PRETRAINING DATA SCALE REVERSES THE MUON–ADAMW FINE-TUNING GAP

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon has recently emerged as a promising optimizer for pretraining large language models, offering a speedup over AdamW in minimizing pretraining loss. However, it remains unclear whether Muon's advantage persists after fine-tuning. In this paper, we study the downstream behavior of Muon and AdamW and uncover a surprising trend: beyond a sufficiently large pretraining data budget, AdamW outperforms Muon after fine-tuning. Moreover, as model size increases, AdamW overtakes Muon at progressively smaller pretraining token–parameter ratios. We observe the same trend under Gaussian perturbations of the pretrained weights, where Muon-pretrained models degrade increasingly more than AdamW-pretrained models as the data budget grows. We complement our empirical findings with a controlled study of Muon and AdamW in a linear model. In this setting, we show theoretically and empirically that the margin-to-noise ratio, a tractable proxy for robustness motivated by a second-order approximation of the loss, shifts from favoring Muon to favoring gradient descent as the data budget grows, mirroring the fine-tuning trend.

1 INTRODUCTION

The Muon optimizer has recently shown strong empirical performance for pretraining LLMs, challenging AdamW's status as the standard optimizer for minimizing pretraining loss (Jordan et al., 2024). Prior work has shown that the pretraining optimization choices that minimize pretraining loss do not necessarily yield the best model after fine-tuning (Watts et al., 2026). What ultimately matters downstream is the *learning–forgetting tradeoff*: how well a fine-tuned model learns the target task while retaining its pretrained knowledge. Motivated by this, we ask:

Does Muon's pretraining advantage over AdamW carry over to the learning–forgetting tradeoff?

To answer this question, we measure the Muon–AdamW *fine-tuning gap*, which compares the two optimizers on this tradeoff; a negative gap indicates that Muon dominates AdamW, and a positive gap indicates the reverse. We show that the answer depends on the pretraining data budget. At small budgets, Muon achieves lower fine-tuning loss with less forgetting than AdamW and the fine-tuning gap is negative. At larger budgets, this trend reverses: AdamW achieves a better learning–forgetting tradeoff than Muon, and the fine-tuning gap becomes positive. This reversal occurs even though Muon achieves lower pretraining loss at every budget we consider.

In Section 3, we pretrain OLMo-style 30M–600M models (Groeneveld et al., 2024; Team OLMo et al., 2025) on 150M to 154B DCLM tokens (Li et al., 2024), comparing AdamW and Muon, and fine-tune each base model on four datasets while sweeping the fine-tuning optimizer and learning rate, for nearly 600 pretraining runs and 29,000 fine-tuning runs in total. We show that **pretraining data scale reverses the Muon–AdamW fine-tuning gap, despite Muon always achieving lower pretraining loss.** At small pretraining budgets, Muon Pareto-dominates AdamW and the fine-tuning gap is negative; beyond a crossover budget, AdamW achieves lower fine-tuning loss with less forgetting than Muon. We then show that **as model size scales, AdamW surpasses Muon at progressively smaller data budgets:** measured by the token–parameter ratio (TPR), the crossover falls from 320 TPR for 30M models to 10 TPR for 600M models. Finally, we observe that **Muon is less robust to Gaussian perturbations than AdamW:** Muon experiences a larger increase in loss under

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

perturbations at every budget, and this difference grows with the pretraining data budget, mirroring the budget trend of the learning–forgetting tradeoff.

In Section 4, we investigate why the Muon–AdamW fine-tuning gap widens with increased pretraining data budget, by studying how robust Muon- and AdamW-trained models are to perturbations of their weights. Our analysis has two steps, supported by both theory and experiments. We first approximate the loss after a perturbation to second order in a linear associative memory, which motivates tracking the **margin-to-noise ratio (MNR)** as a proxy for robustness to perturbations. We then test whether the data-budget trend persists in linear memories trained with gradient descent (GD) and Muon (Section 4.2.1): Muon has the larger MNR at small data budgets, while GD has the larger MNR at large data budgets. Consistent with this, the Muon-trained linear model is the more robust to perturbations at small data budgets and the less robust at large ones, mirroring the widening perturbation degradation gap we observe in language models. Finally, we analyze a single update to understand how this dependence on data budget can arise.

2 Preliminaries

Our goal is to compare AdamW- and Muon-pretrained models *after fine-tuning*. Since fine-tuning can cause forgetting of the pretraining distribution, performance on the fine-tuning task alone is not sufficient: one base model may learn more but also forget more. We therefore compare base models by the tradeoff between the two, and define metrics to track how this comparison changes with model size and pretraining data budget. Let θ_{PT} denote a pretrained model.

Learning–forgetting tradeoff. Fine-tuning improves downstream performance but may degrade pretrained knowledge (Goodfellow et al., 2014; Kirkpatrick et al., 2017; Springer et al., 2025). When θ_{PT} is fine-tuned to obtain θ_{FT}, we measure learning by the fine-tuning loss $\mathcal{L}_{\mathrm{FT}}(\theta_{\mathrm{FT}})$ and forgetting by the pretraining loss $\mathcal{L}_{\mathrm{PT}}(\theta_{\mathrm{FT}})$. Since both depend on the fine-tuning hyperparameters, a single fine-tuned checkpoint does not capture how well θ_{PT} can be adapted. We therefore summarize each base model by the outcomes reachable by fine-tuning it,

$$\mathcal{T}(\theta_{\mathrm{PT}}) \coloneqq \left\{ \left(\mathcal{L}_{\mathrm{FT}}(\theta_{\mathrm{FT}}), \mathcal{L}_{\mathrm{PT}}(\theta_{\mathrm{FT}})\right) \,\middle|\, \theta_{\mathrm{FT}} \in \Theta_{\mathrm{FT}}(\theta_{\mathrm{PT}}) \right\},$$

where $\Theta_{\mathrm{FT}}(\theta_{\mathrm{PT}})$ is the set of models obtained by fine-tuning θ_{PT}, and compare base models through the lower-left Pareto frontier of this set, which we call the *learning–forgetting frontier*. For a collection of base models $\mathcal{B}$, we compare the corresponding sets

$$\mathcal{T}(\mathcal{B}) \coloneqq \{\mathcal{T}(\theta_{\mathrm{PT}}) \mid \theta_{\mathrm{PT}} \in \mathcal{B}\}.$$

Fine-tuning gap. To compare the two optimizers on this tradeoff with a single number, we define the *fine-tuning gap* G_{FT} as the mean signed distance from points sampled uniformly along the AdamW frontier to the Muon frontier. Each distance is negative where the Muon frontier lies closer to the origin, that is, where it is better, and positive otherwise. Thus $G_{\mathrm{FT}} < 0$ when Muon dominates and $G_{\mathrm{FT}} > 0$ when AdamW dominates (Section B.5 gives the details).

Perturbation degradation gap. Gaussian perturbations of the pretrained weights offer a simple and general proxy for fine-tuning updates: although they are not fine-tuning updates themselves, they measure how the pretraining loss degrades under matched weight changes across models (Springer et al., 2025). We add an isotropic Gaussian perturbation to each weight matrix of θ_{PT}, scaled to a fraction δ of the matrix's norm, and measure the *perturbation degradation*

$$\Delta_{\mathrm{noise}}(\theta_{\mathrm{PT}}; \delta) \;\coloneqq\; \mathcal{L}_{\mathrm{PT}}(\widetilde{\theta}_{\delta}) - \mathcal{L}_{\mathrm{PT}}(\theta_{\mathrm{PT}}), \tag{1}$$

where $\widetilde{\theta}_{\delta}$ denotes the perturbed weights, and we average over noise draws. The *perturbation degradation gap* is

$$G_{\mathrm{noise}}(\delta) \;\coloneqq\; \Delta_{\mathrm{noise}}(\theta_{\mathrm{PT}}^{\mathrm{Muon}}; \delta) - \Delta_{\mathrm{noise}}(\theta_{\mathrm{PT}}^{\mathrm{AdamW}}; \delta), \tag{2}$$

which, as for the fine-tuning gap, is positive when AdamW is more robust and negative when Muon is.

Crossover token–parameter ratio (TPR). We measure the pretraining data budget by the *token–parameter ratio* (TPR), the number of pretraining tokens per model parameter. For a fixed model size and fine-tuning dataset, the *crossover TPR* is the smallest TPR at which the fine-tuning gap becomes positive, that is, at which AdamW surpasses Muon.

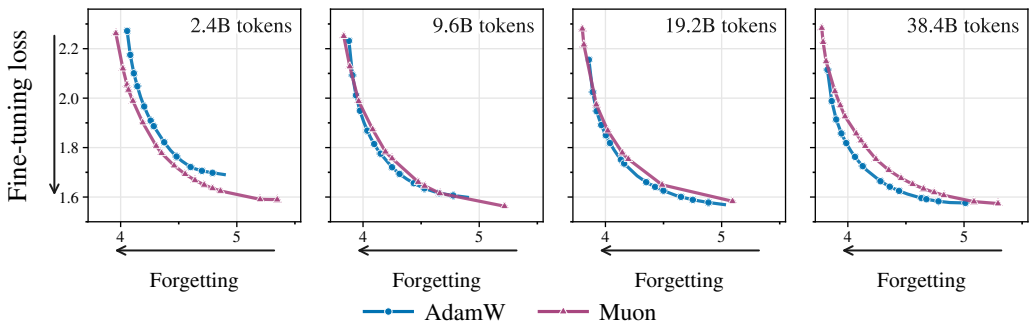

Figure 1: Beyond a sufficiently large data budget, AdamW outperforms Muon. We pretrain OLMo-60M models with Muon and AdamW and fine-tune on StackMathQA. For each pretraining optimizer, we sweep across all pretraining and fine-tuning configurations and plot the best learning–forgetting frontier, from 2.4B to 38.4B pretraining tokens. At 2.4B tokens, Muon's frontier dominates AdamW's; at larger budgets the trend reverses and AdamW surpasses Muon, with the gap growing up to 38.4B tokens. This reversal happens despite Muon achieving lower pretraining loss across all data budgets.

3 Experiments

We now compare how the learning–forgetting tradeoffs of Muon and AdamW scale with pretraining data budget and model size. In Section 3.2, we show that as the pretraining data budget increases, Muon's advantage over AdamW shrinks and eventually reverses in favor of AdamW, even though Muon always achieves lower pretraining loss. This reversal happens at progressively smaller data budgets as model size grows.

3.1 Experimental setup

Pretraining. We pretrain OLMo-style 30M–600M models on 150M to 154B DCLM tokens, comparing AdamW and Muon under a warmup–stable–decay (WSD) schedule. We sweep hyperparameters broadly and report the complete pretraining methodology in Section B.

Fine-tuning. For each base model, we evaluate the learning–forgetting tradeoff on four standard datasets: GSM8K (Cobbe et al., 2021), StackMathQA (Zhang, 2024), Alpaca (Taori et al., 2023), and MusicPile (Yuan et al., 2024), sweeping AdamW learning rates from 10^{-4} to 10^{-2} and Muon learning rates from 6×10^{-4} to 2×10^{-2} (Section B.2). We fine-tune AdamW-pretrained models with AdamW and Muon-pretrained models with both AdamW and Muon: AdamW fine-tuning consistently outperforms Muon fine-tuning on AdamW-pretrained models, whereas neither optimizer is consistently better on Muon-pretrained models (Section C.3). We then plot the learning–forgetting frontier over the resulting runs. We choose a broad learning-rate range to ensure we capture the complete learning–forgetting tradeoff spectrum: higher learning rates favor learning at the expense of forgetting, while lower rates preserve pretraining knowledge but learn less.

Evaluating the learning–forgetting tradeoff. For each fine-tuned checkpoint, we compute the fine-tuning loss and pretraining loss on held-out fine-tuning and pretraining (DCLM) data, and use these checkpoints to approximate the learning–forgetting tradeoff. Although this approximation is constructed from a finite set of fine-tuning runs, we conduct a comprehensive sweep over four datasets and across a wide range of fine-tuning learning rates to obtain a close estimate. For each pretraining optimizer, we sweep across all pretraining and fine-tuning configurations to achieve the best possible learning–forgetting frontier for a given pretraining data budget. We call this the *best-frontier setting* (Section B.1.3). The loss-optimal base model is not always the best model after fine-tuning: models pretrained with other hyperparameters can offer a better learning–forgetting tradeoff (Watts et al., 2026).

3.2 Main result

We now investigate how the data budget and model size affect the Muon–AdamW fine-tuning gap.

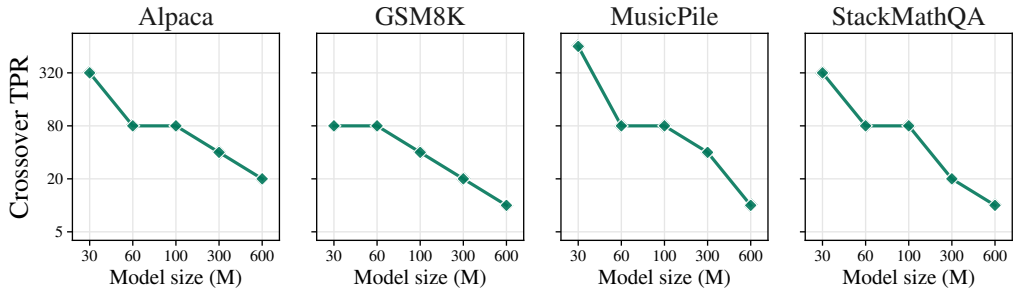

Figure 2: Crossover TPR decreases with model size. In the best-frontier setting, where each pretraining optimizer is swept across all pretraining and fine-tuning configurations, Muon's advantage reverses at substantially smaller pretraining data budgets as model size increases. The exact crossover TPR varies depending on the fine-tuning dataset, but the high-level trend is consistent.

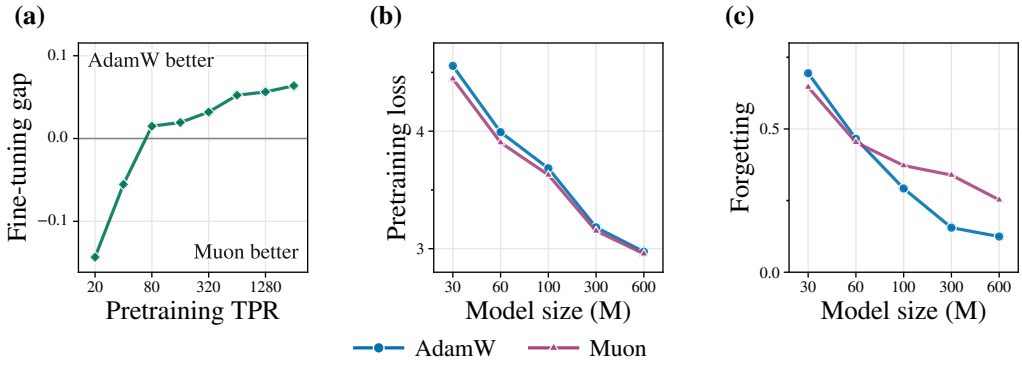

Figure 3: The Muon–AdamW fine-tuning gap widens with increasing pretraining data budget, and Muon's advantage shrinks with model size. (a) The fine-tuning gap after fine-tuning OLMo-60M models on StackMathQA, in the best-frontier setting (each pretraining optimizer swept across all pretraining and fine-tuning configurations). As the pretraining data budget increases, the fine-tuning gap transitions from negative to positive, with AdamW outperforming Muon beyond the crossover TPR. **(b)** Pretraining loss at 40 TPR. Muon's advantage over AdamW shrinks as model size grows. **(c)** Forgetting at 40 TPR, measured at a fine-tuning loss within 10% of the best on StackMathQA. Muon forgets increasingly more than AdamW as model size grows. In (b) and (c), for each optimizer we take the single pretrained model whose fine-tuning runs form the best learning–forgetting frontier, and report that model's pretraining loss and forgetting.

Main result: Pretraining data scale reverses the Muon–AdamW fine-tuning gap, despite Muon always achieving lower pretraining loss. We compare the fine-tuning gap for AdamW and Muon at different pretraining data budgets in the best-frontier setting, using StackMathQA as the fine-tuning dataset (Figure 3a). At small pretraining data budgets (2.4B tokens), Muon Pareto-dominates AdamW (Figure 1) and the fine-tuning gap is negative. At larger pretraining data budgets, this trend reverses: beyond 2.4B tokens, AdamW achieves lower fine-tuning loss with less forgetting than Muon. The reversal does not depend on sweeping pretraining configurations: it also holds when we compare only the loss-optimal base models (Figure 12). We include results for additional fine-tuning datasets in Section C.

We next investigate how the reversal of the Muon–AdamW fine-tuning gap scales with model size. Specifically, we ask how the amount of pretraining data needed for AdamW to overtake Muon changes as model size grows.

As model size scales, AdamW surpasses Muon at progressively smaller data budgets. We find that the crossover TPR decreases with model size. On StackMathQA, the crossover is 320 TPR for 30M models but only 10 TPR for 600M models (Figure 2). The specific crossover TPR depends on the fine-tuning dataset, but the broader trend persists; as model size grows, the crossover TPR decreases, and Muon's advantage reverses with substantially smaller data budgets.

We hypothesize that this scaling trend is driven by two factors (Figure 3b,c). For each optimizer and model size, we take the pretrained model that produces the best learning–forgetting frontier, and compare its base pretraining loss and its forgetting after fine-tuning on StackMathQA. First, for a fixed TPR, the gap in pretraining loss between Muon and AdamW shrinks as model size increases, reducing Muon's advantage as a lower-loss base model. Additionally, for a fixed TPR and fine-tuning loss, Muon forgets increasingly more than AdamW as model size grows. Consequently, as model size scales, Muon's pretraining advantage over AdamW narrows, while its forgetting relative to AdamW increases, allowing AdamW to overtake Muon at progressively smaller data budgets.

Muon is less robust to Gaussian perturbations than AdamW. To understand more generally how robust Muon- and AdamW-pretrained models are to weight updates, we study Gaussian perturbations, a simple and general proxy for fine-tuning updates (Section 2). For each checkpoint, we add isotropic Gaussian noise to each weight matrix, scaled to a fraction δ of its norm, and measure the perturbation degradation Δ_{noise}, the resulting increase in pretraining loss (Equation (1); see Section D.1 for full details). The perturbation degradation gap G_{noise} is positive at every budget and widens as the pretraining data budget grows (Figure 4), matching the budget trend of the learning–forgetting tradeoff. In other words, as the data budget increases, Muon-pretrained models exhibit increasingly greater degradation under weight perturbations than AdamW-pretrained models.

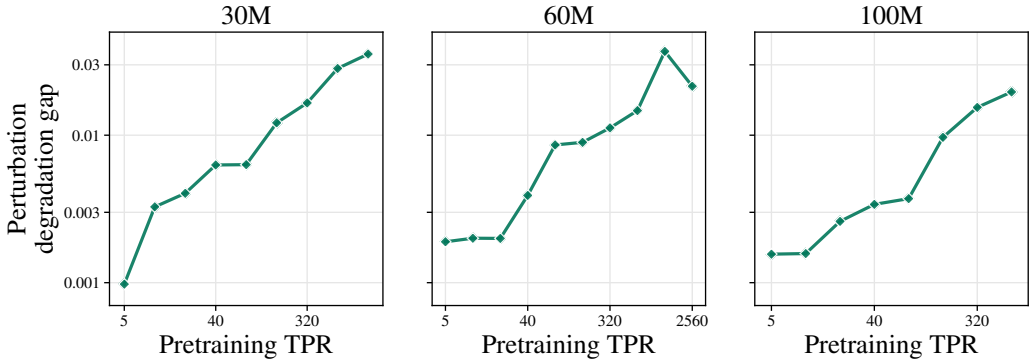

Figure 4: The perturbation degradation gap grows with pretraining data budget. Under a perturbation of magnitude $\delta = 0.02$, Muon degrades increasingly more than AdamW as the pretraining data budget grows. Points show the mean over three noise draws.

4 Analyzing the Muon–AdamW Fine-Tuning Gap

In Section 3, we show that Muon offers a better learning–forgetting tradeoff at small data budgets, but AdamW overtakes it beyond the crossover budget. We observe a similar trend under Gaussian perturbations of the pretrained weights (Section 3.2), where Muon-pretrained models degrade increasingly more than AdamW-pretrained models as the pretraining data budget grows. In this section, we investigate why the Muon–AdamW fine-tuning gap widens with increased pretraining data budget. Our analysis has two steps, supported by both theory and experiments.

Roadmap. We first approximate the loss after a perturbation to second order in a linear associative memory, which motivates tracking the **margin-to-noise ratio (MNR)** as a proxy for robustness to perturbations (Section 4.1). We then test whether the budget trend persists in linear memories trained with GD and Muon, and analyze a single update to understand how this dependence on data budget can arise (Section 4.2).

4.1 Approximating perturbation degradation

To understand which statistics of a checkpoint govern its degradation under a perturbation, we study a linear associative memory. We approximate the loss after a perturbation to second order in the perturbation size, which motivates the MNR.

4.1.1 Theoretical setup

We represent N facts by key–value pairs $(u_i, v_i) \in \mathbb{R}^d \times \mathbb{R}^d$, with all keys and values drawn independently from $\mathcal{N}(0, I_d)$. The memory is a matrix $W \in \mathbb{R}^{d\times d}$ that scores each candidate value as

$s_{ij}(W) := v_j^\top W u_i$ and retrieves the highest-scoring one; fact i is recovered when $s_{ii}(W) > s_{ij}(W)$ for all $j \neq i$. For a query distribution q over the facts, the predicted probabilities and cross-entropy loss are

$$\hat{p}_W(\cdot \mid i) := \mathrm{softmax}\big(s_{i1}(W), \ldots, s_{iN}(W)\big), \qquad \mathcal{L}(W) := -\textstyle\sum_{i=1}^{N} q_i \log \hat{p}_W(i \mid i). \tag{3}$$

Asymptotics and notation. We let d and N grow together with their ratio $\gamma = N/d > 1$ fixed, so $N - 1 \geq d$ and $\log N = o(d)$. **We call γ the data budget:** it measures facts per dimension, playing the role of tokens per parameter in Section 2. For random variables X and a deterministic $a > 0$, $X = O_p(a)$ means that for every $\varepsilon > 0$ there is a C with $\Pr[|X| > Ca] \leq \varepsilon$ eventually, and $X = o_p(a)$ means that $\Pr[|X| > \varepsilon a] \to 0$ for every $\varepsilon > 0$.

4.1.2 Second-order approximation

We perturb a memory W along a fixed direction D, $W_\varepsilon := W + \varepsilon D$, where D can be a fine-tuning update or a draw of Gaussian noise (Section 3.2). For one fact i, the scores are linear in W, so each margin $m_j := s_{ii}(W) - s_{ij}(W)$ changes exactly at rate $d_j := s_{ii}(D) - s_{ij}(D)$, and the loss increase is exactly

$$\Delta L(\varepsilon) = \log\Big(1 + \sum_{j \neq i} e^{-m_j - \varepsilon d_j}\Big) - \log\Big(1 + \sum_{j \neq i} e^{-m_j}\Big). \tag{4}$$

With $p_j := \hat{p}_W(j \mid i)$ the predicted probabilities before the perturbation, the first two derivatives of the loss at $\varepsilon = 0$ are the probability-weighted mean and variance of the margin changes, $h_D := \sum_j p_j d_j$ and $\kappa_D^2 := \sum_j p_j (d_j - h_D)^2$. Keeping only these two terms gives

$$\Delta L(\varepsilon) \approx -h_D \varepsilon + \frac{\kappa_D^2}{2}\varepsilon^2 = -\rho_D x + \frac{x^2}{2}, \qquad \rho_D := \frac{h_D}{\kappa_D}, \quad x := \varepsilon \kappa_D. \tag{5}$$

This approximation drops the higher-order terms and evaluates the curvature at the unperturbed probabilities, so it holds only for small perturbations. At a fixed perturbation size x, the ratio ρ_D of the first- to the second-order term sets the loss degradation.

The ratio h/κ approximates ρ_D. For a step from zero weights along a learned matrix A, as in the one-step setting of Section 4.2.2, the probabilities are uniform and the d_j are the margins of A. With h and κ their mean and standard deviation,

$$h_D = \frac{N-1}{N}\, h, \qquad \kappa_D^2 = \frac{N-1}{N}\, \kappa^2 + \frac{N-1}{N^2}\, h^2, \tag{6}$$

so $\rho_D \approx h/\kappa$ for large N when h^2/N is small compared with κ^2. This correspondence relies on the special structure of the zero-weight setting, where the learned matrix itself defines the perturbation direction. For a trained memory subjected to a random weight perturbation, the expected first-order term vanishes, so the loss degradation is governed by the curvature κ_D^2 rather than by the ratio h/κ. We therefore use h/κ as a tractable proxy for ρ_D motivated by this expansion. Section D.2 gives the full derivations.

4.1.3 Margin-to-noise ratio

Motivated by this approximation, we track h/κ, the ratio of the mean margin to the standard deviation of the margins, as a proxy for robustness to perturbation. We call it the margin-to-noise ratio.

Definition 4.1 (Margin-to-noise ratio). For fact i, define the pairwise margins $m_{ij} := s_{ii}(W) - s_{ij}(W)$ for $j \neq i$, with mean and variance

$$h_i := \operatorname*{mean}_{j \neq i} m_{ij}, \qquad \kappa_i^2 := \operatorname*{mean}_{j \neq i} (m_{ij} - h_i)^2. \tag{7}$$

The *margin-to-noise ratio* (MNR) for fact i is $\rho_i := h_i/\kappa_i$, for $\kappa_i > 0$. At the model level, we average the mean margins and variances uniformly over facts before taking the ratio:

$$\bar{h} := \operatorname*{mean}_i h_i, \qquad \bar{\kappa}^2 := \operatorname*{mean}_i \kappa_i^2, \qquad \rho := \frac{\bar{h}}{\bar{\kappa}}, \tag{8}$$

where ρ is defined for $\bar{\kappa} > 0$.

We average over facts before taking the ratio because the first- and second-order terms of the mean loss are the means of the per-fact terms. We next track the MNR of a linear memory trained with GD and Muon as the data budget grows.

