Proof We proceed by induction on t.

Base Case. At $t = 0$, the initialization $\boldsymbol{Q}_0 = \boldsymbol{0}$ corresponds to $\boldsymbol{\Theta}_0 = \boldsymbol{0}$, so the decomposition $\boldsymbol{Q}_0 = \boldsymbol{V}^* \boldsymbol{\Theta}_0 \boldsymbol{V}^{*\top}$ holds trivially.

Inductive step. Assume $\boldsymbol{Q}_t = \boldsymbol{V}^* \boldsymbol{\Theta}_t \boldsymbol{V}^{*\top}$ for some $t \geq 0$. Substituting into the gradient yields

$$\nabla\mathcal{L}_{\gamma,\nu}(\boldsymbol{Q}_t) = \boldsymbol{V}^* \left(\boldsymbol{\Lambda}_x^3 \boldsymbol{\Lambda}_w \boldsymbol{\Theta}_t - \boldsymbol{\Lambda}_x^2 \boldsymbol{\Lambda}_w\right) \boldsymbol{V}^{*\top}. \tag{11}$$

Applying the SpecGD update and using the fact that msign preserves the eigenbasis, we obtain

$$\begin{aligned}
\boldsymbol{Q}_{t+1} &= \boldsymbol{Q}_t - \eta_t \operatorname{msign}(\nabla\mathcal{L}_{\gamma,\nu}(\boldsymbol{Q}_t)) \\
&= \boldsymbol{Q}_t - \eta_t \operatorname{msign}\Big(\boldsymbol{V}^* \left(\boldsymbol{\Lambda}_x^3 \boldsymbol{\Lambda}_w \boldsymbol{\Theta}_t - \boldsymbol{\Lambda}_x^2 \boldsymbol{\Lambda}_w\right) \boldsymbol{V}^{*\top}\Big) \\
&= \boldsymbol{V}^* \left(\boldsymbol{\Theta}_t - \eta_t \operatorname{diag\text{-}sign}(\boldsymbol{\Lambda}_x^3 \boldsymbol{\Lambda}_w \boldsymbol{\Theta}_t - \boldsymbol{\Lambda}_x^2 \boldsymbol{\Lambda}_w)\right) \boldsymbol{V}^{*\top} \\
&= \boldsymbol{V}^* \left(\boldsymbol{\Theta}_t - \eta_t \operatorname{diag\text{-}sign}(\boldsymbol{\Lambda}_x^2) \operatorname{diag\text{-}sign}(\boldsymbol{\Lambda}_w) \operatorname{diag\text{-}sign}(\boldsymbol{\Lambda}_x \boldsymbol{\Theta}_t - \boldsymbol{I})\right) \boldsymbol{V}^{*\top} \\
&= \boldsymbol{V}^* \left(\boldsymbol{\Theta}_t - \eta_t \operatorname{diag\text{-}sign}(\boldsymbol{\Lambda}_x \boldsymbol{\Theta}_t - \boldsymbol{I})\right) \boldsymbol{V}^{*\top},
\end{aligned}$$

where the final equality uses $\text{diag-sign}(\mathbf{\Lambda}_x^2)\,\text{diag-sign}(\mathbf{\Lambda}_w) = \boldsymbol{I}$, which holds because both $\mathbf{\Lambda}_x^2$ and $\mathbf{\Lambda}_w$ have strictly positive diagonal entries. This establishes $\boldsymbol{Q}_{t+1} = \boldsymbol{V}^* \mathbf{\Theta}_{t+1} \boldsymbol{V}^{*\top}$ with $\mathbf{\Theta}_{t+1} = \mathbf{\Theta}_t - \eta_t\,\text{diag-sign}(\mathbf{\Lambda}_x \mathbf{\Theta}_t - \boldsymbol{I})$, completing the induction. ■

We use the following result from Ma et al. [11].

Lemma 4 (Scalar convergence under SpecGD dynamics) *Let $\{\theta_t\}_{t\geq 0} \subset \mathbb{R}$ be a scalar sequence evolving according to*

$$\theta_{t+1} = \theta_t - \eta_t \, \texttt{sign}(\lambda^\star \theta_t - 1), \tag{12}$$

where $\lambda^\star$ is a scalar bounded below by $\lambda^\star_{\min} > 0$. Choose the learning rate schedule $\eta_t = \frac{C_\eta}{\lambda^\star_{\min}} \rho^t$ with $1/2 \leq \rho < 1$ and $C_\eta \geq 1$, and initialize at $\theta_0 = 0$. Then for every $t \geq 0$,

$$\left| \theta_{t+1} - \frac{1}{\lambda^\star} \right| \leq \eta_t = \frac{C_\eta}{\lambda^\star_{\min}} \rho^t. \tag{13}$$

Invoking Lemma 4 on each eigenmode sequence (Lemma 3), we obtain

$$\left\| \boldsymbol{Q}_{t+1} - \mathbf{\Sigma}_x^{-1} \right\| = \left\| \mathbf{\Theta}_{t+1} - (\mathbf{\Lambda}_x)^{-1} \right\| = \max_{1 \leq i \leq d} \left| \theta_{i,t+1} - \frac{1}{\lambda_i^{(x)}} \right| \leq \eta_t. \tag{14}$$

Consequently, achieving $\left\| \boldsymbol{Q}_T - \mathbf{\Sigma}_x^{-1} \right\| \leq \varepsilon$ requires no more than $T \geq \frac{1}{1-\rho} \log\left(\frac{C_\eta}{\sigma_{\min}(\mathbf{\Sigma}_x)\varepsilon} \right)$ iterations, as claimed.

