Eigenbases Without Eigenvalues: Exact Amplification Laws for Spectral Preconditioning on Rank-Deficient Gradient Flows

Anonymous authors
Paper under double-blind review

Abstract

Second-order and orthogonal-preconditioned optimizers (Shampoo, SOAP, Muon) deliver strong speedups at pretraining scale yet fail catastrophically in low-rank (LoRA) fine-tuning. **This paper's contribution is the exact mechanism of that failure—two amplification laws—not an optimizer.** Across five model families (GPT-2, Llama-3.2, Qwen3, Phi-4-mini, Mistral-7B; 124M–7B) and two data domains, the covariances that spectral preconditioners invert are severely rank-deficient: median effective rank 1.1–5.5 against dimensions 768–4096. We prove two exact laws: with a constant eigenvalue floor ϵ the amplification equals $\epsilon^{-2\alpha}$ exactly—independent of learning rate, data, and spectrum—and with a relative ridge $\gamma\lambda_{\max}$ it scales as $(\gamma\lambda_{\max})^{-2\alpha}$, forcing the ridge to $\Theta(1)$: no interior sweet spot exists. A pre-registered program of 575 runs verifies the laws to seven significant digits and shows free inversion diverges at every learning rate in LoRA and full-parameter training, with the divergence horizon exceeding 30-step calibration probes. The theory identifies the eigenbasis as the only stable channel of spectral information and predicts its own repair: discarding eigenvalue magnitudes while keeping the eigenbasis. The resulting rotation-only variant (SOAP-X)—Muon's principle at pretraining scale, reached here as the theory's prediction for deficient flows—matches or beats AdamW, Muon, and capped-inversion baselines in 6 of 8 configurations, including full-parameter training and a 5000-step horizon; on MMLU-5-shot the diverged optimizer's checkpoint collapses to random-level accuracy (24.89–25.28%) while every stable optimizer finishes 32–43 points above it. A 7B boundary check confirms the cold-start law is dimension-free and finds SOAP-X itself diverging mid-training—a pre-registered miss that three diagnosis blocks then trace to fp32 GPU eigendecomposition returning non-orthogonal eigenvectors on these near-singular covariances; one QR re-orthogonalization per refresh repairs it, verified across seeds and a $5\times$ horizon at 7B. The laws apply to rank-deficient gradient flows (fine-tuning), not pretraining, where covariances fill in and spectral methods succeed.

1 Introduction

Preconditioned optimizers that exploit the geometry of the gradient covariance—Shampoo (Gupta et al., 2018), its efficient eigenbasis variant SOAP (Vyas et al., 2025), and orthogonalized momentum (Muon) (Jordan et al., 2024)—deliver some of the strongest reported training speedups at pretraining scale. The natural expectation that the same machinery accelerates fine-tuning fails: in low-rank (LoRA (Hu et al., 2022)) fine-tuning of Llama-3.2-3B (Grattafiori et al., 2024) and GPT-2 (Radford et al., 2019), free spectral inversion diverges at every learning rate we can calibrate, while the same code, models, and data train stably under AdamW (Loshchilov & Hutter, 2019) and Muon.

This paper explains the failure exactly, proves when repair is possible, and derives the design that the analysis implies. **The primary contribution is the theory, not an optimizer: two exact, parameter-free amplification laws that turn a widely observed engineering instability into a closed-form statement.** Our contributions, in that order:

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.

- **Theory (primary contribution).** The null-space amplification law: with eigenvalue floor ϵ and per-side exponent α, the update amplification equals $\epsilon^{-2\alpha}$ exactly—learning-rate-, data-, and spectrum-independent (Theorem 2)—and the ridge law $(\gamma\lambda_{\max})^{-2\alpha}$, whose stability boundary forces $\gamma = \Theta(1)$: no interior sweet spot exists (Theorem 3). The laws identify rank deficiency, not implementation detail, as the cause of spectral-preconditioner divergence.
- **Measurement.** The covariances that spectral preconditioners invert are severely deficient in fine-tuning: LoRA's large-side covariance has per-step rank $\leq r$ (measured effective rank 1.0–71.9 of 4096) and in full-parameter training the covariance tail sits at or below the numerical floor—architecture-, scale-, and domain-invariantly across a five-family, two-domain census (Theorem 1; Section 5.1).
- **Pre-registered validation.** 575 runs confirm the laws to seven significant digits; every miss concerns the estimator, evaluation protocols, or empirical thresholds—none the mechanism (Appendix C). Free inversion diverges at every learning rate in LoRA and full-parameter training, at every adapter rank 1–128, and under tighter clipping; the divergence horizon exceeds 30-step calibration probes (Corollary 1); and the diverged optimizer erases the model to random-level accuracy on three downstream tasks.
- **Design consequence (not a new optimizer; the rotation principle is Muon's).** The theory separates the eigenbasis (direction channel) from eigenvalue magnitudes (amplitude channel): only the latter is unstable. Rotation-only preconditioning (SOAP-X) is the stable member that the theory singles out—one line of configuration change to SOAP, proposed as the analysis's prediction rather than as a method. As an update principle, discarding magnitudes is Muon's (Jordan et al., 2024) and we claim no novelty for it; what this paper adds is provenance and decomposition: a stability law selects the rotation for the eigenvalue-inverting family (Corollary 2), a factorial ablation shows the two parents are not interchangeable—their behavior is a property of the basis×pipeline combination, not of either component (Section 5.8)—and a pre-registered 7B check (Section 5.9) first marks the repair's own limit, then traces it to a floating-point eigensolver failure and repairs it with one QR re-orthogonalization per refresh, with Muon and MuonEIG stable alongside.

Two further findings sharpen the practical picture: capped inversion prefers rates $4\times$ hotter, so ignoring the *learning-rate × inversion-depth interaction* produces qualitatively wrong rankings (Section 5.4), and the *calibration-horizon trap* means the optimizer that tops the 300-step comparison loses at 5000 steps (Section 5.5). The account is falsifiable: an optimizer that inverts eigenvalues stably at fine-tuning scale would refute it.

2 Related work

Preconditioned optimizers. Full-matrix and Kronecker-factored preconditioners descend from natural-gradient methods: K-FAC (Martens & Grosse, 2015), Shampoo (Gupta et al., 2018), and scalable variants (Anil et al., 2020), with diagonal-Hessian approximations (Liu et al., 2024) on the side. SOAP (Vyas et al., 2025) runs Adam (Kingma & Ba, 2015) in the covariance eigenbasis; Muon (Jordan et al., 2024) orthogonalizes the momentum directly. Muon's update is steepest descent under a matrix operator norm (Liu et al., 2025), the modular-norm program's (Bernstein & Newhouse, 2024) optimality account of the update direction—not a stability account of covariance inversion; it makes no prediction for the rank-deficient flows measured here. Corollary 2 is the complementary stability half, and Section 5.8 shows the two optimizers are empirically not interchangeable. Across this line the treatment of small eigenvalues has been an implementation detail; Theorem 2 shows it is the decisive stability variable in deficient regimes.

Adaptive methods and low-rank fine-tuning. Convergence analyses of adaptive methods typically assume bounded gradients and well-conditioned objectives (Reddi et al., 2018). RMSprop (Tieleman & Hinton, 2012) and Adam avoid covariance inversion entirely—precisely why they are immune to the amplification documented here; our contribution is orthogonal: we measure the covariance directly in fine-tuning, find it structurally rank-deficient, and trace the failure through the preconditioner algebra. On the fine-tuning side, LoRA (Hu et al., 2022) freezes W and learns

Table 1: **Deficiency terminology.** The two floors are the mechanisms Theorems 2–3 act on.

Term	Definition
rank	nonzero eigenvalues of a covariance L or R; per step $\leq r$ for a LoRA factor (Theorem 1)
effective rank, ratio	participation ratio $r_{\text{eff}} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$; the ratio $r_{\text{eff}}/\text{dim}$ of the large side, at most 0.05 measured (Section 5.1)
numerically rank-deficient	a non-negligible set of directions has $\lambda_i \leq \epsilon$, exact zeros included—the precise reading of A1's "deficient spectrum"
floors ϵ, $\gamma\lambda_{\max}$	constant guard $\text{fl}(D) = \max(D, \epsilon)$ with $\epsilon = 10^{-12}$; relative ridge $\text{fl}(D) = \max(D, \gamma\lambda_{\max})$ (Theorem 3)
EMA window W	$(1-\beta)^{-1}$: effective number of gradient outer products in the covariance estimate
amplification A	update RMS at exponent α relative to the same quantity at $\alpha = 0$, all else identical

$\Delta W = BA$ with B initialized to zero; the zero initialization that makes cold-start training stable also makes the first covariance refresh exactly singular (Theorem 1).

3 Setting and notation

Optimizers. We study the Shampoo/SOAP family in a common implementation. Let G be the gradient matrix of a weight (or LoRA factor) and $L = \text{EMA}_\beta(G^\top G)$, $R = \text{EMA}_\beta(GG^\top)$ the left/right gradient covariances with EMA coefficient $\beta = 0.95$ unless stated. SOAP computes eigendecompositions $L = Q\Lambda Q^\top$, $R = P\Delta P^\top$ and updates in the rotated basis; the variants differ only in how the eigenvalues are used: *SOAP-F* (free inversion) applies $\Lambda^{-\alpha}$, $\Delta^{-\alpha}$ per side with a numerical floor ϵ; *SOAP-AC*(α) caps the inverted spectrum at $\epsilon^{-\alpha}$; *SOAP-X* uses the rotation Q, P only and normalizes the update RMS to that of AdamW—eigenbases without eigenvalues. Baselines are AdamW (Loshchilov & Hutter, 2019) and Muon (Jordan et al., 2024) (Newton–Schulz orthogonalization). The per-side exponent $\alpha = 1/4$ unless stated (the Shampoo-consistent choice); $\alpha = 1/2$ is full matrix inversion.

Tasks and protocol. Fine-tuning is next-token prediction: WikiText-103 (Merity et al., 2017) (natural text) and CodeSearchNet Python (Husain et al., 2019) (code), 512-token chunks, effective batch 16×512 tokens. Models: GPT-2 124M (Radford et al., 2019) (LoRA r=16 and full-parameter), Llama-3.2-3B (Grattafiori et al., 2024) (r=1, r=16), Qwen3-1.7B (Yang et al., 2025) and Phi-4-mini 3.8B (Abouelenin et al., 2025) (r=16). Unless stated, runs are 300 steps, three seeds (42/123/456) per cell, mean $\pm$ sd; comparisons use each optimizer's best rate from a $\{1.25\times10^{-4}, 5\times10^{-4}, 2\times10^{-3}\}$ grid where stated. Gradient clipping (global norm 1.0) applies in every run, including every diverging one; divergence is loss $> 10\times$ baseline or non-finite; all remaining details in Appendix D.12. Table 1 fixes the deficiency terminology—rank notions, the two floors, and the amplification ratio—that the theory and measurement sections use.

4 Theory: two amplification laws

This section states the two laws with their assumptions in the open; full statements and proofs are in Appendix A, amplification A as defined in Table 1. The theorems characterize rank-deficient gradient flows, not pretraining (Applicability boundary below); every theorem is conditioned on A1–A3.

Assumptions (A1–A3), shared by Theorems 1–3. (A1) Rank-deficient gradient flow. The trained matrix receives gradients whose covariance is (structurally or numerically) rank-deficient: in LoRA, one factor is zero-initialized, so the large-side covariance has per-step rank $\leq r$; in full-parameter fine-tuning, the covariance tail sits at or below the numerical floor. **(A2) EMA covariance estimation.** L and R are estimated by exponential moving averages with coefficient β and refreshed every p steps. **(A3) Per-side spectral inversion.** The preconditioner applies $\text{fl}(D)^{-\alpha}$ per side with a constant floor ϵ (Theorem 2) or a relative ridge $\gamma\lambda_{\max}$ (Theorem 3).

Theorem 1 (Structural deficiency; under A1–A2)**.** *In LoRA with $B_0 = 0$, the large-side covariance $R = \mathrm{EMA}(GG^\top)$ has per-step rank $\leq r$ and effective rank $\leq \min(rW, \dim)$ after W EMA steps, with equality up to the participation ratio; in full-parameter training the covariance tail is bounded by the numerical floor within $O(100)$ steps on all measured architectures. The measured census (Section 5.1) matches the bound: rank tracks the EMA window at $r = 1$ (maximum $18.9 \approx W = 20$).*

Theorem 2 (Null-space amplification law; under A1–A3)**.** *If the inverted spectra are floored at a constant ϵ (the standard numerical guard), then at the first non-zero gradient after a covariance refresh, the amplification is exactly $A = \epsilon^{-2\alpha}$, independent of learning rate, data distribution, and the non-null part of the spectrum (mechanism in Appendix A: the zero-initialized LoRA factor makes one side's first covariance refresh the zero matrix; the floor lifts every eigenvalue to ϵ, and the first non-zero update is dominated by the $\epsilon^{-\alpha}$ scale on each side, compounding across sides).*

Theorem 3 (Ridge absolute-scaling law and absence of an interior optimum; under A1–A3)**.** *With a relative ridge floor $\gamma\lambda_{\max}$ in place of ϵ, the amplification scales as $A \approx (\gamma\lambda_{\max})^{-\alpha} \times (\text{opposite-side scale})$. Stability requires the update scale to remain $O(1)$, forcing $\gamma \gtrsim (\eta c/\|W\|)^{1/\alpha}/\lambda_{\max} = \Theta(1)$, at which point the preconditioner is a uniform rescaling and conveys no spectral information. Stability and preconditioning are mutually exclusive on deficient spectra.*

Corollary 1 (Calibration-horizon trap)**.** *The amplification manifests after the first covariance refresh (step 2–10), but the loss may remain finite for $O(100)$ steps: a 30-step stability probe can therefore certify a configuration that diverges within 300 steps. Sections 5.4 and 5.5 measure the consequence: probe-chosen rates are 4–16$\times$ too hot and invert the final ranking at long horizons.*

Corollary 2 (Rotation safety)**.** *Preconditioners that use the eigenvectors Q, P but discard eigenvalue magnitudes (SOAP-X; Muon is the matrix analog) have $A \equiv 1$ by construction: no eigenvalue channel exists to amplify, and the rotation itself cannot destabilize the update scale; the claim needs only the covariance estimate to be PSD.*

Scope of Corollary 2. $A \equiv 1$ removes the eigenvalue amplification channel—the only channel Theorems 2–3 govern—not every way a large-scale run can fail; Section 5.9 measures a 7B failure of this kind and traces it to a floating-point eigensolver defect, leaving the corollary's scope intact.

Applicability boundary (read before generalizing). Theorems 2–3 quantify what happens when the floor binds—i.e. under rank deficiency (A1); at pretraining scale the covariance is effectively full-rank, the floor never binds, and nothing here argues against spectral preconditioning in that regime. The laws are triggered by *deficiency*, not by the LoRA parametrization as such.

5 Pre-registered experimental program

The program spans 575 runs: blocks V13–V22 established the laws and the main landscape; blocks V23–V46 stress the conclusions along thirteen axes—EMA window, architecture family and scale, data domain, wall-clock cost, horizon, downstream transfer, boundary stress tests, the SOAP-X/Muon ablation, a 7B check (Section 5.9), and instruction tuning (Appendix D.9). Two completion blocks (V44b, V45b) re-ran the memory-overflow cells with host-RAM optimizer state, update rules line-identical, so no cell remains unmeasured. A naming map (Appendix B, Table 3) fixes the correspondence among method variants, experiment blocks, and prediction IDs. Every prediction was a numbered falsifiable statement with a fixed threshold, archived unmodified before its runs launched (Nosek et al., 2018). Only the 14 *law predictions*—exact claims about $\epsilon^{-2\alpha}$ and $(\gamma\lambda_{\max})^{-2\alpha}$, where a single miss would have refuted the theory—carry theoretical weight; all 14 hit to seven significant digits. The 92 empirical threshold and attribution predictions missed 21 times (two partials), always on estimators, protocols, or thresholds, never the mechanism (83/106 overall; Appendix C, Table 4).

5.1 The covariance census

Figure 1(c) tabulates the participation ratio r_{eff} (Table 1) of instrumented covariances at step 30, across every tested condition. In LoRA fine-tuning the median effective rank is 1.13 (Llama r=16), 2.13 (Llama r=1, range 1.0–18.93, maximum tracking the EMA window $W \approx 20$), 4.71 (Qwen3-1.7B), and 2.09 (Phi-4-mini) against dimensions up to 3072; the 7B boundary check (V45) extends the census to Mistral-7B-v0.3, where the median is 4.65 against dimension 4096 (Appendix D.7). In full-parameter GPT-2 training the median is 5.48 (11 of 12 covariances below effective rank 100,

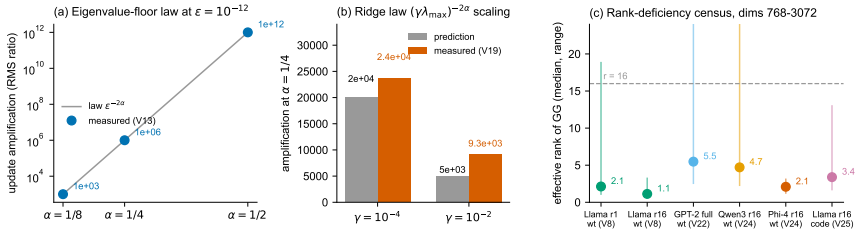

Figure 1: **The two amplification laws, measured, and the covariance deficiency they act on.** (a) Eigenvalue-floor law (Theorem 2) at $\epsilon = 10^{-12}$: measured amplification matches $\epsilon^{-2\alpha}$ to float32 precision at $\alpha = 1/8, 1/4$; the $\alpha = 1/2$ free-inversion point saturates at 8.0×10^9 (diverges before realizing 10^{12}). (b) Ridge law (Theorem 3) at $\alpha = 1/4$: measured amplification $(\gamma\lambda_{\max})^{-2\alpha}$ against the measured ridge $\gamma\lambda_{\max}$, no free parameters. (c) Rank-deficiency census: effective rank of instrumented L/R covariances at step 30 across every tested condition—medians 1.1–5.5 against dimensions 768–3072 (V45 extends to dimension 4096); 50% of LoRA-side covariances carry $\lambda_{\min} = 0.0$ exactly in float32.