4.2 MNR OF GD AND MUON ACROSS DATA BUDGETS

Section 4.1 motivates the MNR as a proxy for robustness to perturbation. Motivated by this and by the budget trend in Section 3.2, we track the MNR of a linear memory trained with GD and Muon as the data budget grows, and ask whether the same shift away from Muon appears. Using GD tests whether this trend also arises without AdamW's coordinatewise adaptive rescaling, while providing a tractable baseline for analyzing Muon's update.

We study the MNR ratio $\mathcal{R} := \rho^{\mathrm{GD}}/\rho^{\mathrm{Muon}}$ rather than the difference, as the ratio arises naturally in the one-step analysis below (Section 4.2.2).

4.2.1 GD AND MUON IN THE LINEAR MEMORY

We train the linear memory for $T = 10^4$ full-batch steps from $W_0 = 0$ at constant learning rates across three seeds. We summarize the saturated MNR by the per-seed median over $t \geq 10^3$ and measure robustness by σ_{50}, the relative perturbation size at which half of the initially recovered facts are lost. We exclude data budgets past the memory's capacity, where the two optimizers recover different sets of facts. The full setup and results for every dimension d appear in Section D.3.

The results in Figure 5 show that $\mathcal{R}$ increases with the data budget. Muon has the larger MNR at small γ, while GD has the larger MNR at large γ. The robustness comparison shifts in the same direction: Muon tolerates stronger perturbations at small γ, and GD does at large γ. Thus, the budget trend persists in the linear setting with GD as the baseline.

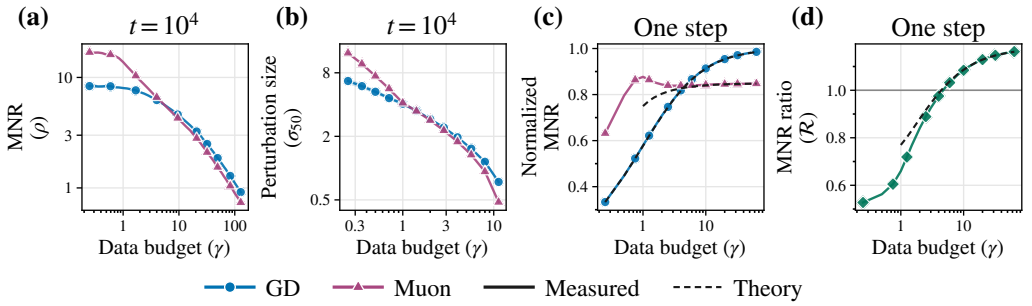

Figure 5: The margin-to-noise ratio of GD gains on Muon's as the data budget grows. (a, b) After $t = 10^4$ steps with model dimension $d = 256$, GD's MNR ρ overtakes Muon's as γ grows, and the ordering of the perturbed robustness σ_{50} turns over with it. **(c, d)** After one step with $d = 512$, GD's normalized MNR $\rho/\sqrt{d/\gamma}$ overtakes Muon's (c), so the MNR ratio $\mathcal{R} = \rho^{\mathrm{GD}}/\rho^{\mathrm{Muon}}$ rises with γ (d), following the theory of Theorem 4.2.

4.2.2 ONE-STEP ANALYSIS

To understand how the budget dependence observed after many training steps can arise, we analyze a single update from zero initialization. The rescaled descent direction is $G := -N\nabla\mathcal{L}(0) = \sum_{k=1}^{N}(v_k - \bar{v})u_k^\top$, where $\bar{v} := N^{-1}\sum_{k=1}^{N} v_k$. Up to positive scalar factors, GD uses G, while Muon uses its *polar factor*: if $G = P\Sigma Q^\top$ is the singular value decomposition, this factor is $\mathrm{Orth}(G) := PQ^\top$. It preserves the singular vectors while replacing the singular values by one.

Theorem 4.2 (One-step MNR ratio across data budgets). *Consider one GD update and one Muon update from $W_0 = 0$ in the setting of Section 4.1.1. For every fixed data budget $\gamma = N/d \geq 16$, their MNR ratio $\mathcal{R} = \rho^{\mathrm{GD}}/\rho^{\mathrm{Muon}}$ satisfies*

$$\left|\mathcal{R} - \frac{3\pi}{8}\left(1 - \frac{9}{10\gamma}\right)\right| \;\leq\; C_1\gamma^{-3/2} + o_p(1) \tag{9}$$

as $d \to \infty$, where C_1 is an absolute constant.

Remark (GD's advantage at large budgets). In the approximation of Equation (9), the term $9/(10\gamma)$ decreases as the data budget γ grows, so the predicted ratio increases toward $\frac{3\pi}{8} \approx 1.18$. Thus, GD has the larger MNR at sufficiently large budgets, and the approximation predicts that its relative advantage grows with budget. The proof explains this trend by comparing how the mean margin and its variability depend on budget under the two updates. This shift already arises after a single update in the Gaussian linear memory and agrees with the many-step trend for GD and Muon (Section 4.2.1). It also parallels the shift toward AdamW in language models (Section 3.2). Together with the second-order motivation for tracking the MNR (Section 4.1), the result provides a simple mechanism consistent with the observed robustness trend.

Proof sketch. Starting from zero, we compare the one-step weights $W^{\mathrm{GD}} = G$ and $W^{\mathrm{Muon}} = O := \mathrm{Orth}(G)$, omitting learning-rate factors because they cancel in the MNR. The mean margin satisfies $\bar{h}(W) = \langle W, G\rangle/(N-1)$, so

$$\bar{h}^{\mathrm{GD}} = \frac{\|G\|_F^2}{N-1}, \qquad \bar{h}^{\mathrm{Muon}} = \frac{\|G\|_*}{N-1}.$$

The limiting singular-value distribution of G under the isotropic Gaussian setup determines these norms, including the factor $8/(3\pi)$ in Muon's margin.

We next estimate $\bar{\kappa}^2$, the mean of the margin variances κ_i^2. For GD, the leading terms in $v_j^\top G u_i$ are $(v_j^\top v_k)(u_k^\top u_i)$, one for each fact k. These terms are uncorrelated; those for facts i and j contain $\|u_i\|^2$ and $\|v_j\|^2$, respectively. Each contributes variance d^3 to leading order, compared with d^2 from each remaining fact. Section E controls the lower-order centering corrections and proves concentration of the empirical average, giving $(\bar{\kappa}^{\mathrm{GD}})^2 = d^3(\gamma+2)(1+o_p(1))$. For Muon, the score is $v_j^\top O u_i$. If O were independent of u_i, v_j, orthogonality would give variance $\|O\|_F^2 = d$. These vectors are used to compute G, and hence O, so the proof isolates independent Gaussian components and bounds the correction due to this dependence, yielding $(\bar{\kappa}^{\mathrm{Muon}})^2 = d[1 + O(\gamma^{-3/2}) + o_p(1)]$. Combining these estimates gives

$$\rho^{\mathrm{GD}} = \sqrt{\frac{d}{\gamma+2}}\,(1+o_p(1)),$$

$$\rho^{\mathrm{Muon}} = \sqrt{\frac{d}{\gamma}}\left[\frac{8}{3\pi}\left(1-\frac{1}{10\gamma}\right) + O(\gamma^{-3/2}) + o_p(1)\right].$$

Taking their ratio, we keep the limiting value and the first budget-dependent term to obtain Equation (9). See Section E for the full proof and Figure 5 and Section D.3 for numerical checks.

Remark (Choice of data distribution). Independent standard Gaussian keys and values provide a canonical isotropic model with no preferred direction. Other data distributions can change both the gradient spectrum and the competing-score variances, so the optimizer ordering may differ. The analogous budget trend in the language-model experiments (Section 3.2) suggests that the behavior captured by this simple model is relevant beyond the Gaussian setting.

5 Related Work

Muon. Muon, an optimizer introduced by Jordan et al. (2024), has recently shown strong empirical performance for pretraining LLMs, challenging AdamW's dominance as the status-quo optimizer. Recent work has sought to explain Muon's pretraining advantage through analyses of simplified theoretical models, associative memory, convergence behavior, and loss landscape geometry (Kim et al., 2026; Li et al., 2026; Sato et al., 2025; Su, 2025; Shen et al., 2025; Wang et al., 2026a). Another line of work investigates the training dynamics of Muon, finding that it learns data components more uniformly than AdamW (Vasudeva et al., 2026; Wang et al., 2026b; Dragutinović et al., 2026), while also characterizing the implicit biases of Muon (Fan et al., 2025; Dragutinović et al., 2026). Yet fewer works have examined the downstream consequences of Muon. Ruan et al. (2026) compare Muon and AdamW after fine-tuning; they find that for small data regimes, Muon outperforms AdamW, consistent with our findings. We show that this advantage does not generalize across pretraining data scale; after a sufficiently large pretraining data budget, the trend reverses, and AdamW outperforms Muon.

Muon versus AdamW on pretraining loss. Muon has been reported to reach lower pretraining loss than AdamW for the same compute across a wide range of settings (Liu et al., 2025; Essential AI et al., 2025; Mehta et al., 2025), and frontier models have since been pretrained with Muon or its variants (Kimi Team, 2025; GLM-4.5 Team et al., 2025). The size of this advantage depends on scale: under controlled tuning, Muon's speedup over AdamW shrinks as model size and data grow, and Muon is overtaken by other matrix-based optimizers at large data-to-model ratios (Wen et al., 2026b;a). All of these works measure pretraining loss. Optimizers can also reach qualitatively different solutions rather than faster or slower routes to the same minimum (Pascanu et al., 2025); we ask whether these differences matter after fine-tuning.

Catastrophic forgetting. Catastrophic forgetting, the degradation of previously learned knowledge during subsequent training, remains a central challenge in deep learning (Goodfellow et al., 2014). Numerous approaches have been proposed to mitigate catastrophic forgetting, including regularization-based methods (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018), replay techniques (Shin et al., 2017), gradient projection methods (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019), and optimization-based approaches (Watts et al., 2026). Our learning–forgetting tradeoff is an instance of the stability–plasticity dilemma, with fine-tuning performance capturing plasticity and knowledge retention capturing stability (Lopez-Paz & Ranzato, 2017). We study forgetting as an intrinsic property of the pretrained checkpoint, rather than as a consequence of the fine-tuning algorithm. Specifically, we investigate how optimization choices during pretraining shape downstream forgetting, showing that Muon and AdamW induce different learning–forgetting tradeoffs.

Pretraining and downstream adaptation. Better pretraining loss does not always yield a better model after fine-tuning. Springer et al. (2025) show that extending pretraining can make language models harder to fine-tune, as the pretrained parameters become increasingly sensitive to modifications, which they measure in part with Gaussian perturbations. Watts et al. (2026) show that sharpness-aware pretraining mitigates forgetting during fine-tuning, and Liu et al. (2026) find that fine-tuning with the same optimizer as pretraining forgets less. We study how the choice of pretraining optimizer interacts with the pretraining data budget, and connect the resulting fine-tuning gap to robustness to weight perturbations.

6 Conclusion

We asked whether Muon's pretraining advantage over AdamW persists after fine-tuning, and found that it depends on the pretraining data budget: beyond a sufficiently large budget, AdamW outperforms Muon after fine-tuning, and it does so at progressively smaller token–parameter ratios as model size grows, even though Muon always achieves lower pretraining loss. Gaussian perturbations of the pretrained weights show the same trend, with Muon-pretrained models degrading increasingly more than AdamW-pretrained models as the data budget grows. In a linear model, the margin-to-noise ratio, a tractable proxy for robustness, shifts from favoring Muon to favoring gradient descent as the data budget grows, which we prove for one step and observe over training. This mirrors the fine-tuning trend and offers a simple lens on why Muon's advantage reverses.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing and editing. Additionally, we used generative AI tools to create and edit the plotting and analysis code for the figures, and to draft and edit parts of the paper for readability. We reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 139–154, 2018.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Sara Dragutinović, Yedi Zhang, and Rajesh Ranganath. To use or not to use Muon: How simplicity bias in optimizers matters, 2026. URL https://arxiv.org/abs/2603.00742.

Essential AI, Ishaan Shah, Anthony M. Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, et al. Practical efficiency of Muon for pretraining, 2025. URL https://arxiv.org/abs/2505.02222.

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

GLM-4.5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, et al. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models, 2025. URL https://arxiv.org/abs/2508.06471.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6211.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, 2024. URL https://aclanthology.org/2024.acl-long.841.

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. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2203.15556.

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

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.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 1091-6490. doi: 10.1073/pnas.1611835114. URL http://dx.doi.org/10.1073/pnas.1611835114.

Binghui Li, Kaifei Wang, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.05725.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Kamal Mohamed Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Joshua P Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah M Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham M. Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander T Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alex Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-LM: In search of the next generation of training sets for language models. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=CNWdWn47IE.

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, 2025. URL https://arxiv.org/abs/2502.16982.

Yuxing Liu, Jianyu Wang, and Tong Zhang. Optimizer-model consistency: Full finetuning with the same optimizer as pretraining forgets less, 2026. URL https://arxiv.org/abs/2605.06654.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Sushant Mehta, Raj Dandekar, Rajat Dandekar, and Sreedath Panat. Muon: Training and trade-offs with latent attention and MoE, 2025. URL https://arxiv.org/abs/2509.24406.

Razvan Pascanu, Clare Lyle, Ionut-Vlad Modoranu, Naima Elosegui Borras, Dan Alistarh, Petar Velickovic, Sarath Chandar, Soham De, and James Martens. Optimizers qualitatively alter solutions and we should leverage this, 2025. URL https://arxiv.org/abs/2507.12224.

Tianyu Ruan, Fengzhuo Zhang, Shuche Wang, and Shihua Zhang. Muon learns more robust and transferable features than Adam, 2026. URL https://arxiv.org/abs/2606.09658.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence bound and critical batch size of Muon optimizer, 2025. URL https://arxiv.org/abs/2507.01598.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon, 2025. URL https://arxiv.org/abs/2505.23737.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1705.08690.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2503.19206.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2501.00656.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YzjS4jcfmS.

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

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. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2509.26030.

Ishaan Watts, Catherine Li, Sachin Goyal, Jacob Mitchell Springer, and Aditi Raghunathan. Sharpness-aware pretraining mitigates catastrophic forgetting. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.02105.

Kaiyue Wen, Xingyu Dang, Kaifeng Lyu, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them II: Hyperball optimization, 2026a. URL https://arxiv.org/abs/2606.16899.

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

Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, Ziyang Ma, Liumeng Xue, Ziyu Wang, Qin Liu, Tianyu Zheng, Yizhi Li, Yinghao Ma, Yiming Liang, Xiaowei Chi, Ruibo Liu, Zili Wang, Pengfei Li, Jingcheng Wu, Chenghua Lin, Qifeng Liu, Tao Jiang, Wenhao Huang, Wenhu Chen, Emmanouil Benetos, Jie Fu, Gus Xia, Roger Dannenberg, Wei Xue, Shiyin Kang, and Yike Guo. ChatMusician: Understanding

and generating music intrinsically with LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6252–6271, 2024. URL https://aclanthology.org/2024.findings-acl.373/.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. PMLR, 2017.

Yifan Zhang. StackMathQA: A Curated Collection of 2 Million Mathematical Questions and Answers Sourced from Stack Exchange. Technical report, Math-AI, 2024.

APPENDIX CONTENTS

A Definitions **15**
A.1 Optimizers . . . 15
A.2 Learning-rate schedules . . . 16
A.3 Token–parameter ratio . . . 16

B Experimental Details **16**
B.1 Pretraining . . . 16
B.2 Fine-tuning . . . 17
B.3 Fine-tuning regularization . . . 18
B.4 Gaussian perturbations . . . 18
B.5 Evaluation . . . 18

C Additional Results for Section 3 **19**
C.1 Best-frontier setting . . . 19
C.2 Loss-optimal base models . . . 20
C.3 Choice of fine-tuning optimizer . . . 22
C.4 Fine-tuning regularization . . . 22

D Additional Results for Section 4 **25**
D.1 Perturbation robustness . . . 25
D.2 Loss under a perturbation: derivations . . . 26
D.3 Margin-to-noise ratio across data budgets . . . 29

E Proof of Theorem 4.2 **32**
E.1 The two one-step iterates . . . 32
E.2 Concentration estimates for the data . . . 33
E.3 The GD step . . . 37
E.4 The Muon step . . . 40
E.5 Combining the GD and Muon estimates . . . 46
E.6 Monotonicity in the data budget . . . 47

A Definitions

A.1 Optimizers

Optimizers update a model's parameters during training so as to decrease a loss. Given the gradient of the loss with respect to the parameters, an optimizer determines the direction and the magnitude of each update, and in doing so shapes training stability, convergence speed and the solution that is finally reached. This section recalls the two optimizers compared in the paper.

A.1.1 AdamW

Adam (Kingma & Ba, 2015) combines momentum with a per-coordinate step size. With $g_t = \nabla_\theta \mathcal{L}(\theta_t)$ the gradient at step t, it maintains exponential moving averages of the first and second moments of the gradient,

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \qquad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2,$$

with the square taken coordinatewise, corrects their bias towards the zero initialization,

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \qquad \hat{v}_t = \frac{v_t}{1-\beta_2^t},$$

and updates

$$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon},$$

where α is the learning rate, $\beta_1, \beta_2 \in [0,1)$ are the decay rates of the two averages, $\epsilon > 0$ guards the division, and the division and the square root are coordinatewise. In effect each coordinate is normalized by the scale of its own recent gradients.

Regularizing Adam with an ℓ_2 penalty $\frac{\lambda}{2}\|\theta\|^2$ adds $\lambda\theta_t$ to g_t, so the penalty is itself divided by $\sqrt{\hat{v}_t}+\epsilon$ and coordinates with large past gradients are regularized less. AdamW (Loshchilov & Hutter, 2019) instead decouples the weight decay from the adaptive step,

$$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} - \alpha\lambda\theta_t,$$

with g_t now the gradient of the unregularized loss and λ the weight decay coefficient, so that every coordinate decays at the same rate. AdamW is the standard optimizer for language-model pretraining and the baseline throughout the paper.

A.1.2 Muon

Muon (Jordan et al., 2024) treats each weight matrix as a whole rather than coordinate by coordinate. For a matrix-valued parameter θ with gradient G_t, it accumulates momentum $M_t = \mu M_{t-1} + G_t$ and updates along the orthogonalized momentum,

$$\theta_{t+1} = \theta_t - \eta \,\mathrm{Orth}(M_t), \tag{10}$$

where η is the learning rate and, for a matrix X with singular value decomposition $X = U\Sigma V^\top$, $\mathrm{Orth}(X) = UV^\top$ is its polar factor: the matrix with the singular vectors of X and all singular values set to one. The update therefore has the same size along every singular direction, and it can be read as the steepest descent direction when the size of a step is measured in the spectral norm (Jordan et al., 2024). An exact singular value decomposition at every step is expensive, so Orth is approximated by a few steps of a Newton–Schulz iteration,

$$X_0 = \frac{M_t}{\|M_t\|_F}, \qquad X_{k+1} = aX_k + b\,(X_k X_k^\top) X_k + c\,(X_k X_k^\top)^2 X_k,$$

whose odd matrix polynomial pushes every singular value of X_0 towards one; the reference implementation uses $(a,b,c) = (3.4445, -4.7750, 2.0315)$ and five steps. Muon is meant for the two-dimensional weight matrices of the hidden layers; embeddings, the output layer and one-dimensional parameters such as normalization gains are updated by AdamW. Large-scale implementations further add decoupled weight decay and rescale the update to the root-mean-square size of AdamW's (Liu et al., 2025). In our experiments the hidden matrices are updated by Muon and all other parameters by AdamW, each with its own learning rate and weight decay (Section 3.1).

A.2 Learning-rate schedules

A.2.1 Warmup–stable–decay

All pretraining runs use a warmup–stable–decay (WSD) schedule: the learning rate increases linearly from zero to its peak value over the warmup steps, stays at the peak for most of training, and is annealed to zero over the final decay phase.

A.2.2 Cosine

All fine-tuning runs use a cosine schedule: after a linear warmup of 20 steps, the learning rate follows $\eta_t = \frac{1}{2}\eta_{\max}\big(1 + \cos(\pi t/T)\big)$ over the remaining T steps.

A.3 Token–parameter ratio

We measure the pretraining data budget of a model with N parameters trained on D tokens by its token–parameter ratio, $\mathrm{TPR} = D/N$. The compute-optimal budget of Hoffmann et al. (2022) corresponds to $\mathrm{TPR} = 20$; our budgets range from 5 to 2,560 TPR.

B Experimental Details

B.1 Pretraining

Models and data. All language models are OLMo-style decoder-only transformers (Groeneveld et al., 2024; Team OLMo et al., 2025) with 30M, 60M, 100M, 300M and 600M parameters, trained on DCLM (Li et al., 2024) with the OLMo 2 tokenizer. Budgets span 5–1,280 TPR at 30M, 5–2,560 TPR at 60M, 5–640 TPR at 100M, 5–160 TPR at 300M and 5–40 TPR at 600M.

Optimization. Every model is pretrained with sequence length 4096, a global batch of 2^{20} tokens, a warmup–stable–decay (WSD) schedule, weight decay 0.1 and gradient clipping at 1.0. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.98)$. Muon updates the hidden weight matrices; the remaining parameters (embeddings, output layer and norms) are updated by AdamW.

B.1.1 Loss-optimal base models

For each model size, budget and optimizer, we sweep the peak learning rate and call the run with the lowest held-out DCLM loss the *loss-optimal* base model. For Muon we sweep the Muon learning rate and fix the learning rate of the AdamW-updated parameters to the tuned AdamW value of the same cell (Section B.1.2). Table 1 lists the resulting learning rates.

Table 1: Tuned peak pretraining learning rates per model size and budget. For Muon, the Muon learning rate; the AdamW-parameter learning rate is fixed to the tuned AdamW value.

Size	Optimizer	TPR									
		5	**10**	**20**	**40**	**80**	**160**	**320**	**640**	**1,280**	**2,560**
30M	AdamW	0.08	0.04	0.04	0.014	0.02	0.014	0.02	0.014	0.01	–
	Muon	0.01	0.01	0.01	0.014	0.02	0.01	0.014	0.01	0.014	–
60M	AdamW	0.056	0.04	0.02	0.02	0.01	0.014	0.01	0.01	0.007	0.01
	Muon	0.014	0.01	0.014	0.01	0.014	0.01	0.014	0.007	0.005	0.005
100M	AdamW	0.04	0.02	0.01	0.01	0.01	0.007	0.007	0.007	–	–
	Muon	0.014	0.014	0.014	0.014	0.01	0.014	0.01	0.005	–	–
300M	AdamW	0.001	0.0025	0.0035	0.0025	0.0025	0.0025	–	–	–	–
	Muon	0.02	0.014	0.02	0.014	0.014	0.02	–	–	–	–
600M	AdamW	0.0025	0.0025	0.0025	0.0025	–	–	–	–	–	–
	Muon	0.014	0.014	0.014	0.014	–	–	–	–	–	–

B.1.2 Learning rate of Muon's AdamW-updated parameters

Muon has a second learning rate for the parameters it leaves to AdamW. Rather than tuning it in every cell, we set it to the tuned AdamW learning rate of the same cell. To check this choice, we pretrain 60M Muon models at 20–160 TPR over a grid of Muon learning rates and three to four AdamW learning rates between $0.5\times$ and $2.8\times$ the tuned value. Figure 6 shows the best loss over Muon learning rates for each AdamW learning rate. It varies by at most 0.021 nats, and the tuned value is within 0.016 nats of the best, well below the 0.03–0.20 nats by which Muon beats AdamW at these budgets.

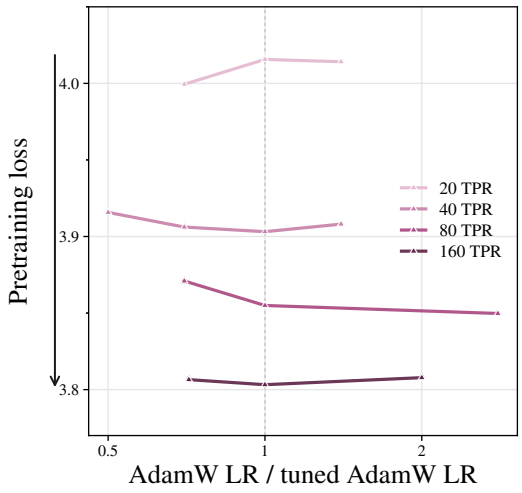

Figure 6: Muon is insensitive to the learning rate of its AdamW-updated parameters. Best pretraining loss over Muon learning rates for each AdamW learning rate, relative to the tuned AdamW value (dashed), for 60M models.

B.1.3 Best-frontier setting

The loss-optimal base model is not always the best starting point for fine-tuning: models pretrained with a larger learning rate can offer a better learning–forgetting tradeoff (Watts et al., 2026). Unless stated otherwise, we therefore compare the two optimizers in the *best-frontier setting*, which sweeps over both pretraining and fine-tuning configurations. For each model size, budget and pretraining optimizer, we pool the fine-tuning runs of the loss-optimal base model and of the base models pretrained at the next five larger learning rates of the sweep grid $\{0.005, 0.007, 0.01, 0.014, 0.02, 0.028, 0.04, 0.056, 0.08, 0.12, 0.16\}$, and compute a single learning–forgetting frontier from the pooled runs. Section C repeats the main results with the loss-optimal base models only.

B.2 Fine-tuning

Each base model is fine-tuned for 20M tokens on each of Alpaca (Taori et al., 2023), GSM8K (Cobbe et al., 2021), MusicPile (Yuan et al., 2024) and StackMathQA (Zhang, 2024), with sequence length 1024, a global batch of 65,536 tokens (about 305 steps), a cosine schedule with 20 warmup steps, weight decay 0.1, $(\beta_1, \beta_2) = (0.9, 0.95)$ and gradient clipping at 1.0. AdamW-pretrained models are fine-tuned with AdamW, and Muon-pretrained models with both AdamW and Muon (Section C.3 motivates this choice). We sweep AdamW learning rates from 10^{-4} to 10^{-2}, and Muon learning rates from 6×10^{-4} to 2×10^{-2} with the learning rate of the AdamW-updated parameters set to $0.25\times$ the Muon learning rate. The sweep ranges from runs that barely move the base model to runs that reach the lowest fine-tuning loss at the cost of more than a nat of forgetting.

