CAN FEATURE LEARNING DECOUPLE FROM LOSS MINIMIZATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Does feature learning stop when the training loss stops improving? We study this question for matrix Muon, whose polar-normalised updates keep a fixed step length regardless of the gradient norm. Near the edge of stability, full-batch Muon on teacher–student problems enters approximately period-2 loss oscillations that persist for thousands of steps: the cycle-mean loss stays flat or rises, yet the weights keep moving and the learned features continue to align with the teacher subspace. For linear teacher–student learning toys, we derive explicit cycle and alignment formulas and conditional plateau and decay bounds. For a population mean-field ReLU model, we prove that, under stated dimension, initialisation and small-head conditions, the leading eigenspace of the average gradient outer product (AGOP) recovers the teacher subspace exactly during a loss plateau, before the loss later drops. Across 33 ReLU, GELU and SiLU teacher configurations, direction-only alignment metrics and projected head refitting show that the directions learned during plateaus are useful for prediction, and further measurements distinguish AGOP alignment from weight-mass concentration. In deep residual ReLU students, freezing the downstream layers while the first layer trains with full-batch exact polar updates recreates a nearly flat cycle-mean loss with improving input-AGOP alignment; freezing and unfreezing switch between this plateau and loss decrease, and the effect is sensitive to momentum and to the choice of orthogonaliser.

1 INTRODUCTION

Large-step training is often non-monotone: gradient descent near the edge of stability oscillates around sharp directions (Cohen et al., 2021; Arora et al., 2022; Li et al., 2022; Damian et al., 2023), and large-learning-rate stochastic gradient descent produces loss spikes, or *catapults* (Lewkowycz et al., 2020; Zhu et al., 2024). When the loss stops improving, has the network also stopped learning useful features? Training loss measures the error of the current predictor, whereas feature learning concerns the representation on which it depends, and the two need not move together. Silent alignment (Atanasov et al., 2022), teacher-direction recovery during oscillatory training (Chen & Bruna, 2023) and catapult-driven alignment of student and teacher Average Gradient Outer Products (AGOPs) (Zhu et al., 2024) make this distinction concrete; the AGOP averages outer products of a predictor's input gradient and also underlies the Neural Feature Ansatz (Radhakrishnan et al., 2024; Beaglehole et al., 2024). We ask *whether alignment with task-relevant directions can continue through a* sustained *plateau, with loss flat or rising, and which training dynamics recreate this separation.*

We study the exact-polar, zero-momentum form of matrix Muon (Jordan et al., 2024), $W_{t+1} = W_t - \eta\ \mathrm{Polar}(\nabla_W L(W_t))$ with $\mathrm{Polar}(g) = AB^\top$ for the compact SVD $g = A\Sigma B^\top$ (Section 3.2). Polar normalisation sets every nonzero gradient singular value to one, so a full-rank $p \times r_s$ step has Frobenius length $\eta\sqrt{M}$, $M = \min(p, r_s)$, whatever the gradient norm: the weights keep moving near a loss floor. Fixed step length alone does not create an oscillation; our learning toys, teacher–student experiments and freezing interventions examine how this motion relates to feature directions.

Main observation. In full-batch one-hidden-layer teacher–student experiments across 33 ReLU, GELU and SiLU teacher configurations (Table 2), we observe thousands of steps of approximately period-2 loss oscillations while the student AGOP aligns with the teacher subspace. The signature $\rho_2 := -\mathrm{Corr}(\Delta L_t, \Delta L_{t+1})$ of the loss increments $\Delta L_t := L_{t+1} - L_t$ has a second-half seed median

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.

rounding to 1.0000 on the 15 configurations of our diagnostic suite; it concerns loss increments, not an exact periodic orbit of the weights, which keep moving. Direction-only AGOP alignment can improve substantially while cycle-mean loss stays nearly constant or increases (Figures 2 and 3). In deep residual students, freezing downstream layers while the first layer trains recreates this regime, and freeze–unfreeze interventions switch between it and loss reduction (Figure 5).

Contributions.

1. *Measurements and interventions:* we use principal-angle AGOP metrics, a no-oracle head-refit diagnostic, ReLU weight-mass measurements, and freezing interventions whose effect depends on momentum and the orthogonaliser (Sections 5.1 and 5.5; Appendix F.4).
2. *Learning toys with closed-form theory:* for normalised-GD linear regression (M1/M2) and matrix-Muon multi-output regression (M3), we derive conditional plateau and drain bounds, exact weight cycles, loss floors and teacher-alignment formulas (Section 4). Symmetric-cycle rates are linearised references, not convergence guarantees from arbitrary starts.
3. *A population mean-field separation:* we prove that, under stated dimension, initialisation and small-head conditions, a Gaussian ReLU mean-field model recovers the teacher subspace exactly in its leading AGOP eigenspace during a 1% loss plateau, followed by a quantified loss decrease (Section 5.4); period-two dynamics are not asserted.
4. *Structural and conditional results:* we present exact polar step-length and update-mass identities, an empirical plateau-height claim and a conditional drift proposition whose assumptions remain unverified for our trajectories and which does not derive the fitted exponent (Sections 5.2 and 5.3).

2 Related work

Feature learning and oscillatory dynamics. Silent alignment describes early NTK alignment before appreciable loss reduction from small initialisation (Atanasov et al., 2022), with related directional changes in ReLU gradient flow (Boursier & Flammarion, 2025). Grokking separates training-data fit from later generalisation (Kumar et al., 2024). Edge-of-stability analyses study non-monotone GD dynamics (Cohen et al., 2021; Li et al., 2022; Damian et al., 2023); normalised GD admits quadratic two-cycles and, under additional assumptions, slow sharpness-reducing drift (Arora et al., 2022). Chen & Bruna (2023) prove teacher-direction recovery with convergence to an oscillatory orbit in a single-neuron ReLU model under specified initialisation and step-size conditions. Catapult experiments link loss spikes to improved student–teacher AGOP alignment (Zhu et al., 2024). We examine AGOP alignment during sustained exact-polar loss plateaus; our supporting mean-field theory uses distinct population and small-head assumptions and does not assert period-two dynamics.

AGOP and weight geometry. Gradient-based dimension reduction has statistical antecedents (Härdle & Stoker, 1989; Trivedi et al., 2014). The Neural Feature Ansatz relates weight geometry to a predictor's own AGOP (Radhakrishnan et al., 2024; Beaglehole et al., 2024). Exact linear-network identities and nonlinear counterexamples delimit this relationship (Tansley et al., 2026); FACT instead derives a weight-geometry identity at stationary points with positive weight decay (Boix-Adserà et al., 2026). We measure alignment to the teacher subspace along the training trajectory. Its separation from weight-mass concentration distinguishes directional from spectral information, but does not itself contradict an identity between weight geometry and the student's own AGOP.

Muon and spectral learning. Practical Muon orthogonalises momentum approximately (Jordan et al., 2024); we isolate its exact-polar, zero-momentum update. Recent work studies balanced spectral acquisition (Vasudeva et al., 2026), feature diversity in a stylised sequential model (Tian, 2026), and robustness and transfer of learned representations (Ruan et al., 2026). AMUSE includes an exact-polar quadratic cycle (Kim et al., 2026), while river-valley theory characterises rapid signal progress and local overshoot (Shen et al., 2026). Our focus is the timing of AGOP alignment relative to loss reduction. The polar map and its norm-constrained optimisation interpretation are background (Higham, 1986; Gawlik & Leok, 2017; Pethick et al., 2025); fixed step length alone does not establish a cycle.

3 PROBLEM SETUP

Conventions. Expectations use $\mathbb{E}[Z]$ or $\mathbb{E}[Z \mid B]$, with variable/law subscripts. Vector $\|\cdot\|$ is Euclidean; matrix $\|\cdot\|_{\mathrm{op}}, \|\cdot\|_F$ are spectral/Frobenius norms. $\mathrm{tr}, \mathrm{rank}, \mathrm{span}, \mathrm{diag}$ mean trace, rank, column span and diagonal matrix; I_k is the identity and e_k a coordinate vector. P_S projects orthogonally onto S, $P_S^\perp = I - P_S$; $A \succ B$ means $A - B$ positive definite. Singular values σ_i decrease; $\cos^2(v, w) = (v^\top w)^2/(\|v\|^2\|w\|^2)$ for $v, w \neq 0$.

We study learning toys (linear regression trained by NGD: **M1**, $p = 2$; **M2**, general p; or matrix Muon: **M3**, multi-output) and a 1-HL teacher–student network (**M4**). The toys isolate a fixed-Hessian mechanism; M4 adds an activation-dependent Hessian. Throughout, η is fixed near the edge of stability, p is the input dimension, r_t the teacher rank, r_s the student width, and

$$M := \min(p, r_s) \tag{1}$$

is the rank scale that sets the Frobenius length of a full-rank matrix-Muon step.

3.1 DATA AND TEACHER–STUDENT MODEL

For M4, the n training inputs are i.i.d. Gaussian, $x_i \sim \mathcal{N}(0, I_p)$, with labels $y_i = f^\star(x_i)$ and $y = (y_1, \ldots, y_n)^\top$. The teacher is a one-hidden-layer network

$$f^\star(x) = \sum_{j=1}^{r_t} \sigma(u_j^\top x), \qquad U = [u_1, \ldots, u_{r_t}] \in \mathbb{R}^{p \times r_t}, \qquad U^\top U = I_{r_t}, \tag{2}$$

with teacher subspace $V := \mathrm{span}(U)$, $P_V := UU^\top$, $P_\perp := I_p - P_V$, and $\sigma \in \{\mathrm{ReLU}, \mathrm{GELU}, \mathrm{SiLU}\}$. The student is a 1-HL (hidden-layer) ReLU network with trainable hidden weights $W = [w_1|\cdots|w_{r_s}] \in \mathbb{R}^{p \times r_s}$ and *fixed* head $a_s = p^{-1/2}\mathbf{1}_{r_s}$:

$$f_W(x) = \sum_{k=1}^{r_s} a_{s,k}\, \mathrm{ReLU}(w_k^\top x), \qquad L_{\mathrm{train}}(W) = \tfrac{1}{n}\sum_{i=1}^{n} \big(f_W(x_i) - f^\star(x_i)\big)^2. \tag{3}$$

Freezing a_s forces all representation changes through W. We use full-batch gradients throughout, and write $\mathcal{L} := L_{\mathrm{train}}$ in M4.

3.2 OPTIMISER: MATRIX MUON

Matrix Muon updates the hidden weights by the polar factor of the gradient:

$$W_{t+1} = W_t - \eta\, \mathrm{Polar}(\nabla_W L(W_t)), \qquad \mathrm{Polar}(g) := AB^\top \text{ for the compact SVD } g = A\Sigma B^\top \tag{4}$$

($\mathrm{Polar}(g) = g(g^\top g)^{-1/2}$ when g has full column rank). The practical optimiser approximates this via Newton–Schulz iterations; we analyse the exact polar update, with $\mathrm{Polar}(0) := 0$. Its M4 iteration map is

$$\Phi(W) := W - \eta\, \mathrm{Polar}(\nabla\mathcal{L}(W)). \tag{5}$$

For $g := \nabla\mathcal{L}(W)$ of rank M, let $\ell_1, \ldots, \ell_M$ be its left singular vectors. The leakage quantities are $\varepsilon_V^2 := \sum_{i \le r_t} \|P_\perp \ell_i\|^2$ and $\varepsilon_\perp^2 := \sum_{i=r_t+1}^{M} \|P_V \ell_i\|^2$ (assuming $r_t \le M$). Two properties drive the analysis. All nonzero singular values of $\mathrm{Polar}(g)$ are one, so

$$\|\mathrm{Polar}(g)\|_F^2 = \mathrm{rank}(g), \qquad \|W_{t+1} - W_t\|_F = \eta\sqrt{M} \text{ when } \mathrm{rank}(\nabla_W L(W_t)) = M, \tag{6}$$

i.e. step length is independent of the gradient norm. And Polar is left/right orthogonal-equivariant: $\mathrm{Polar}(QgR^\top) = Q\ \mathrm{Polar}(g)\ R^\top$ for orthogonal Q, R. When $r_s = 1$ the polar factor reduces to vector normalization $\mathrm{Polar}(g) = g/\|g\|_2$, and matrix Muon becomes NGD (normalized gradient descent),

$$\theta_{t+1} = \theta_t - \eta\, \nabla L(\theta_t)/\|\nabla L(\theta_t)\|_2, \tag{7}$$

which is the update used in M1 and M2.