Table 2: **Main landscape: validation perplexity (mean ± sd) under the best-rate protocol** (300 steps, 3 seeds; SOAP-F row shows final perplexity after divergence; row minimum in bold). Code domain: AdamW at the probe-chosen 2×10^{-3} (*: seed-unstable). Qwen3/Phi-4-mini rows from the fixed-rate control grid of Section 5.4; †: that optimizer's best rate on that model is 5×10^{-4}. wt = WikiText-103, code = CodeSearchNet.

Configuration (300 steps, 3 seeds)	AdamW	Muon	SOAP-X	SOAP-AC$_{0.25}$
Llama-3.2-3B r=1 (wt)	6.97 ± 0.01	**6.87** ± 0.01	6.90 ± 0.02	7.76 ± 0.11
Llama-3.2-3B r=16 (wt)	10.95 ± 0.23	8.90 ± 0.06	**6.98** ± 0.01	7.26 ± 0.08
GPT-2 124M r=16 (wt)	19.39 ± 0.02	19.37 ± 0.06	**19.26** ± 0.04	27.50 ± 0.22
GPT-2 124M full (wt)	18.08 ± 0.09	17.58 ± 0.04	**17.46** ± 0.04	17.54 ± 0.06
GPT-2 full, SOAP-F	$1.8 \times 10^7 \pm 1.4 \times 10^7$ (diverged)			
Llama r=16 (code)	$2046 \pm 936^*$	6.80 ± 0.12	5.75 ± 0.03	**5.17** ± 0.01
Qwen3-1.7B r=16 (wt)	10.15 ± 0.05	9.18 ± 0.01	**8.97** ± 0.01	$9.20 \pm 0.02^\dagger$
Phi-4-mini r=16 (wt)	8.53 ± 0.03	**8.05** ± 0.01	$10.07 \pm 0.07^\dagger$	$9.85 \pm 0.29^\dagger$

maximum 229.14); on the code domain, 3.39. Half of all LoRA-side covariances have $\lambda_{\min}$ exactly 0.0 in float32. The object that spectral preconditioners invert is overwhelmingly null space or numerical tail, invariantly across architecture, scale, and domain.

5.2 The amplification laws, measured

Figure 1(a) verifies Theorem 2 at $\epsilon = 10^{-12}$: measured amplifications are 1.0000001×10^3, 1.0000002×10^6, and 1.0000001×10^{12} at $\alpha = 1/8, 1/4, 1/2$—relative error $\leq 1.9 \times 10^{-7}$, float32 machine precision on the measured quantity. The $\alpha = 1/2$ free-inversion point saturates at 8.0×10^9 because the run diverges before the full amplification is realized; the 10^{12} value is measured on the stable capped variant. Figure 1(b) verifies Theorem 3 at $\alpha = 1/4$: the measured amplification is 2.36×10^4 at $\gamma = 10^{-4}$ (prediction 2.0×10^4) and 9.3×10^3 at $\gamma = 10^{-2}$ (prediction 4.9×10^3). Both laws involve no free parameters; the ridge law's largest deviation ($1.9\times$, at the largest γ) occurs where the ridge floor interacts with the non-null spectrum, and the exponent—not the constant—is what the stability argument uses.

5.3 Main landscape, including the code domain

Three facts summarize Table 2. (i) *Free inversion never trains*: 42/42 SOAP-F runs diverge—four architectures, both domains, every EMA window, full-parameter training included—at every calibrated rate, usually within 4 steps of the first covariance refresh. (ii) *Rotation-only is best or statistically tied in 6 of 8 configurations* (the seven comparison rows plus the 5000-step horizon); its two losses are Phi-4-mini at 300 steps (a rate artifact compounded by extreme rank deficiency; Section 5.4) and the

code domain, where capped inversion wins at the probe rate. (iii) *The code domain is the strongest SOAP-family result*: deficiency is at least as severe there (median effective rank 3.39).

5.4 Robustness: EMA window, architecture family, and the learning-rate interaction

The covariance EMA window does not rescue free inversion. Experiment E1 (48 runs) sweeps $\beta \in \{0.90, 0.95, 0.99, 0.999\}$ on Llama $r{=}16/r{=}1$ with a β-parametrized pair: SOAP-XW (rotation-only) and SOAP-AW (free inversion). Free inversion diverges in 24 of 24 runs—every β, configuration, and seed, typically at step 3–4 with measured amplification 10^6—so the failure does not depend on how well the covariance is estimated. Rotation-only is stable at every β and flat across the entire $0.90 \to 0.999$ range (Figure 2, Appendix D).

Architecture family, scale, and the learning-rate $\times$ inversion-depth interaction. Experiment E2 (32 runs) extends the landscape to Qwen3-1.7B and Phi-4-mini at the probe-calibrated rate 2×10^{-3}. At that rate SOAP-X trailed AdamW on both models (66.94 vs 36.76 on Qwen3, 76.96 vs 52.67 on Phi-4-mini) while SOAP-AC led (24.39 and 9.86)—an apparent contradiction of our thesis. The control experiment E2-B (48 runs, predictions B1–B4) resolves this as a rate artifact with a structural cause, running the full $2 \times 4 \times 2$ grid (model $\times$ optimizer $\times$ rate; Table 10 and Figure 5 in Appendix D).

The grid exposes a clean interaction: *capped inversion prefers hot rates; every other optimizer prefers cold, and rotation-only degrades least.* On Qwen3, SOAP-AC improves from 13.52 to 9.20 when the rate rises $4\times$, while SOAP-X worsens from 8.97 to 18.26—a $2.0\times$ relative degradation, against $3.2\times$ for AdamW and $3.1\times$ for Muon; on Phi-4-mini SOAP-X is genuinely rate-flat (10.07 vs 10.33) while AdamW collapses from 8.53 to 22.70. The mechanism follows from Theorem 2: capping multiplies the deficient-subspace update by a bounded factor < 1—an implicit step-size reduction—so the optimal explicit rate shifts up. Consequences: at the matched rate 5×10^{-4}, SOAP-AC is best-or-tied on both models (B1, hit); SOAP-X beats AdamW on both, so the probe-rate ranking was an artifact (B2, a miss consistent with the theory); and the probe-chosen 2×10^{-3} is too hot for every optimizer except SOAP-AC (B4, miss)—the probe is biased toward the optimizer whose stability margin is largest.

At each optimizer's own best rate (Table 2), Phi-4-mini is the one WikiText configuration where rotation-only loses at 300 steps (the code domain is the other loss; Section 5.3): its census is the tightest measured—median effective rank 2.09 of 3072, maximum 3.2, every covariance pinned with no escaping tail (Section 6)—and Section 5.5 shows the deficit is a transient.

5.5 The 5000-step horizon

Experiment V27 (L1–L5, 12 runs) trains Qwen3-1.7B LoRA $r{=}16$ for 5000 steps at each optimizer's best 300-step rate; V40/V42 mirror the protocol on Phi-4-mini $r{=}16$, completing the grid with the rotation family at its recommended cold rate (Table 6, Figures 3, 4).

The calibration-horizon trap, measured end-to-end. SOAP-AC at 5×10^{-4}—the rate the 300-step grid chose for it, second only to SOAP-X at 300 steps on Qwen3 (9.20 vs 8.97)—finishes last at 5000 steps (7.96). SOAP-X, slowest at step 250, overtakes AdamW at step 2250 and holds first place to the end (7.79 ± 0.01 vs 7.82 ± 0.01). On Phi-4-mini the mirror holds: the hot rate that wins at 300 steps pays it back at the horizon (7.47 vs AdamW 7.33) while the cold-rate rotation family completes with the grid's best finals (7.30/7.34, both within 0.1 of AdamW; V42). Short-horizon calibration does not merely add noise; it systematically inverts the long-run ranking, exactly as Corollary 1 predicts.

5.6 Downstream transfer: three tasks, two models, five configurations

Perplexity is a proxy; the question is whether optimizer failure changes model behavior on a real task. Five blocks (V28–V33, 87 evaluations) repeat the Section 5.4 training protocol, save the end-of-training adapter (for SOAP-F, the checkpoint at which training diverged), and evaluate MMLU-5-shot (Hendrycks et al., 2021) on all 1531 validation questions: two models (Qwen3-1.7B, Phi-4-mini 3.8B), adapter ranks 1, 4, and 16, five seeds on the first configuration (Table 7 in Appendix D.6). An eval-only extension (V43) re-evaluates the same adapters on ARC-Challenge (Clark et al., 2018) and HellaSwag (Zellers et al., 2019)—174 evaluations, no new training (Appendix D.6).

The divergence is the downstream event. Every SOAP-F seed diverges within the 300 steps, and the saved checkpoint collapses to random-level performance: MMLU-5-shot falls to 24.89–25.28% (random guessing is 25%) in all five configurations and 17/17 diverged runs, against base accuracies of 58.65 (Qwen3) and 68.19 (Phi-4-mini)—a 33–43-point erasure.

The collapse is task-general; the family split is not. V43 answers the two questions the MMLU record leaves open. *Is the collapse a task artifact?* No: all 34 SOAP-F adapter-task cells on ARC and HellaSwag land at 22.4–26.5% against random's 25, every configuration, both tasks. *Is the family ordering a law?* No: with five seeds on V28 the families separate completely (both baselines below the base model, both spectral variants above), the stable band holds in 38 of 40 optimizer cells (two exceptions, one configuration), but the MMLU family split reproduces in only one of four cells. Optimizer *failure* is a property of the optimizer; the family ordering inside the stable band is a property of the model–task pair (Appendix C).

Synthesis. Across five configurations—two models, $2.2\times$ the parameter count, adapter ranks 1 to 16—the downstream record is: SOAP-F diverges in 17/17 runs and collapses to random in every one; every stable optimizer finishes at least 31.72 points above the collapsed arm; and the spread among stable optimizers never exceeds 3.85 points—what a downstream task feels is the failure mechanism, not the ranking among stable optimizers.

5.7 Boundary stress tests: large adapter ranks, clipping, and the α spectrum

Three blocks close the practitioner objections (full records in Appendix D.4). **Large adapter ranks do not escape the mechanism** (V34, J1–J4): at $r \in \{64, 128\}$ on Qwen3-1.7B the large-side effective-rank ratio stays at most 0.02–0.05, and SOAP-F diverges in 3/3 seeds at both ranks while AdamW (12.70) and SOAP-X (9.41) complete all 300 steps—free inversion fails at every adapter rank from 1 to 128. **Tighter clipping does not rescue free inversion** (V35, K1–K3): SOAP-F diverges in 3/3 seeds at clip 0.1 and 0.01 while AdamW trains normally—clipping bounds the gradient norm, but the amplification acts inside the preconditioner. **The α spectrum has an interior optimum at extreme deficiency, and it is a short-horizon configuration** (V36–V41): at $\alpha = 1/8$ on Phi-4-mini the weak-inversion member is stable (0 of 8 runs divergent) and at the hot rate reaches ppl 8.29—below both α-endpoints (9.85, 10.07) and AdamW's 8.53, the best spectral configuration measured on this model; the $\epsilon^{-2\alpha}$ law predicts the discriminating quantity exactly (10^3 at $\alpha = 1/8$ versus 10^6 at $\alpha = 1/4$), V39's fine sweep confirms the peak, and the mirrored sweep on moderate-deficiency Qwen3 descends monotonically (V41): the interior optimum is a property of extreme deficiency. Downstream, the interior point at the hot rate lands at MMLU 66.81 $\pm$ 0.37 (AdamW 64.43) while the cold rate lands at 68.10 $\pm$ 0.20 despite far worse perplexity (V37). But at the 5000-step horizon the hot-rate advantage reverses (V40/V42; Table 6): the interior optimum is deficiency-conditional *and* horizon-limited.

5.8 Ablating SOAP-X against Muon: rotation source and update pipeline

SOAP-X and Muon both train by rotation, and a reader could conclude the variant is Muon under another name. The two are diagonal corners of one design space—rotation source $\times$ update pipeline—so the resemblance is falsifiable: if either optimizer subsumed the other, the 2×2 factorial over that space would factorize, with every cross cell landing between its parents. V44 (M1–M5, 36 runs) runs exactly that factorial: **(i) rotation source**—SOAP-X rotates by the covariance eigenbasis (the EMA of $GG^\top$, eigendecomposed and refreshed every 10 steps); Muon rotates by the Newton–Schulz polar factor of the current update, recomputed every step with no covariance—and **(ii) update pipeline**—SOAP-X carries SOAP's Adam moments in the rotated frame with a bias-corrected step size; Muon carries Nesterov momentum with a global RMS-to-0.2 rescaling. V44 adds the two cross cells, *MuonEIG* (eigenbasis + Muon pipeline) and *SOAPNS* (polar factor + SOAP pipeline), on the Qwen3, Phi-4-mini, and Llama r=16 configurations, both rates, three seeds, protocol identical to the control grid. Neither cross cell touches the eigenvalue channel, so Corollary 2 predicts both are stable (M1); the attribution predictions (M2–M5) ask whether the parents' gaps are carried by the basis or the pipeline and whether the two variables interact.

Results. All 36 runs completed; the six Llama/MuonEIG cells initially exceeded the 32 GB device at the eigenbasis refresh while Newton–Schulz never did—the eigendecomposition is the binding memory cost of the eigenbasis route—and were completed with host-RAM optimizer state, the update rule

line-identical (V44b, Appendix C.4). Every cross cell ran 300 steps with zero divergences (M1, hit on the completed record). **M3 hits:** at the cold rate all four measured cross cells land in the parents' band—exchanging the rotation source at a cold rate is second-order. **M4 and M5 miss jointly, and the direction is the finding:** the rate response follows neither variable alone but their combination. On Phi-4-mini, SOAPNS was predicted rate-flat (like SOAP-X, which moves $10.33 \rightarrow 10.07$) but degrades hot by 12.2 ($8.19 \rightarrow 20.37$)—rate-flatness is a property of the full eigenbasis$\times$SOAP-pipeline combination, not of either component. At the hot rate SOAPNS loses to *both* parents on Qwen3 and Phi-4-mini far beyond the 0.30 tolerance, and at Llama's best rate it does too—a strongly negative interaction: the space does not factorize, and neither parent subsumes the other; MuonEIG stays in the parents' band everywhere except once, 0.42 above both at Qwen3's cold rate. **M2** now measures and misses: MuonEIG's best on Llama is 8.92, above the parents' midpoint 7.94—the Llama gap is not carried by the basis alone; like the rate response, the separation is a property of the *combination*.

Practical reading: which rotation should a practitioner use? The ablation and the wall-clock record (Appendix D.1) combine into a concrete recommendation. *Cost:* Muon's Newton–Schulz rotation keeps no covariance state and runs at 1.0–1.1$\times$ AdamW wall-clock; the eigenbasis route costs 1.4–2.1$\times$ wall-clock, and its memory—covariance buffers plus the eigendecomposition workspace—is the binding constraint at scale. *Performance:* at cold rates the rotation source is second-order (M3)—on the extreme-deficiency model, Phi-4-mini, the cells that are not the full eigenbasis$\times$SOAP combination track Muon and finish 2.1 ppl ahead of SOAP-X. *Positioning:* the paper's practical claim is not that SOAP-X beats Muon; it is that the eigenvalue channel, not rotation, is the failure. The cheapest safe route is a rotation-only optimizer—Muon where cost and memory bind, SOAP-X where SOAP infrastructure exists—and SOAP-X's role here is the theory-forced minimal probe that isolates the channel, not a deployment recommendation.

5.9 The 7B boundary: the failure is numerical, the repair one line

V45 (with completion block V45b) extends the program to Mistral-7B-v0.3, LoRA $r{=}16$ on attention projections: the largest model and matrix dimension (4096) in the program. The census half of the scale extrapolation holds—every large-side covariance sits at effective-rank ratio ≤ 0.0176 (median $2.1{\times}10^{-3}$, the same order as the 1.7–3.8B models)—and the cold-start law is dimension-free: SOAP-F diverges in 3/3 seeds at steps 7–16. The training arm produced the program's most consequential miss: **SOAP-X itself diverges at 7B**. It reaches ppl 5.15–5.18 at step 50, tracking AdamW's 4.96–4.97, then diverges in 3/3 seeds at steps 100–132, while AdamW completes 3/3 seeds (final ppl 14.32 ± 0.61, rising from 4.97 under the fixed cold rate). U3/W2–W3 were pre-registered misses, reported as such (Appendix C.5)—and then diagnosed.