B.3 FINE-TUNING REGULARIZATION

A natural question is whether the reversal survives when fine-tuning itself is regularized to limit forgetting. We test two standard methods for mitigating catastrophic forgetting, data replay (Shin et al., 2017) and elastic weight consolidation (EWC; Kirkpatrick et al., 2017), on the loss-optimal base models of every size. Both keep the fine-tuning recipe and learning-rate sweep of Section B.2, and each frontier pools all strengths of the regularizer. Section C.4 reports the results.

Replay. We replace a fraction $r \in \{0.1, 0.2, 0.3\}$ of the fine-tuning tokens with DCLM data drawn from a shard that is disjoint from both the pretraining data and the held-out evaluation shard.

EWC. EWC penalizes movement away from the pretrained weights in the directions that matter for pretraining. It minimizes

$$\mathcal{L}_{\mathrm{FT}}(\theta) + \frac{\lambda}{2} \sum_k F_k \big(\theta_k - \theta_{\mathrm{PT},k}\big)^2, \tag{11}$$

where F is the diagonal empirical Fisher information of the pretraining loss at θ_{PT}, estimated from squared gradients over 100 batches of the disjoint DCLM shard used for replay. Following Watts et al. (2026), F is not normalized, so we sweep $\lambda \in \{10^3, 4 \cdot 10^3, 10^4, 4 \cdot 10^4\}$.

B.4 GAUSSIAN PERTURBATIONS

We perturb each base model by adding independent isotropic Gaussian noise to every parameter tensor W, scaled so that $\|\Delta W\|_F = \delta \|W\|_F$, and measure the perturbation degradation $\Delta_{\mathrm{noise}}(\theta_{\mathrm{PT}}; \delta)$ of Equation (1). We average over three noise draws per checkpoint and magnitude, use the loss-optimal checkpoint of each cell, and consider magnitudes from $\delta = 0.002$ to 0.56.

B.5 EVALUATION

Losses. Forgetting is the loss on a held-out DCLM shard of about 4.19M tokens that no model is trained on, and learning is the loss on the validation split of the fine-tuning dataset. Both are mean next-token cross-entropy in nats.

Frontiers. Each fine-tuned checkpoint θ_{FT} of a base model gives a point $\big(\mathcal{L}_{\mathrm{PT}}(\theta_{\mathrm{FT}}), \mathcal{L}_{\mathrm{FT}}(\theta_{\mathrm{FT}})\big)$. We compute the learning–forgetting frontier as the lower-left arc of the convex hull of these points, treated as a polyline.

Fine-tuning gap and crossover. We estimate the fine-tuning gap G_{FT} of Section 2 from 100 points sampled uniformly by arc length along the AdamW frontier, each assigned its Euclidean distance to the nearest point of the Muon frontier, negative when that point lies closer to the origin. The crossover TPR is the smallest budget at which $G_{\mathrm{FT}} > 0$.

C Additional Results for Section 3

This section extends the fine-tuning results of Section 3. Unless noted otherwise, figures show 60M models, and frontiers are restricted to the region of Section B.5.

C.1 Best-frontier setting

Figures 7 and 8 show the learning–forgetting frontiers and fine-tuning gaps of 60M models on all four fine-tuning datasets in the best-frontier setting of the main text. On every dataset, the Muon frontier is better below the crossover; at the crossover AdamW has caught up with or overtaken Muon, and beyond it the AdamW frontier dominates. The fine-tuning gap turns positive at the crossover and stays positive, broadly growing with the data budget.

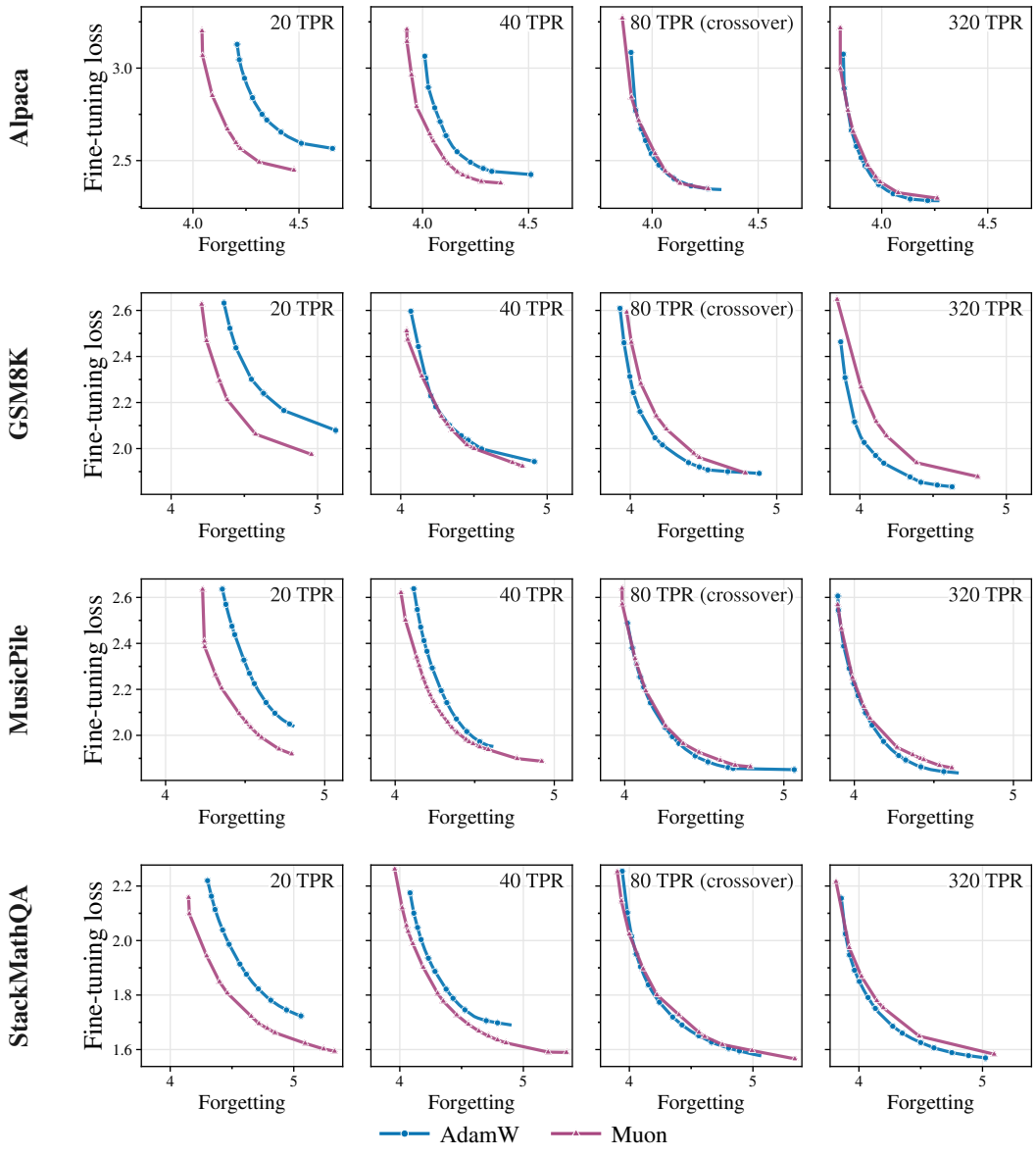

Figure 7: Learning–forgetting frontiers on every fine-tuning dataset, best-frontier setting (60M). Each row shows the budgets around that dataset's crossover TPR.

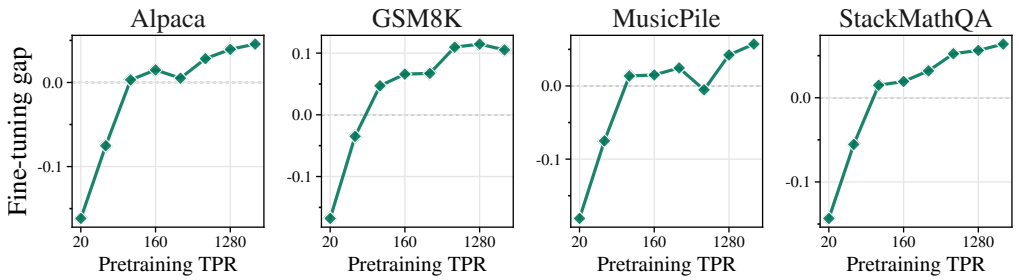

Figure 8: Fine-tuning gap on every fine-tuning dataset, best-frontier setting (60M). Negative values mean the Muon frontier dominates.

C.2 Loss-optimal base models

We repeat the analyses of Section 3 using only each optimizer's loss-optimal base model at each budget. Figure 9 shows the crossover TPR across model sizes, and Figure 10 shows the frontiers of Figure 1 on StackMathQA. Figures 11 and 12 extend the frontiers and gaps to all four datasets, at the budgets around each crossover. The trends match the best-frontier setting: the crossover decreases with model size, and on every dataset Muon is better than or level with AdamW below the crossover and AdamW is better beyond it. Figure 13 confirms that this reversal is not inherited from pretraining loss. The loss-optimal Muon model reaches a lower pretraining loss than its AdamW counterpart at every budget and model size in our best-frontier experiments.

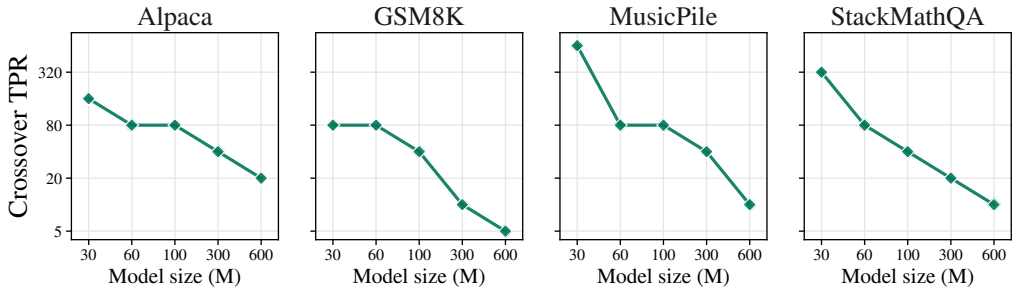

Figure 9: Crossover TPR decreases with model size, for the loss-optimal base models. As model size increases, Muon's advantage reverses with substantially smaller pretraining data budgets.

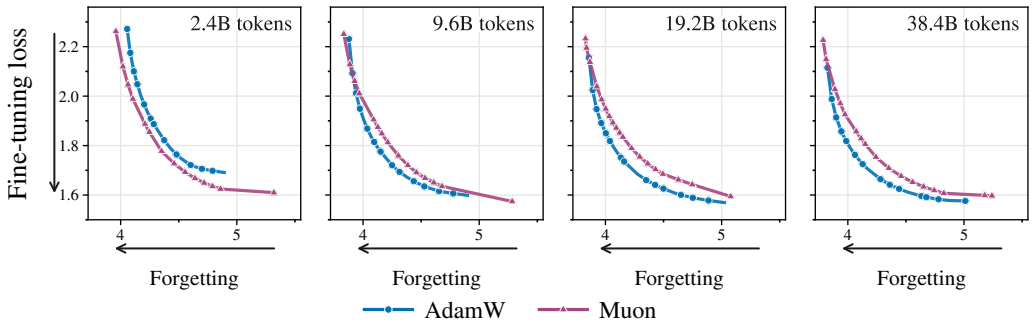

Figure 10: Learning–forgetting frontiers of the loss-optimal base models (60M, StackMathQA). Same budgets, fine-tuning sweep and window as Figure 1, restricted to each optimizer's loss-optimal base model.

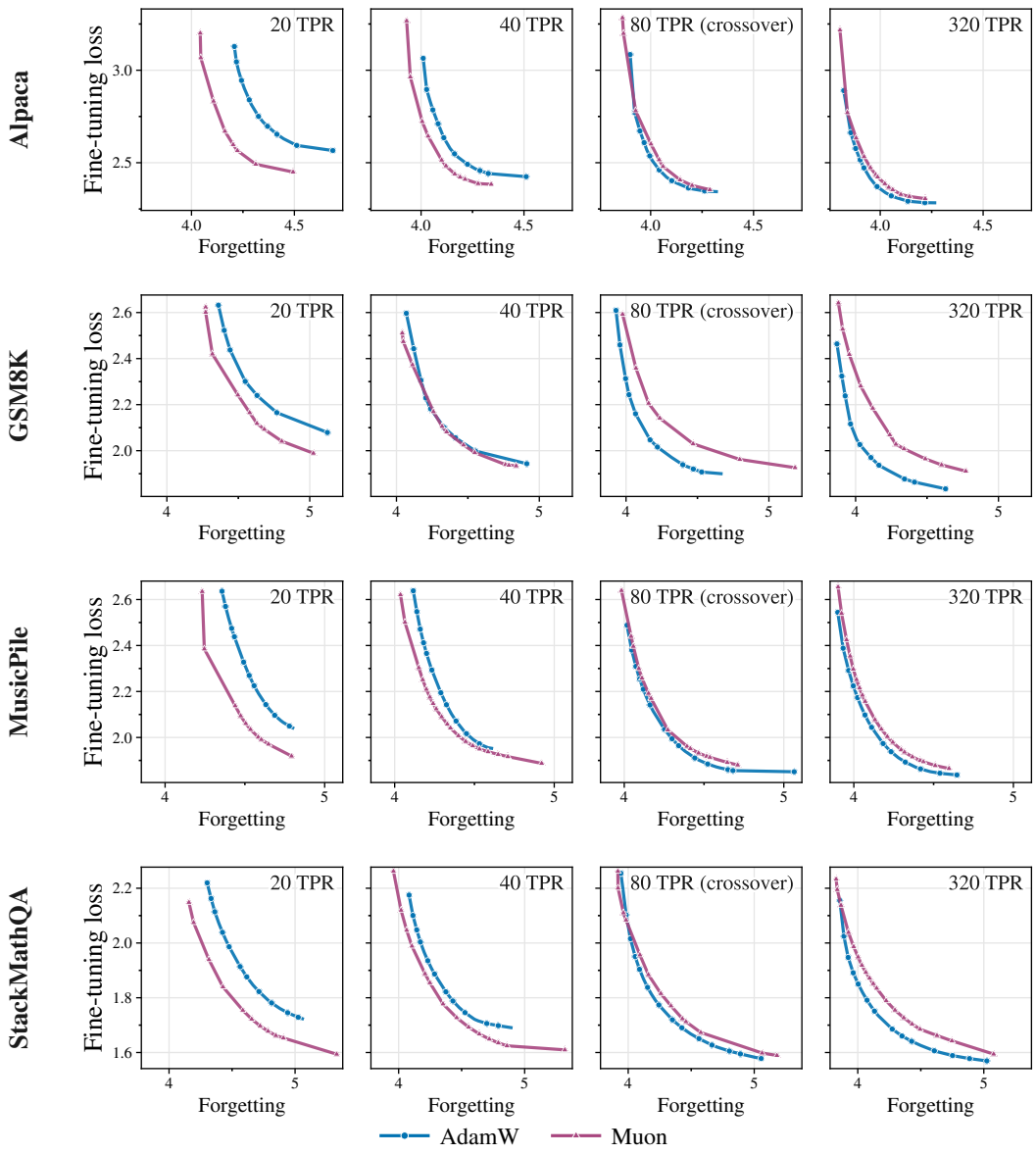

Figure 11: Learning–forgetting frontiers on every fine-tuning dataset, loss-optimal base models (60M). Each row shows the budgets around that dataset's crossover TPR.

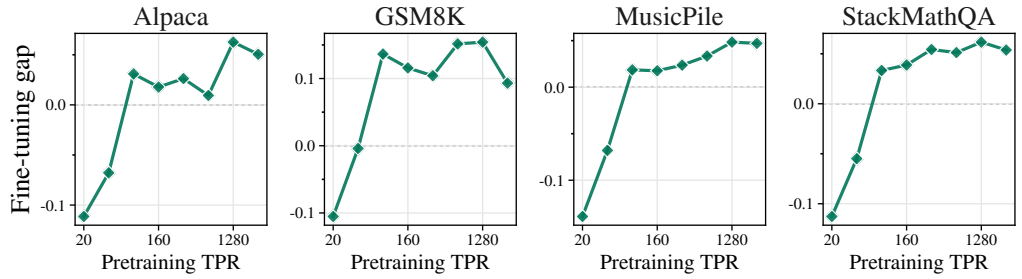

Figure 12: Fine-tuning gap on every fine-tuning dataset, loss-optimal base models (60M). Negative values mean the Muon frontier dominates.

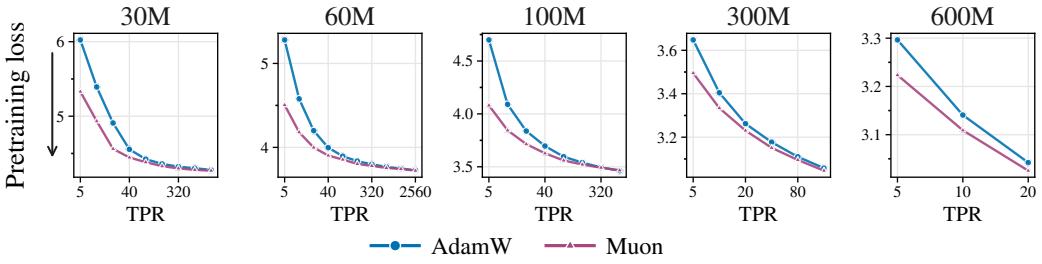

Figure 13: Muon reaches lower pretraining loss than AdamW. DCLM held-out loss of the loss-optimal base models across pretraining data budgets and model sizes.

C.3 Choice of fine-tuning optimizer

Our frontiers fine-tune AdamW-pretrained models with AdamW and Muon-pretrained models with both AdamW and Muon. Figure 14 motivates this choice, and shows that Muon's greater forgetting at large budgets is a property of the pretrained model rather than of the fine-tuning recipe. For each pretraining optimizer and budget, we take the best 60M base model among those fine-tuned with both optimizers, and plot its AdamW- and Muon-fine-tuned frontiers separately. We use GSM8K, the dataset on which AdamW-pretrained models were fine-tuned with both optimizers across the full range of budgets.

Two observations follow. First, the fine-tuning optimizer matters for AdamW-pretrained models but not for Muon-pretrained ones: on AdamW-pretrained models, AdamW fine-tuning dominates Muon fine-tuning at every budget, whereas the two frontiers of a Muon-pretrained model lie close together with no consistent winner. Second, the reversal does not depend on how the Muon-pretrained model is fine-tuned: at 1.2B tokens both of its frontiers lie below those of the AdamW-pretrained model, and from 19.2B tokens onward the AdamW-pretrained, AdamW-fine-tuned frontier dominates both.

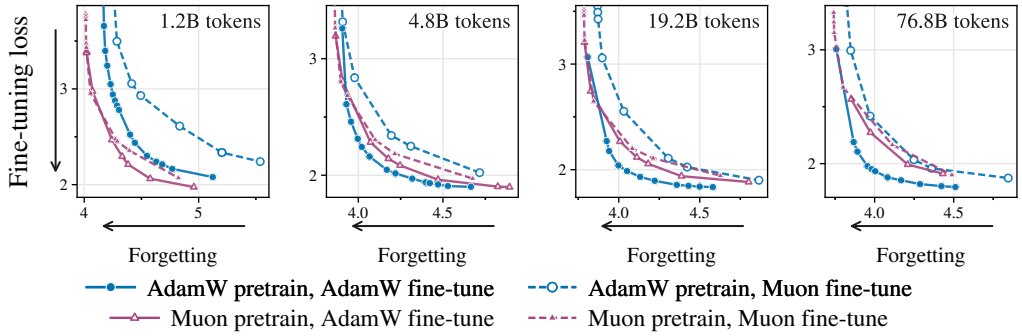

Figure 14: Muon's greater forgetting comes from the pretrained model, not the fine-tuning optimizer. Learning–forgetting frontiers of the best 60M base models on GSM8K, split by fine-tuning optimizer (solid: AdamW; dashed: Muon). AdamW fine-tuning is always better for AdamW-pretrained models, while the two fine-tuning optimizers give nearly the same frontier for Muon-pretrained models. At large budgets, the AdamW-pretrained model dominates under either choice.

C.4 Fine-tuning regularization

We repeat the comparison when fine-tuning is regularized with replay or EWC (setup in Section B.3), using the loss-optimal 60M base models. Figures 15 and 16 show the frontiers on all four datasets; each row shows the budgets around that dataset's crossover under the regularizer, a quarter and half of the crossover TPR, the crossover itself, and four times it. The crossover is the first budget at which the fine-tuning gap turns positive (Section B.5). Both regularizers reduce forgetting and can move the crossover to larger budgets (from 80 to 320 TPR on Alpaca and StackMathQA), but AdamW's frontier is better at large budgets on every dataset under both (Figure 17), with one exception: under EWC on StackMathQA the gap turns negative again at the largest budget, 2,560 TPR. With replay, Muon's frontier is better than or level with AdamW's below the crossover on every dataset. With

EWC the picture below the crossover is mixed: Muon is clearly better at small budgets on GSM8K, while on the other datasets the two frontiers are close.

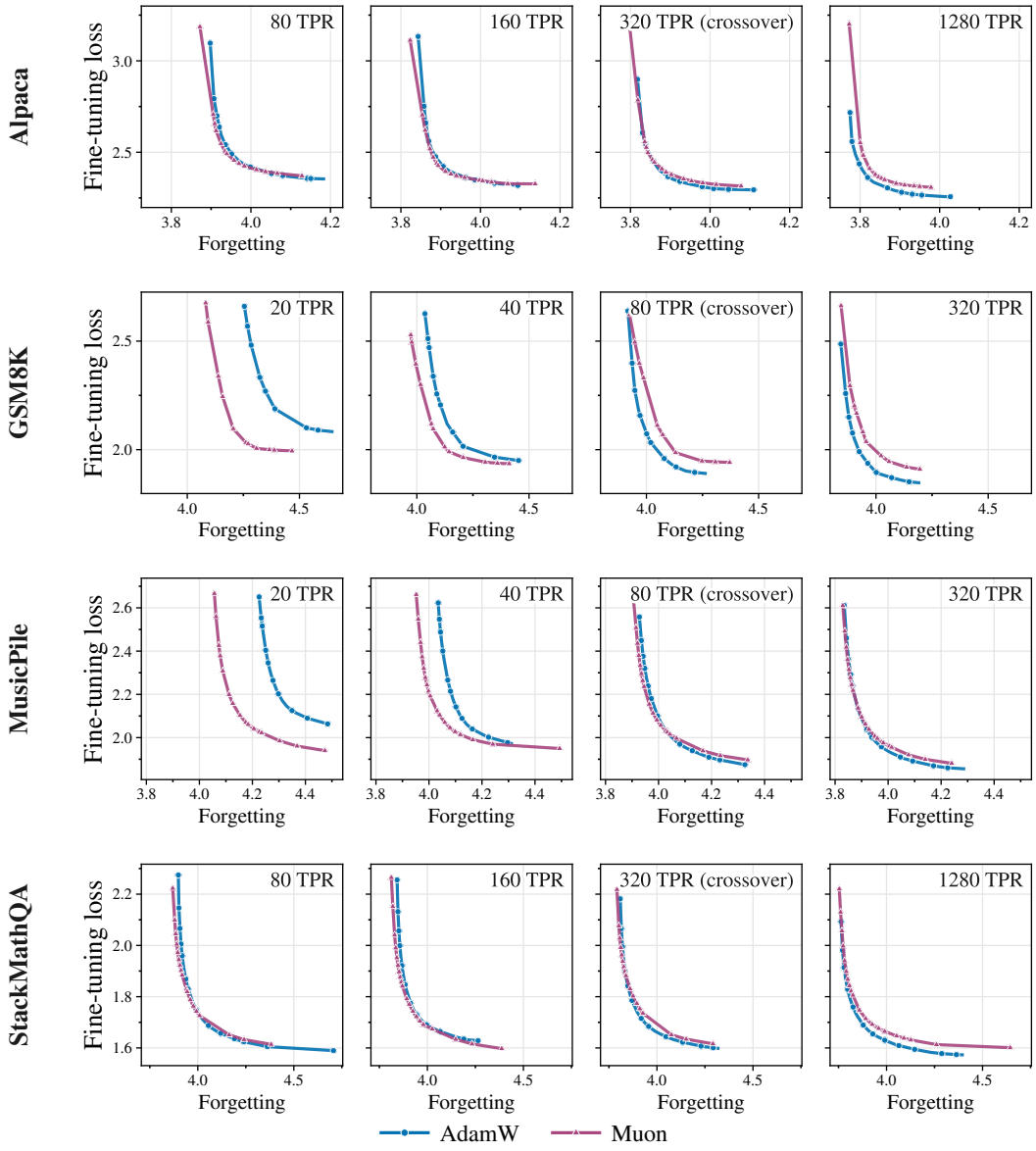

Figure 15: Learning–forgetting frontiers with replay of 10–30% DCLM data, loss-optimal 60M base models. Each row shows the budgets around that dataset's crossover under replay.