3.3 ALIGNMENT METRICS FOR M4

We measure feature alignment through the student AGOP

$$G_s(W) := \tfrac{1}{n}\sum_{i=1}^{n} \nabla_x f_W(x_i) \nabla_x f_W(x_i)^\top \in \mathbb{R}^{p\times p}. \tag{8}$$

Let $\bar{V}_k = [v_1, \dots, v_k]$ collect the top k orthonormal eigenvectors of $G_s(W)$, and let $\theta_1, \dots, \theta_{r_t}$ be the principal angles, in increasing order, between $\mathrm{span}(\bar{V}_{r_t})$ and V ($\cos\theta_i = \sigma_i(\bar{V}_{r_t}^\top U)$). We use the direction-only summaries

$$\overline{\cos^2}(W) = \tfrac{1}{r_t}\sum_{i=1}^{r_t} \cos^2\theta_i, \qquad \cos^2_{\min}(W) = \min_{i\le r_t} \cos^2\theta_i, \tag{9}$$

which separate subspace recovery from AGOP spectrum, and the weight Frobenius mass-in-V

$$A_{\mathrm{sub}}(W) = \|P_V W\|_F^2 / \|W\|_F^2. \tag{10}$$

Our no-oracle head-refit diagnostic projects W onto its top-$\tilde{r}$ AGOP eigenspace: $\widehat{W} := \bar{V}_{\tilde{r}}\bar{V}_{\tilde{r}}^\top W$, with the rank rule in Appendix C. For projected features $\Psi_{ik} := a_{s,k}\mathrm{ReLU}(\widehat{w}_k^\top x_i)$ and $a^\star := \arg\min_a \|\Psi a - y\|_2^2$, define

$$L_{\text{AGOP-opt}}(W) := \tfrac{1}{n}\|\Psi a^\star - y\|_2^2. \tag{11}$$

This minimizes training loss over heads on the student's AGOP-projected features; the oracle $L_{V,\mathrm{opt}}$ replaces $\bar{V}_{\tilde{r}}$ by U. For a loss sequence L_t, its cycle mean is $(L_t + L_{t+1})/2$; $\Delta L_t := L_{t+1} - L_t$ and $\rho_2 := -\mathrm{Corr}(\Delta L_t, \Delta L_{t+1})$, with Pearson correlation over the stated time window.

For the conditional drift result, set $A_t := \cos^2_{\min}(W_t)$, $q := r_t/M$, $e_k := 1 - A_{2k}$ and $v_k := (A_{2k+2} - A_{2k})/2$; k indexes pairs of steps, so v_k is the alignment gain per optimisation step along the even subsequence. The symbol θ collects the remaining fixed family parameters, including the learning-rate rule. We use $\nu > 0$ for the expansion scale q^ν and an integer $j \ge 1$ for its first active dissipative order, with local drift exponent $\alpha = \nu j$ (distinct from the mean-field head scale denoted by α below). For positive quantities, $f \asymp_\theta g$ means $c_\theta g \le f \le C_\theta g$ for positive constants depending only on θ, uniformly over the stated parameter range and trajectory window.

3.4 LEARNING TOYS

The learning toys keep samples, a teacher and the empirical MSE, but use a linear student, so the Hessian is the fixed data matrix $2\hat{\Sigma}$. Draw $x_1, \dots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_x)$ with $\Sigma_x \succ 0$ and $n \ge p \ge 2$, let $X \in \mathbb{R}^{n\times p}$ have rows $x_i^\top$, and let $\hat{\Sigma} := X^\top X/n = Q\,\mathrm{diag}(\mu_1, \dots, \mu_p)\,Q^\top$ with $\mu_1 > \mu_2 > \cdots > \mu_p > 0$ (almost surely) and eigenvectors $q_k = Qe_k$.

M2 (M1 for $p = 2$): NGD on linear regression. Fix $\eta > 0$. Teacher $u \in \mathbb{R}^p \setminus \{0\}$, labels $y_i = u^\top x_i$, student $f_w(x) = w^\top x$, trained by NGD Eq. (7) on

$$\hat{\mathcal{L}}(w) = \tfrac{1}{n}\sum_{i=1}^{n}\big(w^\top x_i - y_i\big)^2 = (w-u)^\top\hat{\Sigma}(w-u) = \sum_{k=1}^{p}\mu_k z_k^2, \qquad z := Q^\top(w-u). \tag{12}$$

We track the teacher alignment $\cos^2(w_t, u)$ — the input AGOP of a linear student is $G_s = ww^\top$, so this is the $r_t = 1$ case of $\cos^2_{\min}$ in Eq. (9) — and the residual alignment $\cos^2(w_t - u, q_1)$; the remaining toy notation is in Appendix A.

M3: matrix Muon on multi-output regression. Teacher $U_\star \in \mathbb{R}^{p\times r}$ with orthonormal columns, labels $y_i = U_\star^\top x_i \in \mathbb{R}^r$, student $f_W(x) = W^\top x$ with $W \in \mathbb{R}^{p\times r}$ with $1 \le r < p$ (so $M = r$), trained by Eq. (4) on $\hat{\mathcal{L}}(W) = \frac{1}{n}\|XW - XU_\star\|_F^2 = \mathrm{tr}(R^\top\hat{\Sigma}R)$, $R := W - U_\star$. The data have a planted block, $\Sigma_x = \lambda_D P_D + \lambda_\perp P_D^\perp$ with $\dim D = r$ and $\lambda_D \gg \lambda_\perp$. The AGOP of f_W (summed over outputs) is $WW^\top$, so the M4 metric becomes $\cos^2_{\min}(\mathrm{range}\,W_t, \mathrm{range}\,U_\star)$. We place the teacher either in the high-variance block, (A) $\mathrm{range}\,U_\star = D$, or adversarially in the low-variance directions, (B) $\mathrm{range}\,U_\star \subset D^\perp$ (requiring $2r \le p$). These are population blocks; their empirical counterparts need not coincide at finite n. Write $\hat{D}$ for the top-r eigenspace of $\hat{\Sigma}$.

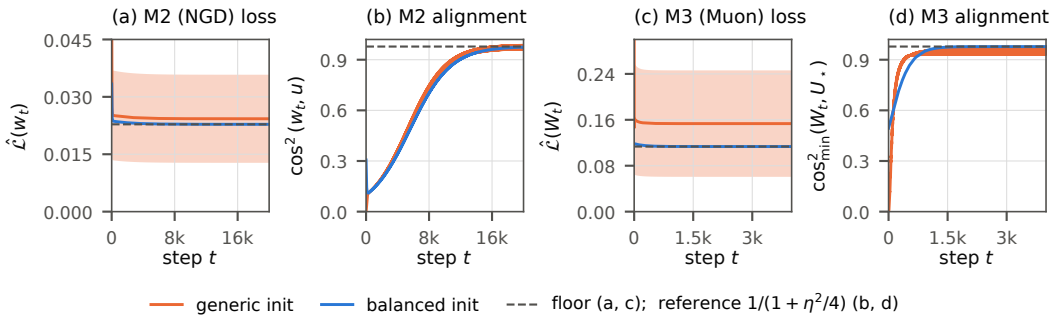


Figure 1: Learning toys: nearly flat cycle-mean loss with improving teacher alignment. (a,b) M2: NGD; (c,d) M3: matrix Muon; $\eta = 0.3$, low-variance teachers. Blue: balanced leading components with nonzero off-mode residuals; orange: generic starts. Bands show adjacent-step loss ranges; dashed lines give symmetric-cycle references. These finite-sample illustrations do not verify all theorem hypotheses. Settings and endpoints: Appendix A.7; optimiser comparison: Figure 8.

3.5 Population mean-field model

We use the Gaussian inputs and teacher directions of Section 3.1, set $r = r_t$, and take $\sigma = \mathrm{ReLU}$ in Eq. (2). For a fresh input X, let $Y = \gamma f^\star(X)$, where $\gamma > 0$ makes $\mathbb{E}[Y^2] = 1$, and write $\sigma_Y := \sqrt{\mathrm{Var}(Y)}$. The student in Eq. (3) now uses fixed head $a_{s,j} = \alpha/r_s$, $\alpha > 0$, and an added intercept b refitted at every step so that $\mathbb{E}[f_W(X)] = \mathbb{E}[Y]$. It takes the polar update Eq. (4) on the population loss $L(W) := \mathbb{E}[(Y - f_W(X))^2]$, starting from $w_j(0) \overset{\text{iid}}{\sim} \nu_0 := \mathcal{N}(0, \sigma_0^2 I_p/p)$, with $\sigma_0, \bar{\eta} > 0$, $\eta = \bar{\eta}\sqrt{r_s}$ and $\kappa := \sigma_0/(\bar{\eta}\sqrt{p})$.

At infinite width $r_s \to \infty$ with $\alpha, \bar{\eta}, \sigma_0$ fixed, the *mean-field model* describes the hidden weights by a probability law ν_t. Starting from ν_0, it follows the distributional polar update in Appendix G, with output $f_t(x) = b_t + \alpha \mathbb{E}_{w\sim\nu_t}[\mathrm{ReLU}(w^\top x)]$, intercept refitted as above, and loss $L(t) := \mathbb{E}[(Y - f_t(X))^2]$. Let $G_s(t)$ be the population version of Eq. (8) for f_t, with decreasing eigenvalues $\lambda_i(t)$ and any top-r spectral projector $P_r(t)$. Then $A_r(t) := r^{-1}\,\mathrm{tr}(P_V P_r(t))$ is the population analogue of $\overline{\cos^2}$ in Eq. (9).

Deep residual students. For depth $L \geq 2$ and width m, the bias-free residual ReLU student is

$$(2 \leq \ell \leq L),\ h_1 = \mathrm{ReLU}(x^\top W_1),\ h_\ell = h_{\ell-1} + \frac{\mathrm{ReLU}(h_{\ell-1}W_\ell)}{L-1}, \quad f_W(x) = h_L a.$$

Here $W_1 \in \mathbb{R}^{p\times m}$, $W_\ell \in \mathbb{R}^{m\times m}$ for $\ell \geq 2$, and the head $a = (c_a/\sqrt{p})\mathbf{1}_m$ is frozen, with $c_a > 0$. Every unfrozen hidden matrix takes the full-batch update $W_\ell \leftarrow W_\ell - \eta_\ell\,\mathrm{Polar}(\nabla_{W_\ell}\mathcal{L})$, where $\eta_\ell = \eta_{\mathrm{deep}}$ for $\ell \geq 2$ and $\mathcal{L}$ is the squared training loss. These runs use all thin-SVD columns in the polar factor; Appendix D gives that convention and the settings. Input-AGOP alignment uses Eq. (9); Appendices E–F specify the freezing interventions.