C.2. Experimental Setup and Results

C.2.1. SETUP

We empirically evaluate and compare the convergence behavior of Muon and SGD on the in-context learning objective $\mathcal{L}_{\gamma,\beta}$. Fixing the matrix dimension at $d = 100$, we sweep the task imbalance strength γ and the input imbalance strength β to probe both regimes. Across all experiments, we adopt an exponential decay learning rate schedule in which the step size is reduced by a factor of 0.3 whenever the loss fails to improve for 50 consecutive iterations.

C.2.2. EXPERIMENTAL RESULTS

Figure 5 provides the full convergence results for Muon and SGD on the in-context linear regression objective $\mathcal{L}_{\gamma,\nu}$ across varying levels of task and input covariance imbalance. Each column fixes the task covariance condition number κ_w, while curves within each subfigure correspond to different input covariance condition numbers κ_x.

Consistent with the theoretical analysis in Section 5, Muon converges rapidly across all spectral imbalance levels and remains largely insensitive to increases in either κ_x or κ_w. In contrast, SGD becomes progressively slower as the covariance spectra become more ill-conditioned, particularly under large input covariance imbalance. These numerical simulations empirically support the theoretical prediction that Muon's matrix-sign update removes dependence on spectral magnitudes, whereas SGD remains sensitive to spectral imbalance.

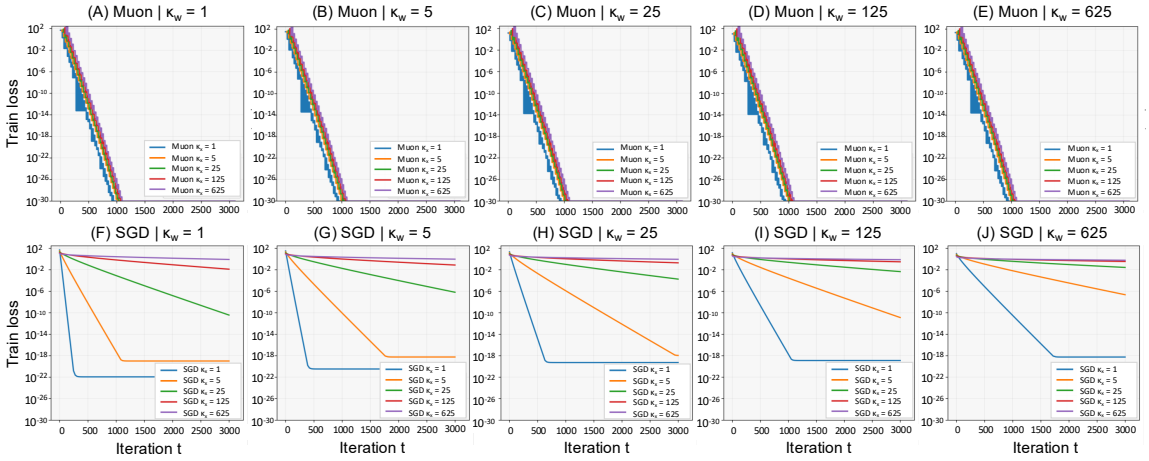

Figure 5: Train loss versus iteration t for Muon (top row, A-E) and SGD (bottom row, F-J) on objective $\mathcal{L}_{\gamma,\beta}$. Each column fixes the task covariance condition number κ_w, while curves within each subfigure correspond to different input covariance condition numbers κ_x.

Appendix D. Discussion

Limitations. Our theory makes the assumption of a shared eigenbasis, which holds only for linear models with specific covariance structures. The multitask sparse parity experiments use relatively small models.

Future work. A natural next step is to connect the parallel acquisition dynamics we observe to a theoretical result for a shift in the step-scaling exponent α_S. The quantization model predicts $\alpha_S^{\mathrm{GD}} = \alpha/(\alpha+1)$ under purely sequential learning; whether Muon's more parallel schedule yields a measurably different α_S requires careful scaling experiments across model sizes, which we leave to future work. Extending the theoretical analysis beyond the shared-eigenbasis setting, for instance to non-linear models or settings where task and input covariances are misaligned, is another important direction. Finally, validating these findings on naturalistic multi-task benchmarks and at larger scales would strengthen the practical relevance of our conclusions.