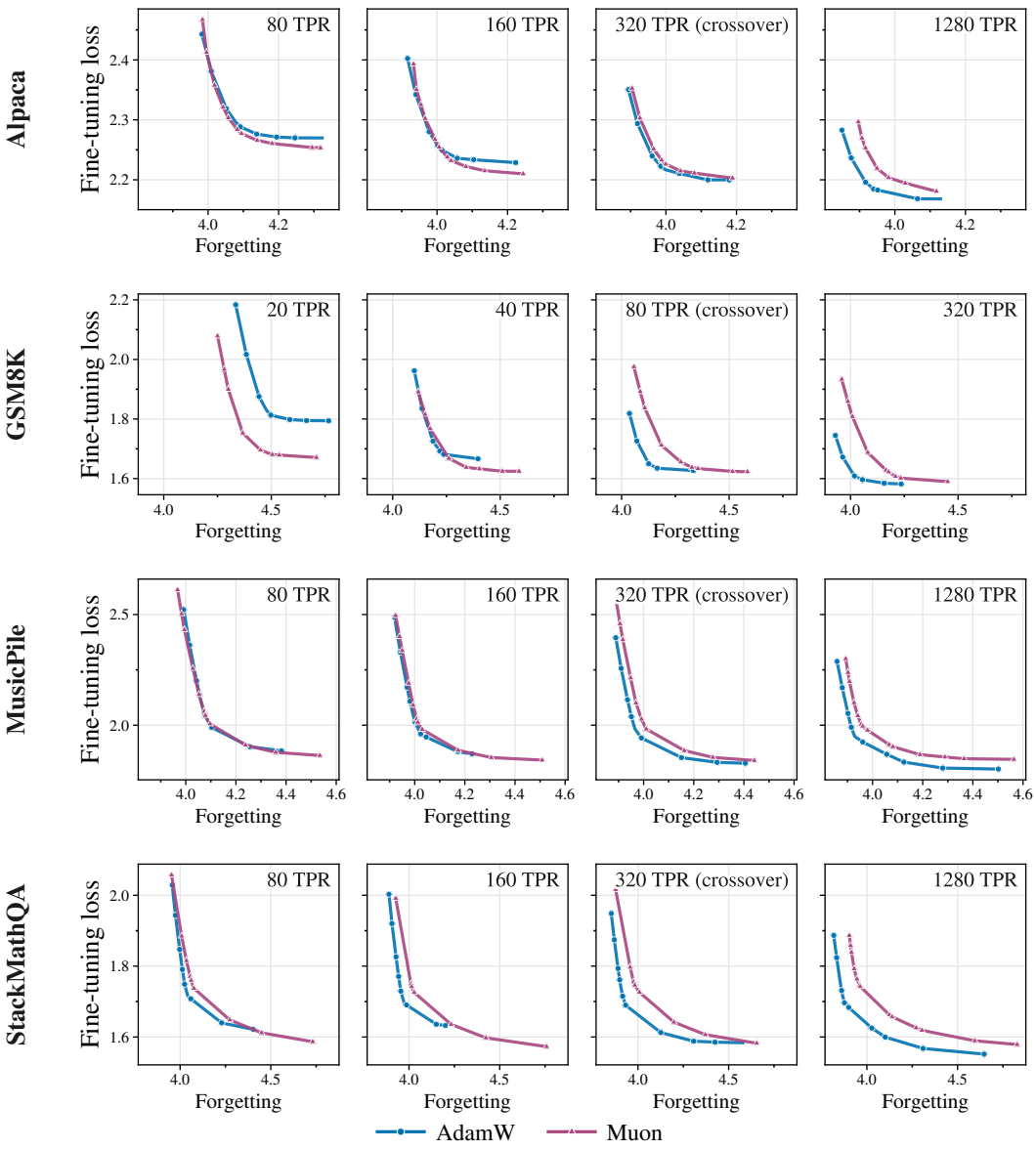

Figure 16: Learning–forgetting frontiers with EWC, $\lambda \in \{10^3, 4 \cdot 10^3, 10^4, 4 \cdot 10^4\}$, loss-optimal 60M base models. Each row shows the budgets around that dataset's crossover under EWC.

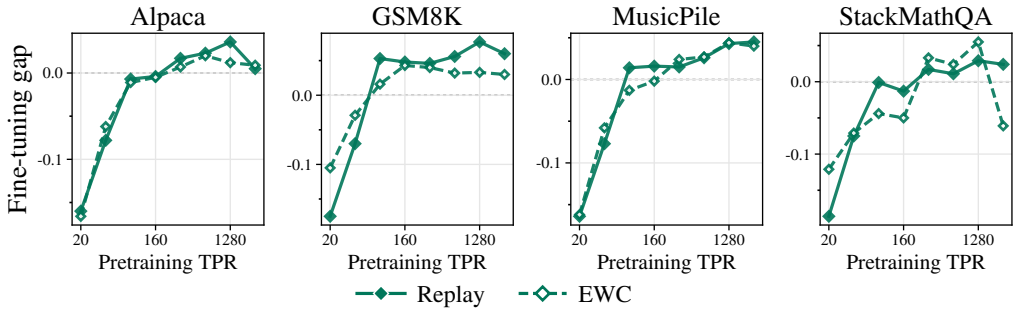

Figure 17: Fine-tuning gap under fine-tuning regularization, with replay (solid) and EWC (dashed), loss-optimal 60M base models. Negative values mean the Muon frontier dominates.

D Additional Results for Section 4

This section extends the analysis of Section 4: perturbation robustness at every magnitude (Section D.1), the loss under a perturbation and its connection to the margin-to-noise ratio (Section D.2), and the data-budget experiments in the linear memory (Section D.3).

D.1 Perturbation robustness

D.1.1 Setup

For every (model size, data budget) cell and each optimizer we take the checkpoint attaining the lowest held-out DCLM loss over the full pretraining learning-rate grid. A Gaussian perturbation is added independently to each parameter tensor of that checkpoint, scaled so that $\|\Delta W\|_F \approx \delta \|W\|_F$, and we report the increase in held-out DCLM loss over the checkpoint without perturbation, averaged over three perturbation draws. The grid spans three model sizes, 30M to 100M, token–parameter ratios from 5 to 2,560 TPR, and perturbation magnitudes δ from 0.005 to 0.28. Magnitudes below $\delta \approx 0.02$ are omitted from the figures because the increase in loss is then at the level of the seed-to-seed variation for both optimizers.

Table 2 and the robustness counts below use the same protocol on the loss-optimal checkpoints of all five model sizes, 30M to 600M: 36 (model size, data budget) cells and perturbation magnitudes δ from 0.002 to 0.56.

D.1.2 Result

Figure 4 fixes the perturbation magnitude; Figure 18 reports the same difference of degradations at every magnitude we sweep. The left three panels fix δ and vary the model size, the right panel fixes the model size at 100M and varies δ. The difference is positive at every budget, model size and magnitude we tested, and grows with the budget at every δ. Magnitudes are not comparable across panels, since the degradation of both optimizers grows steeply with δ; what the panels share is the shape of the curve.

Over the 36 cells, the perturbation magnitude δ^* at which the loss has risen by 0.05, 0.1 or 0.2 nats is larger for AdamW in every cell and at every threshold, by a median factor of 1.4, and Muon's increase in loss exceeds AdamW's in every cell at every δ from 0.005 to 0.2. The ordering does not depend on which checkpoint of a cell is used: giving each optimizer its most perturbation-robust checkpoint within 0.02 nats of the cell's best base loss still leaves Muon degrading 20–400% more.

Loss under perturbation. Table 2 compares the loss under perturbation itself, rather than its increase, and places the result next to the fine-tuning crossover of Figure 9. At every model size, Muon's loss under perturbation is the lower one at small budgets and AdamW's at large ones, apart from the three marked entries. The crossover moves to smaller budgets as δ grows, so its value is not comparable to the fine-tuning crossover; its trend with model size is. From $\delta = 0.1$ on, the crossover budget does not increase with model size, as for every fine-tuning dataset; at $\delta = 0.05$ it is not monotone.

Table 2: Crossover budgets, in TPR (tokens per parameter). Left: the smallest budget at which AdamW's loss under a perturbation of magnitude δ is below Muon's (loss-optimal checkpoints, mean over three perturbation draws); > 20: Muon's is lower up to the largest 600M budget, 20 TPR; ≤ 5: AdamW's is lower already at the smallest budget; †: Muon's is lower again at 160 TPR, by 0.003 nats. Right: the fine-tuning crossover of Figure 9.

	Perturbation, δ				**Fine-tuning**			
Size	**0.05**	**0.1**	**0.13**	**0.2**	**Alpaca**	**GSM8K**	**MusicPile**	**StackMathQA**
30M	160	40	40	20	160	80	640	320
60M	80	40	40	20	80	80	80	80
100M	160	40	20	10	80	40	80	40
300M	$80^\dagger$	40	20	10	40	10	40	20
600M	> 20	20	10	≤ 5	20	5	10	10

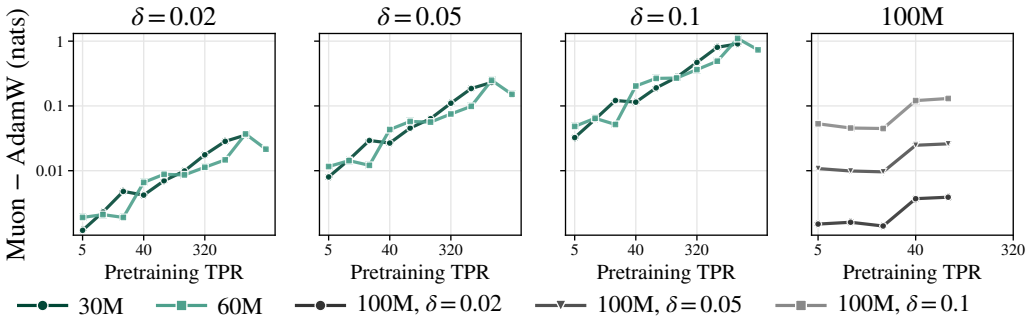

Figure 18: The perturbation degradation gap is positive at every perturbation magnitude. Muon minus AdamW increase in held-out DCLM loss. Left three panels: one δ each; right: 100M across δ.

D.2 Loss under a perturbation: derivations

This section gives the computation behind Section 4.1.2 one step at a time. We fix one fact with target output y (the fact i of Section 4.1.2, so $y = i$), its learned weights W, and a direction D; only the scalar step ε changes. The direction can be a fine-tuning update, a chosen matrix, or a fixed draw of Gaussian noise; an average over noise draws is a different operation (Section D.2.5).

Table 3: Notation for the loss under a perturbation.

Symbol	Description
W, D	Learned weight matrix and direction matrix, of the same size.
ε, W_ε	Scalar step size and perturbed weights $W + \varepsilon D$.
N, j, y	Number of outputs, an output index, and the target index.
$s_j(W)$	Score (logit) of output j at weights W.
m_j, d_j	Current margin $s_y(W) - s_j(W)$ and margin change per unit step $s_y(D) - s_j(D)$.
p_j, $q_j(\varepsilon)$	Probability of output j before and after the perturbation; $q_j(0) = p_j$.
f, ΔL	Loss along the direction and its increase.
h, κ^2	Mean and variance of the current margins over outputs other than y.
h_D, κ_D^2	Mean and variance of the changes d_j, weighted by p_j over all outputs.
ρ_D, x	Directional ratio h_D/κ_D and scaled step $\varepsilon\kappa_D$.

The subscript D denotes a direction, not a derivative. A prime denotes a derivative with respect to ε. A sum $\sum_{j=1}^{N}$ includes the target, while $\sum_{j\neq y}$ does not.

D.2.1 From weights and scores to the loss

Linear scores. In the memory of Section 4.1.1, $s_j(W) = v_j^\top W u$ for a fixed key u and values v_j, so every score is linear in W and $s_j(W + \varepsilon D) = s_j(W) + \varepsilon s_j(D)$. This is the only property of the model used below; D need not be a multiple of W.

Margins and margin changes. Subtracting the scores after the perturbation gives $s_y(W_\varepsilon) - s_j(W_\varepsilon) = m_j + \varepsilon d_j$. A change $d_j > 0$ increases the lead of the target over output j, and $d_j < 0$ reduces it; $m_y = d_y = 0$. The margins m_j and their rates of change d_j are different quantities.

Mean and variance of the margins. Over the $N - 1$ values m_j with $j \neq y$, h and κ^2 are the mean and variance of Definition 4.1; the denominator $N - 1$ is the number of values, not a sample correction. With $\bar{s}_{-y} := (N-1)^{-1} \sum_{j\neq y} s_j(W)$ the mean incorrect score,

$$h = s_y(W) - \bar{s}_{-y}, \qquad \kappa^2 = \frac{1}{N-1} \sum_{j\neq y} \big(s_j(W) - \bar{s}_{-y}\big)^2, \tag{12}$$

so the variance of the margins is also the variance of the incorrect scores.

Probabilities and loss. Softmax maps the perturbed scores to probabilities $q_j(\varepsilon) = e^{s_j(W_\varepsilon)} / \sum_{k=1}^{N} e^{s_k(W_\varepsilon)}$. Dividing the numerator and denominator of the target probability by $e^{s_y(W_\varepsilon)}$

gives

$$f(\varepsilon) = -\log q_y(\varepsilon) = \log\Big(1 + \sum_{j \neq y} e^{-m_j - \varepsilon d_j}\Big), \tag{13}$$

where the 1 is the target term e^0. The increase $\Delta L(\varepsilon) = f(\varepsilon) - f(0)$ is degradation when positive and improvement when negative.

D.2.2 THE FIRST AND SECOND DERIVATIVES

First derivative. Let $Z(\varepsilon) := \sum_{j=1}^{N} e^{-m_j - \varepsilon d_j}$, so $f = \log Z$ and $q_j(\varepsilon) = e^{-m_j - \varepsilon d_j}/Z(\varepsilon)$. Since m_j, d_j are fixed,

$$f'(\varepsilon) = \frac{Z'(\varepsilon)}{Z(\varepsilon)} = -\sum_{j=1}^{N} q_j(\varepsilon) d_j =: -\mu(\varepsilon), \qquad h_D := \mu(0) = \sum_j p_j d_j, \qquad f'(0) = -h_D. \tag{14}$$

Outputs with larger p_j have more influence on this term, so a uniform mean over incorrect outputs would give a different value.

Second derivative. The probabilities also change with ε: the quotient rule gives $q_j'(\varepsilon) = q_j(\varepsilon)\big(\mu(\varepsilon) - d_j\big)$. Differentiating f' and using $\sum_j q_j = 1$ and $\sum_j q_j d_j = \mu$,

$$f''(\varepsilon) = -\sum_j d_j q_j'(\varepsilon) = \sum_j q_j(\varepsilon) d_j^2 - \mu(\varepsilon)^2 = \sum_j q_j(\varepsilon)\big(d_j - \mu(\varepsilon)\big)^2. \tag{15}$$

At zero perturbation this is $f''(0) = \kappa_D^2 := \sum_j p_j (d_j - h_D)^2$, the variance of the changes under the probability weights; the curvature is therefore nonnegative. These identities are exact for the linear model; the approximation enters only when two derivatives are used to predict a finite loss change.

D.2.3 THE QUADRATIC, ITS RATIO, AND THE SIZE OF A STEP

The quadratic. Matching the slope $-h_D$ and curvature κ_D^2 at zero gives the second-order Taylor approximation $\Delta L(\varepsilon) \approx Q(\varepsilon) := -h_D \varepsilon + \frac{\kappa_D^2}{2}\varepsilon^2$. It is local: its accuracy must be checked at the step sizes used, and it is not an upper bound on the loss. If $h_D < 0$, both terms increase the loss; if $h_D > 0$, the linear term improves the loss and the quadratic term reduces that improvement; if $h_D = 0$, the loss increase is $\kappa_D^2 \varepsilon^2/2$, so a zero first derivative does not mean that the perturbation has no effect.

A common scale. For $\kappa_D > 0$, substituting $x = \varepsilon \kappa_D$ and $\rho_D = h_D/\kappa_D$ gives $Q = -\rho_D x + x^2/2$ (Equation (5)). At the same scaled step x, the ratio sets the linear term; at the same original step ε, the value of κ_D also matters, so the ratio alone does not determine that comparison.

The size and sign of D. Replacing D by cD for $c > 0$ gives $h_{cD} = c h_D$, $\kappa_{cD} = c\kappa_D$ and $\rho_{cD} = \rho_D$: the ratio is unchanged, but the same ε is a larger weight change. Replacing D by $-D$ reverses the sign of h_D and ρ_D and preserves κ_D^2, so the same checkpoint can have opposite initial loss changes in two opposite directions.

The minimum of the quadratic. If $h_D > 0$ and $\kappa_D > 0$, setting $Q'(\varepsilon) = 0$ gives

$$\varepsilon_* = \frac{h_D}{\kappa_D^2}, \qquad Q(\varepsilon_*) = -\frac{h_D^2}{2\kappa_D^2} = -\frac{\rho_D^2}{2}, \tag{16}$$

the one-dimensional Newton step, whose predicted loss reduction is half the squared ratio. If $h_D \leq 0$, the minimum over nonnegative steps is at zero.

D.2.4 CONNECTION TO THE MARGIN-TO-NOISE RATIO

One step from zero weights. Set $W = 0$ and $D = A$ for a learned matrix A, as in the one-step analysis of Section 4.2.2. Then $m_j = 0$, every probability is $1/N$, and $d_j = s_y(A) - s_j(A)$ is the margin of A; write $h(A)$ and $\kappa(A)$ for their mean and standard deviation over the $N - 1$ incorrect outputs. Since $d_y = 0$,

$$h_D = \frac{1}{N}\sum_{j=1}^{N} d_j = \frac{N-1}{N} h(A). \tag{17}$$

The factor $(N-1)/N$ appears because the probability average includes the target. With $\alpha = (N-1)/N$, the second moment over all outputs is $\alpha[\kappa(A)^2 + h(A)^2]$ and the mean is $\alpha h(A)$, so

$$\kappa_D^2 = \alpha\kappa(A)^2 + \alpha(1-\alpha)h(A)^2 = \frac{N-1}{N}\kappa(A)^2 + \frac{N-1}{N^2}h(A)^2. \tag{18}$$

For $\kappa(A) > 0$ and $\rho(A) = h(A)/\kappa(A)$, the exact directional ratio is

$$\rho_D = \sqrt{\frac{N-1}{N}}\,\frac{\rho(A)}{\sqrt{1+\rho(A)^2/N}}. \tag{19}$$

Thus $\rho_D \approx \rho(A)$ requires both large N and a small relative term $\rho(A)^2/N$; a large number of outputs alone does not make that term negligible. At a general learned starting point, the probabilities are no longer uniform and the changes depend on D.

Averaging over facts. For K facts indexed by i, Definition 4.1 uses $\bar{h} = K^{-1}\sum_i h_i$ and $\bar{\kappa}^2 = K^{-1}\sum_i \kappa_i^2$. For the mean loss $\bar{f}(\varepsilon) = K^{-1}\sum_i f_i(\varepsilon)$, the derivatives also average, $\bar{f}'(0) = -K^{-1}\sum_i h_{i,D}$ and $\bar{f}''(0) = K^{-1}\sum_i \kappa_{i,D}^2$. The model-level ratio therefore uses the mean first-order term divided by the square root of the mean curvature, which is generally different from the mean of the per-fact ratios.

D.2.5 Conditions and special cases

Exact identities and local approximation. The score changes, margin changes, loss formula and derivatives are exact for the linear model with finite scores. The quadratic predicts a small finite change; its curvature is measured at zero perturbation, and the true curvature changes as the probabilities change. The loss can also change while the highest-scoring output stays the same, so loss degradation and a change in prediction are different measures.

Zero curvature or zero first-order term. If $\kappa_D = 0$, every d_j equals h_D, and $d_y = 0$ forces $d_j = 0$ for all j: the direction leaves every margin fixed and the loss is constant, so the ratio is undefined. If only $h_D = 0$ while $\kappa_D > 0$, the loss increase is quadratic; at a stationary point this holds in every direction, even though the curvatures and loss increases differ, which is a further reason to keep the curvature scale when comparing fixed steps.

Random directions. If D is Gaussian noise with mean zero, independent of W, the changes d_j are linear in D and average to zero, while the p_j do not depend on the draw. Thus

$$\mathbb{E}_D\, h_D = \sum_j p_j\, \mathbb{E}_D\, d_j = 0, \qquad \mathbb{E}_D\, \Delta L(\varepsilon) \approx \frac{\varepsilon^2}{2}\,\mathbb{E}_D\, \kappa_D^2. \tag{20}$$

One draw can have a nonzero first-order term, but the average initial slope is zero. This average does not identify the current margin h with a noise-degradation slope.

Nonlinear models. In a neural network the scores are nonlinear in the weights. Writing the margin along the direction as $m_j(\varepsilon)$, with $d_j = m_j'(0)$ and $a_j = m_j''(0)$, the first derivative is unchanged, $f'(0) = -\sum_j p_j d_j$, but

$$f''(0) = \sum_j p_j (d_j - h_D)^2 - \sum_j p_j a_j. \tag{21}$$

The extra term vanishes in the linear model; otherwise the variance need not equal the full loss curvature, which can be negative.

D.2.6 Gradient and Hessian notation

Writing the entries of W and D as vectors θ and v in the same order, with gradient $g = \nabla_\theta \mathcal{L}$ and Hessian $H = \nabla_\theta^2 \mathcal{L}$ evaluated at θ, the chain rule gives

$$f'(0) = g^\top v = -h_D, \qquad f''(0) = v^\top H v = \kappa_D^2, \qquad \mathcal{L}(\theta + \varepsilon v) - \mathcal{L}(\theta) \approx \varepsilon\, g^\top v + \frac{\varepsilon^2}{2}\, v^\top H v. \tag{22}$$

The first term measures the alignment between the direction and the loss gradient, and the second is the curvature along the direction. The identity $f''(0) = v^\top H v$ holds for any scores; equality with κ_D^2 requires the extra term of Section D.2.5 to be zero, as it is for linear scores.

D.3 MARGIN-TO-NOISE RATIO ACROSS DATA BUDGETS

D.3.1 ONE STEP

This is the regime Theorem 4.2 is about: the two solutions are the exact first step from $W_0 = 0$, so every quantity below can be compared against a closed form.

Setup **Data.** Keys and values are drawn as in Section 4.1.1 at $d = 512$, with $N = \gamma d$ and $\gamma \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \frac{5}{4}, \frac{3}{2}, 2, \frac{5}{2}, 3, 4, 5, 6, 8, 10, 12, 16, 20, 24, 32, 48, 64\}$, five random seeds per data budget. Section 4.1.1 requires $N - 1 \geq d$, so the points with $\gamma \leq 1$ lie outside the range in which Theorem 4.2 is stated and are drawn only for context.

One step. The gradient of the loss of Section 4.1.1 at $W_0 = 0$ is computed exactly, with equal item weights. GD takes it as W^{GD}; Muon takes its polar factor $W^{\mathrm{Muon}} = \mathrm{Orth}(W^{\mathrm{GD}})$, computed exactly by singular value decomposition rather than by the Newton–Schulz iteration of Section A.1.2. As a baseline we also record the first AdamW step, which with zero initialization and the usual ϵ is the entrywise $\mathrm{sign}(W^{\mathrm{GD}})$; Theorem 4.2 says nothing about it.

Observables. Every h_i and κ_i is evaluated in closed form for all i as in Definition 4.1, giving $\bar{h}$, $\bar{\kappa}$, ρ and $\mathcal{R}$.

Plotting. Solid lines are means over the five seeds and bands are $\pm$ one standard deviation; dashed lines are the closed forms behind Theorem 4.2, with $\Phi(\gamma)$ replaced by its limit $\frac{8}{3\pi}$. In the $\bar{h}$ and $\bar{\kappa}$ panels each solution is normalized to $\|W\|_F = 1$, which is needed because the two steps have different scales while ρ does not see them; the $\bar{h}$ and ρ panels are divided by $\sqrt{d/\gamma}$, the unit in which Theorem 4.2 is stated.

Result Figure 19 (top) shows $\mathcal{R}$ rising monotonically from 0.53 at $\gamma = \frac{1}{4}$ to 1.16 at $\gamma = 64$ and crossing 1 near $\gamma = 5$, with the closed form tracking it once $\gamma \gtrsim 2$. The remaining panels separate the two factors. Normalized, GD's margin is constant in γ, which is the identity $\bar{h}^{\mathrm{GD}} = \|W^{\mathrm{GD}}\|_F^2/(N-1)$ read at $\|W\|_F = 1$, while Muon's climbs to about $\frac{8}{3\pi}$; both variances approach the same limit, GD's from above and Muon's from below. The limiting ratio $\frac{3\pi}{8}$ is therefore carried by the margin. Its approach to that limit comes from both factors: Muon's margin contributes the $+\frac{1}{10\gamma}$ correction, while the $-\frac{1}{\gamma}$ part of the $\frac{9}{10\gamma}$ term comes from GD's variance, the factor $\gamma + 2$ in $(\bar{\kappa}^{\mathrm{GD}})^2$ (Section 4.2.2). The bottom row runs the same comparison against the AdamW surrogate: its $\mathcal{R}$ has the same shape but reaches only 0.93 at $\gamma = 64$, so over the data budgets we ran the one-step reversal is a statement about GD, not about AdamW.

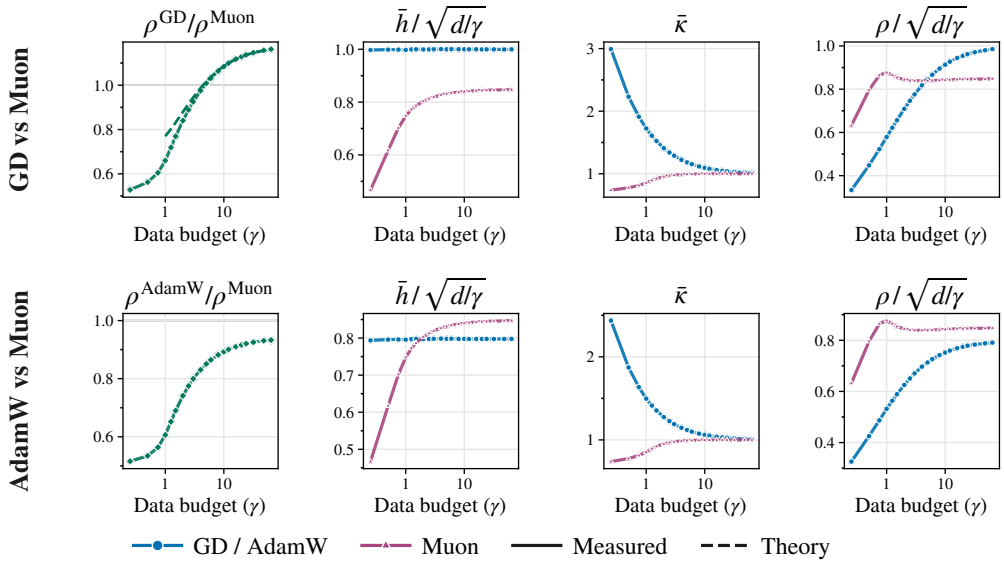

Figure 19: One step at $d = 512$, GD against Muon (top) and the AdamW surrogate against Muon (bottom); solid is experiment, dashed the closed form of Theorem 4.2.

D.3.2 MANY STEPS

The theory covers one step. This experiment runs the same memory to convergence and asks whether the ordering it predicts survives training.

Setup **Data.** Keys and values are drawn as in Section 4.1.1, with item i queried with probability $q_i \propto 1/i$ (Zipf). We use $d \in \{64, 128, 256, 512\}$, γ log-spaced in $[\frac{1}{4}, 128]$, and three seeds per (d, γ) shared by the two optimizers.