4 Learning toys with closed-form theorems

The toys' empirical losses are shifted quadratics. Appendix A gives proofs, exact residual recursions (Lemma 6) and the period-2 theory with teacher alignment; Appendix A.7 gives numerical checks.

Scalar regression (M2). Under the phase-selection and margin hypotheses of Theorem 7, the leading residual coordinate approaches a nearly symmetric two-cycle while the other coordinates decay geometrically. The leading loss contribution is $\mu_1\eta^2/4$, with smaller-mode loss and controlled phase errors; teacher alignment can keep improving as the smaller components disappear. Small-eigenvalue modes can carry substantial misalignment but little loss: at $\eta = 0.3$, a unit residual hidden below 5% of the floor requires at least 443 steps per e-fold to leading order (Corollary 8). The complete bounds and their hypotheses are in Appendix A.2. Generic starts can select asymmetric cycles; their sustained plateaus in Figure 1 are numerical observations, not consequences of the conditional theorem.

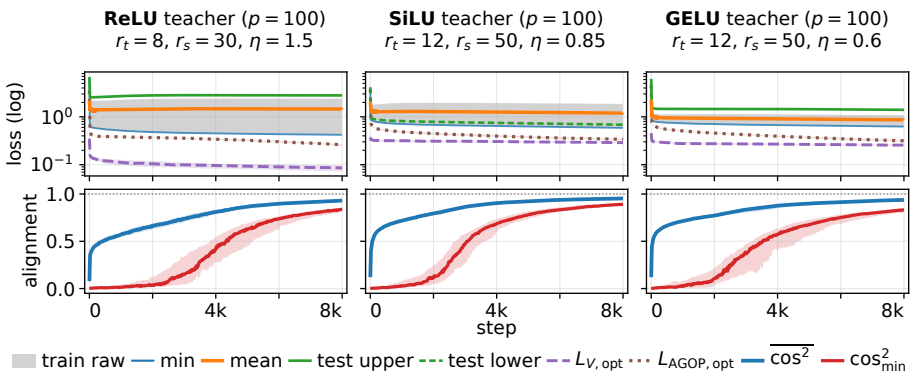


Figure 2: Loss plateaus while AGOP alignment improves. One ReLU, SiLU and GELU teacher configuration; 8,000 matrix-Muon steps. Top: training/test losses and head-refit diagnostics. Bottom: direction-only AGOP alignment. Medians over 53 seeds, shaded interquartile ranges. Full metrics: Figure 13; selection: Appendix C.3.

Proposition 1 (M3: exact weight 2-cycles and teacher alignment)**.** *In M3, let* $U_\star^\top U_\star = I_r$, $R_t := W_t - U_\star$ *and* $\eta > 0$. *Let* $\hat{\Sigma} \succ 0$ *have eigenvalues* $\mu_1 \geq \cdots \geq \mu_p > 0$ *with* $\mu_r > \mu_{r+1}$, *and top-*r *eigenspace* $\hat{D}$. *Suppose* R_T *has full column rank and columns in* $\hat{D}$. *Write* $\hat{\Sigma} R_T = OP$, *where* $O^\top O = I_r$ *and* $P \succ 0$, *and set* $K := O^\top \hat{\Sigma} O$, $P' := \eta K - P$. *If* $P' \succ 0$, *then* $W_{t+2} = W_t$ *for all* $t \geq T$. *The cycle-mean loss is at least* $\frac{\eta^2}{4} \sum_{i \leq r} \mu_i$, *with equality iff* $R_T = \frac{\eta}{2} O$ *(the symmetric cycle* $R_{T+1} = -R_T$*). The two-step transverse linearisation on* $\hat{D}^\perp$ *strictly contracts in Frobenius norm if* $P, P' \succ \frac{\eta \mu_{r+1}}{2} I_r$. *On the symmetric cycle, teacher alignment is* $\cos^2_{\min} = (1 + \eta^2/4)^{-1}$ *if* $\operatorname{range} U_\star \perp \hat{D}$, *and* 1 *if* $\operatorname{range} U_\star = \hat{D}$ *and* $\eta < 2$. *Every step with* $\operatorname{rank} R_t = r$ *has length* $\|W_{t+1} - W_t\|_F = \eta\sqrt{r}$.

Proof, phase losses and linearised drain factors are in Appendix A.5. On an exact weight cycle, loss and alignment (when defined) repeat every two steps; continued alignment change concerns off-cycle trajectories, described here only to first order.

The plateau belongs to the optimiser. All three losses have minimum 0, which plain gradient descent with a stable step reaches (Figure 8); the normalised updates approach nonzero cycles instead, with floors that scale as η^2 (M2 means within 0–4% of $\mu_1 \eta^2/4$ for $\eta \in [0.01, 0.2]$; M3 floors 1.05–$1.39 \times \lambda_D \eta^2 r/4$, reflecting sampling and the asymmetric-cycle excess of Proposition 11(ii)). In M3 the fixed Hessian $2\hat{\Sigma}$ determines which invariant subspaces can support cycles; in M4 the Hessian also depends on the activation pattern and hence on the teacher.

5 The one-hidden-layer ReLU NN under matrix Muon

We study the one-hidden-layer ReLU student defined in Section 3.1.

5.1 Empirical phenomenology

Figure 2 illustrates the behaviour. Across the 33 ReLU/SiLU/GELU teacher configurations listed in Table 2 (Appendix C.1), with step sizes tuned near the edge of stability, we observe two features:

(P1) *Period-2 loss signature with fixed step length.* After a short transient, loss increments alternate almost perfectly: $\rho_2 = 1.000$ to four decimal places (second-half seed medians, all 15 configurations of Figure 3(a)). Each full-rank step has length $\eta\sqrt{M}$; the weights keep moving, allowing continued alignment rather than an exact weight two-cycle.

(P2) *AGOP-into-*V *rotation during the plateau.* Both $\overline{\cos^2}$ and $\cos^2_{\min}$ show large net increases, including configurations with increasing training loss (Figure 2; Appendix C).

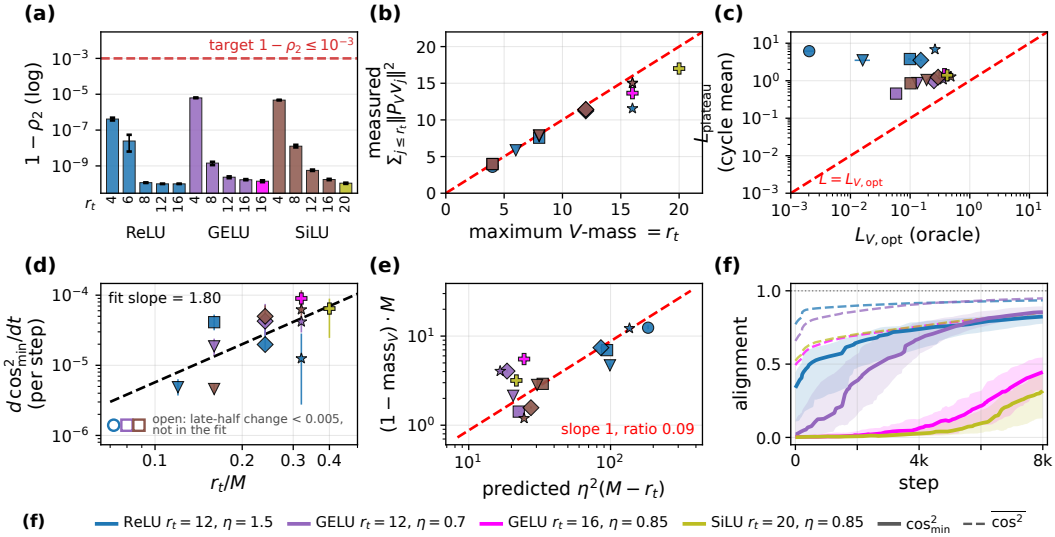


Figure 3: Theory–experiment diagnostics across 15 ReLU, GELU and SiLU configurations. Teachers: ReLU (blue), GELU (purple; magenta r_t=16, η=0.85) and SiLU (brown; olive r_t=20); $p = 100$, $r_s = 50$, $r_t \in \{4, 6, 8, 12, 16, 20\}$, $\eta \in \{0.7, 0.85, 1.5, 2\}$; marker shape indexes r_t (full list: Appendix C.4). Medians and interquartile ranges over 53 seeds. (a) Period-2 loss signature; (b) top-r_t AGOP V-mass versus its maximum; (c) plateau loss versus head oracle; (d) late-half alignment rate versus r_t/M; (e) off-V AGOP-mass proxy with a unit-slope reference line; (f) alignment trajectories. Panel (b) does not measure gradient leakage; (d) is descriptive, not a test of Proposition 4. Definitions, fits and ranges: Appendix C.4.

5.2 Universal polar-update identities

Proposition 2 (polar-update identities)**.** *For any iterate W with full-column-rank gradient g and $r_t \le M$, the matrix Muon update satisfies the three exact identities below (proofs in Appendix B):*

(i) Polar Frobenius identity. $\|\operatorname{Polar}(g)\|_F^2 = M$.

(ii) Per-step displacement. $\|\Phi(W) - W\|_F = \eta\sqrt{M}$, *independent of the loss landscape and of the iterate.*

(iii) $V/V^\perp$ leakage decomposition. $\|P_V \operatorname{Polar}(g)\|_F^2 = r_t - \varepsilon_V^2 + \varepsilon_\perp^2$ *and* $\|P_\perp \operatorname{Polar}(g)\|_F^2 = (M - r_t) + \varepsilon_V^2 - \varepsilon_\perp^2$. *In the small-leakage limit, the polar update splits its squared Frobenius mass between V and $V^\perp$ in the ratio $r_t : (M - r_t)$.*

These identities give the fixed step length in (P1) and the update-mass split; they do not imply AGOP alignment or a periodic network trajectory. The network experiments exhibit approximately period-2 loss oscillations while the weights keep moving. The separate existence result for exact weight cycles concerns only an isotropic quadratic (Theorem 13, Appendix B.2). Figure 3 summarises the 15-configuration diagnostics; Appendix C.4 separates the exact identities from the empirical comparisons.

5.3 Plateau height and alignment drift

Proposition 2 establishes the update length and its decomposition between the teacher subspace and its orthogonal complement. Theorem 13 supplies a separate example of an exact two-cycle on an isotropic quadratic. These results do not establish the observed network oscillations or AGOP alignment. The plateau-height statement below is empirical; Proposition 4 gives sufficient conditions for alignment drift, with its proof in Appendix B.4.

Claim 3 (plateau-height scaling)**.** *During the period-2 loss plateau of matrix Muon on the 1-HL ReLU teacher–student loss, the plateau loss admits the leading-order decomposition*

$$\mathcal{L}(W_t) \;\approx\; \mathcal{L}\big(P_V W_t;\, a_s \text{ fixed}\big) \;+\; c_2\,\eta^2\,(M - r_t), \tag{13}$$

with empirical $c_2 \in [0.5, 1.4]$ (Appendix C, Test C). The decomposition is empirical: the polar identities alone do not establish a loss law.