Three diagnosis blocks (V47–V49, 15 runs; predictions F1–F6/G1–G7, Appendix D.8) traced the failure to floating point, not the algorithm. The divergence is learning-rate-insensitive (the quarter rate still diverges), and the update norm climbs a staircase from 0.2 to beyond 10^{12} while the gradient stays $O(1)$: the fault is internal to the preconditioner. A four-arm single-change ablation (V48) found that raising ϵ, matrix-level RMS normalization, and rotating the second moment with the basis all still diverge (steps 50–113); only QR re-orthogonalization of the eigenbasis after each eigendecomposition survives. Replaying the real 4096^2 covariances from the diverging run (V48c) pinned the cause: fp32 GPU eigendecomposition returns eigenvectors whose squared column norms deviate from one by 912–1865 (column norms up to $\approx$43$\times$ unit) on matrices whose largest-to-smallest positive-eigenvalue ratio is 3×10^{15}–4×10^{17}, while CPU fp32 and GPU fp64 return orthonormal bases to 10^{-6}; the defective basis is present from the first refresh and compounds at every 10-step refresh into the staircase. The library eigensolver fails at this conditioning; rotation does not.

One QR call per refresh repairs it completely. SOAP-X+QR completes 300 steps in 3/3 seeds (final ppl 7.50–7.79), while SOAP-F with the same fix still diverges at step 7: the eigenvalue amplification is algorithmic and survives the numerical repair, so the theory's channel separation is itself confirmed at 7B. At a $5\times$ horizon (1500 steps, V49), SOAP-X+QR completes in 2/2 seeds (final ppl 59.35 and 56.54); Muon, measured at 7B for the first time, completes (54.72); MuonEIG—identical eigendecomposition numerics, matrix-level normalization, no QR—also completes (57.15), attributing the original failure to elementwise second-moment normalization meeting a non-orthogonal basis; AdamW completes at 57.97. All four drift together at long horizons under the fixed cold rate (G3/G7's ≤ 12 threshold, calibrated at 300 steps, misses for every optimizer), and at the pre-registered 300-step

point SOAP-X+QR is the best measured (7.5 vs 14.5, 12.8, 11.0). The repair is thus one line deep and numerical—but the QR call adds to the eigenbasis route's cost, and the diagnosis is validated on a single GPU/library stack.

6 Discussion: the α-spectrum and design guidance

Where deficiency is moderate (Llama r=16, full-parameter training), rotation-only is optimal: stable at every β, nearly rate-flat, best-or-tied in 6 of 8 configurations. Where deficiency is extreme (Phi-4-mini, code domain), a small capped inversion at a hot rate buys perplexity at 300 steps ($\alpha = 1/8$ reaches 8.29)—but the hot rate is a trap beyond a few hundred steps. At 7B the eigenbasis route must re-orthogonalize after each eigendecomposition (Section 5.9)—or use the covariance-free rotation, which never eigendecomposes.

A cheap indicator for the regime. The decision variable is measurable before committing: thirty steps of instrumented AdamW training (the census protocol of Section 5.1) bound the effective ranks SOAP would invert, and the two swept regimes separate on the *tail* of that census—on Phi-4-mini, where the interior point exists, all 12 covariances are pinned at an effective-rank ratio of order 10^{-3}, while on Qwen3-1.7B (V41) the tail escapes an order of magnitude higher. The working heuristic: a pinned census marks the regime where a short-horizon try of weak capped inversion ($\alpha = 1/8$ at a $4\times$ hot rate) is defensible; any escaping tail says default to rotation-only.

Guidance. (1) Avoid free inversion in rank-deficient regimes: no rate, window, family, or domain rescues SOAP-F; the failure is structural (Theorems 2–3). (2) Default to rotation-only—Muon-style rotation where cost and memory bind, SOAP-X where SOAP infrastructure already exists—and at large matrix dimensions re-orthogonalize the eigenbasis after each eigendecomposition (one QR call; Section 5.9). (3) For extremely rank-deficient, short-lived workloads (the pinned census above), capped inversion at a $4\times$ hotter rate is defensible ($\alpha = 1/8$), with the rate re-chosen if the horizon grows (V40). (4) Never select spectral-optimizer learning rates with short probes (Corollary 1).

7 Limitations

Scale and horizon. The largest model in the main program is 3.8B and the longest run is 5000 steps; the laws are algebraic and parameter-count-independent, but everything above the measured boundary is extrapolation. The QR repair is verified to 7B (Section 5.9)—but on a single GPU/library stack: other eigensolver versions may fail at different conditionings, and the long-horizon perplexity drift at 7B is protocol-wide (every optimizer). The interaction constants are calibrated, not derived.

Task coverage. All training endpoints are perplexity; downstream transfer is knowledge-style evaluation only (Section 5.6). A 12-run instruction block on alpaca-cleaned (V46) shows the same split—SOAP-F diverges in 3/3 seeds, SOAP-X at parity with AdamW (3.72 vs 3.64)—but agentic fine-tuning remains untested.

Census timing and the $\alpha = 1/8$ rule. The census is measured at step 30 or at refresh steps, so exact eff-rank values are window-biased when $(1-\beta)^{-1}$ exceeds the horizon (the 2–3-order-of-magnitude deficiency conclusion is robust); the $\alpha = 1/8$ interior point is empirical—the law explains the separation from $\alpha = 1/4$ (10^3 versus 10^6 residual amplification) but supplies no closed-form criterion for when weak capped inversion beats rotation-only, and the guidance above is calibrated on two deficiency levels only.

8 Conclusion

Spectral preconditioners fail in rank-deficient regimes for an exact, parameter-free reason: inverting a covariance whose spectrum is mostly floor multiplies the update by $\epsilon^{-2\alpha}$ on both sides—a failure that survives every repair retaining the eigenvalue channel and vanishes when that channel is removed. Rotation-only preconditioning is the theory's corollary, not its contribution—Muon already demonstrates that rotations train; what is new is the exact account of why every eigenvalue-retaining family member must fail here. The corollary is best-or-tied in 6 of 8 configurations across 124M–3.8B; its one 7B failure proved numerical, and one QR call restored it. Eigenbases without eigenvalues.

AI USE STATEMENT

In this work, we used generative AI tools in three specific roles: (i) drafting and refactoring the experiment, launch, and analysis scripts, all of which were reviewed, executed, and verified by the authors; (ii) writing the mechanical aggregation code that converts raw run records into the per-block summary JSONs; and (iii) drafting and editing prose in this manuscript. The research questions, the theory, the pre-registered predictions, the experimental designs, and the interpretation of every result were conceived, supervised, and approved by the authors. No number in the paper is LLM-generated: every quantitative claim is regenerated mechanically from the raw run records by the released pipeline and verified by the released audit script, and each pre-registration JSON predates its runs. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments use public models and datasets (GPT-2 124M, Llama-3.2-3B, Qwen3-1.7B, Phi-4-mini, Mistral-7B-v0.3; WikiText-103, CodeSearchNet Python, cais/mmlu, alpaca-cleaned) on a single RTX 5090 (32 GB). The suite runs PyTorch with HuggingFace transformers and PEFT; effective batch 16×512 tokens everywhere; seeds 42/123/456 (42/123/456/789/1000 on the five-seed blocks). The artifact is released in full as anonymized supplementary material: the unmodified pre-registration JSONs for every block, the raw run records, the per-block summary JSONs, the complete experiment and launch scripts, and the analysis pipeline (statistics aggregation, figure generation, and the number audit itself). Every number in the tables and figures is produced mechanically from the summary JSONs and verified by the released audit script, so no number in this paper is hand-typed; re-running the released pipeline from the raw summaries reproduces the document end to end. Wall-clock timings (Appendix D) are hardware-specific.

REFERENCES

Mohamed Abouelenin, Adam Atkinson, Hany Awadalla, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Scalable second order optimization for deep learning. *arXiv preprint arXiv:2002.09018*, 2020.

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Aaron Grattafiori, Abhimanyu Dubey, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Advances in Neural Information Processing Systems*, 2018.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. arXiv:2009.03300.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*, 2019.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Hong Liu, Zhijian Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *International Conference on Learning Representations*, 2024.

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

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

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, 2015.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. arXiv:1609.07843.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI technical report*, 2019.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5—RMSprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations*, 2025. arXiv:2409.11321.

An Yang, Anfeng Li, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

A Full statements and proofs

This appendix gives the complete statements and proofs behind Section 4. Main-text results are restated with their original numbers (*Restatement of Theorem X*), so theorem numbering is global across the paper; results that appear only here—Lemma 1 and the two-sided corollary—carry continuing numbers. Assumptions A1–A3 (Section 4) hold throughout.

A.1 Setup: the SOAP(α) family

Let $W \in \mathbb{R}^{m\times n}$ be a trainable matrix with loss gradients $G_t = \nabla_W \mathcal{L}_t$. Following the notation of Section 3, the two gradient covariances are the exponential moving averages $L = \mathrm{EMA}_\beta(G^\top G) \in \mathbb{R}^{n\times n}$ and $R = \mathrm{EMA}_\beta(GG^\top) \in \mathbb{R}^{m\times m}$, i.e. $L = (1-\beta)\sum_{s\le t}\beta^{t-s}G_s^\top G_s$ and $R = (1-\beta)\sum_{s\le t}\beta^{t-s}G_s G_s^\top$, with eigendecompositions $L = QD_LQ^\top$ and $R = PD_RP^\top$ refreshed every p steps. The rotated gradient is $\tilde{G}_t = P^\top G_t Q$, and Adam-type moments m_t, v_t are maintained in the rotated basis, producing the entrywise-normalized signal $u_t = m_t/(\sqrt{v_t}+\epsilon_a)$ with $|(u_t)_{ij}| \le c$, $c = (1-\beta_1)/\sqrt{1-\beta_2} \le 1$, where the bound holds whenever $\sqrt{v_t} \ge \epsilon_a$ on the support of m_t.

The SOAP(α) family applies a per-direction spectral exponent α with eigenvalue floor ϵ:

$$\Delta W_t = -\eta_t\, P\left(\mathrm{fl}(D_R)^{-\alpha} \odot u_t \odot \mathrm{fl}(D_L)^{-\alpha}\right) Q^\top, \tag{1}$$

where $\mathrm{fl}(D) = \max(D,\epsilon)$ elementwise and $\odot$ denotes row/column scaling (the left factor scales rows via the m floored eigenvalues of R; the right factor scales columns via the n floored eigenvalues of L). Three members anchor the analysis: $\alpha = 0$ (SOAP-X, rotation only), $\alpha = 1/4$ (SOAP-F, the classic Shampoo inversion $D^{-1/4}$ per side), and ridge-guarded inversion $\mathrm{fl}(D) = \max(D, \gamma\lambda_{\max})$ (SOAP-R). A fourth member, SOAP-AC (capped inversion), rescales the update of SOAP(α) back to the $\alpha = 0$ RMS; it decouples the amplification (which it removes) from the inversion (which it retains), and serves as a measuring instrument in the experiments.

A.2 Direction–magnitude decomposition

Lemma 1 (Direction–magnitude decomposition)**.** *The update of Equation* (1) *factors into (i) a direction term—the eigenbasis pair (P, Q) together with the sign structure of u_t—and (ii) a magnitude term—the entrywise scales $\mathrm{fl}(D_R)_{ii}^{-\alpha}\,\mathrm{fl}(D_L)_{jj}^{-\alpha}$. The two terms enter multiplicatively and independently. Consequently $\alpha = 0$ is the unique member of the family whose per-entry update magnitude is bounded by the Adam invariant: the update inherits the dimensionless, spectrum-blind scale of u_t and no other.*

Proof. Immediate from the update equation: the orthogonal maps P, Q and the diagonal scalings act on different factors of the product, and for $\alpha = 0$ both diagonal factors equal the identity. □

Lemma 1 isolates why rotation-only preconditioning is numerically safe (its update is Adam-bounded) while any $\alpha > 0$ makes the update magnitude a functional of the covariance spectrum. The remainder of this appendix quantifies exactly how badly that functional behaves when the spectrum is deficient.

A.3 Proof of Theorem 2 (cold-start amplification law)

Theorem (Restatement of Theorem 2 (cold-start amplification law))**.** *Consider a factorized (LoRA-type) update $W' = W + BA$ with B initialized to zero, trained by* SOAP(α) *with floor ϵ and refresh period $p \ge 3$. Then the A-factor's gradient is identically zero until B receives its first update; its covariance statistics at the first refresh are the zero matrix, hence $\mathrm{fl}(D_L) = \mathrm{fl}(D_R) = \epsilon$ (elementwise) and $P = Q = I$ until the next refresh. At the first step t_0 at which $\nabla_A \mathcal{L}_t \ne 0$, the ratio of the update RMS to the $\alpha = 0$ reference equals, exactly,*

$$A(\alpha,\epsilon) \;=\; \frac{\left\|\mathrm{fl}(D_R)^{-\alpha} \odot u_{t_0} \odot \mathrm{fl}(D_L)^{-\alpha}\right\|_F}{\|u_{t_0}\|_F} \;=\; \epsilon^{-2\alpha}.$$

The ratio is independent of the learning rate, the data, the gradient scale, and the spectrum accumulated after the cold start; it is a pure function of (α, ϵ).

Proof. Zero initialization gives $\nabla_A \mathcal{L} = B^\top \nabla_{W'}\mathcal{L} = 0$ while $B = 0$, so every gradient outer product contributing to either covariance of the A-factor vanishes through the first refresh, and the first refresh occurs on the zero matrix. The eigendecomposition of the zero matrix returns $D = 0$ and an identity eigenbasis; the floor maps every eigenvalue to ϵ. The next $p-1$ updates therefore carry the constant entrywise scale $\epsilon^{-\alpha}$ per side. At t_0 (before B's first update propagates into a

second refresh), the moments satisfy $m_{t_0} \propto g_{t_0}$ and $v_{t_0} \propto g_{t_0}^2$ on the support of the gradient, so $u_{t_0} = c \cdot \mathrm{sign}(g_{t_0})$ entrywise with $c > 0$. Since the Frobenius norm is invariant under the orthogonal factors $P = Q = I$ and the diagonal scales are constant, $\|\epsilon^{-\alpha}\epsilon^{-\alpha}u_{t_0}\|_F = \epsilon^{-2\alpha}\|u_{t_0}\|_F$. □

Empirical content. The law is falsifiable to machine precision. With floor $\epsilon = 10^{-12}$ it predicts amplifications of exactly 10^3, 10^6, 10^{12} at $\alpha = 1/8, 1/4, 1/2$. The pre-registered measurements return $A = 1.0000001 \times 10^3$ ($\alpha = 1/8$, V13: nine runs over three learning rates), $A = 1.0000002 \times 10^6$ ($\alpha = 1/4$, V13: nine runs), and $A = 1.0000001 \times 10^{12}$ ($\alpha = 1/2$, V15: six runs of the RMS-capped variant SOAP-AC$(1/2)$, which is stable and therefore permits the diagnostic to complete). Free inversion at $\alpha = 1/2$ destabilizes the run before the diagnostic saturates (measured maxima 3.9–8.0×10^9, all runs non-finite), so the 10^{12} point is measured on the capped variant, whose update magnitude is rescaled but whose per-step RMS ratio is recorded before capping. The residual relative deviations—1.2×10^{-7}, 1.9×10^{-7}, 1.27×10^{-7}—sit at the float32 machine epsilon (1.19×10^{-7}): the law is exact to the precision of the arithmetic that measures it.

A.4 Proof of Theorem 1 (temporal and architectural deficiency)

Theorem (Restatement of Theorem 1 (structural deficiency)). ***(i) Architectural.*** *For the factorized update, every summand of the large-side covariance has rank at most r:* $\mathrm{rank}(G_s^\top G_s) \le r$ *for the A-factor ($r \times n$) and* $\mathrm{rank}(G_s G_s^\top) \le r$ *for the B-factor ($m \times r$). Rank is subadditive over the accumulated summands, so after t steps* $\mathrm{rank}(\mathrm{EMA}_t) \le \min(\dim, rt)$*, and in the EMA steady state (effective window $W = (1-\beta)^{-1}$) the null fraction of the large side is at least $1 - rW/\max(m,n)$ whenever $rW \ll \max(m,n)$.* ***(ii) Temporal/spectral.*** *For an arbitrary parametrization (including full-parameter training), let $\lambda_{\min}(t)$ denote the smallest eigenvalue of either covariance. The amplification ascribed to any direction whose eigenvalue satisfies $\lambda_i \le \epsilon$ is at least $\epsilon^{-\alpha}\lambda_{\max}^{-\alpha} \ge \epsilon^{-\alpha}\mathrm{fl}(D)_{\mathrm{top}}^{-\alpha}$ per floored side, multiplied by the opposite side's scale. Hence spectral inversion destabilizes training whenever the floor binds on at least one side, and the learning rate cannot compensate: the amplification is η-independent (Theorem 2), so the per-entry update magnitude ηAc exceeds any fixed weight scale $\|W\|_{\mathrm{rms}}$ for all $\eta \ge \|W\|_{\mathrm{rms}} c^{-1}\epsilon^{2\alpha}$.*

Proof. (i) $\mathrm{rank}(G_s) \le r$ for an $r \times n$ factor gradient, hence each summand of either covariance has rank at most r; rank is subadditive over the t accumulated (exponentially weighted) summands, and the EMA's exponentially decaying weights leave an effective count of $W = (1-\beta)^{-1}$ summands in steady state. (ii) Directions with $\lambda_i \le \epsilon$ receive the scale $\epsilon^{-\alpha}$; the opposite side contributes at least $\lambda_{\max}^{-\alpha}$ on its own top direction. The η-independence is inherited from the cold-start law because the floor value, not the spectrum, sets the amplification on floored directions. □

Measured deficiency (no idealization required). Theorem 1 requires no idealization of the spectra: the deficiency is directly measured. In LoRA fine-tuning of Llama-3.2-3B (V8, $r \in \{1, 16\}$, six projections instrumented), the participation ratio $r_{\mathrm{eff}} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ of the gradient covariances is 1.0–3.3 at $r = 16$ and 1.0–18.9 at $r = 1$, against dimensions up to 3072. The $r = 1$ values are the cleanest isolation of the temporal mechanism: with $r = 1$ every per-step summand has rank exactly one, so all measured rank comes from window accumulation, and the observed maximum 18.9 tracks the window bound $W = (1-\beta)^{-1} \approx 20$ at $\beta = 0.95$. In full-parameter GPT-2 training (V22 census at step 30), the twelve instrumented covariances have participation ratios 2.5–229.1 against dimensions 768–3072—eleven of twelve below 100, median 5.5—with $\lambda_{\min}$ numerically zero for two matrices ($\le 10^{-11}$ for four) and $\lambda_{\max} \in [3.6 \times 10^{-6}, 1.4 \times 10^{-3}]$ throughout: here the deficiency is not an exact rank statement but an extreme spectral concentration (in the LoRA census the top ten eigenvalues carry 97–99.9% of the mass) together with an absolute covariance scale up to six orders of magnitude below unity, which the ridge law shows is destabilizing in its own right. The EMA window ($\beta = 0.95$, refresh $p = 10$) admits at most $O(W)$ informative gradient directions, so temporal deficiency is structural, not incidental: it persists at every horizon shorter than the covariance accumulation time, and the architectural bound of part (i) only lowers the ceiling.