Training. From $W_0 = 0$ we take $T = 10^4$ full-batch steps of $W_{t+1} = W_t + 0.02\, G_t$ for GD and $W_{t+1} = W_t + 0.005\, \mathrm{Orth}(G_t)$ for Muon, where $G_t = \sum_i q_i \big(1 - \hat{p}_{W_t}(i \mid i)\big) u_i v_i^\top$ is the Hebbian part of the gradient of the loss of Section 4.1.1, at constant learning rate, without momentum or weight decay, and with the polar factor again taken by singular value decomposition.

Observables. h_i and κ_i are as in Definition 4.1, with equal item weights rather than the q_i used in training. The saturated value of an observable is the per-seed median over $t \geq 10^3$; $\mathcal{R}$ is formed per seed and only then averaged. Curves are means over seeds and bands are $\pm$ one standard deviation.

Perturbed robustness. We perturb the trained weights, $W' = W_T + \sigma \|W_T\|_F \xi$ with $\|\xi\|_F = 1$, over 16 perturbation directions shared by the two optimizers, and read off σ_{50}, the σ at which half of the items correct at $\sigma = 0$ are lost. For $N < d$ the Muon solution is rescaled to its component in the span of the keys, since outside that span it is an arbitrary orthogonal completion that carries norm but no information. Data budgets at which either optimizer is already past the capacity of the memory, meaning accuracy below 0.99 at $\sigma = 0$, are dropped from the σ_{50} panels: there the two optimizers recover different sets of items and the two σ_{50} are not comparable.

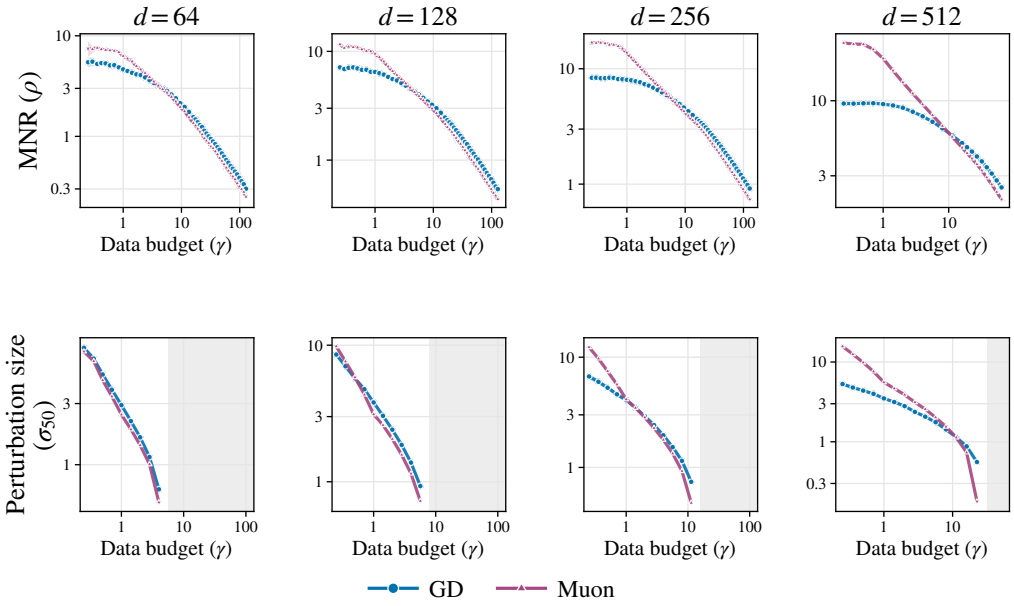

Figure 20: Top: saturated ρ against the data budget, the two curves behind Figure 21. **Bottom: perturbed robustness σ_{50} against the data budget**, below capacity only. One column per d.

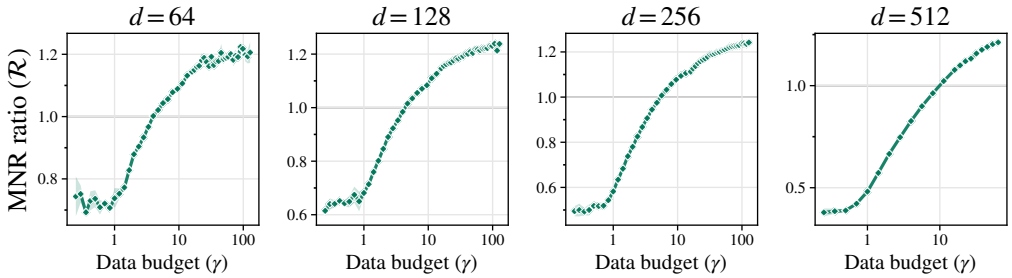

Figure 21: Saturated $\mathcal{R}$ against the data budget, one panel per d.

Result Training does not undo the one-step picture. Figure 21 shows $\mathcal{R}$ increasing with γ at every d, over a range from about 0.4 to about 1.2, and the crossing $\mathcal{R} = 1$ moving to larger data budgets as d grows: $\gamma^* \approx 4.0$, 4.4, 5.4 and 9.9 at $d = 64$, 128, 256 and 512. Figure 20 shows the two ρ separately: both fall with the data budget, Muon's from roughly twice GD's at $\gamma = \frac{1}{4}$ to below it past γ^*. The bottom row shows that the robustness ordering turns over as well, GD's σ_{50} gaining on Muon's at every d. The two turnovers need not coincide, and they do not: the σ_{50} ordering flips at $\gamma \approx 0.5$, 1.3 and 11 at $d = 128$, 256 and 512, and at $d = 64$ GD is already the more robust at the smallest data budget we ran.

E Proof of Theorem 4.2

This section proves Theorem 4.2. We first compute the one-step GD and Muon iterates and collect concentration estimates for the Gaussian data. We then evaluate the mean margin $\bar{h}$ and the margin spread $\bar{\kappa}$ of Definition 4.1 separately for the two iterates, and finally take the ratio of the two margin-to-noise ratios. The GD estimates hold for every $\gamma > 1$; the Muon estimates, and hence the theorem, require $\gamma \geq 16$.

Setup. We work in the setting of Section 4.1.1 with *uniform query weights*, $q_i = 1/N$ for every fact i; this is the setting of Theorem 4.2 and is assumed throughout. The keys and values $u_1, \dots, u_N, v_1, \dots, v_N$ are independent $\mathcal{N}(0, I_d)$ vectors, $N = \gamma d$ with the data budget $\gamma > 1$ held fixed, $N - 1 \geq d$, and $d \to \infty$; O_p and o_p are as in Section 4.1.1.

Notation. We collect the data in $U = [u_1 \cdots u_N] \in \mathbb{R}^{d\times N}$ and $V = [v_1 \cdots v_N] \in \mathbb{R}^{d\times N}$, and write $\mathbf{1} = (1, \dots, 1)^\top$, e_i for the i-th standard basis vector and $\langle X, Y\rangle = \mathrm{tr}(X^\top Y)$. The operator, Frobenius and nuclear norms $\|\cdot\|_{\mathrm{op}}$, $\|\cdot\|_F$ and $\|\cdot\|_*$ satisfy

$$\begin{aligned} &\|X\|_{\mathrm{op}} \leq \|X\|_F \leq \|X\|_* \leq \mathrm{rank}(X)\|X\|_{\mathrm{op}}, \qquad \|X\|_F \leq \sqrt{\mathrm{rank}(X)}\,\|X\|_{\mathrm{op}}, \\ &|\langle X, Y\rangle| \leq \|X\|_*\|Y\|_{\mathrm{op}}. \end{aligned} \tag{23}$$

For $X \in \mathbb{R}^{d\times d}$ with compact SVD $X = P\Sigma Q^\top$ we write $\mathrm{Orth}(X) = PQ^\top$ for its polar factor, so that $\mathrm{Orth}(cX) = \mathrm{Orth}(X)$ for $c > 0$ and $\|\mathrm{Orth}(X)\|_{\mathrm{op}} \leq 1$.

Scores and observables. By Definition 4.1, for any W,

$$\begin{aligned} \bar{h} &= \operatorname*{mean}_i u_i^\top W v_i - \operatorname*{mean}_i \operatorname*{mean}_{j\neq i} u_j^\top W v_i, \\ \bar{\kappa}^2 &= \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j\neq i} \left(u_j^\top W v_i\right)^2 - \operatorname*{mean}_i \Big(\operatorname*{mean}_{j\neq i} u_j^\top W v_i \Big)^2, \end{aligned} \tag{24}$$

where the scores are written as $u_j^\top W v_i$ rather than $v_j^\top W u_i$ as in Section 4.1.1. The two conventions are exchanged by swapping the roles of U and V: reading Section 4.1.1 with (U, V) replaced by (V, U) turns every score $v_j^\top W u_i$ into $u_j^\top W v_i$ and its one-step iterate into Equation (25), leaving the loss and every h_i and κ_i unchanged. Since U and V are i.i.d., the law of ρ is the same under either convention.

E.1 The two one-step iterates

Lemma E.1 (One step). *Up to a positive multiple, one step of GD and one step of Muon from $W_0 = 0$ give*

$$W^{\mathrm{GD}} = UV^\top - \frac{1}{N}(U\mathbf{1})(V\mathbf{1})^\top = U\Big(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N}\Big)V^\top, \qquad W^{\mathrm{Muon}} = \mathrm{Orth}\left(W^{\mathrm{GD}}\right). \tag{25}$$

The positive multiple is immaterial for ρ.

Proof. Gradient. With $\partial_W(u_j^\top W v_i) = u_j v_i^\top$, the gradient of the loss of Section 4.1.1 with uniform weights $q_i = 1/N$ is

$$-\nabla_W \mathcal{L}(W) = \frac{1}{N}\sum_{i=1}^N \Big(u_i - \sum_{j=1}^N \hat{p}_W(j \mid i) u_j\Big) v_i^\top. \tag{26}$$

At $W = 0$ all scores vanish, so $\hat{p}_0(j \mid i) = 1/N$; moreover $\sum_j u_j = U\mathbf{1}$, $\sum_i u_i v_i^\top = UV^\top$ and $\sum_i v_i^\top = (V\mathbf{1})^\top$, whence

$$-\nabla_W \mathcal{L}(0) = \frac{1}{N}\Big(UV^\top - \frac{1}{N}(U\mathbf{1})(V\mathbf{1})^\top\Big) = \frac{1}{N}U\Big(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N}\Big)V^\top. \tag{27}$$

One step. With step size $\eta > 0$, GD gives $-\eta\nabla_W\mathcal{L}(0) = \frac{\eta}{N}W^{\mathrm{GD}}$ and Muon gives $\eta\,\mathrm{Orth}(-\nabla_W\mathcal{L}(0)) = \eta\,\mathrm{Orth}(W^{\mathrm{GD}})$, using $\mathrm{Orth}(cX) = \mathrm{Orth}(X)$ for $c > 0$. *Normalization.*

Replacing W by cW with $c > 0$ multiplies every score $u_j^\top W v_i$ by c, hence $h_i \mapsto ch_i$ and $\kappa_i \mapsto c\kappa_i$, so $\bar{h} \mapsto c\bar{h}$, $\bar{\kappa} \mapsto c\bar{\kappa}$ and ρ is unchanged. □

Lemma E.2 (Full rank and orthogonality). *Almost surely* $\operatorname{rank} W^{\mathrm{GD}} = d$. *Consequently* W^{Muon} *is an orthogonal matrix:* $\|W^{\mathrm{Muon}}x\| = \|x\|$, $\|W^{\mathrm{Muon}}\|_{\mathrm{op}} = 1$ *and* $\|W^{\mathrm{Muon}}\|_F^2 = d$.

Proof. $\det W^{\mathrm{GD}}$ is a polynomial in the entries of (U, V). Taking $U = [\, I_d \;\; 0\,]$ and $V = [\, I_d \;\; -\sum_{k\le d} e_k \;\; 0\,]$, which is possible because $N \ge d+1$, gives $V\mathbf{1} = 0$ and $W^{\mathrm{GD}} = UV^\top = I_d$, so the polynomial is not identically zero; since (U, V) has a density, $\Pr[\det W^{\mathrm{GD}} = 0] = 0$. At full rank the factors P, Q of the compact SVD are $d \times d$ orthogonal matrices, so $W^{\mathrm{Muon}} = PQ^\top$ is orthogonal and $\|W^{\mathrm{Muon}}\|_F^2 = \operatorname{rank} W^{\mathrm{GD}} = d$. □

E.2 Concentration estimates for the data

Lemma E.3 (Concentration for U, V). *As $d \to \infty$ along the sequence above:*

(i) $\operatorname{mean}_k \|u_k\|^2 = d\big(1 + O_p((Nd)^{-1/2})\big)$, *and likewise for* V.

(ii) $\operatorname{mean}_k \big(\|u_k\|\|v_k\| - d\big)^2 = O_p(d)$.

(iii) $\|U\mathbf{1}\|^2/(Nd) \to 1$ *in probability; in particular* $\|U\mathbf{1}\|^2 = O_p(Nd)$. *Likewise for* V.

(iv) $\|U\|_{\mathrm{op}}/\sqrt{N} \to 1 + \gamma^{-1/2}$ *almost surely, hence* $\|U\|_{\mathrm{op}}^2 \le 4N(1 + o_p(1))$; *and for every* $m \ge 1$ *there is* C_m *with* $\mathbb{E}\|U\|_{\mathrm{op}}^{2m} \le C_m N^m$. *Likewise for* V.

(v) $\|UU^\top - NI_d\|_{\mathrm{op}} = O_p(\sqrt{Nd})$. *If* $\gamma \ge 16$, *then with probability tending to one both* $S = \frac{1}{N}VV^\top$ *and* $S = \frac{1}{N}(VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top)$ *satisfy* $\|S - I_d\|_{\mathrm{op}} \le 3\gamma^{-1/2}$ *and* $\lambda_{\min}(S) \ge \frac{1}{8}$.

(vi) $\|VV^\top\|_F^2 = O_p(N^2 d)$, *and* $\mathbb{E}\|VV^\top\|_F^{2m} \le C_{2m}(N^2 d)^m$. *Likewise for* $UU^\top$.

(vii) $\|UV^\top\|_F^2 = Nd^2(1 + O_p(d^{-1}))$ *and* $\|UV^\top\|_F^2 - d\sum_k \|v_k\|^2 = O_p(\sqrt{N}d^{3/2} + Nd)$.

(viii) $UV^\top \overset{d}{=} (UU^\top)^{1/2}\Xi$ *with* $\Xi \in \mathbb{R}^{d\times d}$ *having i.i.d.* $\mathcal{N}(0,1)$ *entries independent of* U; *with probability tending to one,* $\|UV^\top\|_{\mathrm{op}} \le 5\sqrt{Nd}$.

Proof. (i) $\sum_k \|u_k\|^2 \sim \chi^2_{Nd}$ has mean Nd and variance $2Nd$, so Chebyshev gives $\sum_k \|u_k\|^2 = Nd + O_p(\sqrt{Nd})$; divide by N.

(ii) For $x \ge 0$ one has $\sqrt{x} \ge 1 + \frac{x-1}{2} - \frac{(x-1)^2}{2}$: the difference g of the two sides satisfies $g(0) = g(1) = g'(1) = 0$ and $g''(x) = 1 - \frac{1}{4}x^{-3/2}$ changes sign once, at $x_0 = 4^{-2/3}$, so g is concave on $[0, x_0]$ and convex on $[x_0, \infty)$, and the two boundary values force $g \ge 0$. With $u \sim \mathcal{N}(0, I_d)$, $\mathbb{E}\|u\|^2/d = 1$ and $\operatorname{var}(\|u\|^2/d) = 2/d$, so $\mathbb{E}\|u\| \ge \sqrt{d}(1 - \frac{1}{d})$ and $(\mathbb{E}\|u\|)^2 \ge d - 2$. Hence for independent u, v,

$$\mathbb{E}\big(\|u\|\|v\| - d\big)^2 = \mathbb{E}\|u\|^2\,\mathbb{E}\|v\|^2 - 2d\,\mathbb{E}\|u\|\,\mathbb{E}\|v\| + d^2 \le 2d^2 - 2d(d-2) = 4d, \tag{28}$$

and Markov's inequality gives the claim.

(iii) $U\mathbf{1} = \sum_k u_k \sim \mathcal{N}(0, NI_d)$, so $\|U\mathbf{1}\|^2/N \sim \chi^2_d$ has mean d and variance $2d$; Chebyshev gives $\|U\mathbf{1}\|^2/(Nd) = 1 + O_p(d^{-1/2})$.

(iv) $\|U\|_{\mathrm{op}}^2 = \lambda_{\max}(UU^\top)$, and by the Bai–Yin law $\lambda_{\max}(UU^\top/N) \to (1 + \gamma^{-1/2})^2 \le 4$ almost surely, since $\gamma \ge 1$. For the moments, $U \mapsto \|U\|_{\mathrm{op}}$ is 1-Lipschitz for $\|\cdot\|_F$ and $\mathbb{E}\|U\|_{\mathrm{op}} \le \sqrt{N} + \sqrt{d} \le 2\sqrt{N}$ by Gordon's theorem, so Gaussian concentration gives $\Pr[\|U\|_{\mathrm{op}} \ge 2\sqrt{N} + t] \le e^{-t^2/2}$; integrating the tail bounds all moments.

(v) By Bai–Yin, $\lambda_{\max}(UU^\top/N) \to (1 + \gamma^{-1/2})^2$ and $\lambda_{\min}(UU^\top/N) \to (1 - \gamma^{-1/2})^2$ almost surely, so $\|UU^\top/N - I_d\|_{\mathrm{op}} \to \max\{(1+\gamma^{-1/2})^2 - 1, 1 - (1-\gamma^{-1/2})^2\} = 2\gamma^{-1/2} + \gamma^{-1}$, which is at most $3\gamma^{-1/2} < 4\gamma^{-1/2}$; multiplying by N gives the first claim. For $\gamma \ge 16$ we have $\gamma^{-1/2} \le \frac{1}{4}$,

so in probability $\|VV^\top/N - I_d\|_{\mathrm{op}} \to 2\gamma^{-1/2} + \gamma^{-1} \le \frac{9}{4}\gamma^{-1/2} < 3\gamma^{-1/2}$ and $\lambda_{\min}(VV^\top/N) \to (1-\gamma^{-1/2})^2 \ge \frac{9}{16} > \frac{1}{8}$, and the inequalities being strict they hold with probability tending to one. For the centred version, write $\frac{1}{N}\big(VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top\big) = \frac{VV^\top}{N} - \frac{1}{N^2}(V\mathbf{1})(V\mathbf{1})^\top$, where $\frac{1}{N^2}(V\mathbf{1})(V\mathbf{1})^\top \succeq 0$ has operator norm $\|V\mathbf{1}\|^2/N^2 = \frac{\|V\mathbf{1}\|^2}{Nd}\cdot\frac{d}{N} \to \gamma^{-1}$ in probability by (iii). By Weyl's inequality,

$$\begin{aligned}\lambda_{\min}\Big(\frac{1}{N}\big(VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top\big)\Big) &\ge \lambda_{\min}\Big(\frac{VV^\top}{N}\Big) - \frac{\|V\mathbf{1}\|^2}{N^2},\\ \Big\|\frac{1}{N}\big(VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top\big) - I_d\Big\|_{\mathrm{op}} &\le \Big\|\frac{VV^\top}{N} - I_d\Big\|_{\mathrm{op}} + \frac{\|V\mathbf{1}\|^2}{N^2},\end{aligned} \tag{29}$$

and in probability the two right-hand sides converge to

$$(1-\gamma^{-1/2})^2 - \gamma^{-1} \ge \frac{9}{16} - \frac{1}{16} = \frac{1}{2} > \frac{1}{8}, \qquad 2\gamma^{-1/2} + 2\gamma^{-1} \le \frac{5}{2}\gamma^{-1/2} < 3\gamma^{-1/2}, \tag{30}$$

using $\gamma^{-1/2} \le \frac{1}{4}$; the inequalities being strict, they hold with probability tending to one.

(vi) $VV^\top$ has rank at most d, so $\|VV^\top\|_F \le \sqrt{d}\|VV^\top\|_{\mathrm{op}} = \sqrt{d}\|V\|_{\mathrm{op}}^2$ by Equation (23); now use (iv).

(vii) Expanding, $\|UV^\top\|_F^2 = \sum_{k,l}(u_k^\top u_l)(v_k^\top v_l) = \sum_k \|u_k\|^2\|v_k\|^2 + \sum_{k\ne l}(u_k^\top u_l)(v_k^\top v_l)$.

Mean. $\mathbb{E}\|u_k\|^2\|v_k\|^2 = d^2$ and, for $k \ne l$, $\mathbb{E}(u_k^\top u_l)(v_k^\top v_l) = 0$, so $\mathbb{E}\|UV^\top\|_F^2 = Nd^2$.

Variance. The variables $\|u_k\|^2\|v_k\|^2$, $k \in [N]$, are independent with $\mathrm{var}(\|u_k\|^2\|v_k\|^2) = (d^2+2d)^2 - d^4 \le 8d^3$. Next, $\sum_{k\ne l}(u_k^\top u_l)(v_k^\top v_l) = 2\sum_{k<l}(u_k^\top u_l)(v_k^\top v_l)$; for unordered pairs $\{k,l\} \ne \{k',l'\}$ one of the indices k,l,k',l' occurs exactly once, and integrating over the u with that index gives $\mathbb{E}\big[(u_k^\top u_l)(v_k^\top v_l)(u_{k'}^\top u_{l'})(v_{k'}^\top v_{l'})\big] = 0$; since $\mathbb{E}(u_k^\top u_l)^2(v_k^\top v_l)^2 = d^2$, this gives $\mathrm{var}\sum_{k\ne l}(u_k^\top u_l)(v_k^\top v_l) = 4\cdot\frac{N(N-1)}{2}\cdot d^2 \le 2N^2d^2$. For $k' \in [N]$ and $k \ne l$, one of k,l differs from k', and integrating over the u with that index gives $\mathbb{E}\big[(\|u_{k'}\|^2\|v_{k'}\|^2 - d^2)(u_k^\top u_l)(v_k^\top v_l)\big] = 0$, so the two sums are uncorrelated. Therefore

$$\mathrm{var}\|UV^\top\|_F^2 \le 8Nd^3 + 2N^2d^2, \qquad \frac{\big(\mathrm{var}\|UV^\top\|_F^2\big)^{1/2}}{\mathbb{E}\|UV^\top\|_F^2} \le \sqrt{\frac{8}{Nd} + \frac{2}{d^2}} \le \frac{4}{d}, \tag{31}$$

and Chebyshev gives the first claim.

Centred version. $\|UV^\top\|_F^2 - d\sum_k\|v_k\|^2 = \sum_k(\|u_k\|^2 - d)\|v_k\|^2 + \sum_{k\ne l}(u_k^\top u_l)(v_k^\top v_l)$ has mean zero. The variables $(\|u_k\|^2 - d)\|v_k\|^2$ are independent with variance $\mathbb{E}(\|u\|^2-d)^2\,\mathbb{E}\|v\|^4 = 2d(d^2+2d) \le 6d^3$; the second sum has variance at most $2N^2d^2$ as above; and for $k' \in [N]$ and $k \ne l$, integrating over the u whose index among k,l differs from k' gives $\mathbb{E}\big[(\|u_{k'}\|^2 - d)\|v_{k'}\|^2(u_k^\top u_l)(v_k^\top v_l)\big] = 0$. Hence the variance is at most $6Nd^3 + 2N^2d^2$, its square root is at most $\sqrt{6}\sqrt{N}d^{3/2} + \sqrt{2}Nd$, and Chebyshev gives the second claim.

(viii) Conditionally on U, the columns of $UV^\top$ are i.i.d. $\mathcal{N}(0, UU^\top)$, i.e. of the form $(UU^\top)^{1/2}g$ with $g \sim \mathcal{N}(0, I_d)$; stacking the d vectors g as the columns of Ξ gives the stated equality in distribution, and the conditional law of Ξ does not depend on U, so Ξ is independent of U. For the norm, $\|(UU^\top)^{1/2}\Xi\|_{\mathrm{op}} \le \|U\|_{\mathrm{op}}\|\Xi\|_{\mathrm{op}}$, and almost surely $\|U\|_{\mathrm{op}}/\sqrt{N} \to 1 + \gamma^{-1/2} \le 2$ by (iv) while $\|\Xi\|_{\mathrm{op}}/\sqrt{d} \to 2$ by Bai–Yin with aspect ratio 1. □

Lemma E.4 (Norms of W^{GD}).

(i) $\|W^{\mathrm{GD}}\|_F^2 = (N-1)d^2(1+O_p(d^{-1}))$, *and* $\|W^{\mathrm{GD}}\|_F^2 - d\,\mathrm{tr}\left(VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top\right) = O_p(\sqrt{N}d^{3/2} + Nd)$.

(ii) $\|W^{\mathrm{GD}}\|_{\mathrm{op}} \le \|U\|_{\mathrm{op}}\|V\|_{\mathrm{op}}$, *and with probability tending to one* $\|W^{\mathrm{GD}}\|_{\mathrm{op}} \le 7\sqrt{Nd}$.

(iii) With probability tending to one, $\frac{1}{8}d\sqrt{Nd} \le \|W^{\mathrm{GD}}\|_* \le 7d\sqrt{Nd}$.

(iv) In probability, $\|W^{\mathrm{GD}}\|_*/(Nd) \to \Phi(\gamma)/\sqrt{\gamma}$, *where*

$$\Phi(\gamma) := \frac{1}{\pi}\int_0^1 \left[\sqrt{1-\frac{p}{\gamma}}\,\frac{1+p}{\sqrt{p}} + \frac{\sqrt{p}\,(1-p)}{\gamma\sqrt{1-p/\gamma}}\right]\mathrm{d}p, \tag{32}$$

and for $\gamma \ge 16$

$$\Phi(\gamma) = \frac{8}{3\pi}\Big(1 - \frac{1}{10\gamma} + r(\gamma)\Big), \qquad |r(\gamma)| \le \gamma^{-2}. \tag{33}$$

Part (iv) is where the constant $3\pi/8$ enters: Muon replaces W^{GD} by its polar factor, which turns the diagonal score into $\|W^{\mathrm{GD}}\|_*$, and $8/(3\pi)$ is the limiting Marčenko–Pastur integral at $\gamma = \infty$.