Proposition 4 (Conditional secular drift)**.** *Let $A_t = \cos^2_{\min}(W_t)$, $q = r_t/M$, $e_k = 1 - A_{2k}$ and $v_k = \frac{1}{2}(A_{2k+2} - A_{2k})$, and fix the remaining family parameters θ, including the learning-rate rule. Suppose that, on a trajectory window, the leading AGOP subspace is spectrally separated and near the teacher subspace, and its two-step transverse dynamics satisfy the uniform expansion of Appendix B.4: positive operator-norm dissipation of finite order $q^{\nu j}$ with a correspondingly scaled nonlinear remainder. Then, for sufficiently small $q > 0$ and $e_k > 0$, uniformly on the window,*

$$v_k \asymp_\theta q^\alpha e_k, \qquad \alpha = \nu j > 0. \tag{14}$$

Under the additional stationary slow-mode conditions there, $v_k = C(\theta) q^\alpha e_k (1 + o(1))$ with $C(\theta) > 0$ as $q, e_k \to 0$ and $kq^\alpha \to \infty$ within the window.

These conditions are unverified for our sweep and do not fix the exponent in Figure 3(d); plateau entry and residence are separate assumptions.

5.4 A MEAN-FIELD THEOREM: PLATEAU, THEN FEATURE LEARNING, THEN LOSS DROP

In Section 3.5's mean-field model, ordinary initialisation means $\kappa \gg 1$: the initial weight scale is large relative to the effective step size. The theorem separates the feature-learning time t_F, a chosen plateau horizon t_P, and the later loss-drop time t_2. All times count discrete updates.

Theorem 5 (Plateau, then feature learning, then loss drop; informal version of Theorem 16)**.** *Let $r \geq 2$, $p \geq 64r$ and $p - r \geq 400$, with head scale $\alpha > 0$. There are $\tau_F \in (0, 0.101)$, $\kappa_{\min} < \infty$ and $\delta > 0$, depending only on (r, p), such that the following holds for every $\kappa \geq \kappa_{\min}$, with $t_F := \lfloor \tau_F \kappa \rfloor$, any integer $t_P \geq t_F$ and $\alpha^*(t_P) := 0.00845\, \sigma_Y/(\bar{\eta}(\kappa + t_P))$. There is a threshold $a_0 = a_0(r, p, \kappa) > 0$ such that:*

- *(P)* Plateau. *If $\alpha \leq \alpha^*(t_P)$, the loss stays within 1% of its initial value, and above $0.99\, \sigma_Y^2$, for $0 \leq t \leq t_P$.*
- *(F)* Feature learning. *The AGOP is isotropic at $t = 0$. If $\alpha\bar{\eta} \leq a_0$, then at t_F its top-r eigenspace is exactly V, so $A_r(t_F) = 1$, with $\lambda_r(t_F) \geq (1 + \delta/2)\lambda_{r+1}(t_F)$.*
- *(R)* Loss drop. *If $\alpha \leq \alpha^*(t_P)$, then $L(0) - L(t_2) \geq 0.03\, \sigma_Y^2$ at $t_2 := \lceil 0.05086\, \sigma_Y/(\alpha\bar{\eta}) \rceil > t_P$; at $\alpha = \alpha^*(t_P)$, $t_2 \leq 6.02(\kappa + t_P) + 1$.*

All three hold together for $0 < \alpha \leq \min\{\alpha^*(t_P),\, a_0/\bar{\eta}\}$.

The plateau and loss-drop constants are explicit; a_0 is not, since (F) transfers from zero to small positive heads by continuity of the normalized AGOP G_s/α^2. Its proof is analytic except on a finite (r, p) region covered by 80 outward-rounded interval certificates. Appendix G gives the precise theorem and complete proof.

5.5 DEEP RESIDUAL STUDENTS

Depth-6–16 residual ReLU students show the same pattern with a large first-layer and small downstream exact-polar steps (Figure 4; Appendix D).

A freezing principle. In these students, continued feature learning on a loss plateau appears when W_1 trains against a fixed downstream map (Appendices E and F). Across depths 4, 8, 16 and 24, freezing every layer after W_1 lets exact polar steps keep the cycle-mean loss within 5% in almost all runs while input-AGOP $\cos^2_{\min}$ rises by 0.38–0.53 over the window. Training the downstream layers at $\eta_1/50$ or faster lowers loss by 7–14%; mid-run freezing or unfreezing switches the plateau on or off while alignment keeps rising (Figure 5). After partial pretraining with slow GD or Adam, training W_1 alone usually aligns features rapidly; GD or Adam then lower the cycle-mean loss, whereas a large polar step holds it near a floor (Appendix F.2). In the deep student, Muon's Newton–Schulz quintic and the tested momentum settings disrupt the combination of flat loss and continued alignment, whereas a convergent iteration preserves it. The one-hidden-layer student's small-head plateau is more robust to both (Appendix F.3). Outcomes for transformers, vision transformers and a CNN remain limited (Appendix F.4).

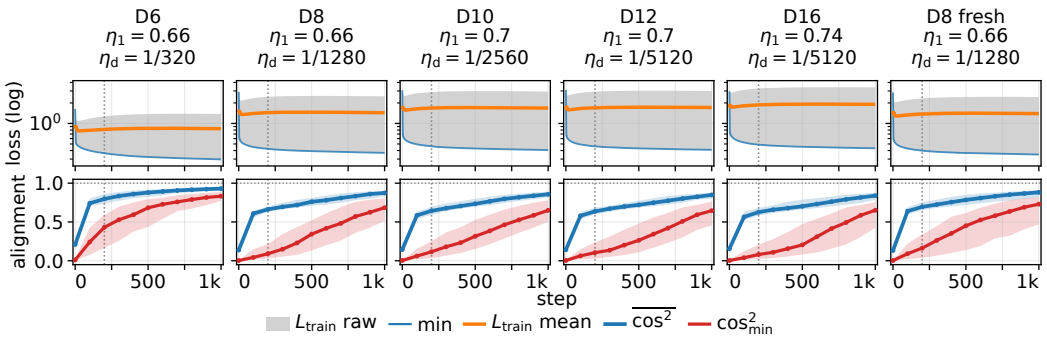


Figure 4: Deep residual students: oscillating loss and rising alignment. Medians/IQRs over 100 runs per column (20 problems $\times$ 5 initialisations). Dotted lines mark scoring windows; across columns, the median per-run cycle-mean loss change is at most about 2.6%. Settings and metrics: Appendix D.

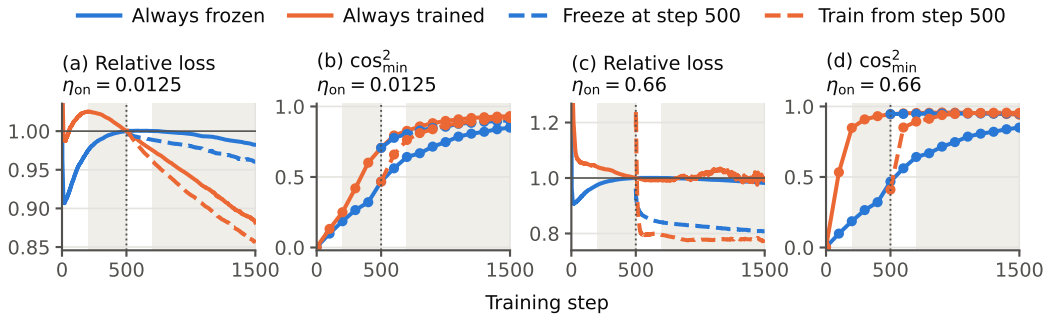


Figure 5: Freezing downstream layers can restore a loss plateau. Depth 8; the first layer always trains. Dashed curves switch the deep layers at step 500; solid controls keep their state. Their step while training is η_{on}. Panels (a,c): relative cycle-mean loss; (b,d): $\cos^2_{\text{min}}$. Pooled medians over 48 runs; protocol: Appendix E.

6 Discussion and Conclusion

At long-run ReLU checkpoints, 83–90% of AGOP eigenvalue mass lies in V ($\overline{\cos^2} \in [0.85, 0.96]$), versus 29–50% of weight Frobenius mass (Appendix C.4). The weight Gram matrix is therefore not proportional to the AGOP, the simplest Neural Feature Ansatz (Radhakrishnan et al., 2022); power-law variants are untested.

Catapult experiments connect loss spikes to AGOP-alignment improvements (Lewkowycz et al., 2020; Zhu et al., 2024); here alignment improves over thousands of plateau steps, even with rising loss. On the six cleanest ReLU configurations, $L_{\text{AGOP-opt}}$ is about 5–11 times below cycle-mean training loss over the last 30% of training (seed medians; Appendix C). The learned directions support a better predictor; the projection changes features, so this does not isolate the head as the limitation. Direction-only AGOP metrics complement the loss.

Our exact results concern learning toys, polar identities and the mean-field model. The plateau law and fitted drift exponent are empirical, and Proposition 4's conditions remain unverified. Results for depth, Newton–Schulz, momentum and minibatches are empirical (Appendices D–F); the full-batch exact-polar analysis is a mechanistic study, not a theory of Muon at scale.

Conclusion. Under polar updates, a loss plateau need not be a pause in feature learning. The separation is proved for the learning toys and the mean-field model, measured in one-hidden-layer and deep residual students, and switched on and off by freezing. The training loss alone can therefore understate representation learning; direction-only AGOP metrics and head refitting are inexpensive complementary checks. Extending the analysis to minibatch, momentum and Newton–Schulz Muon at scale is the main open direction.

AI USE STATEMENT

We used large language model assistants to support the development of the learning-toy models, the formulation of mathematical claims and hypotheses, and aspects of proof development and writing. These tools also assisted with experimental design, implementation (including synthetic-data sampling), and interpretation of results. Additionally, we used generative AI tools to create and edit software code and scientific figures, to check literature and references, and to draft and edit parts of the paper. We have reviewed all AI-assisted work. Numerical consistency checks for the learning toys accompany the analytic proofs (Appendix A.7); they do not substitute for those proofs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies optimisation dynamics on synthetic problems and one small public dataset. Apart from the 8×8 optical-digit images of Alpaydin & Kaynak (1998), used in the vision-transformer and CNN runs of Appendix F.4, all data are Gaussian inputs labelled by planted teachers. The work involves no human subjects, no personal or sensitive data and no dataset release. We are not aware of harmful applications specific to this work. The authors have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Complete proofs of the theorems and propositions, whose assumptions are stated in the results themselves, are in Appendix A (learning toys, Section 4), Appendix B (Section 5) and, for the mean-field theorem, Appendix G. Section 3 and Appendix C give the models, step sizes, widths and seeds of the one-hidden-layer experiments; Appendices D–F give the models, step sizes, seeds and scoring protocol of the deep-network experiments. Code for all experiments, figures and tables, including the interval-arithmetic checks of the mean-field proof, is at https://anonymous.4open.science/r/plateaus_and_feature_learning_iclr2027-674D/. The supplementary material reproduces the learning-toy figures (Figure 1 and Appendix A.8) and the numerical consistency checks of Appendix A.7, whose configurations are stated there. Unless stated otherwise, training is full batch with the exact polar factor and fixed seeds; Appendix F also uses Newton–Schulz iterations, momentum, fresh minibatches, gradient descent, Adam and AdamW, and states each setting where it is used; Table 10 lists the settings of its architecture runs.

REFERENCES

Ethem Alpaydin and Cenk Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. URL https://archive.ics.uci.edu/dataset/80/optical+recognition+of+handwritten+digits.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 948–1024, 2022. URL https://proceedings.mlr.press/v162/arora22a.html.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=1NvflqAdoom.