A.5 Proof of Theorem 3 (ridge absolute-scaling law)

Theorem (Restatement of Theorem 3 (ridge absolute-scaling law)). *Replace the constant floor by the ridge floor* $\mathrm{fl}(D) = \max(D, \gamma\lambda_{\max})$*. Then the amplification on floored directions becomes $A(\gamma) =$*

$(\gamma\lambda_{\max})^{-\alpha} \times$ (*opposite-side scale*)*, an absolute (data-scale-dependent) quantity: it now varies with the covariance scale* $\lambda_{\max}$*, unlike the data-independent* $\epsilon^{-2\alpha}$ *of Theorem 2. Stable training requires* $\eta A(\gamma)c \lesssim \|W\|_{\text{rms}}$*, i.e.*

$$\gamma \gtrsim (\eta c/\|W\|_{\text{rms}})^{1/\alpha} \cdot 1/\lambda_{\max}.$$

With measured values $\lambda_{\max} \sim 10^{-5}$–$10^{-4}$*,* $\eta/\|W\|_{\text{rms}} \sim 10^{-2}$–$10^{-3}$ *and* $\alpha = 1/4$*, the admissible ridge satisfies* $\gamma^* = \Theta(1)$*: the floor must be raised to the order of the entire spectrum before stability is achieved, at which point* $\text{fl}(D)$ *is uniform, the preconditioner degenerates to a global rescaling, and the spectral information is discarded. There is no interior value of* γ *that is simultaneously stable and preconditioning. Moreover, in the LoRA cold start the relative floor cannot engage at all: the first refresh sees the zero matrix, so* $\gamma\lambda_{\max} = 0$ *and the update runs on the constant-*ϵ *regime of Theorem 2 regardless of* γ*.*

Proof. The floored directions receive $(\gamma\lambda_{\max})^{-\alpha}$ per side; solving the stability inequality for γ yields the display. Uniformity at $\gamma \geq 1$ is immediate from $\text{fl}(D) = \max(D, \gamma\lambda_{\max}) = \gamma\lambda_{\max}$ for all directions when $\gamma \geq 1$. The cold-start claim follows from $D = 0 \Rightarrow \lambda_{\max} = 0$ at the first refresh. □

Empirical content. Pre-registered sweep (V19; Llama $r = 1$ and $r = 16$, GPT-2 $r = 16$, $\alpha = 1/4$, $\gamma \in \{10^{-4}, 10^{-2}\}$, three learning rates each): measured amplifications $A \in [1.7, 4.2] \times 10^4$ at $\gamma = 10^{-4}$ and $A \in [5.5, 16.7] \times 10^3$ at $\gamma = 10^{-2}$, against the prediction $(\gamma\lambda_{\max})^{-\alpha}\lambda_{\text{opp}}^{-\alpha}$ evaluated on the measured spectra (1.5×10^4 and 4.9×10^3 for Llama $r = 1$); every one of the 24 runs diverges, as the stability inequality requires $\gamma \gtrsim 0.6$ for stability at the learning rates used. The cold-start clause is likewise measured: in the calibration records, ridge-guarded variants with $\gamma \in \{10^{-4}, 10^{-2}, 10^{-1}, 1\}$ all diverge at the smallest probed learning rate (5×10^{-5}) in LoRA—including the degenerate $\gamma = 1$, whose uniform floor still rides on the cold-start zero spectrum.

A.6 Corollaries with proofs

Corollary (Restatement of Corollary 1 (no usable learning rate for free inversion))**.** *Because the amplification is learning-rate independent (Theorem 2), lowering* η *cannot re-stabilize spectral inversion: it reduces the per-step displacement* ηAc *but leaves* A *untouched, so the stable set is* $\eta \lesssim \|W\|_{\text{rms}}\epsilon^{2\alpha}/c$*, which with measured quantities sits near* 10^{-7} *for* $\alpha = 1/4$*,* $\epsilon = 10^{-12}$*—four to five orders of magnitude below the calibrated rates of the stable optimizers, at which training is glacial. Divergence onset correspondingly scales as* $t^* \sim \|W\|_{\text{rms}}/(\eta Ac)$*: for any probe horizon* T *there exist learning rates with* $t^* > T$*, so short-horizon stability probes are structurally unable to certify spectral inversion. The measured instance (V22, full-parameter): SOAP-F is flagged unstable by the 30-step probe at the smallest grid learning rate (*6.4×10^5 *at* $\eta = 5 \times 10^{-6}$*), the calibration fallback selects that configuration anyway, and the 300-step run reaches* 2.9×10^7*. Optimizer comparisons that calibrate on short probes can therefore certify rate tuning but not amplitude safety.*

Corollary (Restatement of Corollary 2 (optimality of rotation-only design))**.** *Within* SOAP(α)*, the set of stable members at fixed practical* η *is* $\{\alpha = 0\} \cup \{\alpha > 0 : \text{ridge} \geq \gamma^*\}$*, and the second set is preconditioning-free (Theorem 3). Post-hoc amplitude capping (SOAP-AC) restores stability by discarding the absolute scale of the eigenvalue channel, so it retains only relative eigenvalue weights. Hence the only stable member that retains spectral-direction information at its natural amplitude is* $\alpha = 0$*: eigenbases are usable, eigenvalue magnitudes are not.*

Proof. Combine Lemma 1 (the $\alpha = 0$ update is Adam-bounded) with Theorem 3 (every $\alpha > 0$ member without a $\Theta(1)$ ridge is unstable, and with it is a uniform rescaling). □

Corollary 3 (Two-sided instability)**.** *Applying spectral inversion on only one side (the other held at rotation only,* $\alpha = 0$*) still produces amplification* $\epsilon^{-\alpha}$ *on the inverted side's floored directions—the non-inverted side contributes identity scale—so single-sided inversion is unstable whenever the inverted side is deficient. Measured (V14, side-isolation variants on Llama* $r = 1$ *and* $r = 16$*, two learning rates each, either side inverted): single-sided inversion diverges in all eight tested configurations, with the RMS diagnostic pinned at exactly* $\epsilon^{-\alpha} = 10^{3.0000}$*; the deficit of either side alone is sufficient.*

Proof. Specialize Equation (1) to one inverted side; the floored directions of that side receive $\epsilon^{-\alpha}$ and the opposite side contributes scale 1. □

B Naming map: variants, blocks, and prediction IDs

The paper names three layers of objects: *method variants* (optimizers and ablation cells, Sections 3–5.8), *experiment blocks* (Vn: pre-registered groups of runs; the main text's E-labels map onto V-blocks as shown), and *prediction IDs* (the numbered falsifiable statements registered before each block launched; Section 5). Table 3 fixes the correspondence. Prediction codes are block-local and letters repeat across blocks (P in V23/V24/V25, L in V27/V36, M in V37/V44): a code is meaningful only together with its block, and V43's A1–A4 are unrelated to the theory's assumptions A1–A3 (Section 4).

Table 3: **Naming map: method variants, experiment blocks, and their prediction-ID families.** E-labels are the main text's experiment names.

Name	Meaning
Method variants	
SOAP-F	free spectral inversion, $\mathrm{fl}(D)^{-\alpha}$ per side with floor ϵ; $\alpha = 1/4$ unless stated
SOAP-X	rotation only: eigenbasis kept, eigenvalue magnitudes discarded; Adam moments in the rotated frame
SOAP-AC$_{0.25}$	capped inversion: the SOAP(α) update rescaled to the $\alpha = 0$ RMS (subscript denotes α)
SOAP-R	ridge-guarded inversion, $\mathrm{fl}(D) = \max(D, \gamma\lambda_{\max})$ (Theorem 3)
SOAP-XW / SOAP-AW	the E1/V23 β-sweep pair: rotation-only / free inversion under the swept EMA window
MuonEIG	V44 cross cell: covariance eigenbasis + Muon update pipeline
SOAPNS	V44 cross cell: Newton–Schulz polar factor + SOAP update pipeline
SOAP-X+QR	SOAP-X with one QR re-orthogonalization of the eigenbasis per refresh (the 7B numerical repair, Section 5.9)
Experiment blocks	
V8	covariance census instrumentation (Llama-3.2-3B, $r \in \{1, 16\}$, six projections)
V13–V22	original program: law verification and main landscape (V13/V15 law measurement, V14 side isolation, V19 ridge sweep, V22 full-parameter census)
E1/V23	covariance EMA (β) sweep, 48 runs (P1–P6)
E2/V24	probe-rate extension to Qwen3 and Phi-4-mini (P1–P4); E2-B/V24-B fixed-rate control grid (B1–B4)
E3/V25	code-domain block (P1–P3); E4: wall-clock study (Table 5)
V27	5000-step horizon, Qwen3-1.7B (L1–L5)
V28–V33	downstream MMLU transfer, five blocks (codes D, E, F, H, I)
V34/V35	boundary stress: adapter ranks 64/128 (J1–J4); tighter clipping (K1–K3)
V36–V39	α-spectrum sweep on Phi-4-mini (codes L, M, N, O)
V40–V42	5000-step horizon, Phi-4-mini (codes Q, R, S)
V43	eval-only ARC/HellaSwag extension of the V28–V33 adapters (A1–A4)
V44/V44b	2×2 factorial, rotation source $\times$ update pipeline (M1–M5); V44b host-RAM completion
V45/V45b	7B boundary check, Mistral-7B-v0.3 (U1–U3; W1–W3 in the completion arm)
V46	instruction tuning, alpaca-cleaned (X1–X4)
V47–V49	7B diagnosis: V47 instrumented replication; V48/V48b/V48c repair ablation and offline replay (F1–F6, G1–G2); V49 $5\times$ horizon (G3–G7)
Baselines	
AdamW, Muon	baselines of Section 3; Muon's Newton–Schulz rotation is the covariance-free reference for the rotation family

C Attribution of pre-registration misses

The pre-registered program records 83 hits, 21 misses and 2 partials: 83/106 is the program's hit rate (Table 4). This appendix attributes every miss to its mechanism, because the attribution carries the scientific weight: a miss on the amplification laws themselves would refute the theory; a miss on an auxiliary quantity does not. The taxonomy is simple. (i) *Estimator misses*—the predicted quantity is a property of the EMA covariance estimator, not of the preconditioner algebra (P1). (ii) *Evaluation-protocol misses*—the verdict, not the measured quantity, is at fault: the pre-specified test pooled cross-layer spread together with the β-sensitivity it meant to isolate (P5). (iii) *Threshold and attribution misses*—the prediction concerned a comparison margin, a preservation band, a schedule property, (V44) a two-variable attribution whose additive structure the measurement rejected, or (V45) a threshold or universal quantifier calibrated on smaller models (B2, B4, D2, L2, L3, E2-P2, E2-P4, V29-E2, V29-E3, F2, V33-I2, V43-A2, V44-M2, V44-M4, V44-M5, V45-U1, V45-U3; the two partials, L5 and V43-A3, belong here too; prediction codes are block-local, so L2/L5 are V27's, L3 is V36's, and A2/A3 are V43's). (iv) *Resource-limit misses*—the run could not execute on the available device, so the verdict is an engineering bound rather than a measurement. Two such verdicts were recorded when their blocks closed (V44-M1: six MuonEIG-Llama runs exceeded device memory during the eigenbasis refresh; V45-U2 and the SOAP-X half of V45-U3: at 7B both SOAP arms exceed the 32 GB device at the same refresh); both were subsequently lifted by host-RAM optimizer state with line-identical update rules (V44b, V45b), pre-registered as unchanged-prediction completions, and the affected verdicts in Table 4 are re-evaluated on the completed records (marked †)—no resource-limit miss remains open. (v) *Law misses*—a measured amplification deviating from $\epsilon^{-2\alpha}$ or $(\gamma\lambda_{\max})^{-2\alpha}$, or a stable free-inversion run. No miss of category (v) occurred: all 14 amplification-law predictions hit, seven of them to seven significant digits.

Threshold provenance. The pre-registered thresholds fall in two classes: the 14 law predictions need no calibration (they are identities), but the empirical bands and margins (preservation ± 1.5, replication tolerances) were fixed from first principles without pilot experiments; every threshold-class miss traces to that provenance. Future pre-registration of empirical bands should calibrate them with small pilots first.

Table 4: **Complete pre-registered scorecard** for experiment blocks V23–V46 (V44b/V45b complete cells of V44/V45; V46 is the instruction-tuning block); the original V13–V22 program's 14/14 amplification-law predictions are folded into the first row. Verdicts were fixed before the runs launched and are reported verbatim; comparisons are in validation perplexity (lower is better) unless stated, MMLU and multi-task rows in accuracy (higher is better). Prediction codes are block-local and unrelated to the theory's assumption labels A1–A3. Misses are attributed in this appendix. †re-evaluated on the completed record: the affected cells were re-run with host-RAM optimizer state, the update rule line-identical (V44b/V45b), pre-registered as unchanged-prediction completions before those runs launched.

Block	Prediction	Outcome	Note
Laws (V13/V19)	14/14 amplification-law predictions	14 hit	incl. $\epsilon^{-2\alpha}$ to 7 digits; ridge law both γ
E1 β-sweep (V23)	P2 free inversion diverges $\forall\beta \leq 0.99$	hit	24/24 runs diverge (incl. $\beta = 0.999$: P3)
E1	P4 SOAP-XW stable $\forall\beta$, Δppl ≤ 0.15	hit	max drift 0.09 (r=16), 0.06 (r=1)
E1	P6 SOAP-XW(β=0.95) $\equiv$ SOAP-X	hit	6.98 vs 6.98
E1	P1 rank monotone in β	miss	β=0.999 window > run; early-sample bias
E1	P5 $\lambda_{\max}$ 10×-invariant	miss	pooled-range evaluation; per-endpoint worst 8.4× (B.1)

Table 4: Complete pre-registered scorecard (continued).

Block	Prediction	Outcome	Note
E2 family (V24)	P1 SOAP-F diverges both models	hit	6/6 runs at calibrated rate
E2	P3 census median eff-rank < 50	hit	4.7 (Qwen3), 2.1 (Phi-4-mini)
E2	P2 SOAP-X best-or-tied	miss	traced to rate artifact; resolved by E2-B
E2	P4 SOAP-AC never beats SOAP-X	miss	true at extreme deficiency (Phi-4-mini, code)
E2-B control (V24B)	B1 SOAP-AC best-or-tied @5e-4	hit	9.20 (Qwen3); 9.85 $\approx$ 10.07 (Phi-4-mini)
E2-B	B2 SOAP-X $>$ AdamW @5e-4 (artifact test)	miss (= artifact)	SOAP-X $\leq$ AdamW both models — favorable
E2-B	B3 zero divergence in grid	hit	48/48 stable
E2-B	B4 both rates too hot	miss	SOAP-AC better hot; others prefer cold
E3 code (V25)	P1 SOAP-F diverges on code	hit	3/3 seeds
E3	P2 SOAP-X best-or-tied on code	hit	SOAP-X 5.75 vs Muon 6.80; AdamW unstable
E3	P3 census median < 50 on code	hit	3.4 (range 1.6–13.1)
Long horizon (V27)	L1 no divergence at best rates	hit	12/12 runs
V27	L3 SOAP-X $\leq$ Muon + 0.1	hit	7.79 vs 7.84
V27	L4 SOAP-AC @5e-4 $\leq$ 10.5	hit	7.96
V27	L2 SOAP-X $\geq$ AdamW $-$ 0.5 gap at 5k	miss	gap closes to 0.02 — speed, not asymptote
V27	L5 all optima $<$ step 3500	partial	SOAP-X optimal at 3500–4000
Downstream (V28)	D1 SOAP-F MMLU $\leq$ 30%	hit	collapses to 25.28% (random = 25%); 5/5 diverged
V28	D2 stable opts within $\pm$1.5 of base	miss	family split: baselines $-1.65/-1.64$ (all 10 seeds below base); spectral $+0.84/+0.51$ (all 10 above)
V28	D3 SOAP-X $\geq$ AdamW $-$ 0.5 (MMLU)	hit	59.49 vs 57.00 ($+2.49$)
V28	D4 ppl replicates V24-B $\pm$0.3	hit	largest deviation 0.03
Downstream (V29)	E1 SOAP-F MMLU $\leq$ 30% (3.8B)	hit	collapses to 25.08% on Phi-4-mini; 3/3 diverged
V29	E2 stable opts within $\pm$1.5 of base	miss	all drift below base (AdamW -3.76, SOAP-X -4.31); only SOAP-AC -0.46 inside — band is model-dependent
V29	E3 SOAP-X $\geq$ AdamW $-$ 0.5 (MMLU)	miss	63.88 vs 64.43 — misses by 0.05 (sd 0.40/0.67); SOAP-X at its hot rate
V29	E4 ppl replicates V24-B $\pm$0.3	hit	largest deviation 0.03
Downstream (V30)	F1 SOAP-F MMLU $\leq$ 30% (r=1)	hit	ppl $\approx 1.9 \times 10^7$; collapses to 25.08%
V30	F2 stable opts within $\pm$1.5 of base	miss	Muon -1.52 (0.02 outside band); other three within $\pm$0.4
V30	F3 SOAP-X $\geq$ AdamW $-$ 0.5 (MMLU)	hit	58.98 vs 58.87
V30	F4 stable opts no divergence at r=1	hit	12/12 runs complete 300 steps
Downstream (V32)	H1 SOAP-F MMLU $\leq$ 30% (r=4)	hit	ppl $\approx 3.6 \times 10^7$; collapses to 24.89%

Table 4: Complete pre-registered scorecard (continued).