Proof. (i) By Equation (25) and $\|X-Y\|_F^2 = \|X\|_F^2 - 2\langle X, Y\rangle + \|Y\|_F^2$,

$$\|W^{\mathrm{GD}}\|_F^2 = \|UV^\top\|_F^2 - \frac{2}{N}(U\mathbf{1})^\top UV^\top(V\mathbf{1}) + \frac{1}{N^2}\|U\mathbf{1}\|^2\|V\mathbf{1}\|^2, \tag{34}$$

where the cross term is $O_p(N^2d)/N = O_p(Nd)$ and the last term is $O_p(d^2)$ by Lemma E.3(iii),(iv),(viii). Substituting $\|UV^\top\|_F^2 = Nd^2(1+O_p(d^{-1}))$ from Lemma E.3(vii) and using $d^2 \le Nd$ gives the first claim. For the second, $\mathrm{tr}(VV^\top) = \sum_k \|v_k\|^2$ and $\mathrm{tr}((V\mathbf{1})(V\mathbf{1})^\top) = \|V\mathbf{1}\|^2$, so the same expansion gives

$$\begin{aligned}\|W^{\mathrm{GD}}\|_F^2 - d\,\mathrm{tr}\left(VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top\right) &= \Big(\|UV^\top\|_F^2 - d\sum_k \|v_k\|^2\Big) + \frac{d\|V\mathbf{1}\|^2}{N} \\ &\quad - \frac{2}{N}(U\mathbf{1})^\top UV^\top(V\mathbf{1}) + \frac{\|U\mathbf{1}\|^2\|V\mathbf{1}\|^2}{N^2},\end{aligned} \tag{35}$$

whose four terms are $O_p(\sqrt{N}d^{3/2} + Nd)$ by the second half of Lemma E.3(vii), $O_p(d^2)$, $O_p(Nd)$ and $O_p(d^2)$ respectively; since $d^2 \le Nd$, the claim follows.

(ii) $\|I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N}\|_{\mathrm{op}} = 1$ gives the first inequality. For the second, the triangle inequality and the operator norm of a rank-one matrix give $\|W^{\mathrm{GD}}\|_{\mathrm{op}} \le \|UV^\top\|_{\mathrm{op}} + \|U\mathbf{1}\|\|V\mathbf{1}\|/N \le 5\sqrt{Nd} + \frac{3}{2}d \le 7\sqrt{Nd}$ with probability tending to one, by Lemma E.3(iii),(viii) and $d \le \sqrt{Nd}$.

(iii) For any X, $\|X\|_F^2 = \sum_a s_a^2 \le s_1 \sum_a s_a = \|X\|_{\mathrm{op}}\|X\|_*$, and $\|X\|_* \le \mathrm{rank}(X)\|X\|_{\mathrm{op}}$. Combining with (i) and (ii),

$$\begin{aligned}\|W^{\mathrm{GD}}\|_* &\ge \frac{\|W^{\mathrm{GD}}\|_F^2}{\|W^{\mathrm{GD}}\|_{\mathrm{op}}} \ge \frac{(N-1)d^2(1-o_p(1))}{7\sqrt{Nd}} \ge \frac{1}{8}d\sqrt{Nd}, \\ \|W^{\mathrm{GD}}\|_* &\le d\|W^{\mathrm{GD}}\|_{\mathrm{op}} \le 7d\sqrt{Nd}.\end{aligned} \tag{36}$$

(iv) *Removing the centring.* By the triangle inequality for the nuclear norm and the nuclear norm of a rank-one matrix, $|\|UV^\top\|_* - \|W^{\mathrm{GD}}\|_*| \le \|U\mathbf{1}\|\|V\mathbf{1}\|/N = O_p(d)$, which is $o(Nd)$; so it suffices to prove the limit for $\|UV^\top\|_*$.

Limit. By Lemma E.3(viii), $\{s_a(UV^\top)^2/(Nd)\}_{a\le d}$ has the same law as the eigenvalues of $\frac{1}{d}(\frac{UU^\top}{N})^{1/2}\Xi\Xi^\top(\frac{UU^\top}{N})^{1/2}$ with Ξ independent of U: conditionally on U this is a sample covariance matrix with population covariance $\frac{1}{N}UU^\top$ and aspect ratio 1, and the spectral distribution of $\frac{1}{N}UU^\top$ converges almost surely to the Marčenko–Pastur law H with ratio $1/\gamma$ and mean 1. By the

Marčenko–Pastur–Silverstein theorem the empirical distribution of $\{s_a(UV^\top)^2/(Nd)\}$ converges weakly in probability to a law ν whose Stieltjes transform m satisfies

$$z = -\frac{1}{m} + \int \frac{t}{1+tm}\,\mathrm{d}H(t). \tag{37}$$

By Lemma E.3(viii) all $s_a^2/(Nd) \le 25$ with probability tending to one, so both the empirical law and ν are supported in $[0,25]$; taking the limit for the bounded continuous function $\lambda \mapsto \min(\sqrt{\lambda}, 5)$ gives $\frac{1}{d}\sum_a s_a/\sqrt{Nd} \to \int \sqrt{\lambda}\,\mathrm{d}\nu$ in probability, that is

$$\frac{\|UV^\top\|_*}{Nd} = \sqrt{\frac{d}{N}}\cdot\frac{1}{d}\sum_a \frac{s_a}{\sqrt{Nd}} \longrightarrow \frac{1}{\sqrt{\gamma}}\int \sqrt{\lambda}\,\mathrm{d}\nu(\lambda). \tag{38}$$

Closed form. Let $\psi(x) := \int \frac{\lambda x}{1-\lambda x}\mathrm{d}\nu$. From $\frac{\lambda x}{1-\lambda x} = -1 + \frac{1/x}{1/x-\lambda}$ we get $\psi(x) = -1 - \frac{1}{x}m(1/x)$, i.e. $m(1/x) = -x(1+\psi(x))$. Taking $z = 1/x$ in Equation (37) and multiplying by $x(1+\psi(x))$,

$$\psi(x) = \int \frac{x(1+\psi(x))t}{1-x(1+\psi(x))t}\,\mathrm{d}H(t). \tag{39}$$

On the other hand the Stieltjes transform of the Marčenko–Pastur law with ratio $1/\gamma$ and mean 1 satisfies $\frac{1}{\gamma}zm_H(z)^2 - (1-\frac{1}{\gamma}-z)m_H(z) + 1 = 0$, and $\frac{ut}{1-ut} = -1 + \frac{1/u}{1/u-t}$ gives $m_H(1/u) = -u\big(1+\int \frac{ut}{1-ut}\mathrm{d}H(t)\big)$; substituting $z = 1/u$ into the quadratic yields, for all u,

$$\frac{u}{\gamma}\Big(1+\int\frac{ut}{1-ut}\mathrm{d}H(t)\Big)^2 + \Big(1-\frac{1}{\gamma}\Big)u\Big(1+\int\frac{ut}{1-ut}\mathrm{d}H(t)\Big) - \int\frac{ut}{1-ut}\mathrm{d}H(t) = 0. \tag{40}$$

Take $u = x(1+\psi(x))$. By Equation (39) the integral then equals $\psi(x)$, so

$$\frac{1}{\gamma}x(1+\psi)^3 + \Big(1-\frac{1}{\gamma}\Big)x(1+\psi)^2 - \psi = 0, \qquad \text{i.e.} \qquad x = \frac{\psi(x)}{(1+\psi(x))^2\,(1+\psi(x)/\gamma)}. \tag{41}$$

Using $\pi\sqrt{\lambda} = \int_0^\infty \frac{\lambda}{\lambda+w}\frac{\mathrm{d}w}{\sqrt{w}}$ and Tonelli's theorem, and then substituting $p = -\psi(-1/w) \in (0,1)$,

$$\begin{aligned}
\int \sqrt{\lambda}\,\mathrm{d}\nu &= \frac{1}{\pi}\int_0^\infty \frac{-\psi(-1/w)}{\sqrt{w}}\,\mathrm{d}w, \\
w &= \frac{(1-p)^2(1-p/\gamma)}{p}, \qquad -\frac{\mathrm{d}w}{\mathrm{d}p} = w\Big[\frac{2}{1-p} + \frac{1/\gamma}{1-p/\gamma} + \frac{1}{p}\Big], \\
\int \sqrt{\lambda}\,\mathrm{d}\nu &= \frac{1}{\pi}\int_0^1 \sqrt{1-\frac{p}{\gamma}}\Big[2\sqrt{p} + \frac{\sqrt{p}(1-p)/\gamma}{1-p/\gamma} + \frac{1-p}{\sqrt{p}}\Big]\mathrm{d}p;
\end{aligned} \tag{42}$$

then $2\sqrt{p} + \frac{1-p}{\sqrt{p}} = \frac{1+p}{\sqrt{p}}$ and $\sqrt{1-p/\gamma}\cdot\frac{\sqrt{p}(1-p)/\gamma}{1-p/\gamma} = \frac{\sqrt{p}(1-p)}{\gamma\sqrt{1-p/\gamma}}$ give Equation (32).

Expansion. For $0 \le s \le \frac{1}{16}$ one has

$$\Big|\sqrt{1-s} - 1 + \frac{s}{2}\Big| \le \frac{s^2}{5}, \qquad 0 \le (1-s)^{-1/2} - 1 \le \frac{2s}{3}. \tag{43}$$

For the first, $g(s) := 1 - \frac{s}{2} - \sqrt{1-s}$ satisfies $g(0) = 0$, $g'(s) = \frac{1}{2}((1-s)^{-1/2} - 1) \ge 0$ and $g''(s) = \frac{1}{4}(1-s)^{-3/2} \le \frac{1}{4}(1-\frac{1}{16})^{-3/2} < \frac{2}{5}$ on $[0, \frac{1}{16}]$, so $g(s) \le \frac{s^2}{2}\max g'' < \frac{s^2}{5}$. For the second, $f(s) := (1-s)^{-1/2} - 1 - \frac{2s}{3}$ satisfies $f(0) = 0$ and $f'(s) = \frac{1}{2}(1-s)^{-3/2} - \frac{2}{3} < 0$ for $s < 1 - (\frac{3}{4})^{2/3} \approx 0.175$, hence on $[0, \frac{1}{16}]$, so $f \le 0$ there. Take $s = p/\gamma$ with $\gamma \ge 16$. From

$$\frac{1}{\pi}\int_0^1 \frac{1+p}{\sqrt{p}}\,\mathrm{d}p = \frac{8}{3\pi}, \qquad -\frac{1}{2\gamma\pi}\int_0^1 \sqrt{p}(1+p)\,\mathrm{d}p + \frac{1}{\gamma\pi}\int_0^1 \sqrt{p}(1-p)\,\mathrm{d}p = \frac{8}{3\pi}\Big(-\frac{1}{10\gamma}\Big), \tag{44}$$

and Equation (32), the remainder is bounded by

$$\begin{aligned}
\Big|\Phi(\gamma) - \frac{8}{3\pi}\Big(1-\frac{1}{10\gamma}\Big)\Big| &\le \frac{1}{\pi\gamma^2}\Big[\frac{1}{5}\int_0^1 p^{3/2}(1+p)\,\mathrm{d}p + \frac{2}{3}\int_0^1 p^{3/2}(1-p)\,\mathrm{d}p\Big] \\
&= \frac{1}{\pi\gamma^2}\Big[\frac{24}{175} + \frac{8}{105}\Big] < \frac{8}{3\pi}\gamma^{-2},
\end{aligned} \tag{45}$$

which is Equation (33). □

E.3 THE GD STEP

By Equation (24) with $W = W^{\mathrm{GD}}$,

$$\bar{h}^{\mathrm{GD}} = \underbrace{\operatorname*{mean}_i u_i^\top W^{\mathrm{GD}} v_i}_{\text{diagonal}} - \underbrace{\operatorname*{mean}_i \operatorname*{mean}_{j\neq i} u_j^\top W^{\mathrm{GD}} v_i}_{\text{off-diagonal}}. \tag{46}$$

Lemma E.5 (Diagonal-sum identity)**.** *The diagonal scores of W^{GD} sum to*

$$\sum_{i=1}^{N} u_i^\top W^{\mathrm{GD}} v_i = \|W^{\mathrm{GD}}\|_F^2 + \frac{1}{N}(U\mathbf{1})^\top W^{\mathrm{GD}}(V\mathbf{1}).$$

Proof. A scalar equals its own trace and $\sum_i u_i v_i^\top = UV^\top$, so $\sum_i u_i^\top W^{\mathrm{GD}} v_i = \operatorname{tr}\big((W^{\mathrm{GD}})^\top \sum_i u_i v_i^\top\big) = \langle W^{\mathrm{GD}}, UV^\top\rangle$. Rearranging Equation (25) gives $UV^\top = W^{\mathrm{GD}} + \frac{1}{N}(U\mathbf{1})(V\mathbf{1})^\top$, and bilinearity together with $\langle Z, ab^\top\rangle = a^\top Z b$ gives the claim. □

Lemma E.6 (Rank-one cross term)**.** $\frac{1}{N^2}|(U\mathbf{1})^\top W^{\mathrm{GD}}(V\mathbf{1})| = O_p(d)$.

Proof. By Lemma E.4(ii) and Lemma E.3(iii),(iv),

$$\frac{|(U\mathbf{1})^\top W^{\mathrm{GD}}(V\mathbf{1})|}{N^2} \le \frac{\|U\mathbf{1}\|\,\|W^{\mathrm{GD}}\|_{\mathrm{op}}\,\|V\mathbf{1}\|}{N^2} = \frac{O_p(\sqrt{Nd})\cdot O_p(N)\cdot O_p(\sqrt{Nd})}{N^2} = O_p(d). \tag{47}$$

□

Lemma E.7. $\operatorname{mean}_i u_i^\top W^{\mathrm{GD}} v_i = d^2(1 + O_p(d^{-1}))$.

Proof. Divide Lemma E.5 by N. By Lemma E.4(i), $\|W^{\mathrm{GD}}\|_F^2/N = \frac{N-1}{N} d^2(1 + O_p(d^{-1})) = d^2(1 + O_p(d^{-1}))$, using $\frac{1}{N} \le \frac{1}{d}$; by Lemma E.6 the second term is $O_p(d) = d^2 \cdot O_p(d^{-1})$. □

Lemma E.8 (Column-sum identity)**.** *For every $W \in \mathbb{R}^{d\times d}$ and every i,* $(N-1)\operatorname{mean}_{j\neq i} u_j^\top W v_i = (U\mathbf{1})^\top W v_i - u_i^\top W v_i$.

Proof. $(N-1)\operatorname{mean}_{j\neq i} u_j^\top W v_i = \sum_j u_j^\top W v_i - u_i^\top W v_i$ and $\sum_j u_j = U\mathbf{1}$. □

Lemma E.9 (Two sums of squares)**.** $\sum_i \big((U\mathbf{1})^\top W^{\mathrm{GD}} v_i\big)^2 = O_p(N^4 d)$ *and* $\sum_i \big(u_i^\top W^{\mathrm{GD}} v_i\big)^2 = O_p(N^4 d)$.

Proof. The numbers $(U\mathbf{1})^\top W^{\mathrm{GD}} v_i$ are the entries of $V^\top (W^{\mathrm{GD}})^\top U\mathbf{1}$, so by Lemma E.4(ii) and Lemma E.3(iii),(iv),

$$\big\|V^\top (W^{\mathrm{GD}})^\top U\mathbf{1}\big\|^2 \le \|V\|_{\mathrm{op}}^2 \|W^{\mathrm{GD}}\|_{\mathrm{op}}^2 \|U\mathbf{1}\|^2 = O_p(N\cdot N^2\cdot Nd) = O_p(N^4 d). \tag{48}$$

For the second sum, $u_i^\top W^{\mathrm{GD}} v_i = [U^\top W^{\mathrm{GD}} V]_{ii}$ and $\sum_i X_{ii}^2 \le \|X\|_F^2$; by Equation (25), $U^\top W^{\mathrm{GD}} V = (U^\top U)(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})(V^\top V)$, and since the nonzero eigenvalues of $V^\top V$ and $VV^\top$ coincide, $\|V^\top V\|_F = \|VV^\top\|_F$, so by Lemma E.3(iv),(vi),

$$\sum_i \big(u_i^\top W^{\mathrm{GD}} v_i\big)^2 \le \|U^\top W^{\mathrm{GD}} V\|_F^2 \le \|U\|_{\mathrm{op}}^4 \|VV^\top\|_F^2 = O_p(N^2\cdot N^2 d) = O_p(N^4 d). \tag{49}$$

□

Lemma E.10. $\operatorname{mean}_i \big(\operatorname{mean}_{j\neq i} u_j^\top W^{\mathrm{GD}} v_i\big)^2 = O_p(\gamma^2 d^2)$*, hence* $|\operatorname{mean}_i \operatorname{mean}_{j\neq i} u_j^\top W^{\mathrm{GD}} v_i| = O_p(\gamma d)$.

Proof. Square Lemma E.8 with $W = W^{\mathrm{GD}}$, use $(x-y)^2 \le 2x^2 + 2y^2$, sum over i and apply Lemma E.9:

$$\sum_i \Big((N-1)\operatorname*{mean}_{j\neq i} u_j^\top W^{\mathrm{GD}} v_i\Big)^2 = O_p(N^4 d) \le O_p(N^5), \tag{50}$$

using $d \le N$. Dividing by $N(N-1)^2 \ge \frac{1}{4}N^3$ and substituting $N = \gamma d$ gives the first claim; the second follows from Cauchy–Schwarz, $|\operatorname{mean}_i x_i| \le (\operatorname{mean}_i x_i^2)^{1/2}$. □

Lemma E.11 ($\bar{h}$ for GD)**.** $\bar{h}^{\mathrm{GD}} = d^2\big(1 + O_p(\gamma/d)\big)$.

Proof. By Equation (46), Lemma E.7 and Lemma E.10, $\bar{h}^{\mathrm{GD}} = d^2(1 + O_p(d^{-1})) - O_p(\gamma d) = d^2(1 + O_p(d^{-1}) + O_p(\gamma/d))$, and $\gamma \geq 1$ absorbs the first error term into the second. □

We turn to $\bar{\kappa}^{\mathrm{GD}}$. By Equation (24),

$$(\bar{\kappa}^{\mathrm{GD}})^2 = \underbrace{\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} \big(u_j^\top W^{\mathrm{GD}} v_i\big)^2}_{\text{second moment}} - \underbrace{\operatorname*{mean}_i \Big(\operatorname*{mean}_{j \neq i} u_j^\top W^{\mathrm{GD}} v_i \Big)^2}_{\text{squared mean}}. \tag{51}$$

Lemma E.12 (Expectation of the second moment)**.**

$$\begin{aligned} \mathbb{E} \sum_{i=1}^{N} \sum_{j \neq i} \big(u_j^\top W^{\mathrm{GD}} v_i\big)^2 &= N(N-1)d^2(N+2d+2) \\ &\quad - d^2\Big[\Big(1 + \frac{1}{N}\Big)(N+d+1)^2 - 2(d+2)^2 - 2(N-1)\Big] \\ &= N(N-1)d^2(N+2d+2)\big(1 + O(N^{-1})\big). \end{aligned} \tag{52}$$

Proof. Write $u_j^\top W^{\mathrm{GD}} v_i = [U^\top W^{\mathrm{GD}} V]_{ji}$; by Equation (25), $U^\top W^{\mathrm{GD}} V = (U^\top U)(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})(V^\top V)$, i.e. $u_j^\top W^{\mathrm{GD}} v_i = \sum_{k,l} (u_j^\top u_k)(\delta_{kl} - \frac{1}{N})(v_l^\top v_i)$, with U and V independent. Two second moments, identical for u and v, are used:

$$\mathbb{E}\big[(u_k^\top u_l)(u_l^\top u_m)\big] = \begin{cases} d^2 + 2d, & k = l = m \\ d, & k = m \neq l \\ 0, & \text{otherwise,} \end{cases} \qquad \mathbb{E}\big[(u_i^\top u_k)(u_i^\top u_{k'})\big] = \delta_{kk'} d\big[1 + (d+1)\delta_{ki}\big]. \tag{53}$$

The first of these gives $\big[\mathbb{E}(U^\top U)^2\big]_{km} = \sum_l \mathbb{E}\big[(u_k^\top u_l)(u_l^\top u_m)\big] = \delta_{km}\big[(d^2+2d) + (N-1)d\big] = \delta_{km} d(d+N+1)$, i.e. $\mathbb{E}(U^\top U)^2 = d(d+N+1)I_N$, and likewise for V. *All pairs* (i,j). Since $\|X\|_F^2 = \operatorname{tr}(X^\top X)$ and by cyclicity,

$$\sum_{i,j} \big(u_j^\top W^{\mathrm{GD}} v_i\big)^2 = \|U^\top W^{\mathrm{GD}} V\|_F^2 = \operatorname{tr}\Big[(U^\top U)^2 \Big(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N}\Big)(V^\top V)^2 \Big(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N}\Big)\Big], \tag{54}$$

and by independence and $\operatorname{tr}[(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})^2] = N - 1$,

$$\mathbb{E} \sum_{i,j} \big(u_j^\top W^{\mathrm{GD}} v_i\big)^2 = \big[d(d+N+1)\big]^2 (N-1). \tag{55}$$

Diagonal $j = i$. Using the independence of U and V, the second display above for both factors, and then $(\delta_{kl} - \frac{1}{N})^2 = \delta_{kl}(1 - \frac{2}{N}) + \frac{1}{N^2}$ together with $\sum_k [1 + (d+1)\delta_{ki}] = N + d + 1$ and $\sum_k [1 + (d+1)\delta_{ki}]^2 = N - 1 + (d+2)^2$,

$$\begin{aligned} \mathbb{E}\big(u_i^\top W^{\mathrm{GD}} v_i\big)^2 &= \sum_{k,l,k',l'} \mathbb{E}\big[(u_i^\top u_k)(u_i^\top u_{k'})\big]\, \mathbb{E}\big[(v_l^\top v_i)(v_{l'}^\top v_i)\big] \Big(\delta_{kl} - \frac{1}{N}\Big)\Big(\delta_{k'l'} - \frac{1}{N}\Big) \\ &= d^2 \sum_{k,l} \big[1 + (d+1)\delta_{ki}\big]\big[1 + (d+1)\delta_{li}\big]\Big(\delta_{kl} - \frac{1}{N}\Big)^2 \\ &= d^2\Big[\Big(1 - \frac{2}{N}\Big)\big(N - 1 + (d+2)^2\big) + \frac{(N+d+1)^2}{N^2}\Big]. \end{aligned} \tag{56}$$

Subtracting. $\mathbb{E}\sum_i \sum_{j \neq i} = \mathbb{E}\sum_{i,j} - N\mathbb{E}(u_i^\top W^{\mathrm{GD}} v_i)^2$; rearranging with $(N+d+1)^2 - (d+2)^2 - (N-1) = (N-1)(N+2d+2)$ gives the first equality. Since $d \leq N$, the discarded terms are bounded by $38N^2d^2 \leq \frac{38}{N-1} \cdot N(N-1)d^2(N+2d+2)$, which is the stated relative error. □

Lemma E.13 (Variance of the second moment)**.** *There is an absolute constant C with* $\operatorname{var}\big[\sum_i \sum_{j\neq i}(u_j^\top W^{\mathrm{GD}} v_i)^2\big] \le CN^7 d$.

Proof. By Equation (25), $\sum_i \sum_{j\neq i}(u_j^\top W^{\mathrm{GD}} v_i)^2 = \|U^\top W^{\mathrm{GD}} V\|_F^2 - \sum_i [U^\top W^{\mathrm{GD}} V]_{ii}^2$, a smooth function of the $2Nd$ independent standard Gaussian entries of (U, V). By the Gaussian Poincaré inequality the variance is at most $\mathbb{E}\|\nabla_U \cdot\|_F^2 + \mathbb{E}\|\nabla_V \cdot\|_F^2$.

The gradient of $M \mapsto \|M\|_F^2 - \sum_i M_{ii}^2$ is $2M - 2\operatorname{diag}(M)$, of Frobenius norm at most $2\|M\|_F$; the directional derivative of $U \mapsto (U^\top U)(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})(V^\top V)$ along H has Frobenius norm at most $2\|H\|_F \|U\|_{\mathrm{op}} \|V\|_{\mathrm{op}}^2$. With $\|U^\top W^{\mathrm{GD}} V\|_F \le \|U\|_{\mathrm{op}}^2 \|VV^\top\|_F$, the chain rule gives

$$\Big\|\nabla_U \sum_i \sum_{j\neq i}(u_j^\top W^{\mathrm{GD}} v_i)^2\Big\|_F \le 4\|U\|_{\mathrm{op}}^3 \|V\|_{\mathrm{op}}^2 \|VV^\top\|_F. \tag{57}$$

Squaring, taking expectations and applying Hölder together with Lemma E.3(iv),(vi) gives $\mathbb{E}\|\nabla_U \cdot\|_F^2 \le 16(\mathbb{E}\|U\|_{\mathrm{op}}^{18})^{1/3}(\mathbb{E}\|V\|_{\mathrm{op}}^{12})^{1/3}(\mathbb{E}\|VV^\top\|_F^6)^{1/3} \le 16(C_9 C_6^2)^{1/3} N^7 d$. The expression is invariant under transposition combined with swapping U and V, so the same bound holds for ∇_V; adding the two gives the claim with $C = 32(C_9 C_6^2)^{1/3}$. □

Lemma E.14 (Second moment)**.** *The second moment of the off-diagonal GD scores satisfies*

$$\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j\neq i} \big(u_j^\top W^{\mathrm{GD}} v_i\big)^2 = d^3(2+\gamma)\big(1 + O_p(\sqrt{\gamma}/d)\big).$$

Proof. By Equation (52) and $N(N-1)d^2(N+2d+2) \ge N^2(N-1)d^2 \ge \frac{1}{2}N^3 d^2$ for $N \ge 2$, the expectation is at least $\frac{1}{4}N^3 d^2$ for N large. Combined with Lemma E.13, the relative standard deviation is at most $\sqrt{CN^7 d}/(\frac{1}{4}N^3 d^2) = 4\sqrt{C}\sqrt{N/d^3} = 4\sqrt{C}\sqrt{\gamma}/d$, so Chebyshev gives

$$\sum_i \sum_{j\neq i} \big(u_j^\top W^{\mathrm{GD}} v_i\big)^2 = \Big(\mathbb{E}\sum_i \sum_{j\neq i} \big(u_j^\top W^{\mathrm{GD}} v_i\big)^2\Big)\big(1 + O_p(\sqrt{\gamma}/d)\big). \tag{58}$$

Since $N^{-1} = \gamma^{-1} d^{-1} \le \sqrt{\gamma}/d$, the relative error $O(N^{-1})$ of Lemma E.12 is absorbed; dividing by $N(N-1)$ leaves $d^2(N+2d+2) = d^3(\gamma + 2 + \frac{2}{d}) = d^3(2+\gamma)(1 + \frac{2}{d(2+\gamma)})$ and $\frac{2}{d(2+\gamma)} \le \frac{1}{d} \le \sqrt{\gamma}/d$. □

Lemma E.15 ($\bar\kappa$ for GD)**.** $(\bar\kappa^{\mathrm{GD}})^2 = d^3(2+\gamma)\big(1 + O_p(\gamma/d)\big)$.