Daniel Beaglehole, Ioannis Mitliagkas, and Atish Agarwala. Feature learning as alignment: A structural property of gradient descent in non-linear neural networks. *Transactions on Machine Learning Research*, November 2024. doi: 10.48550/arXiv.2402.05271. URL https://openreview.net/forum?id=JXCe2ZcUXr.

Enric Boix-Adserà, Neil Mallinar, James B. Simon, and Mikhail Belkin. FACT: a first-principles alternative to the neural feature ansatz for how networks learn representations. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6e6c5da107b46b9f433cdc1f3ba4b995-Abstract-Conference.html.

Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research*, 26(183):1–75, 2025. URL https://jmlr.org/papers/v26/24-1523.html.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4330–4391, 2023. URL https://proceedings.mlr.press/v202/chen23b.html.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nhKHA59gXz.

Evan S. Gawlik and Melvin Leok. Iterative computation of the Fréchet derivative of the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1354–1379, 2017. doi: 10.1137/16M108971X.

Wolfgang Härdle and Thomas M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408):986–995, 1989. doi: 10.1080/01621459.1989.10478863.

Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986. doi: 10.1137/0907079.

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

Jueun Kim, Baekrok Shin, Jihun Yun, Beomhan Baek, Minhak Song, and Chulhee Yun. AMUSE: Anytime Muon with stable gradient evaluation. *arXiv preprint arXiv:2605.22432*, 2026. URL https://arxiv.org/abs/2605.22432.

Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/63ed15a46a143ff57484b38cd6b85d91-Abstract-Conference.html.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: The catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020. doi: 10.48550/arXiv.2003.02218. URL https://arxiv.org/abs/2003.02218.

Zhouzi Li, Zixuan Wang, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2207.12678.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49069–49104, 2025. doi: 10.48550/arXiv.2502.07529. URL https://proceedings.mlr.press/v267/pethick25a.html.

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism of feature learning in deep fully connected networks and kernel machines that recursively learn features. *arXiv preprint arXiv:2212.13881*, 2022. doi: 10.48550/arXiv.2212.13881. URL https://arxiv.org/abs/2212.13881.

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383(6690):1461–1467, 2024. doi: 10.1126/science.adi5639. URL https://www.science.org/doi/10.1126/science.adi5639.

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

Tianqi Shen, Jinji Yang, Runze Shi, Jianhao Ma, Jiaye Teng, and Ziye Ma. Towards understanding the power and limits of the Muon optimizer: A river-valley perspective. *arXiv preprint arXiv:2606.21514*, 2026. URL https://arxiv.org/abs/2606.21514.

Edward Tansley, Estelle Massart, and Coralia Cartis. On the neural feature ansatz for deep neural networks. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 5077–5085. PMLR, 2026. URL https://proceedings.mlr.press/v300/tansley26a.html.

Yuandong Tian. Li_2: A framework on dynamics of feature emergence and delayed generalization. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7ce5da35e01cfa8d303c2dc71e61a470-Abstract-Conference.html.

Shubhendu Trivedi, Jialei Wang, Samory Kpotufe, and Gregory Shakhnarovich. A consistent estimator of the expected gradient outerproduct. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pp. 819–828, 2014. URL https://www.columbia.edu/~skk2175/Papers/GOP-UAI.pdf.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e6231c5f46598cfd09ff1970524e0436-Abstract-Conference.html.

Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Catapults in SGD: Spikes in the training loss and their impact on generalization through feature learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62476–62509, 2024. doi: 10.48550/arXiv.2306.04815. URL https://proceedings.mlr.press/v235/zhu24h.html.

APPENDIX: CONTENTS

A. Learning toys: proofs, cycle classification and additional results p. 13
A.1 Proof of Lemma 6 (exact reduction) p. 14
A.2 Proof of Theorem 7 and Corollary 8 p. 14
A.3 The normalised-gradient recursion on a quadratic p. 15
A.4 The scalar cycle family and generic initialisation p. 20
A.5 Matrix Muon (M3): full statement and proof p. 21
A.6 Loss envelopes p. 22
A.7 Numerical verification of the statements p. 23
A.8 Additional figures p. 24
B. Proofs and supporting derivations for Section 5 p. 26
B.1 Proof of Proposition 2 (polar-update identities) p. 26
B.2 Period-2 orbits on the isotropic quadratic (Theorem 13) p. 27
B.3 Block-orthogonal equivariance p. 27
B.4 Assumptions and proof of conditional secular drift p. 28
C. Additional experiments and theory–experiment alignment p. 31
C.1 Plateau-search grid p. 31
C.2 Per-teacher configurations and the full ReLU grid p. 31
C.3 Configuration selection p. 31
C.4 Theory–experiment overlay and test-by-test detail p. 33
C.5 How theory and experiment align p. 38
D. Deep Gaussian teacher–student experiments p. 39
D.1 Data, teacher and student p. 39
D.2 Metrics and scoring protocol p. 39
D.3 Configurations and seeds p. 40
D.4 Depth results p. 40
E. Freezing the deep layers switches the plateau on p. 41
E.1 Setup p. 41
E.2 Results p. 42
E.3 Mechanism p. 44
F. A freezing principle for the plateau with feature learning p. 44
F.1 Depth and width p. 45
F.2 Training only W_1 after other schemes p. 45
F.3 Newton–Schulz and momentum: two regimes p. 46
F.4 Other architectures and scope p. 49
G. The mean-field theorem: complete proof p. 51
G.1 Main result p. 51
G.2 The mean-field map p. 54
G.3 The plateau and the loss drop p. 65
G.4 The $\alpha = 0$ trajectory from the Gaussian initialization p. 73
G.5 The good-set certificate at $\alpha = 0$ p. 76
G.6 Evaluating the certificate p. 87
G.7 Proof of the main theorem p. 95
G.8 Finite verification details p. 97
G.9 Complete certificate data p. 98

A LEARNING TOYS: PROOFS, CYCLE CLASSIFICATION AND ADDITIONAL RESULTS

Notation. We use the learning models of Sections 3.4 and 4. The sample matrix $X \in \mathbb{R}^{n\times p}$ has rows $x_i^\top$, and $\hat{\Sigma} := X^\top X/n = Q\,\mathrm{diag}(\mu_1,\dots,\mu_p)Q^\top$, where $Q = [q_1,\dots,q_p]$ is orthogonal and $\mu_1 > \cdots > \mu_p > 0$. The step size is $\eta > 0$. For M1/M2 the teacher is $u \neq 0$, the label

vector is $y = Xu$, and $z_t := Q^\top(w_t - u)$; $z_{t,\geq 2} := (z_{t,2}, \ldots, z_{t,p})$ and $\hat{G}_t := (\sum_j \mu_j^2 z_{t,j}^2)^{1/2}$. For M3, $1 \leq r < p$ and $W_t, U_\star \in \mathbb{R}^{p\times r}$ with $U_\star^\top U_\star = I_r$, $Y = XU_\star$, and $R_t := W_t - U_\star$. Write $\hat{D} := \mathrm{span}(q_1, \ldots, q_r)$, $R_{\hat{D}} := P_{\hat{D}} R$ and $R_\perp := (I_p - P_{\hat{D}})R$, where P_S is the orthogonal projector onto a subspace S. When R has full column rank with columns in $\hat{D}$, write $\hat{\Sigma}R = OP$ ($O^\top O = I_r$, $P \succ 0$) and $K := O^\top \hat{\Sigma} O$; K has eigenvalues $\mu_1, \ldots, \mu_r$. The empirical losses are $\hat{\mathcal{L}}(w) = n^{-1}\|Xw - y\|_2^2$ and $\hat{\mathcal{L}}(W) = n^{-1}\|XW - Y\|_F^2$; their symmetric-cycle floors are $\mu_1\eta^2/4$ and $\frac{\eta^2}{4}\sum_{i\leq r}\mu_i$, respectively. Here $\|\cdot\|_2$ on vectors is the Euclidean norm, $\|\cdot\|_{\mathrm{op}}$ on matrices is the spectral norm, and $\|\cdot\|_F$ is the Frobenius norm. We write $\sigma_i(B)$ for the singular values of B, $\mathrm{tr}(B)$ for its trace, and $B \succ 0$ for positive definiteness; $B \succ C$ means $B - C \succ 0$. For a compact SVD $B = U_B \Sigma_B V_B^\top$, the polar factor is $\mathrm{Polar}(B) := U_B V_B^\top$. For nonzero vectors, $\cos^2(a, b) := (a^\top b)^2/(\|a\|_2^2\|b\|_2^2)$; for equal-dimensional subspaces, $\cos^2_{\min}$ is the smallest squared principal-angle cosine, as in Eq. (9).

A.1 Exact reduction and its proof

Lemma 6 (exact reduction). *In M2 the residual coordinates $z_t = Q^\top(w_t - u)$ obey, under NGD,*

$$z_{t+1,k} = z_{t,k}\Big(1 - \frac{\eta\mu_k}{\hat{G}_t}\Big), \qquad \hat{G}_t = \Big(\sum_j \mu_j^2 z_{t,j}^2\Big)^{1/2}, \tag{15}$$

i.e. the normalised-gradient recursion Eq. (17) on a diagonal quadratic with $\lambda_k = \mu_k$, and $\hat{\mathcal{L}}(w_t) = \sum_k \mu_k z_{t,k}^2$ at every step. In M3 the residual obeys $R_{t+1} = R_t - \eta\,\mathrm{Polar}(\hat{\Sigma}R_t)$.

Positive gradient rescaling leaves NGD and Polar unchanged, so the gradient's factor 2 only doubles the reference quadratic loss $\frac{1}{2}\sum_k \lambda_k y_k^2$.

Proof. Since $y_i = u^\top x_i$, $Xw - y = X(w-u)$, hence $\hat{\mathcal{L}}(w) = \frac{1}{n}\|X(w-u)\|_2^2 = (w-u)^\top\hat{\Sigma}(w-u) = \sum_k \mu_k z_k^2$ and $\nabla\hat{\mathcal{L}}(w) = 2\hat{\Sigma}(w-u)$. The NGD step is therefore $w_{t+1} - u = (w_t - u) - \eta\,\hat{\Sigma}(w_t - u)/\|\hat{\Sigma}(w_t - u)\|_2$; multiplying by the orthogonal $Q^\top$ gives $z_{t+1,k} = z_{t,k} - \eta\mu_k z_{t,k}/(\sum_j \mu_j^2 z_{t,j}^2)^{1/2}$, which is Eq. (15) and coincides with Eq. (17) for $\lambda_k = \mu_k$. The loss $\frac{1}{2}\sum_k \lambda_k y_k^2$ evaluated at $y = z$ equals $\frac{1}{2}\hat{\mathcal{L}}$. For M3, $XW - Y = X(W - U_\star)$, so $\hat{\mathcal{L}}(W) = \mathrm{tr}(R^\top\hat{\Sigma}R)$, $\nabla\hat{\mathcal{L}} = 2\hat{\Sigma}R$ and $\mathrm{Polar}(2\hat{\Sigma}R) = \mathrm{Polar}(\hat{\Sigma}R)$. □

A.2 Scalar regression: plateau and teacher alignment

Theorem 7 (M2: loss plateau with teacher alignment). *Let $\rho_t^2 := \sum_{k\geq 2}\mu_k^2 z_{t,k}^2/(\mu_1\eta)^2$ and $q := \max_{k\geq 2}|1 - 2\mu_k/\mu_1| < 1$. Assume the phase-selection and margin hypotheses (H), (C) of Lemma 9 for z_t with $\lambda_k = \mu_k$. Then for all $t \geq T_1$:*

(i) Cycle: $z_{t,1} = (-1)^{t-T_1}\frac{\eta}{2} + \xi_\infty + O(\eta\rho_t^2)$ *for a constant* $\xi_\infty = O\big(\eta(\varepsilon_{\mathrm{phase}} + \rho_{T_1}^2)\big)$.