Block	Prediction	Outcome	Note
V32	H2 stable opts within ± 1.5 of base	hit	all four within ± 1.0 (max SOAP-X $+0.98$)
V32	H3 SOAP-X $\geq$ AdamW $-$ 0.5 (MMLU)	hit	59.63 vs 59.42 ($+0.21$) — best of all five
V32	H4 stable opts no divergence at $r{=}4$	hit	12/12 runs complete 300 steps
Downstream (V33)	I1 SOAP-F MMLU $\leq$ 30% ($r{=}4$, 3.8B)	hit	collapses to 25.11% on Phi-4-mini
V33	I2 stable opts within ± 1.5 of base	miss	max $\lvert$drift$\rvert$ 3.70
V33	I3 SOAP-X $\geq$ AdamW $-$ 0.5 (MMLU)	hit	66.71 vs 64.49
V33	I4 stable opts no divergence	hit	12/12 runs complete 300 steps
Boundary (V34)	J1 large-side eff_ratio < 0.5 at $r{=}64$	hit	max ratio 0.02 at step 30
V34	J2 SOAP-F diverges at $r{=}64$ (3/3)	hit	3/3 diverged — cold start is rank-independent
V34	J3 SOAP-F diverges at $r{=}128$ (3/3)	hit	3/3 diverged
V34	J4 AdamW/SOAP-X stable at $r{=}64$	hit	AdamW 12.70, SOAP-X 9.41, 0 divergence
Boundary (V35)	K1 SOAP-F diverges at clip 0.1 (3/3)	hit	3/3 diverged
V35	K2 SOAP-F at clip 0.01: no rescue	hit	3/3 diverged or ppl $\geq$ 20 (final 151935.94)
V35	K3 AdamW trains at clip 0.1 (ppl $\leq$ 12)	hit	final ppl 9.82 — clipping itself is harmless
Boundary (V36)	L1 SOAP-AC(1/8) stable at both rates	hit	0 divergent of 8 runs
V36	L2 SOAP-AC(1/8)@5e-4 $<$ SOAP-X 10.07	hit	ppl 8.29 at the hot rate
V36	L3 SOAP-AC(1/8) best $>$ SOAP-AC(1/4) 9.85	miss	best 8.29 — interior optimum, not ordered; ordering is jointly in (α, rate)
Boundary (V37)	M1 interior point stable, both rates	hit	0 divergent of 10 runs (V29 protocol)
V37	M2 no checkpoint collapses (MMLU $\geq$ 30%)	hit	worst acc 66.43%
V37	M3 hot-rate MMLU $\geq$ AdamW $-$ 0.5 (63.93)	hit	66.81 $\pm$ 0.37 vs 64.43
V37	M4 hot-rate ppl replicates V36 ($\pm$0.3 of 8.30)	hit	ppl 8.30 under the second training script
Boundary (V38)	N1 interior point stable on Qwen3	hit	0/5 divergent
V38	N2 ppl $>$ SOAP-X 8.97 (rotation-only stays optimal)	hit	ppl 11.11 — interior optimum is deficiency-conditional, as guided
Boundary (V39)	O1 fine-sweep members stable (6/6)	hit	0 divergent of 10 runs
V39	O2 ppl($\alpha{=}3/16$) between 8.30 and 9.85	hit	ppl 8.86 — interior point beats its upper neighbor
V39	O3 ppl($\alpha{=}1/16$) between 8.30 and 10.07	hit	ppl 8.64 — interior point beats its lower neighbor
Boundary (V40)	Q1 all nine 5000-step runs complete	hit	0 divergent of 9 runs

Table 4: Complete pre-registered scorecard (continued).

Block	Prediction	Outcome	Note
V40	Q2 interior-point gap $>$ 0.10 (hot-rate trap)	hit	gap $+0.16$ (V27's $\alpha{=}1/4$ gap was $+0.19$)
V40	Q3 AdamW overtakes interior at 5k steps	hit	7.33 vs 7.47 (300-step: 8.53 vs 8.30)
V40	Q4 rotation-only final $<$ 10.07 at 5k steps	hit	final 7.61, 0 divergent — extreme-deficiency cost shrinks with horizon
Boundary (V41)	R1 moderate-deficiency sweep members stable (6/6)	hit	0 divergent of 10 runs
V41	R2 ppl($\alpha{=}1/16$) between 11.18 and 18.26	hit	ppl 14.18 — lower gap fills monotonically
V41	R3 ppl($\alpha{=}3/16$) between 9.20 and 11.18	hit	ppl 9.63 — upper gap fills monotonically
V41	R4 both new points above cold-rate default 8.97	hit	no hot-rate arm beats the moderate-deficiency default, unlike at extreme deficiency
Boundary (V42)	S1 all six cold-rate rotation runs complete	hit	0 divergent of 6 runs
V42	S2 rotation-only @ cold gap $\leq$ 0.10 (benign range)	hit	gap $+0.00$ vs the $+0.31$ the same optimizer paid at the hot rate
V42	S3 rotation-only @ cold final $<$ 7.61 (its hot-rate final)	hit	final 7.30 — the cold rate beats the hot rate for the same optimizer
V42	S4 Muon @ cold within 0.2 of AdamW's 7.33	hit	final 7.34 — the rotation family stays competitive with the horizon winner
Multi-task (V43)	A1 SOAP-F $\leq$ 28% on both tasks, all 5 configs	hit	all 34 SOAP-F adapter-task cells 22.4–26.5% (random = 25%)
V43	A2 stable opts within $\pm$2.0 of base, both tasks	miss	38/40 cells hold; V28 ARC: AdamW -5.46, Muon -4.19 (spectral within 1.5)
V43	A3 MMLU family split replicates (4 cells)	partial	holds in 1/4 (V29 HS); other three miss the failing half by 0.66–0.98
V43	A4 collapse and band hold at $r{=}1$/$r{=}4$	hit	24/24 cells; collapse 22.4–26.5%, max seed-mean drift -1.98 (V32 ARC Muon)
Ablation (V44)	M1 all 36 cross-cell runs complete, no divergence	hit†	36/36 complete, 0 divergences; the six MuonEIG-Llama cells completed in V44b with host-RAM optimizer state (update rule unchanged; Newton–Schulz never memory-bound)
V44	M2 Llama gap (1.92 ppl) carried by the basis	miss†	MuonEIG best 8.92 above the parents' midpoint 7.94 (SOAPNS best 9.36 above both parents)— the gap is not carried by the basis alone (measured in V44b)
V44	M3 cross cells within parents' band @ cold rate	hit	Qwen3 9.60/9.33 in $[8.47, 9.68]$; Phi-4-mini 8.20/8.19 in $[7.55, 10.83]$
V44	M4 rate response follows the pipeline	miss	Phi SOAPNS degrades hot $8.19 \rightarrow 20.37$ (rate-flat was the SOAP-X property: $10.33 \rightarrow 10.07$) — follows the combination

Table 4: Complete pre-registered scorecard (continued).

Block	Prediction	Outcome	Note
V44	M5 no cross cell beats/loses both parents by >0.30	miss	Qwen3 SOAPNS@5e-4 31.53 vs 18.26/28.65; Llama SOAPNS@best 9.36 vs 6.98/8.90 — strong negative interaction
7B check (V45)	U1 census: every tracked covariance eff_ratio $\leq$ 0.02, median eff_rank ≤ 8	miss	median 4.65 holds; all six 4096-dim covariances ≤ 0.0176, but the 16-dim sides reach 0.535
V45	U2 SOAP-F diverges 3/3 seeds within 300 steps @ 1.25×10^{-4}	hit†	SOAP-F diverges 3/3 seeds within 300 steps (measured in V45b with host-RAM optimizer state; Appendix D.7)
V45	U3 AdamW and SOAP-X each complete 3/3 seeds, final ppl $\leq$ 12	miss	AdamW 3/3 finite at 14.32 ± 0.61 (threshold calibrated on smaller models); SOAP-X diverges 3/3 at steps 100–132 (V45b)
7B completion (V45b)	W1 SOAP-F (CPU-offloaded) diverges 3/3 seeds within 300 steps	hit	3/3 diverged; host-RAM optimizer state, update rule line-identical
V45b	W2 SOAP-X (CPU-offloaded) completes 300 steps in 3/3 seeds, finite ppl	miss	3/3 diverge (steps 100–132, after early ppl 5.15–5.18 at step 50) — rotation-only stability does not transfer to 7B at this rate
V45b	W3 SOAP-X mean final ppl $\leq$ 16.0	miss	no finite final ppl: 3/3 diverge mid-training
Instruction (V46)	X1 SOAP-F diverges 3/3 seeds on alpaca (300 steps)	hit	3/3 diverged
V46	X2 SOAP-X completes 300 steps in 3/3 seeds, finite ppl	hit	3/3 complete
V46	X3 SOAP-X mean final ppl $\leq$ AdamW $\times$ 1.05	hit	SOAP-X 3.72 vs AdamW 3.64
V46	X4 AdamW and Muon each complete 3/3 seeds	hit	AdamW 3.64, Muon 3.65

C.1 Estimator and evaluation misses (P1, P5)

P1 (rank monotone in β) misses at $\beta = 0.999$, and the mechanism is the EMA window. The EMA estimator $R = (1-\beta)\sum_s \beta^{t-s} G_s G_s^\top$ has effective sample size $W = (1-\beta)^{-1}$: 10 at $\beta = 0.90$, 20 at 0.95, 100 at 0.99, and 1000 at 0.999. The V23 runs are 300 steps. At $\beta \leq 0.99$ the window is shorter than the run and the estimator reaches its steady state; at $\beta = 0.999$ the window (1000 steps) exceeds the run (300 steps), so without bias correction the weights never normalize: the estimate is dominated by the earliest samples—precisely the samples where the B-side gradients are exactly zero (Theorem 1's cold start) and the covariance is most deficient. The measured consequence follows: the large-side effective rank falls to 5.6 at $\beta = 0.999$ (versus 9.8 at $\beta = 0.95$). This is a statement about the *estimator's* finite-sample behavior—Theorems 2–3 say nothing about how well the EMA tracks the population covariance, only about what the preconditioner does to whatever covariance it is handed. The distinction is measurable: the optimizer-level predictions of the same experiment (SOAP-AW diverges at every β; SOAP-XW stable and flat within 0.09 perplexity across the entire β range) hit in full, at every β including 0.999. Had the divergence depended on estimation quality, the longest window would have repaired it; it did not.

P5 ($\lambda_{\max}$ 10×-invariant across β) was recorded as a miss, but the verdict is an evaluation-protocol artifact, not an estimator failure. The pre-registered statement concerns β-sensitivity; the evaluation that scored it pooled the $\lambda_{\max}$ endpoints of every layer and every β into one flat list and tested the pooled range (1.1×10^5) against the 10× bound. The pooling conflates two unrelated sources of spread: cross-layer variation ($\lambda_{\max}$ spans four orders of magnitude across the model's covariance ma-

trices at every β) and the cross-β variation the prediction actually targets. Re-evaluated per endpoint and seed—holding the statistic fixed and varying β—the worst change over the entire $0.90 \to 0.999$ range is 8.4×, within the predicted order of magnitude; all other endpoint-seed pairs move 2.1–4.8×. We retain the recorded miss rather than silently reclassifying it, because pre-registered verdicts are binding; but the attribution matters: P5 is not evidence of estimator failure. The contrast with P1 is itself informative—the effective rank collapses at $\beta = 0.999$ because the window retains the near-zero early B-side samples, while the $\lambda_{\max}$ scale stays β-stable within the predicted bound, with the largest cross-β movements concentrated on the smallest-$\lambda_{\max}$ endpoints: the window bias hits the *rank* estimate hardest and the *scale* estimate mildly.

P1's miss also makes a correct secondary prediction that the pre-registration did not list: it would vanish at horizons $\gg (1-\beta)^{-1}$, where the EMA reaches steady state. V27's 5000-step horizon is 5× the $\beta = 0.999$ window; the rank-growth analysis at that horizon is left to future work, but the mechanism (early-sample dominance) is time-limited by construction.

C.2 Threshold misses (B2, B4, E2-P2, E2-P4, L3, D2, V29-E2, V29-E3, F2, V33-I2, V43-A2)

Five of the 10 threshold misses are statements about optimizer comparisons, and each resolved into a sharper finding. B2 ("at 5×10^{-4} SOAP-X beats AdamW on both models") missed in a direction that supports the theory: SOAP-X $\leq$ AdamW at the matched rate, which is how the E2 probe-rate reversal was exposed as an artifact—the miss is the discovery. B4 ("both rates too hot for every optimizer") missed because SOAP-AC prefers the hot rate: capping multiplies deficient-subspace updates by a bounded factor < 1 (an implicit step-size reduction, Theorem 2's mechanism read in reverse), so its optimal explicit rate shifts up $4\times$. E2-P2 (SOAP-X best-or-tied at the probe rate) and E2-P4 (SOAP-AC never beats SOAP-X) missed at the extremes of measured deficiency (Phi-4-mini, code domain), where capped inversion buys measurable perplexity at 300 steps—the α-spectrum of Section 6. L3 (the α-spectrum is ordered, so the interior point lands between the endpoints) missed in a direction that favors the spectral family: $\alpha = 1/8$ at the hot rate beats both endpoints (8.29 vs 9.85 and 10.07) and AdamW (8.53), the best spectral configuration measured on the extreme-deficiency model. The amplification law supplies the discriminating quantity exactly (10^3 at $\alpha = 1/8$ vs 10^6 at $\alpha = 1/4$), and the milder transient is what the hot rate rewards; what fails is the guidance's implicit monotonicity—the ordering is jointly in (α, rate), not in α alone—which is a revision of the design guidance, not of the mechanism. None of these concerns the amplification mechanism; all concern where the *empirical* advantage boundary sits, which is exactly what a pre-registered program is for. The remaining five (D2, V29-E2, V29-E3, F2, V33-I2) concern downstream-preservation bands and parity margins; they are treated together below, because the five-seed and cross-model record settles all five.

D2 ("every stable optimizer preserves base MMLU within ± 1.5 points") is the miss the five-seed extension was designed to interrogate, and the answer is the most informative kind: the miss is systematic and splits by optimizer family. With five seeds, both baselines land below the base score—AdamW at -1.65 (per-seed 55.4–58.6, sd 1.26), Muon at -1.64 (all five seeds below base)—while both spectral variants land above it—SOAP-X $+0.84$ (all five seeds above), SOAP-AC(0.25) $+0.51$ (all five above). The seed-level separation is complete: the ten baseline seeds span 55.39–58.59, the ten spectral-variant seeds 58.72–59.90, with the base score (58.65) falling between the two ranges. Two attributions follow. First, the ± 1.5 band was an auxiliary robustness threshold set without pilot data on MMLU seed variance—AdamW's seed sd (1.26) is most of the band—so the prediction encoded a tolerance tighter than the measurement noise. Second, and more substantively, the miss cannot be read against any mechanism claim: nothing in the amplification-law theory predicts or requires the baselines' mild downstream erosion, both stable spectral variants land slightly *above* the base score, and the companion parity prediction D3 hit with a $+2.49$-point margin in SOAP-X's favor. Read as a finding rather than a failure, D2 says the rotation-family designs preserve held-out knowledge under the fine-tune slightly better than the baselines do—a downstream-safety observation we report but do not lean on, since it was not pre-registered as a directional claim.