Proof. By Lemma E.10 the squared mean in Equation (51) is $O_p(\gamma^2 d^2) = d^3(2+\gamma) \cdot O_p\big(\frac{\gamma^2}{d(2+\gamma)}\big) = d^3(2+\gamma) \cdot O_p(\gamma/d)$. Subtracting it from Lemma E.14 and using $\sqrt{\gamma} \le \gamma$ gives the claim. □

Lemma E.16 (ρ for GD). *The margin-to-noise ratio of the GD iterate is*

$$\rho^{\mathrm{GD}} = \sqrt{\frac{d}{2+\gamma}}\big(1 + O_p(\gamma/d)\big) = \sqrt{\frac{d}{\gamma}}\Big(1 - \frac{1}{\gamma} + O(\gamma^{-2}) + o_p(1)\Big).$$

Both $\bar h^{\mathrm{GD}}$ and $\bar\kappa^{\mathrm{GD}}$ are a power of d times a function of γ alone, so ρ^{GD} depends on d only through the prefactor $\sqrt{d}$.

Proof. By Lemma E.11, Lemma E.15 and $\sqrt{1 + O_p(\gamma/d)} = 1 + O_p(\gamma/d)$,

$$\rho^{\mathrm{GD}} = \frac{\bar h^{\mathrm{GD}}}{\bar\kappa^{\mathrm{GD}}} = \frac{d^2\big(1 + O_p(\gamma/d)\big)}{d^{3/2}\sqrt{2+\gamma}\big(1 + O_p(\gamma/d)\big)} = \sqrt{\frac{d}{2+\gamma}}\big(1 + O_p(\gamma/d)\big). \tag{59}$$

With γ fixed and $d \to \infty$, $O_p(\gamma/d) = o_p(1)$; moreover $\sqrt{\gamma/(2+\gamma)} = (1 + 2/\gamma)^{-1/2} = 1 - \gamma^{-1} + O(\gamma^{-2})$. □

E.4 THE MUON STEP

Throughout this subsection $\gamma \geq 16$. By Equation (24) with $W = W^{\mathrm{Muon}}$ the same two decompositions hold as in Equation (46) and Equation (51), with W^{GD} replaced by W^{Muon}. We use throughout that W^{Muon} is orthogonal (Lemma E.2).

Lemma E.17 (Diagonal-sum identity)**.** *The diagonal scores of W^{Muon} sum to*

$$\sum_{i=1}^{N} u_i^\top W^{\mathrm{Muon}} v_i = \|W^{\mathrm{GD}}\|_* + \frac{1}{N}(U\mathbf{1})^\top W^{\mathrm{Muon}}(V\mathbf{1}).$$

Proof. As in Lemma E.5, $\sum_i u_i^\top W^{\mathrm{Muon}} v_i = \langle W^{\mathrm{Muon}}, UV^\top \rangle = \langle W^{\mathrm{Muon}}, W^{\mathrm{GD}} \rangle + \frac{1}{N}(U\mathbf{1})^\top W^{\mathrm{Muon}}(V\mathbf{1})$. Writing the compact SVD $W^{\mathrm{GD}} = P\Sigma Q^\top$, so that $W^{\mathrm{Muon}} = PQ^\top$, and using $P^\top P = Q^\top Q = I$,

$$\langle W^{\mathrm{Muon}}, W^{\mathrm{GD}} \rangle = \operatorname{tr}\left(QP^\top P\Sigma Q^\top\right) = \operatorname{tr}\left(\Sigma Q^\top Q\right) = \operatorname{tr}\Sigma = \|W^{\mathrm{GD}}\|_*. \tag{60}$$

□

Lemma E.18 (Rank-one cross term)**.** $\frac{1}{N^2}|(U\mathbf{1})^\top W^{\mathrm{Muon}}(V\mathbf{1})| = O_p(d/N)$.

Proof. By $\|W^{\mathrm{Muon}}\|_{\mathrm{op}} \leq 1$ and Lemma E.3(iii), $|(U\mathbf{1})^\top W^{\mathrm{Muon}}(V\mathbf{1})|/N^2 \leq \|U\mathbf{1}\|\|V\mathbf{1}\|/N^2 = O_p(Nd)/N^2 = O_p(d/N)$. □

Lemma E.19. $\operatorname{mean}_i u_i^\top W^{\mathrm{Muon}} v_i = \frac{\|W^{\mathrm{GD}}\|_*}{N}\left(1 + O_p((Nd)^{-1/2})\right)$.

Proof. Divide Lemma E.17 by N and use Lemma E.18. By Lemma E.4(iii), with probability tending to one $\|W^{\mathrm{GD}}\|_*/N \geq d\sqrt{Nd}/(8N) = d^{3/2}/(8\sqrt{N})$, so the error term satisfies $O_p(d/N) = \frac{\|W^{\mathrm{GD}}\|_*}{N} O_p\left(\frac{d}{N}\frac{8\sqrt{N}}{d^{3/2}}\right) = \frac{\|W^{\mathrm{GD}}\|_*}{N} O_p((Nd)^{-1/2})$. □

Lemma E.20 (Two sums of squares)**.** $\sum_i \left((U\mathbf{1})^\top W^{\mathrm{Muon}} v_i\right)^2 = O_p(N^2 d)$ *and* $\sum_i \left(u_i^\top W^{\mathrm{Muon}} v_i\right)^2 = O_p(Nd^2)$.

Proof. For the first, W^{Muon} is orthogonal, so $\|(W^{\mathrm{Muon}})^\top U\mathbf{1}\| = \|U\mathbf{1}\|$, and $\sum_i v_i v_i^\top = VV^\top$, whence by Lemma E.3(iii),(iv)

$$\sum_i \left((U\mathbf{1})^\top W^{\mathrm{Muon}} v_i\right)^2 = (U\mathbf{1})^\top W^{\mathrm{Muon}} VV^\top (W^{\mathrm{Muon}})^\top (U\mathbf{1}) \leq \|V\|_{\mathrm{op}}^2 \|U\mathbf{1}\|^2 = O_p(N \cdot Nd). \tag{61}$$

For the second, Cauchy–Schwarz and $\|W^{\mathrm{Muon}} v_i\| = \|v_i\|$ give $(u_i^\top W^{\mathrm{Muon}} v_i)^2 \leq \|u_i\|^2\|v_i\|^2$, and $x^2 \leq 2(x-d)^2 + 2d^2$ with Lemma E.3(ii) gives $\sum_i \|u_i\|^2\|v_i\|^2 \leq 2\sum_i (\|u_i\|\|v_i\| - d)^2 + 2Nd^2 = N \cdot O_p(d) + 2Nd^2 = O_p(Nd^2)$. □

Lemma E.21. $\operatorname{mean}_i \left(\operatorname{mean}_{j\neq i} u_j^\top W^{\mathrm{Muon}} v_i\right)^2 = O_p(d/N)$, *hence* $|\operatorname{mean}_i \operatorname{mean}_{j\neq i} u_j^\top W^{\mathrm{Muon}} v_i| = O_p(\sqrt{d/N})$.

Proof. Square Lemma E.8 with $W = W^{\mathrm{Muon}}$, use $(x-y)^2 \leq 2x^2 + 2y^2$, sum over i and apply Lemma E.20:

$$\sum_i \left((N-1)\operatorname*{mean}_{j\neq i} u_j^\top W^{\mathrm{Muon}} v_i\right)^2 \leq 2O_p(N^2 d) + 2O_p(Nd^2) = O_p(N^2 d), \tag{62}$$

using $d \leq N$. Dividing by $N(N-1)^2 \geq \frac{1}{4}N^3$ gives the first claim, and Cauchy–Schwarz the second. □

Lemma E.22 ($\bar{h}$ for Muon)**.** *The mean margin of the Muon iterate is*

$$\bar{h}^{\mathrm{Muon}} = \frac{\|W^{\mathrm{GD}}\|_*}{N}\left(1 + O_p(d^{-1})\right) = \frac{d\,\Phi(\gamma)}{\sqrt{\gamma}}\left(1 + o_p(1)\right).$$

Proof. By Lemma E.19 and Lemma E.21, $\bar{h}^{\mathrm{Muon}} = \frac{\|W^{\mathrm{GD}}\|_*}{N}(1 + O_p((Nd)^{-1/2})) - O_p(\sqrt{d/N})$. By Lemma E.4(iii), with probability tending to one $\frac{\|W^{\mathrm{GD}}\|_*}{N} \geq \frac{d^{3/2}}{8\sqrt{N}}$, so

$$O_p\big(\sqrt{d/N}\big) = \frac{\|W^{\mathrm{GD}}\|_*}{N} \cdot O_p\Big(\sqrt{\frac{d}{N}} \cdot \frac{8\sqrt{N}}{d^{3/2}}\Big) = \frac{\|W^{\mathrm{GD}}\|_*}{N} \cdot O_p(d^{-1}), \tag{63}$$

and $(Nd)^{-1/2} \leq d^{-1}$, which gives the first equality. For the second, Lemma E.4(iv) gives $\frac{\|W^{\mathrm{GD}}\|_*}{Nd} = \frac{\Phi(\gamma)}{\sqrt{\gamma}} + o_p(1)$, and $\frac{\Phi(\gamma)}{\sqrt{\gamma}} \geq \frac{1}{8\sqrt{\gamma}} > 0$ by Lemma E.4(iii), so the additive $o_p(1)$ may be written multiplicatively. □

We now turn to $\bar{\kappa}^{\mathrm{Muon}}$. The orthogonality of W^{Muon} turns the second moment into three terms.

Lemma E.23 (Three-term identity)**.** *Almost surely*

$$\begin{aligned}\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} \big(u_j^\top W^{\mathrm{Muon}} v_i\big)^2 = \frac{N}{N-1} \operatorname*{mean}_i \|v_i\|^2 - \frac{\operatorname{mean}_i \big(u_i^\top W^{\mathrm{Muon}} v_i\big)^2}{N-1} \\ + \frac{\operatorname{tr}\big[(UU^\top - NI_d) W^{\mathrm{Muon}} VV^\top (W^{\mathrm{Muon}})^\top\big]}{N(N-1)}.\end{aligned} \tag{64}$$

Proof. Since $(U^\top x)_j = u_j^\top x$ and $\|W^{\mathrm{Muon}} v_i\| = \|v_i\|$,

$$\begin{aligned}\sum_{j \neq i} \big(u_j^\top W^{\mathrm{Muon}} v_i\big)^2 &= \|U^\top W^{\mathrm{Muon}} v_i\|^2 - \big(u_i^\top W^{\mathrm{Muon}} v_i\big)^2, \\ \|U^\top W^{\mathrm{Muon}} v_i\|^2 &= N\|v_i\|^2 + \big(W^{\mathrm{Muon}} v_i\big)^\top \big(UU^\top - NI_d\big)\big(W^{\mathrm{Muon}} v_i\big).\end{aligned} \tag{65}$$

Summing over i and using $x^\top Z x = \operatorname{tr}(Z x x^\top)$ with $\sum_i v_i v_i^\top = VV^\top$ turns the middle term into the stated trace; dividing by $N(N-1)$ and using $\sum_i \|v_i\|^2 = N \operatorname{mean}_i \|v_i\|^2$ gives the claim. □

Lemma E.24 (Norm term)**.** $\frac{N}{N-1} \operatorname{mean}_i \|v_i\|^2 = d\big(1 + O(N^{-1}) + O_p((Nd)^{-1/2})\big)$.

Proof. Lemma E.3(i) gives $\operatorname{mean}_i \|v_i\|^2 = d(1 + O_p((Nd)^{-1/2}))$, and $\frac{N}{N-1} = 1 + \frac{1}{N-1} = 1 + O(N^{-1})$. □

The trace term requires two auxiliary facts.

Lemma E.25 (Decomposition of $UU^\top$)**.** *Almost surely $VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top$ is invertible and there is $\Gamma \in \mathbb{R}^{(N-1-d)\times d}$ with*

$$UU^\top - \frac{1}{N}(U\mathbf{1})(U\mathbf{1})^\top = W^{\mathrm{GD}}\Big(VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top\Big)^{-1}(W^{\mathrm{GD}})^\top + \Gamma^\top \Gamma, \tag{66}$$

where, conditionally on V, the entries of Γ are i.i.d. $\mathcal{N}(0,1)$ and independent of W^{GD}. Consequently, for every symmetric $Y \in \mathbb{R}^{d\times d}$ that is a function of (V, W^{GD}),

$$\begin{aligned}\mathbb{E}\big[\operatorname{tr}(\Gamma^\top \Gamma Y) \mid V, W^{\mathrm{GD}}\big] &= (N-1-d)\operatorname{tr} Y, \\ \operatorname{var}\big[\operatorname{tr}(\Gamma^\top \Gamma Y) \mid V, W^{\mathrm{GD}}\big] &= 2(N-1-d)\|Y\|_F^2.\end{aligned} \tag{67}$$

Proof. Invertibility. $VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top = V(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})V^\top \succeq 0$ and its determinant is a polynomial in V which is nonzero at $V = [\,I_d \quad -\sum_{k\leq d} e_k \quad 0\,]$, where it equals $\det(I_d + (\sum_k e_k)(\sum_k e_k)^\top) = 1 + d \neq 0$; so it is nonzero almost surely.

Basis. If $c_0 \mathbf{1}^\top + c^\top V = 0$ for some $c_0 \in \mathbb{R}$ and $c \in \mathbb{R}^d$, then multiplying on the right by $(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})V^\top c$, and using $\mathbf{1}^\top(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N}) = 0$, gives $c^\top(VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top)c = 0$, so $c = 0$ and then $c_0 = 0$. Hence $\{x \in \mathbb{R}^N : \mathbf{1}^\top x = 0, Vx = 0\}$ has dimension $N - 1 - d$ almost surely; stack an orthonormal basis of it as the columns of $R \in \mathbb{R}^{N\times(N-1-d)}$, so that $R^\top R = I$, $\mathbf{1}^\top R = 0$ and $VR = 0$, and set $\Gamma := R^\top U^\top$.

Projection identity. Because $\mathbb{R}^N = \operatorname{span}\mathbf{1} \oplus \operatorname{col}\big((I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})V^\top\big) \oplus \operatorname{col} R$ with the three subspaces pairwise orthogonal and of dimensions $1 + d + (N - 1 - d) = N$,

$$I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N} = \Big(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N}\Big)V^\top\Big(VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top\Big)^{-1}V\Big(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N}\Big) + RR^\top. \tag{68}$$

Both sides are symmetric, so it suffices to check them on the three subspaces. Write $P := I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N}$ and $G := VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top$, and use $P^2 = P$, $VPV^\top = G$ and $PR = R - \frac{1}{N}\mathbf{1}(\mathbf{1}^\top R) = R$. For $x = \mathbf{1}$: $Px = 0$, $VPx = 0$ and $R^\top x = 0$, so both sides vanish. For $x = PV^\top y$: $Px = x$, while $VPx = VPV^\top y = Gy$, so the first term of the right-hand side is $PV^\top G^{-1}Gy = x$, and $R^\top x = (VR)^\top y = 0$, so the second vanishes. For $x = Rz$: $Px = Rz = x$, while $VPx = VRz = 0$ kills the first term, and $RR^\top Rz = Rz = x$. In all three cases both sides agree. Multiplying Equation (68) on the left by U and on the right by $U^\top$, and using $U(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})U^\top = UU^\top - \frac{1}{N}(U\mathbf{1})(U\mathbf{1})^\top$ and $U(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})V^\top = W^{\mathrm{GD}}$, gives the decomposition with $URR^\top U^\top = \Gamma^\top\Gamma$.

Distribution. Conditionally on V, the rows of U are i.i.d. $\mathcal{N}(0, I_N)$; the a-th column of Γ is $R^\top U^\top e_a \sim \mathcal{N}(0, R^\top R) = \mathcal{N}(0, I_{N-1-d})$ with independent columns, so the entries of Γ are i.i.d. standard Gaussian. The a-th row of W^{GD} is $\big(V(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})U^\top e_a\big)^\top$ and $\operatorname{Cov}\big(V(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})U^\top e_a, R^\top U^\top e_a\big) = V(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})R = 0$; being jointly Gaussian conditionally on V, Γ and W^{GD} are conditionally independent.

Moments. The rows of Γ are i.i.d. $\mathcal{N}(0, I_d)$ and $\operatorname{tr}(\Gamma^\top\Gamma Y) = \sum_m (\Gamma^\top e_m)^\top Y(\Gamma^\top e_m)$; for $g \sim \mathcal{N}(0, I_d)$ and fixed symmetric Y, $\mathbb{E}[g^\top Yg] = \operatorname{tr} Y$ and $\operatorname{var}[g^\top Yg] = 2\|Y\|_F^2$; summing over the $N - 1 - d$ rows gives the claim. □

Lemma E.26 (Trace of a commutator)**.** *Almost surely, with* $S := \frac{1}{N}\big(VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top\big)$,

$$\Big|\operatorname{tr}\Big[W^{\mathrm{GD}}(NS)^{-1}(W^{\mathrm{GD}})^\top W^{\mathrm{Muon}}(NS)(W^{\mathrm{Muon}})^\top\Big] - \|W^{\mathrm{GD}}\|_F^2\Big| \le \frac{2\|S - I_d\|_{\mathrm{op}}^2}{\lambda_{\min}(S)}\|W^{\mathrm{GD}}\|_F^2. \tag{69}$$

Proof. A matrix inequality. For symmetric $S \succ 0$ and symmetric T,

$$\big|\operatorname{tr}[S^{-1}TST] - \|T\|_F^2\big| \le \frac{2\|S - I_d\|_{\mathrm{op}}^2\|T\|_F^2}{\lambda_{\min}(S)}. \tag{70}$$

Indeed, cyclicity gives $\operatorname{tr}[S^{-1}STT] = \|T\|_F^2$ and $\operatorname{tr}[(TS - ST)T] = 0$, so with $S^{-1} = I_d - S^{-1}(S - I_d)$,

$$\operatorname{tr}[S^{-1}TST] - \|T\|_F^2 = \operatorname{tr}\big[S^{-1}(TS - ST)T\big] = -\operatorname{tr}\big[S^{-1}(S - I_d)(TS - ST)T\big], \tag{71}$$

and $|\operatorname{tr}(ABC)| \le \|A\|_{\mathrm{op}}\|B\|_F\|C\|_F$, $\|S^{-1}(S - I_d)\|_{\mathrm{op}} \le \|S - I_d\|_{\mathrm{op}}/\lambda_{\min}(S)$ and $\|TS - ST\|_F \le 2\|S - I_d\|_{\mathrm{op}}\|T\|_F$ give Equation (70).

Polar decomposition. With $W^{\mathrm{GD}} = P\Sigma Q^\top$ and $W^{\mathrm{Muon}} = PQ^\top$ we have $(W^{\mathrm{GD}})^\top W^{\mathrm{Muon}} = Q\Sigma Q^\top = (W^{\mathrm{Muon}})^\top W^{\mathrm{GD}}$ and $\|Q\Sigma Q^\top\|_F = \|\Sigma\|_F = \|W^{\mathrm{GD}}\|_F$.

Substitution. Moving the leading W^{GD} to the end by cyclicity,

$$\operatorname{tr}\Big[W^{\mathrm{GD}}(NS)^{-1}(W^{\mathrm{GD}})^\top W^{\mathrm{Muon}}(NS)(W^{\mathrm{Muon}})^\top\Big] = \operatorname{tr}\Big[(NS)^{-1}\big(Q\Sigma Q^\top\big)(NS)\big(Q\Sigma Q^\top\big)\Big], \tag{72}$$

and applying Equation (70) with $T = Q\Sigma Q^\top$ (the factors N cancel) gives the claim. □

Lemma E.27 (Trace term). *If $\gamma \ge 16$, then with probability tending to one*

$$\frac{1}{N(N-1)d}\Big|\operatorname{tr}\Big[\big(UU^\top - NI_d\big)W^{\mathrm{Muon}}VV^\top(W^{\mathrm{Muon}})^\top\Big]\Big| \le 145\gamma^{-2}. \tag{73}$$

The assumption $\gamma \ge 16$ is tailored to this estimate: through Lemma E.3(v) it controls how far $\frac{1}{N}VV^\top$ is from the identity, so that the fluctuation of $UU^\top$ around NI_d, seen through the Muon iterate, contributes only $O(\gamma^{-2})$ to the second moment.

Proof. Write $B := W^{\mathrm{Muon}}\big(VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top\big)(W^{\mathrm{Muon}})^\top$. By Lemma E.2 orthogonal conjugation preserves $\|\cdot\|_F$, $\|\cdot\|_{\mathrm{op}}$ and the trace; moreover $VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top = V(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N})V^\top$ satisfies $0 \preceq VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top \preceq VV^\top$. Hence, by Lemma E.3(iii),(iv),(vi)

$$\begin{aligned} \|B\|_F &\le \|VV^\top\|_F + \frac{\|V\mathbf{1}\|^2}{N} = O_p(N\sqrt{d}), \\ \|B\|_{\mathrm{op}} &\le \|V\|_{\mathrm{op}}^2 = O_p(N), \\ \operatorname{tr} B &= \sum_k \|v_k\|^2 - \frac{\|V\mathbf{1}\|^2}{N} = O_p(Nd). \end{aligned} \tag{74}$$

Rank-one replacement. $VV^\top - (VV^\top - \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top) = \frac{1}{N}(V\mathbf{1})(V\mathbf{1})^\top$, so by Lemma E.3(iii),(v),

$$\begin{aligned} &\Big|\operatorname{tr}\big[(UU^\top - NI_d)W^{\mathrm{Muon}}VV^\top(W^{\mathrm{Muon}})^\top\big] - \operatorname{tr}\big[(UU^\top - NI_d)B\big]\Big| \\ &\qquad \le \frac{1}{N}\|UU^\top - NI_d\|_{\mathrm{op}}\|V\mathbf{1}\|^2 = N(N-1)d\cdot O_p\big(\sqrt{d}N^{-3/2}\big). \end{aligned} \tag{75}$$

Splitting. By Lemma E.25 and $d + (N-1-d) + 1 = N$,

$$UU^\top - NI_d = \Big[W^{\mathrm{GD}}\big(NS\big)^{-1}(W^{\mathrm{GD}})^\top - dI_d\Big] + \Big[\Gamma^\top\Gamma - (N-1-d)I_d\Big] + \Big[\frac{1}{N}(U\mathbf{1})(U\mathbf{1})^\top - I_d\Big], \tag{76}$$

and we bound the trace of each bracket against B separately.

First bracket. Since $\operatorname{tr} B$ appears, write

$$\begin{aligned} \operatorname{tr}\big[(W^{\mathrm{GD}}(NS)^{-1}(W^{\mathrm{GD}})^\top - dI_d)B\big] = \big(&\operatorname{tr}[W^{\mathrm{GD}}(NS)^{-1}(W^{\mathrm{GD}})^\top B] - \|W^{\mathrm{GD}}\|_F^2\big) \\ &+ \big(\|W^{\mathrm{GD}}\|_F^2 - d\operatorname{tr} B\big). \end{aligned} \tag{77}$$

By Lemma E.26, Lemma E.3(v) ($\gamma \ge 16$) and Lemma E.4(i), with probability tending to one the first difference is at most

$$\frac{2(3\gamma^{-1/2})^2}{1/8}\|W^{\mathrm{GD}}\|_F^2 = 144\gamma^{-1}(N-1)d^2\big(1+O_p(d^{-1})\big) = N(N-1)d\cdot 144\gamma^{-2}\big(1+O_p(d^{-1})\big), \tag{78}$$

while the second is $O_p(\sqrt{N}d^{3/2} + Nd) = N(N-1)d\cdot O_p(\sqrt{d}N^{-3/2} + N^{-1})$ by Lemma E.4(i).

Second bracket. B is symmetric and a function of (V, W^{GD}), so Lemma E.25 gives $\mathbb{E}\big[\operatorname{tr}[(\Gamma^\top\Gamma - (N-1-d)I_d)B] \mid V, W^{\mathrm{GD}}\big] = 0$ and, by Equation (74), conditional variance $2(N-1-d)\|B\|_F^2 = O(N)\cdot O_p(N^2d) = O_p(N^3d)$. The conditional Chebyshev inequality then gives $\operatorname{tr}[(\Gamma^\top\Gamma - (N-1-d)I_d)B] = O_p(N^{3/2}d^{1/2}) = N(N-1)d\cdot O_p((Nd)^{-1/2})$.

Third bracket. By $\operatorname{tr}(aa^\top Y) = a^\top Ya \le \|a\|^2\|Y\|_{\mathrm{op}}$, Equation (74) and Lemma E.3(iii),

$$\begin{aligned} \Big|\operatorname{tr}\Big[\big(\frac{1}{N}(U\mathbf{1})(U\mathbf{1})^\top - I_d\big)B\Big]\Big| &\le \frac{\|U\mathbf{1}\|^2}{N}\|B\|_{\mathrm{op}} + |\operatorname{tr} B| \\ &= O_p(d)\cdot O_p(N) + O_p(Nd) = N(N-1)d\cdot O_p(N^{-1}). \end{aligned} \tag{79}$$

Total. Adding the four contributions and dividing by $N(N-1)d$, with probability tending to one the left-hand side is at most $144\gamma^{-2}(1+O_p(d^{-1})) + o_p(1) = 144\gamma^{-2} + o_p(1)$, which is at most $145\gamma^{-2}$ since γ is fixed. □

The remaining term of Lemma E.23, the diagonal one, needs the following sandwich.