(ii) Plateau: $\hat{\mathcal{L}}(w_t) = \frac{\mu_1\eta^2}{4} + \sum_{k\geq 2}\mu_k z_{t,k}^2 + O\big(\mu_1\eta^2(\varepsilon_{\mathrm{phase}} + \rho_{T_1}^2 + \rho_t^2)\big)$.

(iii) Drain: *for $k \geq 2$, $z_{t+1,k} = \big(1 - \frac{2\mu_k}{\mu_1}\big)z_{t,k} + O\big(\frac{\mu_k}{\mu_1}(\varepsilon_{\mathrm{phase}} + \rho_{T_1}^2 + \rho_t^2)|z_{t,k}|\big)$; thus $\|z_{t,\geq 2}\|$ decays geometrically with some factor $\bar{q} \in (q, 1)$; q is the symmetric-cycle linearised reference.*

(iv) Teacher alignment: *with the cycle point $w_t^{\mathrm{cyc}} := u + z_{t,1}q_1$ (assumed nonzero, as is w_t),*

$$\big|\cos^2(w_t, u) - \cos^2(w_t^{\mathrm{cyc}}, u)\big| \;\leq\; \|z_{t,\geq 2}\|/\|w_t^{\mathrm{cyc}}\|, \tag{16}$$

where $\cos^2(w_t^{\mathrm{cyc}}, u) = 1$ if $u \parallel q_1$, and $\cos^2(w_t^{\mathrm{cyc}}, u) = \|u\|^2/(\|u\|^2 + z_{t,1}^2)$ if $u \perp q_1$, which for a unit teacher is the cycle cap $(1 + \eta^2/4)^{-1}$ *up to $O(\eta|\xi_\infty| + \eta^2\rho_t^2)$.*

Items (i)–(iii) follow from Lemmas 9 and 6; (iv) converts residual drain into teacher alignment. A small-μ_k mode can carry substantial misalignment but little loss, at the cost of slow decay:

Corollary 8 (invisible misalignment drains slowly)**.** *In the setting of Theorem 7, let $k \geq 2$ and $f \in (0,1)$. If off-mode k carries amplitude $|z_{t,k}| = Z > 0$ while contributing at most a fraction f of the floor, $\mu_k Z^2 \leq f\mu_1\eta^2/4$, then its leading-order contraction factor in (iii) is at least $1 - f\eta^2/(2Z^2)$, so to leading order it needs at least $2Z^2/(f\eta^2) - 1$ steps to shrink by a factor e (a vacuous bound when $f\eta^2 \geq 2Z^2$).*

At $\eta = 0.3$, a unit misalignment hidden below 5% of the floor needs at least 443 steps per e-fold: on the symmetric cycle, feature learning that the loss cannot see is necessarily a long phase, and it happens *inside* the plateau (Figure 1a,b). Without (H) the cycle can settle on another member $\{a, a-\eta\}$ of the period-2 family: Proposition 10 gives its plateau, raised by $2(a^2 + (\eta - a)^2)/\eta^2$, and its first-order drain (modes with $\eta\mu_k/\mu_1 < \min(a, \eta - a)$ drain faster, e-fold ratio $4a(\eta - a)/\eta^2$), but that the plateau then stays flat is a numerical finding (orange in Figure 1).

Here $\rho_t^2 := \sum_{k\geq 2} \mu_k^2 z_{t,k}^2/(\mu_1\eta)^2$ and $q := \max_{k\geq 2} |1 - 2\mu_k/\mu_1|$. The handoff time T_1 and phase tolerance $\varepsilon_{\text{phase}}$ are those of hypotheses Eq. (21) and Eq. (22) in Lemma 9; ξ_∞ is its limiting phase shift. The comparison point is $w_t^{\text{cyc}} := u + z_{t,1}q_1$.

Proof of Theorem 7. By Lemma 6, (z_t) is a trajectory of Eq. (17) for the diagonal quadratic with $\lambda_k = \mu_k$ (the rotation Q is irrelevant to Lemma 9, which is stated in eigen-coordinates), and $\hat{\mathcal{L}}(w_t) = 2L(z_t)$. Items (i) and (iii) are Lemma 9(i),(iii) verbatim; (ii) is Lemma 9(ii) multiplied by 2. By (iii) and the margin hypothesis (C), $|z_{t+1,k}| \leq \bar{q}|z_{t,k}|$ for some $\bar{q} \in (q, 1)$, uniformly over $k \geq 2$ and $t \geq T_1$. The factor q describes the symmetric-cycle linearisation, not an exact asymptotic rate for every phase-shifted trajectory.

For (iv), $w_t - w_t^{\text{cyc}} = \sum_{k\geq 2} z_{t,k}q_k$ has norm $\|z_{t,\geq 2}\|_2$. For nonzero a, b, u write $\hat{a} = a/\|a\|_2$ and let $\Pi_a^\perp$ be the orthogonal projector onto $a^\perp$. Since $\cos^2(a,u) = \hat{u}^\top \hat{a}\hat{a}^\top \hat{u}$,

$$\begin{aligned} |\cos^2(a,u) - \cos^2(b,u)| &\leq \|\hat{a}\hat{a}^\top - \hat{b}\hat{b}^\top\|_{\text{op}} = \sin\angle(a,b) \\ &= \frac{\|\Pi_a^\perp b\|_2}{\|b\|_2} = \frac{\|\Pi_a^\perp(b-a)\|_2}{\|b\|_2} \leq \frac{\|a-b\|_2}{\|b\|_2}. \end{aligned}$$

which gives Eq. (16) with $a = w_t$, $b = w_t^{\text{cyc}}$. If $u \parallel q_1$ then $w_t^{\text{cyc}} \parallel u$ and the cosine is 1 (when $w_t^{\text{cyc}} \neq 0$). If $u \perp q_1$ then $\langle w_t^{\text{cyc}}, u\rangle = \|u\|_2^2$ and $\|w_t^{\text{cyc}}\|_2^2 = \|u\|_2^2 + z_{t,1}^2$, so $\cos^2(w_t^{\text{cyc}}, u) = \|u\|_2^2/(\|u\|_2^2 + z_{t,1}^2)$; inserting (i), $z_{t,1}^2 = \eta^2/4 + O(\eta|\xi_\infty| + \eta^2\rho_t^2)$. □

Proof of Corollary 8. The hypothesis gives $\delta := 2\mu_k/\mu_1 \leq f\eta^2/(2Z^2)$, so the leading-order factor $1 - 2\mu_k/\mu_1 = 1 - \delta$ of Theorem 7(iii) is at least $1 - f\eta^2/(2Z^2)$. If $\delta < 1$, the number of steps to shrink by e at the constant factor $1 - \delta \in (0,1)$ is $1/(-\log(1-\delta))$, and $-\log(1-\delta) \leq \delta/(1-\delta)$ gives $1/(-\log(1-\delta)) \geq 1/\delta - 1 \geq 2Z^2/(f\eta^2) - 1$; if $\delta \geq 1$ the bound is ≤ 0 and there is nothing to prove. □

A.3 The normalised-gradient recursion on a quadratic

Theorem 7 is proved by transporting Lemma 9 below through Lemma 6. This appendix states the lemma and proves it.

Setting. Let $w \in \mathbb{R}^p$, $L(w) = \frac{1}{2}w^\top Aw$ with $A = QDQ^\top$, $D = \text{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_p > 0$. Here Q is orthogonal, $q_k := Qe_k$ is the kth eigenvector, and e_k is the kth standard coordinate vector. The step size is $\eta > 0$; all vector norms in this subsection are Euclidean. This local D is the diagonal eigenvalue matrix, not the population data subspace used for M3. In eigenbasis coordinates $y_t := Q^\top w_t$, NGD decouples component-wise:

$$y_{t+1,k} = y_{t,k}\big(1 - \tfrac{\eta\lambda_k}{G_t}\big), \qquad G_t = \big(\textstyle\sum_{j=1}^p \lambda_j^2 y_{t,j}^2\big)^{1/2}. \tag{17}$$

The gradient norm is $G_t = \|\nabla L(w_t)\|_2 = \|Dy_t\|_2$. We write $y_{t,\geq 2} := (y_{t,2}, \ldots, y_{t,p})$ for the off-leading coordinates. The target direction is the leading eigenvector $q_1 = Qe_1$, and

$$\text{align}_t = \cos^2\angle(w_t, q_1) = y_{t,1}^2/\|y_t\|_2^2. \tag{18}$$

Lemma 9 is stated conditionally from a phase-selected handoff T_1, because the scalar normalized map has a continuum of period-2 cycles $\{a, a-\eta\}$.

Statement. With this notation, define

$$\rho_t^2 := \frac{\sum_{k\geq 2} \lambda_k^2 y_{t,k}^2}{(\lambda_1 \eta)^2}. \tag{19}$$

We state the result conditionally from a phase-selected handoff time T_1. This is necessary because the scalar normalized map has a continuum of period-2 cycles.

Lemma 9 (Conditional plateau with alignment for normalised gradient descent on a quadratic)**.** *Let*

$$q \;:=\; \max_{k\geq 2} \left|1 - \tfrac{2\lambda_k}{\lambda_1}\right| \;<\; 1 \tag{20}$$

(strict from the spectral gap $\lambda_1 > \lambda_2$ and $\lambda_k > 0$ for $k \geq 2$). Fix $C_H > 0$. There exists $c_\star > 0$, depending only on C_H and the fixed eigenvalue ratios λ_k/λ_1, for which the following holds. Suppose at some time T_1, for $\varepsilon_{\text{phase}} \geq 0$,

$$\text{(H)} \qquad \left| y_{T_1,1} - \frac{\eta}{2} \right| \leq C_H \eta \varepsilon_{\text{phase}}, \tag{21}$$

and that the perturbation lies inside the contraction margin,

$$\text{(C)} \qquad \varepsilon_{\text{phase}} + \rho_{T_1}^2 + \rho_t^2 \;\leq\; c_\star(1-q) \qquad \text{for all } t \geq T_1. \tag{22}$$

All $O(\cdot)$ constants below may depend on C_H and these eigenvalue ratios, but not on t or η. Then for all $t \geq T_1$:

(i) *There exists a constant phase shift $\xi_\infty = O\big(\eta(\varepsilon_{\text{phase}} + \rho_{T_1}^2)\big)$ such that*

$$y_{t,1} = (-1)^{t-T_1}\frac{\eta}{2} + \xi_\infty + O(\eta \rho_t^2). \tag{23}$$

(ii) *The loss satisfies*

$$L(w_t) = \frac{\lambda_1 \eta^2}{8} + \frac{1}{2}\sum_{k\geq 2} \lambda_k y_{t,k}^2 + O\big(\lambda_1 \eta^2 (\varepsilon_{\text{phase}} + \rho_{T_1}^2 + \rho_t^2)\big)\,. \tag{24}$$

(iii) *For each $k \geq 2$,*

$$y_{t+1,k} = y_{t,k}\left(1 - \frac{2\lambda_k}{\lambda_1}\right) + O\big((\lambda_k/\lambda_1)(\varepsilon_{\text{phase}} + \rho_{T_1}^2 + \rho_t^2)|y_{t,k}|\big)\,. \tag{25}$$

(iv) *For some $\bar{q} \in (q, 1)$,*

$$\|y_{t,\geq 2}\|_2 \leq \bar{q}^{\,t-T_1}\|y_{T_1,\geq 2}\|_2, \qquad 1 - \frac{y_{t,1}^2}{\|y_t\|_2^2} \leq \frac{16\|y_{T_1,\geq 2}\|_2^2}{\eta^2}\,\bar{q}^{\,2(t-T_1)}.$$

The worst squared multiplier of the symmetric-cycle linearisation is q^2; this need not be the exact asymptotic rate of an individual trajectory.