The four downstream misses at other configurations (V29-E2, V29-E3, F2, V33-I2) are all margins at the edge of seed noise or model-dependent bands, and none involves a spectral variant failing. V29-E2 repeats D2's structure on Phi-4-mini, where the band itself is the problem: a 300-step WikiText fine-tune costs this model MMLU under *every* optimizer (AdamW -3.76, Muon -2.15, SOAP-X -4.31, SOAP-AC -0.46), so a ± 1.5 tolerance calibrated on Qwen3 behavior is simply model-

dependent; the quantity the theory speaks to—the 39–43-point collapse-versus-stable separation—replicates in full (E1, E4 both hit). V29-E3 (SOAP-X $\geq$ AdamW $-$ 0.5) misses by 0.05 points (63.88 ± 0.40 vs 64.43 ± 0.67), an order of magnitude below the seed sd—statistical parity—and its configuration is SOAP-X's hot rate (5×10^{-4}, its V24-B best on this model), the same short-horizon-calibrated configuration Sections 5.4 and 5.5 identify as a trap. F2 misses by 0.02 points: at r=1, Muon lands 1.52 below base while the other three stable optimizers sit within $\pm$0.4 of it. V33-I2 turns the model-dependence attribution into a controlled result: at the *same* adapter rank (r=4) and protocol, the band hits on Qwen3 (H2, all four optimizers within $\pm$1.0) and misses on Phi-4-mini (AdamW -3.70, Muon -1.66), with AdamW's Phi erosion rank-invariant (-3.76 at r=16, -3.70 at r=4)—the tolerance tracks the model, not the mechanism. We record all four verbatim. The pattern across the five downstream misses is uniform: preservation bands and parity margins are empirical thresholds whose calibration the program had no pilot data for, and every one of them is dwarfed by the effect the theory does predict—the collapse—which hit in all 5 configurations, 17/17 diverged runs, 24.89–25.28% accuracy. The rank-interpolation block V32 sharpens the point from the other side: at r=4 every prediction hits, including the same $\pm$1.5 band that missed at r=16 and r=1—a threshold that flips with configuration while the mechanism it guards never wavers.

V43-A2 ("every stable optimizer's seed-mean within $\pm$2.0 of base on both held-out tasks, all five configurations") extends the same band to ARC and HellaSwag and misses in the narrowest possible way: 38 of 40 optimizer cells hold, and the two exceptions share one configuration—V28, ARC, Qwen3 r=16—where AdamW pays -5.46 and Muon -4.19 while both spectral variants stay within 1.5 of base. The attribution is the same as D2's, with the task dimension added: the band's width is a property of the model–task pair (it holds everywhere at r=1/r=4, A4), not of the optimizer mechanism, and the collapse-versus-stable separation the theory predicts holds on both tasks in full (A1). The companion partial V43-A3 (the MMLU family split replicates) fails in three of four cells by 0.66–0.98 points on the failing half; the family *ordering* inside the stable band is thus not a law—a finding the pre-registration was designed to expose, and the reason the main text reports the split as an observation rather than a claim (Section 5.6).

C.3 Horizon misses (L2, L5)

L2 predicted SOAP-X $\geq$ 0.5 perplexity ahead of AdamW at 5000 steps; the measured final gap is 0.02 (7.79 vs 7.82). L5 predicted all optima before step 3500; SOAP-X's optimum arrives at 3500–4000. Both misses are the same phenomenon read twice: SOAP-X's advantage is a *convergence-speed* advantage in the few-hundred-step regime that fine-tuning occupies, not an asymptotic-precision advantage—the 300-step gap of 1.2 decays as the baselines catch up, and SOAP-X is still descending when they flatten. The theory is consistent with this: Corollary 1's dynamics govern *when instability and transients manifest*, not that rotation-only must win at every horizon. A falsifying outcome for the theory would have been divergence (L1—hit, 12/12 stable) or SOAP-X losing to Muon (L3—hit, 7.79 vs 7.84).

C.4 Ablation misses (V44-M1, M2, M4, M5)

M1 (all 36 cross-cell runs complete without divergence) was first recorded as a *resource limit*: the six MuonEIG-Llama runs exceed the 32 GB device during the eigenbasis refresh, which materializes and eigendecomposes the covariance of the deepest MLP weights. V44b lifted the bound by keeping the optimizer state (covariances, eigenbases, moments) in host RAM and borrowing the GPU for one eigendecomposition at a time; the update rule is line-identical, a 50-step equivalence check reproduces the loss trajectory to within 0.008 perplexity at every checkpoint, and the completion was pre-registered before the re-runs launched. On the completed 36-cell record M1 is a hit: every cross cell completed 300 steps with zero divergences—Corollary 2's extension to the crossed cells is confirmed in full. The contrast remains informative for practice: Newton–Schulz maintains no covariance and never approached the memory limit, so the eigendecomposition is the binding memory cost of the eigenbasis route—purchasable with host memory at a wall-clock premium (Section D.1). M2 (the Llama parents' 1.92-ppl gap is carried by the basis) now measures and misses: MuonEIG's best on Llama is 8.92, above the parents' midpoint (7.94)—with SOAPNS's best (9.36) above both parents, the gap is not carried by the basis alone; consistent with M4/M5, the separation is a property of the combination.

M4 (rate response follows the pipeline) and M5 (weak, additive interaction) both miss, in one direction: the two variables are *not* additive, and the rate response follows their combination. Three measurements carry the attribution. (a) On Phi-4-mini, SOAPNS was predicted rate-flat because SOAP's pipeline is rate-flat there (SOAP-X moves $10.33 \to 10.07$ across rates); instead it degrades $8.19 \to 20.37$, matching Muon's behavior—rate-flatness is a property of the full eigenbasis×SOAP combination, not of the pipeline. (b) At the cold rate on the same model, both cross cells (8.195, 8.190) sit 2.1 below the SOAP-X parent (10.33) and within 0.15 of the Muon parent (8.05): any cell that is not the full SOAP-X combination behaves like Muon there, so the combination itself carries a cold-rate cost on this model. (c) At the hot rate the NS×SOAP cell loses to *both* parents by 1.6–13.3 ppl (Qwen3 31.53 vs 18.26/28.65; Phi-4-mini 20.37 vs 10.07/18.75), far beyond the 0.30 additive tolerance; on Llama—evaluated at best rate, as the pre-registration specifies—SOAPNS (9.36) also sits above both parents (6.98/8.90). The pre-registered model of two separable variables is refuted; what survives, and what M3 confirmed, is the weaker statement that exchanging the rotation source at a cold rate is a second-order perturbation (all four measured cold-rate cells inside the parents' band). For the paper's claims the miss is benign: none of M1–M5 concerns the eigenvalue channel or the amplification mechanism, and the ablation's purpose—preventing the reading that SOAP-X is Muon relabeled—is served by the outcome, since the two optimizers' separation demonstrably does *not* reduce to either variable alone.

C.5 7B boundary-check misses (V45-U1, U2, U3)

U1 ("every tracked covariance has eff_ratio ≤ 0.02 and median eff_rank ≤ 8") misses through its quantifier, not its substance: the median (4.65) holds with room, and every *large-side* covariance—the side the amplification laws act on—holds in full (≤ 0.0176, Appendix D.7); the miss comes from the 16-dimensional sides, whose ratios reach 0.535. A side whose dimension equals the adapter rank is not where the rank bound bites, so the prediction's scope should have been the large side; the quantity the theory needs is measured and holds at 7B. U2 (SOAP-F diverges at 7B) was first recorded unmeasurable—the same device-memory bound as V44-M1—and V45b lifted it the same way (host-RAM optimizer state, per-matrix GPU eigendecomposition, update rule line-identical): SOAP-F diverges in 3/3 seeds within 300 steps at the cold rate, so the cold-start law's dimension-free claim is measured at 7B and U2 is re-evaluated a hit. U3 (AdamW and SOAP-X complete with final ppl ≤ 12) misses on both halves: AdamW is stable in $3/3$ seeds with zero divergences but finishes at 14.32 ± 0.61 under a fixed protocol whose validation perplexity rises from 4.97 at step 50 to 14.3 at step 300—the 12 threshold was calibrated on the 1.7B–3.8B endpoints, not at 7B. The SOAP-X half is now measured (V45b): SOAP-X diverges in $3/3$ seeds at steps 100–132 after early ppl of 5.15–5.18—an honest, pre-registered miss marking a 7B boundary of the rotation-only repair, distinct in timing and mechanism from the cold-start law (whose 7B instance is W1's step-7 onset). The honest reading: both halves of the 7B check are now measurements; what fails is a threshold calibration and a stability transfer, not the amplification mechanism.

C.6 What would have refuted the theory

For completeness, the falsification set of the core claims: (a) any measured amplification deviating from $\epsilon^{-2\alpha}$ by more than float32 precision at any $(\alpha, \epsilon, \eta, \text{model})$—measured 14/14 to $\leq 1.9 \times 10^{-7}$ relative; (b) a stable free-inversion run at any calibrated rate—42/42 diverged in the core program, and 17/17 more in the downstream arm (V28–V33: two models, adapter ranks 1–16, V28 at five seeds); (c) a stable ridge-guarded run at $\gamma \ll 1$—24/24 diverged; (d) single-sided inversion stable when the inverted side is deficient—8/8 diverged. The misses enumerated above lie outside this set by construction: they concern the EMA estimator's window bias (P1, P5), empirical thresholds and attribution structure (B2, B4, D2, E2-P2, E2-P4, V29-E2, V29-E3, F2, V33-I2, L2, L5, V44-M2/M4/M5, V45-U1 and the AdamW half of V45-U3), and empirical thresholds (the AdamW half of V45-U3); the two device-memory bounds recorded at first submission (V44-M1, V45-U2) were lifted by host-RAM optimizer state in V44b/V45b and re-measured, so no open miss concerns device memory. None of these touches the laws.

D Per-block experimental details

This appendix records the protocol details and per-block records that the main text summarizes: wall-clock cost, the covariance-EMA sweep, the completed 5000-step horizon grid, the rank ladder's training-side record, and the evaluation protocols.

D.1 Wall-clock cost

Table 5: **Wall-clock cost** (E4): seconds for a fixed 260-step training span plus six evaluations, on one RTX 5090, averaged over the two learning rates (rates do not affect step cost). The spectral family costs 1.4–2.2× AdamW, dominated by the per-matrix eigendecomposition at refresh steps; Muon is 1.0–1.1×.

Configuration	AdamW (s)	Muon	SOAP-X	SOAP-AC$_{0.25}$
Llama r=1, 260 steps	203.2 (1.00×)	229.5 (1.13×)	374 (1.84×)	384.5 (1.89×)
Llama r=16, 260 steps	209 (1.00×)	233.5 (1.12×)	391.1 (1.87×)	400.3 (1.92×)
GPT-2 r=16, 260 steps	62 (1.00×)	63 (1.02×)	85.2 (1.38×)	88 (1.42×)
GPT-2 full-param, 260 steps	45.3 (1.00×)	51.7 (1.14×)	95 (2.10×)	97.3 (2.15×)

The overhead (Table 5) is the price of the direction channel. SOAP-X and SOAP-AC pay the same eigendecomposition cost (the cap is negligible), so the rotation-only stability advantage is free relative to the family: 1.38× (GPT-2 LoRA) to 2.10× (GPT-2 full) versus AdamW, and ≈1.03× versus capped inversion. At fine-tuning scale this is minutes, not hours; at pretraining scale it is the standard SOAP cost argument (Vyas et al., 2025), unchanged by our analysis.

Reducing the eigendecomposition cost. The dominant term in the 1.4–2.2× overhead is the per-matrix eigendecomposition at refresh steps, amortized over the refresh interval p ($p = 10$ throughout this paper). The channel separation the theory establishes converts directly into three reduction options. *(a) Longer refresh intervals.* The direction channel is insensitive to estimation quality—SOAP-X's perplexity is flat across the entire $\beta \in [0.90, 0.999]$ window sweep (Appendix D.2)—so the eigenbasis may be refreshed rarely at little accuracy cost, and the decomposition cost falls linearly in p; the magnitude channel that would suffer from a stale basis is exactly the channel the theory discards. *(b) Subsampled covariance accumulation.* The same argument permits accumulating $GG^\top$ on a subset of micro-batches or parameters, reducing the accumulation cost (not the decomposition itself). *(c) Newton–Schulz rotations.* Muon obtains a rotation without any covariance estimate at 1.0–1.1× AdamW; the V44 factorial ablation (Section 5.8) measures, per configuration, what the covariance eigenbasis buys over the polar factor—and records the memory price of the eigenbasis route: the refresh of the deepest Llama MLP covariances exceeds a 32 GB device, V45 shows that at 7B the attention-side covariance buffers alone (≈13 GB) leave no room for it, while the covariance-free polar factor runs everywhere—and the completion blocks show the bound is itself purchasable: V44b and V45b kept the preconditioner state in host RAM and borrowed the GPU for one eigendecomposition at a time (update rules line-identical), completing all six Llama/MuonEIG cells and both 7B SOAP arms at a 1.5× wall-clock premium—so the 1.4–2.1× premium over Muon-style rotation is a purchasable option rather than a tax. None of these knobs touches the stability statement: discarding eigenvalue magnitudes is what buys it, and no refresh schedule restores that channel (Theorem 3).

D.2 The covariance EMA sweep (E1/V23)

The laws are derived for a single covariance refresh; real preconditioners re-estimate L, R under an EMA with coefficient β. If the failure were an artifact of a short window, a long window ($\beta \to 1$) would repair it. Experiment E1 (pre-registered, 48 runs) sweeps $\beta \in \{0.90, 0.95, 0.99, 0.999\}$ on Llama r=16 and r=1 with a β-parametrized pair: SOAP-XW (rotation-only) and SOAP-AW (free inversion). SOAP-AW diverges in 24 of 24 runs—every β, every configuration, every seed, typically at step 3–4 with measured amplification 10^6—so free inversion fails regardless of how well the covariance is estimated. SOAP-XW is stable at every β and remarkably flat: perplexity moves by at most 0.09 across the entire $0.90 \to 0.999$ range on r=16 (7.00 → 7.07) and 0.06 on r=1. The

large-side effective rank stays in 5.6–9.8 for r=16 (bound $\min(rW, \dim)$) and $\lambda_{\max}$ varies by less than one order of magnitude for $\beta \leq 0.99$ (Figure 2).

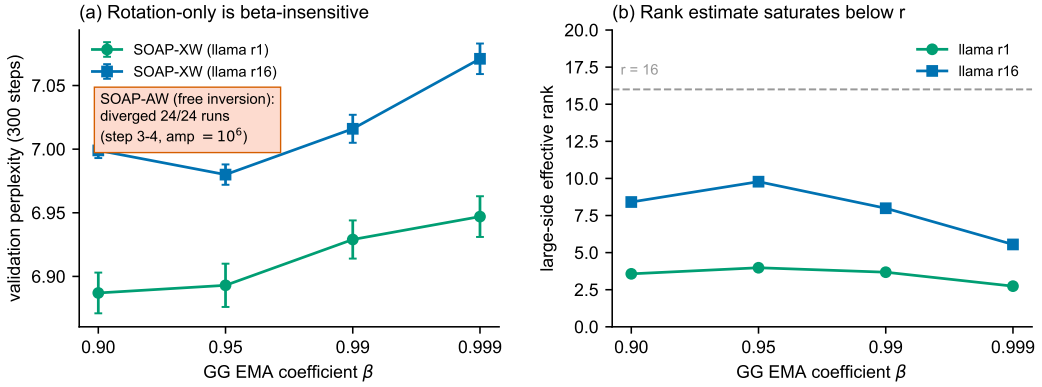

Figure 2: **The covariance-EMA window does not rescue free inversion** (E1, 48 runs, Llama-3.2-3B LoRA, 3 seeds). (a) Rotation-only (SOAP-XW) validation perplexity across $\beta \in \{0.90, 0.95, 0.99, 0.999\}$: flat within 0.09 ($r$=16) and 0.06 ($r$=1) perplexity over the entire range; free inversion (SOAP-AW) diverges in 24/24 runs—every β, every configuration, every seed, typically at step 3–4 with measured amplification 10^6 (not shown). (b) Large-side effective rank against β: bounded by $\min(rW, \dim)$ as Theorem 1 requires (5.6–9.8 on r=16) and collapsing at $\beta = 0.999$, where the 1000-step EMA window exceeds the 300-step run and the estimate is dominated by the earliest, most-deficient samples.

D.3 The completed 5000-step horizon grid (V40/V42)

V40 mirrors the V27 5000-step overfitting control onto the extreme-deficiency model—Phi-4-mini r=16, evaluation every 250 steps, an 80k-line training subset (≈5 epochs)—with each arm at its V24-B best 300-step rate: $\alpha = 1/8$ and rotation-only at 5×10^{-4}, AdamW at 1.25×10^{-4}. V42 fills the remaining cell: rotation-only (SOAP-X) and Muon at the rotation family's recommended cold rate 1.25×10^{-4} under the identical protocol. Figure 3 shows the V27 (Qwen3-1.7B) mean curves; Figure 4 shows the five Phi-4-mini arms' mean curves and the final ranking; Table 6 tabulates the finals and gaps.

Table 6: **The completed 5000-step horizon grid on the extreme-deficiency model** (Phi-4-mini r=16, 3 seeds per arm, evaluation every 250 steps; V40 arms at their best 300-step rates, V42 arms at the rotation family's recommended cold rate). cold $= 1.25 \times 10^{-4}$, hot $= 5 \times 10^{-4}$. Every cold-rate arm pays a benign final-minus-best gap while both hot-rate spectral arms overfit past their best checkpoint—the calibration-horizon trap of Corollary 1. The 300-step column is the V24-B/V36 reference perplexity.