Lemma E.28 (Nuclear-norm sandwich)**.** *For all $a, b \in \mathbb{R}^d$,*

$$a^\top \operatorname{Orth}\big(W^{\mathrm{GD}} - ab^\top\big)b \;\le\; \|W^{\mathrm{GD}}\|_* - \|W^{\mathrm{GD}} - ab^\top\|_* \;\le\; a^\top W^{\mathrm{Muon}}b. \tag{80}$$

Proof. For all X, Y, $\langle X, \operatorname{Orth}(Y)\rangle \le \|X\|_*\|\operatorname{Orth}(Y)\|_{\mathrm{op}} \le \|X\|_*$ by Equation (23), and writing the compact SVD $Y = P\Sigma Q^\top$ gives $\langle Y, \operatorname{Orth}(Y)\rangle = \operatorname{tr}(Q\Sigma P^\top PQ^\top) = \|Y\|_*$; also $\langle ab^\top, Z\rangle = a^\top Zb$. Hence

$$\begin{aligned} a^\top \operatorname{Orth}\big(W^{\mathrm{GD}} - ab^\top\big)b &= \big\langle W^{\mathrm{GD}}, \operatorname{Orth}(W^{\mathrm{GD}} - ab^\top)\big\rangle - \big\langle W^{\mathrm{GD}} - ab^\top, \operatorname{Orth}(W^{\mathrm{GD}} - ab^\top)\big\rangle \\ &\le \|W^{\mathrm{GD}}\|_* - \|W^{\mathrm{GD}} - ab^\top\|_*, \end{aligned}$$

and symmetrically

$$\|W^{\mathrm{GD}}\|_* - \|W^{\mathrm{GD}} - ab^\top\|_* = \langle W^{\mathrm{GD}}, W^{\mathrm{Muon}}\rangle - \|W^{\mathrm{GD}} - ab^\top\|_* \\ \le \langle W^{\mathrm{GD}}, W^{\mathrm{Muon}}\rangle - \langle W^{\mathrm{GD}} - ab^\top, W^{\mathrm{Muon}}\rangle = a^\top W^{\mathrm{Muon}} b. \qquad \square$$

Lemma E.29 (Self term). *If $\gamma \ge 16$, then with probability tending to one*

$$0 \le \frac{1}{(N-1)d} \operatorname*{mean}_i \left(u_i^\top W^{\mathrm{Muon}} v_i\right)^2 \le 15\gamma^{-3/2}. \tag{81}$$

This bounds the diagonal term of Lemma E.23; it is what forces the $\gamma^{-3/2}$ error in $\bar\kappa^{\mathrm{Muon}}$, and the only place where Lemma E.28 is used.

Proof. Nonnegativity is clear.

Step 1 (reduction to a first absolute moment). By $\|W^{\mathrm{Muon}}\|_{\mathrm{op}} \le 1$, $|u_i^\top W^{\mathrm{Muon}} v_i| \le \|u_i\|\|v_i\|$, so

$$\operatorname*{mean}_i \left(u_i^\top W^{\mathrm{Muon}} v_i\right)^2 \le d \operatorname*{mean}_i \left|u_i^\top W^{\mathrm{Muon}} v_i\right| \\ + \left(\operatorname*{mean}_i \left(u_i^\top W^{\mathrm{Muon}} v_i\right)^2\right)^{1/2} \left(\operatorname*{mean}_i \left(\|u_i\|\|v_i\| - d\right)^2\right)^{1/2}. \tag{82}$$

For $x, c_0, c_1 \ge 0$, $x^2 \le c_0 + c_1 x$ forces $x \le \frac{c_1 + \sqrt{c_1^2 + 4c_0}}{2} \le c_1 + \sqrt{c_0}$ and hence $x^2 \le 2c_1^2 + 2c_0$; taking $x = (\mathrm{mean}_i (u_i^\top W^{\mathrm{Muon}} v_i)^2)^{1/2}$, $c_0 = d\,\mathrm{mean}_i |u_i^\top W^{\mathrm{Muon}} v_i|$ and $c_1 = (\mathrm{mean}_i (\|u_i\|\|v_i\| - d)^2)^{1/2}$, and using Lemma E.3(ii),

$$\operatorname*{mean}_i \left(u_i^\top W^{\mathrm{Muon}} v_i\right)^2 \le O_p(d) + 2d \operatorname*{mean}_i \left|u_i^\top W^{\mathrm{Muon}} v_i\right|. \tag{83}$$

Step 2 (a signed mean). By Lemma E.17 and Lemma E.18, $\mathrm{mean}_i\, u_i^\top W^{\mathrm{Muon}} v_i = \frac{\|W^{\mathrm{GD}}\|_*}{N} + O_p(d/N)$. For real $x \ge \ell$ one has

$$|x| \le x + 2|\ell| : \tag{84}$$

this is clear for $x \ge 0$, and for $x < 0$ it follows from $|x| = -x \le -\ell \le |\ell| \le x + 2|\ell|$, the last step because $x + 2|\ell| \ge \ell + 2|\ell| \ge |\ell|$.

Step 3 (conditioning). Fix i and set

$$a := \frac{u_i - \frac{1}{N} U\mathbf{1}}{\sqrt{1 - 1/N}}, \qquad b := \frac{v_i - \frac{1}{N} V\mathbf{1}}{\sqrt{1 - 1/N}}.$$

Since $\mathbf{1}^\top (e_i - \frac{1}{N}\mathbf{1}) = 0$ and $\|e_i - \frac{1}{N}\mathbf{1}\|^2 = 1 - \frac{1}{N}$, Equation (25) gives

$$W^{\mathrm{GD}} - ab^\top = U\Big[\Big(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N}\Big) - \frac{(e_i - \frac{1}{N}\mathbf{1})(e_i - \frac{1}{N}\mathbf{1})^\top}{1 - 1/N}\Big] V^\top, \tag{85}$$

where the bracket, call it M, is the orthogonal projection onto $\{x \in \mathbb{R}^N : \mathbf{1}^\top x = 0,\ (e_i - \frac{1}{N}\mathbf{1})^\top x = 0\}$. Indeed $M = P - \Pi$ with $P := I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N}$ and $\Pi := (e_i - \frac{1}{N}\mathbf{1})(e_i - \frac{1}{N}\mathbf{1})^\top / \|e_i - \frac{1}{N}\mathbf{1}\|^2$ the orthogonal projections onto $\mathbf{1}^\perp$ and onto $\mathrm{span}(e_i - \frac{1}{N}\mathbf{1})$; since $\mathbf{1}^\top (e_i - \frac{1}{N}\mathbf{1}) = 0$ the range of Π lies in that of P, so $P\Pi = \Pi P = \Pi$ and hence $M^\top = M$, $M^2 = M$ and $M(e_i - \frac{1}{N}\mathbf{1}) = 0$. Consequently a and b are standard Gaussian and independent of $W^{\mathrm{GD}} - ab^\top$. For a: its k-th coordinate is $(e_i - \frac{1}{N}\mathbf{1})^\top U^\top e_k / \sqrt{1 - 1/N}$ and the k-th row of UM is $MU^\top e_k$; conditionally on V the vectors $U^\top e_k$ are i.i.d. $\mathcal{N}(0, I_N)$, and

$$\mathrm{Cov}\left((e_i - \frac{1}{N}\mathbf{1})^\top U^\top e_k,\ MU^\top e_k\right) = M(e_i - \frac{1}{N}\mathbf{1}) = 0, \tag{86}$$

so, being jointly Gaussian, a is independent of UM, and $\|(e_i - \frac{1}{N}\mathbf{1})/\sqrt{1 - 1/N}\| = 1$ gives $a \sim \mathcal{N}(0, I_d)$. The same argument applied to the rows of V gives $b \sim \mathcal{N}(0, I_d)$ independent of

VM, and $U \perp V$ makes a, b independent of each other; since $W^{\mathrm{GD}} - ab^\top = (UM)(VM)^\top$ by idempotence of M, the pair (a, b) is independent of $W^{\mathrm{GD}} - ab^\top$.

Step 4 (two mean-square bounds). For independent $a, b \sim \mathcal{N}(0, I_d)$ and Z independent of them, $\mathbb{E}[(a^\top Z b)^2 \mid Z] = \|Z\|_F^2$; taking $Z = \mathrm{Orth}(W^{\mathrm{GD}} - ab^\top)$, whose squared Frobenius norm is its rank, at most d, and averaging over i,

$$\Big(\operatorname*{mean}_i \big(a^\top \mathrm{Orth}(W^{\mathrm{GD}} - ab^\top) b\big)^2\Big)^{1/2} = O_p(\sqrt{d}). \tag{87}$$

Substituting $u_i = \sqrt{1 - \frac{1}{N}}\, a + \frac{1}{N} U\mathbf{1}$ and $v_i = \sqrt{1 - \frac{1}{N}}\, b + \frac{1}{N} V\mathbf{1}$,

$$\begin{aligned} u_i^\top W^{\mathrm{Muon}} v_i - \Big(1 - \frac{1}{N}\Big) a^\top W^{\mathrm{Muon}} b = \sqrt{1 - \frac{1}{N}}\, a^\top W^{\mathrm{Muon}} \frac{V\mathbf{1}}{N} + \sqrt{1 - \frac{1}{N}}\, \frac{(U\mathbf{1})^\top}{N} W^{\mathrm{Muon}} b \\ + \frac{(U\mathbf{1})^\top}{N} W^{\mathrm{Muon}} \frac{V\mathbf{1}}{N}, \end{aligned} \tag{88}$$

so by $(x+y+z)^2 \le 3(x^2+y^2+z^2)$, $\|W^{\mathrm{Muon}}\|_{\mathrm{op}} \le 1$, $\mathbb{E}\|a\|^2 = \mathbb{E}\|b\|^2 = d$ (so that $\operatorname{mean}_i \|a\|^2 = O_p(d)$ by Markov) and Lemma E.3(iii),

$$\begin{aligned} &\operatorname*{mean}_i \Big(u_i^\top W^{\mathrm{Muon}} v_i - \big(1 - \frac{1}{N}\big) a^\top W^{\mathrm{Muon}} b\Big)^2 \\ &\qquad \le 3\Big[\operatorname*{mean}_i \|a\|^2 \frac{\|V\mathbf{1}\|^2}{N^2} + \frac{\|U\mathbf{1}\|^2}{N^2} \operatorname*{mean}_i \|b\|^2 + \frac{\|U\mathbf{1}\|^2 \|V\mathbf{1}\|^2}{N^4}\Big] = O_p\Big(\frac{d^2}{N}\Big), \end{aligned} \tag{89}$$

that is

$$\Big(\operatorname*{mean}_i \Big(u_i^\top W^{\mathrm{Muon}} v_i - \big(1 - \frac{1}{N}\big) a^\top W^{\mathrm{Muon}} b\Big)^2\Big)^{1/2} = O_p\big(d/\sqrt{N}\big). \tag{90}$$

Step 5 (combining). By Lemma E.28 and $1 - \frac{1}{N} > 0$, $u_i^\top W^{\mathrm{Muon}} v_i \ge (1 - \frac{1}{N}) a^\top \mathrm{Orth}(W^{\mathrm{GD}} - ab^\top) b - |u_i^\top W^{\mathrm{Muon}} v_i - (1 - \frac{1}{N}) a^\top W^{\mathrm{Muon}} b|$. Using Equation (84) of Step 2 with this lower bound, averaging over i, and applying Equation (87), Equation (90) and $\|W^{\mathrm{GD}}\|_* \le 7d\sqrt{Nd}$ from Lemma E.4(iii),

$$\operatorname*{mean}_i \big|u_i^\top W^{\mathrm{Muon}} v_i\big| \le \frac{\|W^{\mathrm{GD}}\|_*}{N} + O_p(d/N) + O_p(\sqrt{d}) + O_p(d/\sqrt{N}) \le \frac{7d^{3/2}}{\sqrt{N}} + O_p(\sqrt{d}). \tag{91}$$

Substituting into Equation (83) and dividing by $(N-1)d$,

$$\begin{aligned} \frac{1}{(N-1)d} \operatorname*{mean}_i \big(u_i^\top W^{\mathrm{Muon}} v_i\big)^2 &\le \frac{O_p(d)}{(N-1)d} + \frac{2}{N-1}\Big(\frac{7d^{3/2}}{\sqrt{N}} + O_p(\sqrt{d})\Big) \\ &= 14\gamma^{-3/2}\Big(1 + \frac{1}{N-1}\Big) + o_p(1), \end{aligned} \tag{92}$$

which is at most $15\gamma^{-3/2}$ with probability tending to one. □

Lemma E.30 (Second moment for Muon). *If $\gamma \ge 16$, then $\frac{1}{N(N-1)d} \sum_{i=1}^N \sum_{j \ne i} \big(u_j^\top W^{\mathrm{Muon}} v_i\big)^2 = 1 + O(\gamma^{-3/2}) + o_p(1)$, where with probability tending to one the $O(\gamma^{-3/2})$ term has absolute value at most $52\gamma^{-3/2}$.*

Proof. Divide Lemma E.23 by d and subtract 1. By Lemma E.24 the norm term contributes $O(N^{-1}) + O_p((Nd)^{-1/2}) = o_p(1)$; by Lemma E.27 the trace term is at most $145\gamma^{-2} \le \frac{145}{4}\gamma^{-3/2}$ in absolute value for $\gamma \ge 16$; by Lemma E.29 the diagonal term lies in $[-15\gamma^{-3/2}, 0]$. Adding, $\frac{145}{4} + 15 \le 52$. □

Lemma E.31 ($\bar\kappa$ for Muon). *If $\gamma \ge 16$, then $(\bar\kappa^{\mathrm{Muon}})^2 = d\big(1 + O(\gamma^{-3/2}) + o_p(1)\big)$ with the same constant as in Lemma E.30.*

Proof. By Equation (51) with $W = W^{\mathrm{Muon}}$, Lemma E.30 and Lemma E.21, $(\bar\kappa^{\mathrm{Muon}})^2 = d(1 + O(\gamma^{-3/2}) + o_p(1)) - O_p(d/N) = d(1 + O(\gamma^{-3/2}) + o_p(1))$. □

Lemma E.32 (ρ for Muon). *If $\gamma \geq 16$, then*

$$\rho^{\mathrm{Muon}} = \sqrt{\frac{d}{\gamma}}\Big[\frac{8}{3\pi}\Big(1 - \frac{1}{10\gamma}\Big) + O(\gamma^{-3/2}) + o_p(1)\Big]. \tag{93}$$

Unlike for GD, the data budget enters ρ^{Muon} a second time, through $\Phi(\gamma)$.

Proof. By Lemma E.22, $\bar{h}^{\mathrm{Muon}}/\sqrt{d} = \sqrt{d/\gamma}\,\Phi(\gamma)(1 + o_p(1))$. Set $y := (\bar{\kappa}^{\mathrm{Muon}})^2/d$. By Lemma E.31 and $\gamma \geq 16$, with probability tending to one

$$|y - 1| \leq 52\gamma^{-3/2} + o_p(1) \leq \frac{52}{64} + o_p(1) \leq \frac{9}{10},$$

so $y \geq \frac{1}{10}$. For $y \geq \frac{1}{10}$,

$$|y^{-1/2} - 1| = \frac{|1 - y|}{\sqrt{y}\,(1 + \sqrt{y})} \leq \sqrt{10}\,|y - 1|,$$

which gives $\sqrt{d}/\bar{\kappa}^{\mathrm{Muon}} = 1 + O(\gamma^{-3/2}) + o_p(1)$. Multiplying the two,

$$\rho^{\mathrm{Muon}} = \frac{\bar{h}^{\mathrm{Muon}}}{\sqrt{d}} \cdot \frac{\sqrt{d}}{\bar{\kappa}^{\mathrm{Muon}}} = \sqrt{\frac{d}{\gamma}}\,\Phi(\gamma)\big(1 + O(\gamma^{-3/2}) + o_p(1)\big), \tag{94}$$

and Lemma E.4(iv) gives $\Phi(\gamma) = \frac{8}{3\pi}(1 - \frac{1}{10\gamma} + r(\gamma))$ with $|r(\gamma)| \leq \gamma^{-2} \leq \frac{1}{4}\gamma^{-3/2}$, so $0 < \Phi(\gamma) \leq \frac{8}{3\pi}(1 + \gamma^{-2}) < 1$ and $\Phi(\gamma)(1 + O(\gamma^{-3/2}) + o_p(1)) = \frac{8}{3\pi}(1 - \frac{1}{10\gamma}) + O(\gamma^{-3/2}) + o_p(1)$.

Two by-products of this computation are used in the proof of Theorem 4.2. First, the constant implied by the $O(\gamma^{-3/2})$ in the displayed multiplicative form is absolute, so there is an absolute $\gamma_1 \geq 16$ such that for $\gamma \geq \gamma_1$ that factor lies in $[\frac{1}{2}, \frac{3}{2}]$ with probability tending to one. Second, for every $\gamma \geq 16$ the same display bounds ρ^{Muon} away from zero without any expansion: $y \leq \frac{19}{10}$ gives $\sqrt{d}/\bar{\kappa}^{\mathrm{Muon}} \geq (19/10)^{-1/2} > \frac{7}{10}$, and $\Phi(\gamma) \geq \frac{1}{8}$ by Lemma E.4(iii),(iv), so with probability tending to one

$$\rho^{\mathrm{Muon}} \;\geq\; \frac{7}{80}\sqrt{d/\gamma}\,(1 - o_p(1)) \;\geq\; \frac{1}{16}\sqrt{d/\gamma}. \tag{95}$$

□

E.5 Combining the GD and Muon estimates

Proof of Theorem 4.2. Reduction to large γ. By Lemma E.16 and $2 + \gamma \geq \gamma$, $\rho^{\mathrm{GD}} \leq \sqrt{d/\gamma}\,(1 + o_p(1))$, so Equation (95) gives $\mathcal{R} \leq 17$ with probability tending to one, for every $\gamma \geq 16$; together with $0 \leq \frac{3\pi}{8}(1 - \frac{9}{10\gamma}) \leq \frac{3\pi}{8}$ this makes the left-hand side of Equation (9) at most 19 there. On a bounded range $16 \leq \gamma \leq \gamma_1$ we have $19 \leq 19\gamma_1^{3/2}\gamma^{-3/2}$, so Equation (9) holds on that range as soon as $C_1 \geq 19\gamma_1^{3/2}$. It therefore suffices to prove it for $\gamma \geq \gamma_1$, with γ_1 the absolute constant of Lemma E.32; on that range the error factor of Lemma E.32 lies in $[\frac{1}{2}, \frac{3}{2}]$ with probability tending to one, so its reciprocal is again $1 + O(\gamma^{-3/2}) + o_p(1)$.

Expansion. By Lemma E.16 and Lemma E.32, and since $\Phi(\gamma) > 0$,

$$\begin{aligned}\mathcal{R} = \frac{\rho^{\mathrm{GD}}}{\rho^{\mathrm{Muon}}} &= \sqrt{\frac{d}{2+\gamma}}\big(1 + o_p(1)\big) \cdot \sqrt{\frac{\gamma}{d}}\,\frac{1}{\Phi(\gamma)\big(1 + O(\gamma^{-3/2}) + o_p(1)\big)} \\ &= \frac{\sqrt{\gamma/(2+\gamma)}}{\Phi(\gamma)}\big(1 + O(\gamma^{-3/2}) + o_p(1)\big).\end{aligned} \tag{96}$$

The deterministic prefactor is expanded in $1/\gamma$: $\sqrt{\gamma/(2+\gamma)} = (1 + 2/\gamma)^{-1/2} = 1 - \gamma^{-1} + O(\gamma^{-2})$, while Lemma E.4(iv) gives $\Phi(\gamma)^{-1} = \frac{3\pi}{8}\big(1 + \frac{1}{10\gamma} + O(\gamma^{-2})\big)$, so

$$\frac{\sqrt{\gamma/(2+\gamma)}}{\Phi(\gamma)} = \frac{3\pi}{8}\Big(1 - \frac{1}{\gamma} + O(\gamma^{-2})\Big)\Big(1 + \frac{1}{10\gamma} + O(\gamma^{-2})\Big) = \frac{3\pi}{8}\Big(1 - \frac{9}{10\gamma}\Big) + O(\gamma^{-2}). \tag{97}$$

For $\gamma \geq 16$ the prefactor is bounded, $0 < \sqrt{\gamma/(2+\gamma)}/\Phi(\gamma) \leq \frac{3\pi}{8}(1-\frac{1}{160}-\frac{1}{256})^{-1}$, so multiplying by $1+O(\gamma^{-3/2})+o_p(1)$ leaves the error of the same order; and $\gamma^{-2} \leq \frac{1}{4}\gamma^{-3/2}$ in that range absorbs the $O(\gamma^{-2})$ term. Hence

$$\mathcal{R} = \frac{3\pi}{8}\Big(1 - \frac{9}{10\gamma}\Big) + O(\gamma^{-3/2}) + o_p(1), \tag{98}$$

where, with probability tending to one, the $O(\gamma^{-3/2})$ term has absolute value at most $C_1\gamma^{-3/2}$ with C_1 independent of γ, d, N. This is Equation (9). □

E.6 Monotonicity in the data budget

The expansion Equation (9) has two consequences for how $\mathcal{R}$ changes with γ. The first compares a fixed pair of data budgets; the second concerns the limit $d \to \infty$.

Lemma E.33 (Fixed ratio)**.** *Let $\lambda > 0$ and*

$$\gamma > \gamma_0(\lambda) := \max\Big\{16, \Big(\frac{160C_1(1+\lambda)}{27\pi\lambda}\Big)^2\Big\}.$$

For the same d, let $N_1 = \gamma d$ and $N_2 = (1+\lambda)\gamma d$, and let $\mathcal{R}_{N_1,d}$ and $\mathcal{R}_{N_2,d}$ be the corresponding values of $\mathcal{R}$. Then, with probability tending to one, $\mathcal{R}_{N_2,d} > \mathcal{R}_{N_1,d}$.

Proof. Apply Equation (9) once at γ and once at $(1+\lambda)\gamma$, both of which are at least 16: for every $\varepsilon > 0$, with probability tending to one,

$$\begin{aligned}\mathcal{R}_{N_2,d} - \mathcal{R}_{N_1,d} &\geq \frac{3\pi}{8}\Big(1 - \frac{9}{10(1+\lambda)\gamma}\Big) - \frac{3\pi}{8}\Big(1 - \frac{9}{10\gamma}\Big) \\ &\quad - C_1\big((1+\lambda)\gamma\big)^{-3/2} - C_1\gamma^{-3/2} - 2\varepsilon \\ &\geq \frac{27\pi}{80}\cdot\frac{\lambda}{(1+\lambda)\gamma} - 2C_1\gamma^{-3/2} - 2\varepsilon.\end{aligned} \tag{99}$$

Moreover $\frac{27\pi}{80}\cdot\frac{\lambda}{(1+\lambda)\gamma} > 2C_1\gamma^{-3/2}$ is equivalent to $\gamma^{1/2} > \frac{160C_1(1+\lambda)}{27\pi\lambda}$, which holds by assumption; taking $\varepsilon = \frac{1}{4}\big(\frac{27\pi}{80}\cdot\frac{\lambda}{(1+\lambda)\gamma} - 2C_1\gamma^{-3/2}\big)$ gives the claim. □

Lemma E.34 (Pointwise monotonicity)**.** *Suppose that for every sufficiently large γ the limit in probability $\mathcal{R}_\infty(\gamma) := \lim_{d\to\infty}\mathcal{R}$ exists, and that $\mathcal{R}_\infty$ is analytic as a function of $1/\gamma$ in a neighbourhood of 0, i.e. $\mathcal{R}_\infty(\gamma) = \sum_{k\geq 0} c_k\gamma^{-k}$ for all sufficiently large γ. Then there is γ_2 such that $\mathcal{R}'_\infty(\gamma) > 0$ for $\gamma > \gamma_2$.*

Proof. By Equation (9) and convergence in probability, $|\mathcal{R}_\infty(\gamma) - \frac{3\pi}{8}(1-\frac{9}{10\gamma})| \leq C_1\gamma^{-3/2}$, that is

$$\Big|\sum_{k\geq 0} c_k\gamma^{-k} - \frac{3\pi}{8} + \frac{27\pi}{80}\gamma^{-1}\Big| \leq C_1\gamma^{-3/2}. \tag{100}$$

Letting $\gamma \to \infty$ gives $c_0 = \frac{3\pi}{8}$; dividing both sides by γ^{-1} and letting $\gamma \to \infty$ gives $c_1 = -\frac{27\pi}{80}$. Hence

$$\mathcal{R}'_\infty(\gamma) = -\sum_{k\geq 1} k\,c_k\gamma^{-k-1} = \gamma^{-2}\Big(\frac{27\pi}{80} - \sum_{k\geq 2} k\,c_k\gamma^{-(k-1)}\Big), \tag{101}$$

and the series in the parentheses tends to 0 as $\gamma \to \infty$, so the parenthesized quantity is positive for all sufficiently large γ. □

Remark E.35. Neither hypothesis of Lemma E.34 is proved here. The estimates above pin $(\bar{\kappa}^{\mathrm{Muon}})^2/d$ down only to within $O(\gamma^{-3/2})$ (Lemma E.31), so they do not by themselves give the limit in probability; $\rho^{\mathrm{GD}}/\sqrt{d}$ and $\bar{h}^{\mathrm{Muon}}/d$, by contrast, do converge, by Lemma E.16 and Lemma E.22. Analyticity does not follow from Equation (9) alone either: with $x = 1/\gamma$, the function $1 - x + x^{3/2}\sin(1/x)$ admits an expansion of the same form and is not monotone near $x = 0$.

Numerical consequences. Since $\frac{3\pi}{8} \approx 1.178 > 1$, Equation (9) puts $\mathcal{R}$ above one for all large γ: past some data budget the GD iterate has the larger margin-to-noise ratio. The leading term $\frac{3\pi}{8}(1 - \frac{9}{10\gamma})$ is increasing in γ; Lemma E.33 turns this into a statement about $\mathcal{R}$ itself, and Lemma E.34 into one about its limit, under the hypotheses discussed in Remark E.35. The leading term crosses one at $\gamma \approx 5.95$, below the range $\gamma \geq 16$ in which Lemma E.32 is proved, so the location of the crossing is established empirically rather than by Theorem 4.2.