The contraction-margin hypothesis Eq. (22) ensures the *perturbed* per-step factor $|1 - 2\lambda_k/\lambda_1| + O(\varepsilon_{\text{phase}} + \rho_{T_1}^2 + \rho_t^2)$ stays uniformly below 1; the strict spectral gap $q < 1$ is necessary, the margin-relative smallness of the perturbation in (C) is what makes the geometric off-mode decay (used both in the bootstrap of (i)–(iii) and the conclusion of (iv)) uniform in t. The absolute value in the definition of q accounts for sign-flipping off-modes with $\lambda_k > \lambda_1/2$. The handoff hypothesis (H) is assumed; no entry-time result is asserted.

Proof of Lemma 9.

Setup. Work in the eigenbasis of A. With $y_t := Q^\top w_t$, the NGD update is

$$y_{t+1,k} = y_{t,k}\left(1 - \frac{\eta\lambda_k}{G_t}\right), \qquad G_t := \|Dy_t\|_2 = \left(\sum_j \lambda_j^2 y_{t,j}^2\right)^{1/2}. \tag{26}$$

Recall that

$$\rho_t^2 = \frac{\sum_{k\geq 2} \lambda_k^2 y_{t,k}^2}{(\lambda_1 \eta)^2}. \tag{27}$$

The proof starts from the handoff hypothesis Eq. (21); it does not require or assert convergence into this neighbourhood from arbitrary initialisation.

Remark on phase selection. If $\rho_t = 0$, the leading-mode map reduces to

$$y_{t+1,1} = y_{t,1} - \eta \operatorname{sgn}(y_{t,1}). \tag{28}$$

This scalar map has the continuum of period-2 cycles $\{a, a-\eta\}$, $a \in (0, \eta)$. Only the symmetric cycle $a = \eta/2$ has constant leading loss $\lambda_1\eta^2/8$. For a generic cycle $\{a, a-\eta\}$, the leading loss alternates between

$$\frac{1}{2}\lambda_1 a^2 \qquad \text{and} \qquad \frac{1}{2}\lambda_1(a-\eta)^2. \tag{29}$$

Thus hypothesis Eq. (21) is a phase-selection condition.

Step A: Leading-mode residual recurrence. Let

$$s_t := (-1)^{t-T_1}, \qquad y_{t,1} = s_t\frac{\eta}{2} + \sigma_t. \tag{30}$$

By Eq. (21),

$$\sigma_{T_1} = O(\eta\varepsilon_{\text{phase}}). \tag{31}$$

Assume temporarily that $|\sigma_t|/\eta \ll 1$, so that $\operatorname{sgn}(y_{t,1}) = s_t$. Then

$$y_{t,1} = s_t\frac{\eta}{2}\left(1 + 2s_t\sigma_t/\eta\right). \tag{32}$$

Using Eq. (27), factor G_t as

$$G_t = \frac{\lambda_1\eta}{2}\left(1 + 2s_t\sigma_t/\eta\right)\sqrt{1+v_t}, \qquad v_t := \frac{4\rho_t^2}{(1+2s_t\sigma_t/\eta)^2}. \tag{33}$$

The cancellation used below is exact:

$$\frac{y_{t,1}}{1+2s_t\sigma_t/\eta} = \frac{s_t\eta}{2}. \tag{34}$$

Hence

$$\frac{\eta\lambda_1 y_{t,1}}{G_t} = \frac{s_t\eta}{\sqrt{1+v_t}}. \tag{35}$$

Substituting into the leading-coordinate update gives

$$\begin{aligned} y_{t+1,1} &= y_{t,1} - \frac{s_t\eta}{\sqrt{1+v_t}} \\ &= s_t\frac{\eta}{2} + \sigma_t - s_t\eta\left(1 - \frac{1}{2}v_t + O(v_t^2)\right) \\ &= -s_t\frac{\eta}{2} + \sigma_t + \frac{s_t\eta}{2}v_t + O(\eta v_t^2). \end{aligned} \tag{36}$$

Since

$$v_t = 4\rho_t^2 + O\big(\rho_t^2|\sigma_t|/\eta\big), \tag{37}$$

and $s_{t+1} = -s_t$, identifying $y_{t+1,1} = s_{t+1}\eta/2 + \sigma_{t+1}$ yields

$$\sigma_{t+1} = \sigma_t + 2s_t\eta\rho_t^2 + E_t, \qquad E_t = O\big(\eta\rho_t^4 + \rho_t^2|\sigma_t|\big). \tag{38}$$

Here E_t denotes the deterministic remainder in this recurrence. Every error term carries a factor of ρ_t^2. In particular, if $\rho_t = 0$, then $\sigma_{t+1} = \sigma_t$ exactly.

Step B: Bootstrap and convergence of the phase shift. Set

$$\varepsilon_\star := \varepsilon_{\text{phase}} + \rho_{T_1}^2. \tag{39}$$

We show that $|\sigma_t| = O(\eta\varepsilon_\star)$. Assume on an interval $T_1 \le s \le T$ that

$$|\sigma_s| \le B\eta\varepsilon_\star. \tag{40}$$

Then

$$G_s = \frac{\lambda_1\eta}{2}\left(1 + O(\varepsilon_\star + \rho_s^2)\right). \tag{41}$$

Therefore, for each $k \geq 2$,

$$y_{s+1,k} = y_{s,k}\left(1 - \frac{\eta\lambda_k}{G_s}\right) = y_{s,k}\left(1 - \frac{2\lambda_k}{\lambda_1} + O\big((\lambda_k/\lambda_1)(\varepsilon_\star + \rho_s^2)\big)\right). \tag{42}$$

Under the contraction-margin hypothesis Eq. (22), the perturbed per-step factor in the off-mode update satisfies $|1 - 2\lambda_k/\lambda_1| + O(\varepsilon_\star + \rho_s^2) \leq q'$ for some $q' \in (q, 1)$ uniformly over $k \geq 2$ and $s \geq T_1$, hence

$$\rho_{s+1} \leq q'\rho_s. \tag{43}$$

Consequently,

$$\sum_{s=T_1}^{T} \rho_s^2 \leq C\rho_{T_1}^2, \qquad \sum_{s=T_1}^{T} \rho_s^4 \leq C\rho_{T_1}^4, \tag{44}$$

where C may depend on the spectral contraction margin but not on T.

Telescoping Eq. (38),

$$\sigma_t = \sigma_{T_1} + 2\eta \sum_{s=T_1}^{t-1} s_s\rho_s^2 + \sum_{s=T_1}^{t-1} E_s. \tag{45}$$

The main sum is alternating with geometrically decreasing magnitudes, hence

$$\left|\sum_{s=T_1}^{t-1} s_s\rho_s^2\right| \leq C\rho_{T_1}^2. \tag{46}$$

Using Eq. (40) and Eq. (44),

$$\begin{aligned}\sum_{s=T_1}^{t-1} |E_s| &\leq C\eta \sum_{s=T_1}^{t-1} \rho_s^4 + C \sum_{s=T_1}^{t-1} \rho_s^2|\sigma_s| \\ &\leq C\eta\rho_{T_1}^4 + CB\eta\varepsilon_\star\rho_{T_1}^2 \\ &\leq C_B\eta\varepsilon_\star^2.\end{aligned} \tag{47}$$

Combining Eq. (31), Eq. (45), Eq. (46), and Eq. (47),

$$|\sigma_t| \leq C_1\eta\varepsilon_\star + C_2B\eta\varepsilon_\star^2. \tag{48}$$

Choose $B > 2C_1$, and then take $\varepsilon_\star$ sufficiently small so that $C_2\varepsilon_\star \leq 1/2$. The usual first-exit argument closes the bootstrap:

$$|\sigma_t| \leq B\eta\varepsilon_\star \qquad \text{for all } t \geq T_1. \tag{49}$$

The same estimates show that $\sum_{s \geq T_1} E_s$ converges absolutely. Define

$$\xi_\infty := \sigma_{T_1} + 2\eta \sum_{s=T_1}^{\infty} s_s\rho_s^2 + \sum_{s=T_1}^{\infty} E_s. \tag{50}$$

Then

$$\xi_\infty = O(\eta\varepsilon_{\text{phase}}) + O(\eta\rho_{T_1}^2) = O(\eta\varepsilon_\star). \tag{51}$$

Applying the same argument to the tail from t to ∞ gives

$$|\sigma_t - \xi_\infty| = O(\eta\rho_t^2). \tag{52}$$

Therefore

$$y_{t,1} = (-1)^{t-T_1}\frac{\eta}{2} + \xi_\infty + O(\eta\rho_t^2), \tag{53}$$

which proves Eq. (23).

Step C: Off-mode update. From Eq. (53),

$$G_t = \frac{\lambda_1 \eta}{2}\left(1 + O(\varepsilon_\star + \rho_t^2)\right). \tag{54}$$

Substituting into Eq. (26), for $k \geq 2$,

$$\begin{aligned} y_{t+1,k} &= y_{t,k}\left(1 - \frac{\eta\lambda_k}{G_t}\right) \\ &= y_{t,k}\left(1 - \frac{2\lambda_k}{\lambda_1}\right) + O\big((\lambda_k/\lambda_1)(\varepsilon_\star + \rho_t^2)|y_{t,k}|\big). \end{aligned} \tag{55}$$

This proves Eq. (25).

Step D: Loss expansion. Using Eq. (53),

$$y_{t,1}^2 = \frac{\eta^2}{4} + O\big(\eta^2(\varepsilon_\star + \rho_t^2)\big). \tag{56}$$

Hence

$$\begin{aligned} L(w_t) &= \frac{1}{2}\lambda_1 y_{t,1}^2 + \frac{1}{2}\sum_{k\geq 2}\lambda_k y_{t,k}^2 \\ &= \frac{\lambda_1\eta^2}{8} + \frac{1}{2}\sum_{k\geq 2}\lambda_k y_{t,k}^2 + O\big(\lambda_1\eta^2(\varepsilon_\star + \rho_t^2)\big). \end{aligned} \tag{57}$$

This proves Eq. (24).