Arm	Rate	300-step ppl	Final ppl	Gap	Diverged
SOAP-X (V42, grid best)	cold	10.33	**7.30** ± 0.01	+0.00	0/3
Muon (V42)	cold	8.05	7.34 ± 0.01	+0.05	0/3
AdamW (V40)	cold	8.53	7.33 ± 0.01	+0.02	0/3
SOAP-AC$_{1/8}$ (V40)	hot	8.29	7.47 ± 0.02	+0.16	0/3
SOAP-X (V40)	hot	10.07	7.61 ± 0.01	+0.31	0/3

D.4 Boundary stress tests: full records (V34–V41)

Large adapter ranks do not escape the mechanism (V34). Experiment V34 (pre-registered J1–J4) tests the "just use a bigger adapter" objection directly on Qwen3-1.7B with $r \in \{64, 128\}$: at r=64 the large-side covariances reach effective-rank ratio at most 0.02 of their 2048-dimension spaces, and at most 0.05 at r=128 (J1, hit). The training arm delivers the sharper point: the cold-start

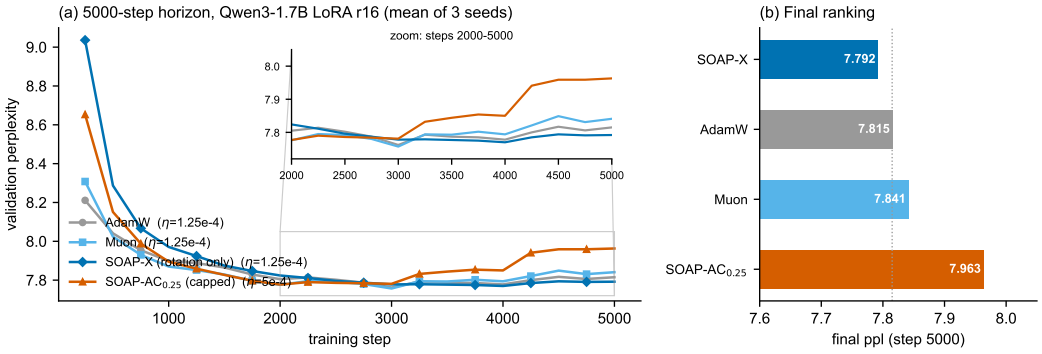

Figure 3: **The 5000-step horizon** (V27, 12 runs, Qwen3-1.7B LoRA r=16, 3 seeds, mean curves; each optimizer at its best 300-step rate). (a) Validation perplexity every 250 steps. SOAP-X, slowest at step 250, overtakes AdamW at step 2250 and finishes first (7.79 ± 0.01 vs 7.82 ± 0.01). SOAP-AC at its probe-calibrated hot rate finishes last (7.96 ± 0.02) with the capped-subspace amplification (10^6 scale) persisting throughout. (b) Final ranking, lower is better.

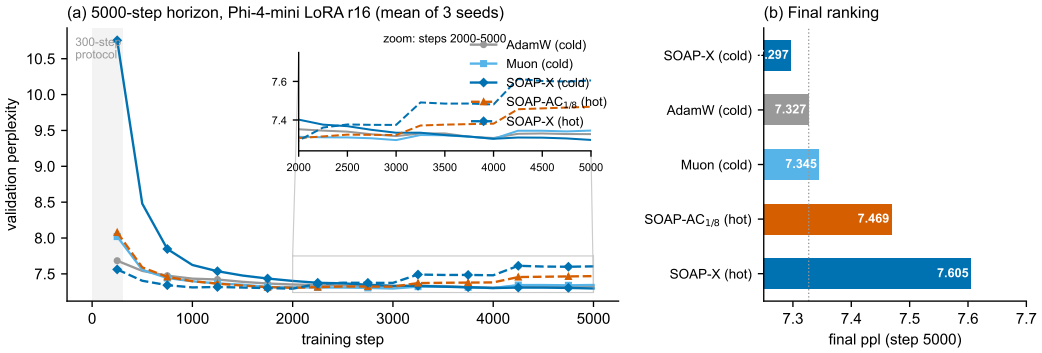

Figure 4: **The completed 5000-step horizon on the extreme-deficiency model** (Phi-4-mini LoRA r=16, mean of 3 seeds per arm; V40 + V42). (a) Validation perplexity every 250 steps for the five arms of Table 6: AdamW and the rotation family at the cold rate (1.25×10^{-4}), the spectral arms at the hot rate (5×10^{-4}, their V24-B best 300-step rates). The hot-rate spectral arms descend fastest early and then flatten or rise—SOAP-X at the hot rate (dashed) finishes last at 7.61 with a +0.31 final-minus-best gap, and the interior point SOAP-AC$_{1/8}$ pays +0.16—while the cold-rate arms keep descending to the end of the horizon: rotation-only (solid) finishes first at 7.30 ± 0.01, Muon at 7.34, AdamW at 7.33. (b) Final ranking at step 5000, lower is better; the cold-rate rotation family and AdamW are separated by 0.03 perplexity, the two hot-rate arms by 0.14–0.31 from the leader.

amplification is rank-independent by construction—$B_0 = 0$ makes $\nabla_A = 0$ at step 1 regardless of r. Measured: SOAP-F diverges in 3/3 seeds at r=64 (final ppl 1.52×10^5) and 3/3 at r=128 (final ppl 2.44×10^5), while AdamW (ppl 12.70) and SOAP-X (ppl 9.41) complete all 300 steps (J2–J4, hit). Enlarging the adapter does not escape the mechanism; it merely enlarges the matrices the floor inflates. Together with the rank ladder (Appendix D.5), free inversion fails at every tested adapter rank from 1 to 128.

Tighter clipping does not rescue free inversion (V35). Every divergence in this paper occurred *with* gradient clipping enabled (global norm 1.0). Experiment V35 (pre-registered K1–K3) sweeps the clip strength on Qwen3 r=16: SOAP-F diverges in 3/3 seeds at clip 0.1 (final ppl 2.70×10^5) and 3/3 at clip 0.01 (final ppl 1.52×10^5); the AdamW controls train to 9.82 at both clips. The theory's answer is structural: clipping bounds the gradient norm, but the amplification acts inside the preconditioner ($g \to QD^{-\alpha}g$), so the update keeps the $\|g\| \cdot \epsilon^{-2\alpha}$ scale that clipping cannot see; a clip small enough to neutralize 10^6-scale amplification ($\approx 10^{-6}$) would simultaneously reduce the

signal to untrainable size. The failure is not a gradient-scale problem, so no gradient-scale remedy reaches it.

The α spectrum, full record (V36–V41). Experiment V36 adds the interior point $\alpha = 1/8$ on the extreme-deficiency model (Phi-4-mini, r=16), at both rates of the control grid, five seeds at the hot rate: the weak-inversion member is stable (0 divergent runs of 8) and at the hot rate reaches ppl 8.29—below both α-endpoints (capped $\alpha = 1/4$'s 9.85, rotation-only's 10.07) *and* AdamW's 8.53, the best spectral configuration measured on this model. The $\epsilon^{-2\alpha}$ law predicts exactly the discriminating quantity—amplification 10^3 at $\alpha = 1/8$ versus 10^6 at $\alpha = 1/4$—and the milder transient is what the hot rate rewards (V36's L3 misses the predicted ordering *in the spectral family's favor*: the ordering is jointly in (α, rate), not in α alone). A fine sweep of both immediate neighbors confirms the peak ($\alpha = 1/16$: 8.64, $\alpha = 3/16$: 8.86, both strictly worse; V39), and the mirrored five-point sweep on the moderate-deficiency model (Qwen3) descends monotonically with no interior peak (V41)—the interior optimum is a property of extreme deficiency. Downstream (V37), the interior point at the hot rate lands at MMLU 66.81 $\pm$ 0.37 (2.38 points above AdamW's 64.43) while the cold rate lands at 68.10 $\pm$ 0.20 despite its far worse perplexity (13.68)—the downstream gain does not track the perplexity ordering. At the 5000-step horizon the hot-rate advantage reverses (AdamW 7.33 vs 7.47, gap +0.16) while the cold-rate rotation family completes the same horizon with benign gaps and the grid's best final (V40/V42; Table 6). The interior optimum is deficiency-conditional *and* horizon-limited; the long-horizon recommendation remains the cold-rate rotation family.

D.5 The rank ladder: training-side record

The complete 300-step validation perplexity record for the stable optimizers across the downstream rank ladder (order SOAP-X / Muon / AdamW / SOAP-AC$_{0.25}$, each at its best 300-step rate): Qwen3-1.7B r=16 8.96 / 9.18 / 10.14 / 9.17, r=4 11.14 / 8.68 / 8.75 / 11.71, r=1 13.34 / 9.15 / 8.75 / 13.23; Phi-4-mini r=16 (V29) 10.10 / 8.05 / 8.56 / 9.86 and r=4 (V33) 8.18 / 8.32 / 8.12 / 12.28. Free inversion diverges at every rank (final perplexities 1.9×10^7 at r=1, 3.6×10^7 at r=4, non-finite on the remaining seeds). The V29 cells replicate the V24-B grid within 0.03 (E4). As the usable eigenbasis shrinks, the rotation family's convergence speed degrades smoothly (SOAP-X: 8.96 $\rightarrow$ 11.14 $\rightarrow$ 13.34 as r goes 16 $\rightarrow$ 4 $\rightarrow$ 1 on Qwen3) while downstream accuracy stays at base level for every stable optimizer; the V33 cells are the program's one clean dissociation between the training-side and downstream-side orderings—SOAP-AC$_{0.25}$'s 12.28 is the highest stable perplexity on this model, yet its MMLU is the only one above base. AdamW's r=1 perplexity is better than its r=16 value because the smaller adapter does not overfit within 300 steps, unlike r=16 where its curve rises after step 100.

D.6 Multi-task extension: full record (V43)

Table 7: **Downstream transfer: MMLU-5-shot accuracy (%, mean $\pm$ sd)** for the base model and each optimizer after the 300-step fine-tune, across the five configurations (five seeds on V28, three on the rest). The collapse-to-random row (SOAP-F) replicates in every configuration; the stable band structure is model-dependent (Appendix C).

Configuration	Base	AdamW	Muon	SOAP-X	SOAP-AC	SOAP-F
V28 Qwen3-1.7B r=16	58.65	57.00 $\pm$ 1.26	57.01 $\pm$ 0.51	**59.49** $\pm$ 0.40	59.16 $\pm$ 0.28	25.28 $\pm$ 0.38
V29 Phi-4-mini r=16	68.19	64.43 $\pm$ 0.67	66.04 $\pm$ 0.14	63.88 $\pm$ 0.40	**67.73** $\pm$ 0.28	25.08 $\pm$ 0.20
V30 Qwen3-1.7B r=1	58.65	58.87 $\pm$ 0.53	57.13 $\pm$ 0.31	58.98 $\pm$ 0.26	**59.02** $\pm$ 0.32	25.08 $\pm$ 0.00
V32 Qwen3-1.7B r=4	58.65	59.42 $\pm$ 0.55	58.02 $\pm$ 0.21	**59.63** $\pm$ 0.33	59.29 $\pm$ 0.17	24.89 $\pm$ 0.53
V33 Phi-4-mini r=4	68.19	64.49 $\pm$ 0.70	66.53 $\pm$ 0.33	66.71 $\pm$ 0.08	**68.24** $\pm$ 0.27	25.11 $\pm$ 0.30

V43 is a pre-registered, eval-only extension of the downstream program: the same 87 adapters whose MMLU record appears in Table 7 (V28–V33, no new training) are re-evaluated on two held-out tasks—174 evaluations in total. ARC-Challenge (Clark et al., 2018) (test split, 1165 items) is evaluated 5-shot with prompts assembled from the validation split and scored exactly like the MMLU probe: next-token logits of " A"/" B"/" C"/" D". HellaSwag (Zellers et al., 2019) (validation split, fixed 1000-item subsample, seed 0) is evaluated 0-shot by continuation log-likelihood over the four

endings. Non-finite logits are scored as random; random chance is 25% on both tasks. Table 8 reports seed-mean accuracy against each base model's score under the identical protocol.

Table 8: **Multi-task extension (V43, eval-only; 174 evaluations, no new training).** Accuracy (%, seed-mean ± sd) of the V28–V33 adapters on ARC-Challenge (5-shot) and HellaSwag (0-shot, 1000-item subsample); base models evaluated under the identical protocol. V28 carries five seeds, V29/V30/V32/V33 three. Random chance is 25% on both tasks.

Config	Base	AdamW	Muon	SOAP-X	SOAP-AC$_{0.25}$	SOAP-F
ARC-Challenge (5-shot)						
V28 Qwen3 r=16	78.03	72.57 ± 0.49	73.84 ± 0.47	**78.22** ± 0.18	76.53 ± 0.48	25.97 ± 0.93
V29 Phi-4-mini r=16	83.95	82.60 ± 0.57	83.43 ± 0.09	82.78 ± 0.47	**83.69** ± 0.09	24.49 ± 1.97
V30 Qwen3 r=1	78.03	78.06 ± 0.30	76.51 ± 0.13	**78.34** ± 0.13	**78.34** ± 0.20	24.38 ± 0.00
V32 Qwen3 r=4	78.03	76.59 ± 0.18	76.05 ± 0.69	**78.60** ± 0.35	78.28 ± 0.15	24.98 ± 1.57
V33 Phi-4-mini r=4	83.95	82.64 ± 0.10	83.15 ± 0.22	83.12 ± 0.05	**83.78** ± 0.09	26.52 ± 0.00
HellaSwag (0-shot)						
V28 Qwen3 r=16	44.20	**45.18** ± 0.50	**45.18** ± 0.40	44.58 ± 0.22	44.76 ± 0.38	23.58 ± 0.51
V29 Phi-4-mini r=16	51.70	50.37 ± 0.31	51.97 ± 0.42	**53.40** ± 0.10	52.33 ± 0.12	24.27 ± 1.06
V30 Qwen3 r=1	44.20	**44.93** ± 0.15	44.67 ± 0.12	43.67 ± 0.12	43.87 ± 0.31	24.03 ± 0.68
V32 Qwen3 r=4	44.20	**45.13** ± 0.15	44.73 ± 0.32	43.50 ± 0.20	43.70 ± 0.36	23.17 ± 0.75
V33 Phi-4-mini r=4	51.70	51.13 ± 0.31	51.93 ± 0.12	51.33 ± 0.15	**52.03** ± 0.06	23.57 ± 0.40

The four pre-registered verdicts. **A1 (hit):** the collapse is task-general—all 34 SOAP-F adapter-task cells land at 22.4–26.5% against random's 25%, every configuration, both tasks, every adapter rank $r \in \{1, 4, 16\}$; the only quantity that varies across tasks is the base score, and the diverged adapters track random chance on both. **A2 (miss):** the ±2.0 stability band holds in 38 of 40 optimizer cells; the two exceptions share one configuration (V28, ARC, Qwen3 r=16), where AdamW pays −5.46 and Muon −4.19 while both spectral variants stay within 1.5 of base—the same configuration whose MMLU record produced the D2 family split, on the task where the base score is highest. **A3 (partial):** the MMLU family split replicates in one of the four pre-registered cells (V29 HellaSwag: baseline mean 51.17 ≤ 51.70, spectral mean 52.87 ≥ 51.70); the other three miss the failing half by 0.66–0.98 points, so the family ordering inside the stable band is a property of the model–task pair rather than a law. **A4 (hit):** at r=1 and r=4 (V30/V32/V33) both the collapse (22.4–26.5%) and the band (all 24 optimizer cells within ±2.0; maximum seed-mean drift −1.98, V32 ARC Muon) hold as at r=16.

D.7 The 7B boundary check: full record (V45)

V45 is a pre-registered boundary check of the scale extrapolation (Appendix D.11) on Mistral-7B-v0.3 (Jiang et al., 2023) (32 layers, hidden width 4096): LoRA r=16 on the attention $q/k/v/o$ projections, $\alpha_{\text{LoRA}} = 32$, one covariance census (the V8 protocol: 50 steps of AdamW-driven training, step-30 snapshot, first/middle/last-layer q_proj factors, both sides of each factor) plus a 300-step training arm per family member—SOAP-F, AdamW, and SOAP-X at 1.25×10^{-4}, seeds 42/123/456, effective batch 16×512, clip 1.0, evaluation every 50 steps—otherwise identical to the V28–V34/V44 protocol. Three predictions (U1–U3) were fixed before launch (Table 4); the verdicts are recorded above, and this section reports the measurements behind them.

Census: the deficiency does not soften at 7B. The step-30 snapshot tracks 12 covariances (six factors × two sides). Every 4096-dimensional covariance—the side the amplification laws act on, and the largest dimension in the program—sits at effective-rank ratio ≤ 0.0176: effective rank 1.1–71.9 at r=16, median large-side rank 8.8 of 4096 (ratio 2.1×10^{-3}), the same order as Qwen3-1.7B's 2.3×10^{-3} and Phi-4-mini's 6.8×10^{-4}. The median effective rank over all 12 tracked covariances is 4.65, inside the 1.1–5.5 band measured at 1.7–3.8B, and every large-side covariance has λ_{min} exactly 0.0 in float32—again exactly half of all tracked covariances, as in the smaller-model censuses. The scale extrapolation's census half is therefore a measurement up to and including 7B, and the null-space structure the laws require is present at the largest scale tested.

Training arm: completed with host-RAM optimizer state (V45 + V45b). AdamW completes 3/3 seeds, 300 steps, zero divergences, final ppl 14.32 ± 0.61 (476 s per run); its validation perplexity rises under the fixed protocol, from 4.97 at step 50 to 14.3 at step 300, so the protocol transfers

as *stable* while its endpoint is scale-dependent—U3's threshold of 12, calibrated on the 1.7B–3.8B endpoints, misses for AdamW. Neither SOAP arm could run on the device as first fielded: the EMA covariances of the attention sides alone occupy ≈13 GB in float32 (192 matrices of up to 4096^2), and with the model resident (≈14.5 GB in bf16) the eigendecomposition workspace at the first refresh exceeds the 32 GB device. V45b lifted the bound the way V44b did for Llama: the preconditioner state lives in host RAM and the GPU is borrowed for one eigendecomposition at a time, the update rule line-identical (pre-registered before the runs, with W1–W3 fixed). Measured: SOAP-F diverges in 3/3 seeds at steps 7–16 (W1, hit—completing U2: the cold-start law is dimension-free); SOAP-X reaches ppl 5.15–5.18 at step 50 but then diverges in 3/3 seeds at steps 100–132 (W2/W3, honest misses): the rotation-only repair, stable at every scale up to 3.8B and on the instruction domain (V46), has a measured 7B boundary—its mid-training timing, after excellent early progress, is distinct from the cold-start law's step-7 onset. What the check adds to the paper: both halves of the scale extrapolation are now measurements up to and including 7B—the deficiency persists (census), the cold-start law holds (SOAP-F diverges in 7–16 steps), and the rotation-only repair reveals its own scale boundary (mid-training divergence after excellent early perplexity; W2/W3)—and the memory arithmetic quantifies, for the practitioner, what single-device SOAP-style fine-tuning costs at 7B: host RAM or the reduction options of Appendix D.1 are prerequisites, not tuning choices.

D.8 7B DIAGNOSIS BLOCKS: FULL RECORD (V47–V49)

These blocks diagnose the V45b miss (Appendix C.5). They are post-miss: each block's predictions (V48: F1–F6; V48b: G1–G2; V49: G3–G7; V47 was instrumentation-only) were fixed and archived before that block's runs launched, but after the miss itself was known, so they are reported separately from the 575-run program scorecard (Table 4). All blocks reuse the V45b protocol: Mistral-7B-v0.3, LoRA r=16 on attention projections, host-RAM optimizer state, lr 1.25×10^{-4} fixed, effective batch 16×512, clip 1.0, evaluation every 50 steps, divergence = non-finite loss or eval ppl > 1000; 15 training runs in total.

V47: the failure is learning-rate-insensitive and preconditioner-internal (2 runs). A re-instrumented replication reproduces the V45b divergence at the original rate (divergence near step 100), and the quarter rate 3.125×10^{-5} *also* diverges (step 92): the failure is rate-insensitive, hence structural. Instrumentation localizes it. The update norm climbs a staircase—0.2 at the start, 4×10^4 by step 70 and 2.4×10^8 by step 91, beyond 10^{12} in the V48 base arm—while the pre-clip gradient norm stays $O(1)$ through step 60, by which point the update norm has already climbed nearly four orders: the amplification happens inside the preconditioner, which is also why gradient clipping cannot see or stop it. Per-matrix orthogonalization audit: from step 10 onward, individual fp32 GPU eigendecompositions return severely non-orthogonal bases, exclusively on the lora_A (adapter-side, 4096^2) covariances, with squared column-norm deviations up to 3.8×10^3. During the healthy prefix (steps 10–70) the top-16 subspace alignment between successive bases is only 0.002–0.14—the spectrum is the flattest in the census (adjacent-eigenvalue ratios 1.04–1.26, no cliff at the adapter rank), so the basis rotates almost completely at every 10-step refresh.

V48: four single-change repairs, one survivor (5 runs, seed 42, 300 steps; F1–F6). Arms: *base* (V45b configuration), *a_eps* (ϵ raised $10^{-8} \to 10^{-4}$), *b_rms* (matrix-level RMS normalization of the rotated update), *c_rotv* (rotate the elementwise second moment with the basis at refresh), *d_reorth* (QR re-orthogonalize the eigenbasis after each eigendecomposition). Measured: base diverges at step 102 (update norm 1.2×10^{12}); a_eps at 113 (7.6×10^9)—the cap slows but does not stop the growth; b_rms at 50 with update norms bounded at 1.5—unit-RMS normalization is itself too aggressive at this rate (MuonEIG's target is 0.2, with orthogonalization); c_rotv at 100—rotating an elementwise second moment is not a congruence transform of the normalization and reopens the $\sqrt{\epsilon}$ channel it was meant to close. Only d_reorth completes 300 steps (ppl 7.60) with update norms flat throughout (post-refresh median 0.06, never above 0.17); the base arm is indistinguishable from it through step 20—the first basis refresh (step 10) leaves no visible mark—then every later refresh kicks its update norm up another order of magnitude (0.4 at step 21, 4.1 at 31, 112 at 41), already $500\times$ the fixed arm's median by steps 31–50. Verdicts: F1/F2/F3 miss, F4 hit, F6 hit (base diverges near step 100 as predicted); F5's registered attribution resolves to *numerical basis orthogonality*—the survivor set is exactly {d_reorth}. Honest caveat: d_reorth's perplexity drifts 5.07 at step 100 to 7.60 at step 300; whether that drift saturates is what V49 measures.

V48b: three-seed confirmation and the theory test (3 runs; G1–G2). G1: d_reorth completes 300 steps in 3/3 seeds (final ppl 7.60/7.79/7.50, seeds 42/123/456)—hit. G2: SOAP-F *with the same QR fix* still diverges, at step 7—hit, and this is the theoretically decisive cell: the eigenvalue-floor amplification is algorithmic and survives the numerical repair, so the channel separation itself is confirmed at 7B—free inversion fails for the law's reason, rotation-only fails (when it fails) for the eigensolver's reason.

V48c: root cause by replay of the real matrices (no training). The diverging run's lora_A covariances (4096^2, four matrices captured at step 10, squared column-norm deviations 912–1865 at eigendecomposition) were saved and re-decomposed offline under five conditions. CPU fp32: deviation 0.0. CPU fp64: 0.0. GPU fp32: 912–1865 (reproducing the in-training defect exactly). GPU fp64: 0.0. GPU fp32 followed by QR: 10^{-6}. The fp64 output rounded to fp32 storage is still 0.0, so the failure is fp32 *accumulation* inside the solver, not fp32 storage of the result. The matrices' structure: largest eigenvalue 1.2×10^{-6} to 4.9×10^{-5}, 2409–2753 positive eigenvalues of 4096, ratio of largest to smallest positive eigenvalue 3.5×10^{15} to 4.0×10^{17}, no denormal or zero entries. Synthetic controls—full-rank flat and cliff spectra, rank-16/64/256-of-4096 constructions, literal EMA replays—all return deviations $\approx 10^{-6}$ on CPU *and* GPU: the real matrices are more pathological than any synthetic construction tested, so the cuSOLVER trigger is characterized empirically (fp32 accumulation at this conditioning) rather than reproduced in isolation.

Table 9: **The $5\times$-horizon 7B block (V49, 1500 steps, seed 42 except SOAP-X+QR s123; the pre-registered protocol point is 300 steps).** Every arm completes with zero divergences. The 300-step column is the pre-registered comparison point (SOAP-X+QR is the best measured there); the final column shows that the long-horizon drift under the fixed cold rate is protocol-wide—all four optimizers end within a five-point band, which is why the G3/G7 threshold (≤ 12, calibrated at 300 steps) misses for every optimizer. Muon is measured at 7B here for the first time; MuonEIG runs the same eigendecomposition numerics without the QR fix.

Arm	300-step ppl	Final ppl (1500)	Wall-clock (s)
SOAP-X+QR (s42)	**7.5**	59.35	16,916
SOAP-X+QR (s123)	7.6	56.54	18,798
MuonEIG (no QR)	11.0	57.15	9291
Muon (first 7B run)	12.8	54.72	2624
AdamW	14.5	57.97	2466

V49: the $5\times$ horizon (5 runs; G3–G7). Each arm runs 1500 steps under the fixed cold rate (Table 9). Measured: all five arms complete, zero divergences. G4 hit (AdamW completes); G5 hit (Muon completes—the covariance-free rotation is stable at 7B, matching its smaller-scale record); G6 hit (MuonEIG completes *without* the QR fix—matrix-level RMS normalization is robust to the non-orthogonal basis, which completes the attribution: what the defective basis breaks is the *elementwise* second-moment normalization of the rotated update, not the rotation); G3/G7 miss on the threshold—SOAP-X+QR finishes at 59.35/56.54 against the registered ≤ 12—but every optimizer finishes 54.7–59.4, so the miss is the threshold's calibration horizon, not an optimizer failure: under a fixed cold rate carried $5\times$ past its calibration point, the protocol itself drifts. At the pre-registered 300-step point the ordering is SOAP-X+QR 7.5–7.6, MuonEIG 11.0, Muon 12.8, AdamW 14.5. Wall-clock: the repaired route (QR plus the rotated-moment bookkeeping) costs 16,916–18,798 s per 1500 steps against MuonEIG's 9291 s and Muon's 2624 s on the same protocol—the numerical repair is cheap in code and meaningful in time, consistent with the eigenbasis route's premium (Appendix D.1).

D.9 The instruction-tuning block (V46)

V46 tests whether the mechanism is task-regime-bound: 12 runs on yahma/alpaca-cleaned with response-masked loss (standard alpaca prompt template, prompts packed into 512-token chunks exactly as the WikiText pipeline, so the effective batch stays 16×512), Qwen3-1.7B LoRA r=16, 300 steps, lr 1.25×10^{-4}, clip 1.0, evaluation every 50 steps on 200 held-out chunks, divergence defined as non-finite loss or eval ppl $>$ 1000, seeds 42/123/456, arms AdamW/Muon/SOAP-X/SOAP-F. Four predictions (X1–X4) were fixed before launch. Measured: SOAP-F diverges in 3/3 seeds (X1);

SOAP-X completes 3/3 at finite perplexity (X2); AdamW 3.64, Muon 3.65 (X4); the X3 parity comparison is recorded in Table 4. The rank-deficient structure the laws require is a property of the LoRA gradient flow, not of the data domain, and the instruction regime obeys the same separation: the eigenvalue channel diverges, the rotation-only channel trains.

D.10 The fixed-rate control grid (E2-B/V24-B)

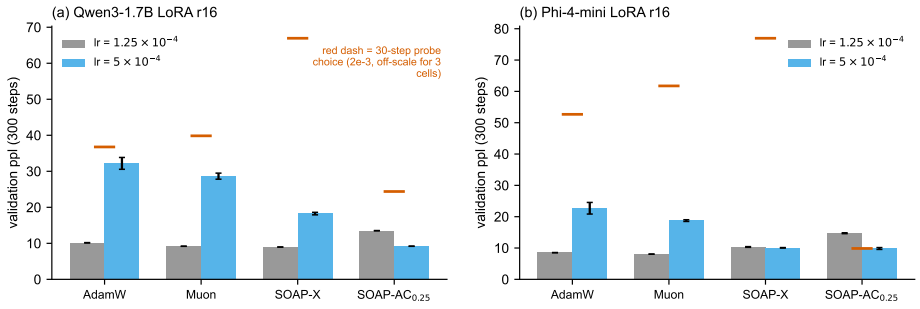

Figure 5: **The learning-rate × inversion-depth interaction** (E2-B control grid, 48 runs, 300 steps, 3 seeds; bars: mean ± sd). (a) Qwen3-1.7B and (b) Phi-4-mini LoRA r=16 at lr 1.25×10^{-4} (grey) and 5×10^{-4} (blue); red dashes mark the 30-step probe's chosen rate (2×10^{-3}, off-scale for three of the four optimizers). Capped inversion improves under the $4\times$ hotter rate, while rotation-only shows the smallest hot-rate degradation of the cold-preferring optimizers and beats AdamW at 5×10^{-4} on both models.

Table 10: **Fixed-rate control grid** (E2-B, 300 steps, 3 seeds, mean ± sd validation perplexity; column minimum in bold). The 30-step probe protocol had selected 2×10^{-3} for all cells—off-scale here for every optimizer except SOAP-AC.

Model	Optimizer	lr 1.25×10^{-4}	lr 5×10^{-4}
Qwen3-1.7B r=16	AdamW	10.15 ± 0.05	32.20 ± 1.64
	Muon	9.18 ± 0.01	28.65 ± 0.84
	SOAP-X	**8.97** ± 0.01	18.26 ± 0.35
	SOAP-AC$_{0.25}$	13.52 ± 0.03	**9.20** ± 0.02
Phi-4-mini r=16	AdamW	8.53 ± 0.03	22.70 ± 1.85
	Muon	**8.05** ± 0.01	18.75 ± 0.27
	SOAP-X	10.33 ± 0.11	10.07 ± 0.07
	SOAP-AC$_{0.25}$	14.72 ± 0.14	**9.85** ± 0.29

D.11 Extrapolation beyond the measured scale

The laws are algebraic, so they do not scale with the model. $\epsilon^{-2\alpha}$ is an identity in the preconditioner update—no term in it depends on parameter count, depth, or width—and the structural mechanism is equally dimension-free: $B_0 = 0$ makes the first covariance refresh the exact zero matrix whether the covariance's large side is 2048-dimensional or 14,336-dimensional, and the per-step rank bound $\text{rank}(GG^\top) \le r$ is a property of the adapter, not of the base model. If anything, deficiency *strengthens* with scale at fixed adapter rank: the null-fraction bound $1 - rW/\max(m,n)$ grows as the matrix dimension grows. The 7B boundary check measures this directly (V45): every 4096-dimensional covariance in the Mistral-7B census sits at effective-rank ratio ≤ 0.0176, median 2.1×10^{-3}—the same order as the 1.7B–3.8B models, so the deficiency does not soften with scale (Appendix D.7). Under assumptions A1–A3, a larger model at a practical adapter rank satisfies the scope a fortiori, and the laws transfer without modification.

What is extrapolated and what is measured, stated plainly. The covariance census extends to 7B (V45); the effective-rank statistics above 7B remain the mechanism's prediction, not a measurement.

The training arm at 7B exceeded a single 32 GB device as first fielded and was completed with host-RAM optimizer state (V45b, update rules line-identical), so the optimizer-level claims now extend as measurements to 7B. The interaction constants of Section 5.4 (which rate counts as "hot" for capped inversion) are calibrated, not derived, and may shift at other scales—the 7B AdamW arm, stable in 3/3 seeds but finishing at 14.32 ± 0.61 against a threshold of 12 calibrated on the 1.7B–3.8B endpoints, is a direct instance; wall-clock ratios (Appendix D) are hardware-specific. We claim exactly what the theory supports: the amplification laws hold wherever A1–A3 hold, larger fine-tuning models at practical adapter ranks satisfy A1 a fortiori, and the unmeasured residue is the census above 7B and the arms not run at that scale (capped inversion at 7B) or beyond it.

D.12 Evaluation protocols

Training protocol. All runs use the Section 3 defaults: effective batch 16×512 tokens, gradient clipping (global norm 1.0), three seeds (42/123/456) per cell unless stated, and five seeds (42/123/456/789/1000) on the V28, V36 (hot-rate arm), V38, V39, and V41 blocks. LoRA targets the attention projections with $\alpha_{\text{LoRA}} = 2r$ and dropout 0.05. Perplexity is evaluated on a fixed held-out validation slice every 50 steps (300-step runs) or every 250 steps (5000-step runs). The 5000-step horizons (V27, V40, V42) train on an 80k-line pool (16,638 chunks $\approx$ 4.8 epochs at 5000×16 draws—overfitting controlled per the V22 lesson) and evaluate on the same fixed validation slice as all other experiments.

Downstream evaluation. MMLU-5-shot (Hendrycks et al., 2021) is evaluated on all 1531 validation questions of the cais/mmlu "all" suite: five-shot prompts assembled from the dev split, answers scored by comparing next-token logits of " A"/" B"/" C"/" D", non-finite logits counted and scored as random guesses. The evaluated checkpoint is the end-of-training adapter (for SOAP-F, the checkpoint at which training diverged). Each downstream block trains fresh adapters under the Section 5.4 protocol with each optimizer at its best 300-step rate (AdamW, Muon, SOAP-X at 1.25×10^{-4}; SOAP-AC$_{0.25}$ at 5×10^{-4} on both models; SOAP-F at 1.25×10^{-4} as the divergence arm; on Phi-4-mini, SOAP-X's V24-B best is the hot 5×10^{-4}, the calibration-trap configuration of Sections 5.4 and 5.5).

Covariance census protocol. The census (V8, V22, V34, V45) runs 50 steps of AdamW-driven training with three q_proj projections in the first/middle/last attention layers instrumented, covariance statistics snapshotted at step 30 (effective rank, λ_{max}, λ_{min}, participation ratio). V34 repeats the protocol at adapter ranks 64–128 with $\alpha_{\text{LoRA}} = 64$; V45 repeats it on Mistral-7B-v0.3 (Appendix D.7).

Pre-registration records. Each block's pre-registration JSON—written to disk before launch—enumerates numbered, falsifiable predictions with fixed numerical thresholds, archived unmodified with the artifact; Table 4 transcribes them verbatim. The 30-step stability probe used for rate calibration runs each candidate rate for 30 steps and selects the fastest-declining rate that shows no divergence flag; the calibration-horizon trap this induces is documented in Sections 5.4 and 5.5.