Step E: Alignment bound and linearised rate. By Eq. (49), taking $c_\star$ sufficiently small ensures $|y_{t,1}| \geq \eta/4$ for all $t \geq T_1$. The componentwise bound used in Eq. (43) also gives $\|y_{t,\geq 2}\|_2 \leq \bar{q}^{\,t-T_1}\|y_{T_1,\geq 2}\|_2$ for some $\bar{q} \in (q, 1)$. Hence

$$1 - \frac{y_{t,1}^2}{\|y_t\|_2^2} = \frac{\|y_{t,\geq 2}\|_2^2}{y_{t,1}^2 + \|y_{t,\geq 2}\|_2^2} \leq \frac{16}{\eta^2}\,\bar{q}^{\,2(t-T_1)}\|y_{T_1,\geq 2}\|_2^2. \tag{58}$$

This proves (iv). On the exactly symmetric limiting cycle, the linearised k-th off-mode multiplier is $1 - 2\lambda_k/\lambda_1$, whose worst squared magnitude is q^2. A nonzero limiting phase shift instead gives alternating multipliers; even on the symmetric cycle, a mode absent from the trajectory need not determine its asymptotic rate. The rigorous conclusion here is the uniform geometric bound above. □

Numerical verification. $p = 8$, $\lambda_1 = 1$, off-mode tail $[0.30, 0.10, 0.05, 0.02, 0.01, 0.005, 0.001]$, symmetric init $y_{0,1} = (k + \frac{1}{2})\eta$, off-modes $y_{0,j} = c\,\eta$ for $j \geq 2$, $n_{\text{steps}} = 4000$. The measured loss is the mean of the final 200 iterates; decay_8 is $\exp(b)$, where b is the least-squares slope of $\log|y_{t,8}|$ over $t = 2000, \ldots, 4000$.

Table 1: Finite-run comparison with the symmetric-cycle reference: the averaged loss is within 1.1% of the reference floor, and the fitted slow-mode multiplier is within 10^{-4} of 0.998. These initialisations give $\rho_0^2 \leq 0.092$ and the displayed starting alignments; they do not certify the conditional handoff hypotheses. Final residual alignment rounds to 1.000. Small nonzero phase shifts remain, so 0.998^2 is a linearised squared-rate reference, not an exact rate established by this table.

η	k	c	align(0)	$\lambda_1\eta^2/8$ pred	L_{late} measured	$1 - 2\lambda_8/\lambda_1$ pred	decay_8 measured
0.10	50	0.70	0.999	0.001250	0.001251	0.998	0.998
0.20	8	0.94	0.921	0.005000	0.005033	0.998	0.998
0.50	2	0.94	0.503	0.031250	0.031583	0.998	0.998

A.4 THE SCALAR CYCLE FAMILY AND GENERIC INITIALISATION

Proposition 10 (the scalar cycle family)**.** *In M2 (or M1):*

(a) *If $z_{T,k} = 0$ for all $k \geq 2$, the off-modes stay zero and $z_{t+1,1} = z_{t,1} - \eta \operatorname{sgn}(z_{t,1})$. Every orbit that does not hit 0 is, after at most $\lceil |z_{T,1}|/\eta \rceil$ steps, the period-2 orbit $\{a, a-\eta\}$ for some $a \in (0,\eta)$. The loss alternates between $\mu_1 a^2$ and $\mu_1(\eta-a)^2$, and its two-step mean is the floor $\mu_1\eta^2/4$ times $2(a^2+(\eta-a)^2)/\eta^2 \geq 1$, with equality iff $a = \eta/2$.*

(b) *Once mode 1 is cycling, i.e. along any stretch of trajectory with $a_t := |z_{t,1}| \in [c\eta, (1-c)\eta]$ for some $c > 0$, write $\chi_k := \eta\mu_k/\mu_1$ and $\tilde{\rho}_t := (\sum_{j\geq 2} \mu_j^2 z_{t,j}^2)^{1/2}/(\mu_1 a_t)$. As $\tilde{\rho}_t \to 0$ with c fixed,*

$$z_{t+2,k} = \Big(1 - \frac{\chi_k}{a_t}\Big)\Big(1 - \frac{\chi_k}{\eta - a_t}\Big) z_{t,k} + O_c\big(\tilde{\rho}_t^2 |z_{t,k}|\big), \qquad k \geq 2.$$

If $\chi_k < \min(a_t, \eta - a_t)$, the displayed leading-order factor lies in $\big(0, (1-2\mu_k/\mu_1)^2\big]$, with equality iff $a_t = \eta/2$. Thus the linearised map at fixed cycle amplitude drains these modes at least as fast as the symmetric-cycle linearisation; the e-fold time is shortened by $4a_t(\eta - a_t)/\eta^2$ to leading order in χ_k. If $\chi_k \geq \min(a_t, \eta - a_t)$ the factor can be negative, or exceed 1 in absolute value.

(c) *(Selection at frozen off-mode content.) For a fixed $g \in (0, \mu_1\eta/2)$, the map $T_g(y) = y - \eta\mu_1 y/\sqrt{\mu_1^2 y^2 + g^2}$ (the mode-1 update with the off-mode gradient norm frozen at g) has the unique symmetric 2-cycle $\{a_g, -a_g\}$, $a_g^2 = (\eta/2)^2 - g^2/\mu_1^2$, which is locally attracting with two-step multiplier $(1 - s'(a_g))^2$, $s'(a_g) = 8g^2/(\mu_1^2\eta^2) \in (0,2)$.*

Proof. (a) With $z_{t,\geq 2} = 0$ the gradient is $2\mu_1 z_{t,1} q_1$ and its normalisation is $\operatorname{sgn}(z_{t,1}) q_1$; the off-mode coordinates are multiplied by $1 - \eta\mu_k/\hat{G}_t$ and stay 0. The sign map decreases $|z|$ by η until $|z| < \eta$ (unless it lands on 0), after which $a \mapsto a - \eta \mapsto a$. The loss values follow from $\hat{\mathcal{L}} = \mu_1 z_1^2$, and $a^2 + (\eta - a)^2 = \eta^2/2 + 2(a - \eta/2)^2$. (b) Exactly, $z_{t+1,k} = z_{t,k}(1 - \eta\mu_k/\hat{G}_t)$ with $\hat{G}_t = \mu_1 a_t (1+\tilde{\rho}_t^2)^{1/2}$, and $z_{t+1,1} = z_{t,1} - \eta \operatorname{sgn}(z_{t,1})(1+\tilde{\rho}_t^2)^{-1/2}$; for $a_t < \eta(1+\tilde{\rho}_t^2)^{-1/2}$ the sign flips, which gives the exact *sum rule* $|z_{t,1}| + |z_{t+1,1}| = \eta(1+\tilde{\rho}_t^2)^{-1/2}$, hence $a_{t+1} = \eta - a_t + O(\eta\tilde{\rho}_t^2)$. Multiplying the two one-step factors gives the display. For sufficiently small $\tilde{\rho}_t$ depending on c, the sign-flip condition holds and the next amplitude remains bounded below by $c\eta/2$; the off-mode multipliers are then bounded in terms of c, so $\tilde{\rho}_{t+1} = O_c(\tilde{\rho}_t)$. This also justifies the second-step expansion. Writing $\chi = \chi_k$ and $b = \eta - a_t$, $(1 - \frac{\chi}{a_t})(1 - \frac{\chi}{b}) = 1 - \frac{\chi(\eta-\chi)}{a_t b}$ because $a_t + b = \eta$; since $a_t b \leq \eta^2/4$ with equality iff $a_t = b$, this is at most $1 - 4\chi(\eta - \chi)/\eta^2 = (1 - 2\chi/\eta)^2$, and it is positive when $\chi < \min(a_t, b)$. For small χ its logarithm is $-\chi\eta/(a_t b) + O(\chi^2)$ against $-4\chi/\eta + O(\chi^2)$ on the symmetric cycle, which gives the e-fold ratio $4a_t b/\eta^2$. (c) A symmetric 2-cycle requires $T_g(a) = -a$, i.e. $s(a) = 2a$ with $s(y) = \eta\mu_1 y/\sqrt{\mu_1^2 y^2 + g^2}$; $s(a)/a$ decreases strictly from $\eta\mu_1/g > 2$ to 0, so the solution is unique and $\mu_1^2 a_g^2 + g^2 = \mu_1^2\eta^2/4$. Since s is odd, $T_g'(\pm a_g) = 1 - s'(a_g)$ with $s'(y) = \eta\mu_1 g^2/(\mu_1^2 y^2 + g^2)^{3/2}$, which at a_g equals $8g^2/(\mu_1^2\eta^2)$; it lies in $(0,2)$ iff $g < \mu_1\eta/2$. □

Part (c) explains the two behaviours of Figure 1. The off-mode gradient norm $g_t^2 = \sum_{k\geq 2} \mu_k^2 z_{t,k}^2$ carries a *second* power of μ_k, so a mode that is invisible in the loss (small $\mu_k z_k^2$) exerts almost no symmetrising pressure, and a generic start freezes an asymmetric member of the family of (a); loss-visible slow modes, by contrast, symmetrise the cycle on their way out (the $\eta = 0.1$ and M1 "slow" runs, Figure 6). Writing $m_t = (z_{t,1} + z_{t+1,1})/2$ for the cycle midpoint, summing its frozen-gradient linearised contraction $1 - s_t'$ over a symmetric-rate drain gives, in the additional slow-mode regime $\mu_k/\mu_1 \ll 1$ for the modes carrying the off-mode loss, the heuristic *energy budget* $-\log(m_\infty/m_{t_0}) \approx \sum_{t\geq t_0} s_t' \approx E_{\text{off}}(t_0)/(2L_{\text{floor}})$, with $E_{\text{off}}(t_0) = \sum_{k\geq 2} \mu_k z_{t_0,k}^2$ and $L_{\text{floor}} = \mu_1\eta^2/4$: the cycle symmetrises by the off-mode loss it still has to dissipate, in floor units. Here the logarithm concerns nonzero midpoints of the same sign. This uses the symmetric-cycle rate, the frozen-g multiplier and the slow-mode approximation; a hidden-eigenvalue sweep gives fitted slope 1.09, a numerical comparison rather than a proof.

A.5 MATRIX MUON (M3): FULL STATEMENT AND PROOF

This is the full version of Proposition 1.

Proposition 11 (M3: exact weight 2-cycles, floor, drain and teacher alignment)**.** *Let $R_t := W_t - U_\star$, with fixed teacher $U_\star$, and assume $\mu_r > \mu_{r+1}$. Then:*

(i) Step and loss: $\|W_{t+1} - W_t\|_F = \eta\sqrt{r}$ *whenever R_t has full column rank, and* $\hat{\mathcal{L}}(W_t) = \operatorname{tr}(R_{\hat{D},t}^\top \hat{\Sigma} R_{\hat{D},t}) + \operatorname{tr}(R_{\perp,t}^\top \hat{\Sigma} R_{\perp,t})$.

(ii) Exact weight 2-cycles: *if R_T has columns in $\hat{D}$ and $0 \prec P \prec \eta K$, then $W_{t+2} = W_t$ for all $t \geq T$, with $W_{T+1} = W_T - \eta O$. The two phase losses are* $\operatorname{tr}(PK^{-1}P)$ *and* $\operatorname{tr}(P'K^{-1}P')$ *(possibly equal), where $P' := \eta K - P$; their mean is at least $\frac{\eta^2}{4}\sum_{i\leq r}\mu_i$, with equality iff $W_T = U_\star + \frac{\eta}{2}O$ and $W_{T+1} = U_\star - \frac{\eta}{2}O$ (the symmetric residual cycle, with all $\sigma_i(R_T) = \eta/2$).*

(iii) Drain of the complement (linearised): *to first order in $R_\perp$, around a cycle of (ii) each row $j > r$ of $R_\perp$ in the eigenbasis of $\hat{\Sigma}$ is multiplied per two steps by $(I - \eta\mu_j P^{-1})(I - \eta\mu_j P'^{-1})$, a strict Euclidean contraction on that row when $P, P' \succ \frac{\eta\mu_{r+1}}{2} I$, while the cycle moves only at $O(\|R_\perp\|_F^2)$. On the symmetric cycle the per-step map is $R_\perp \mapsto R_\perp - 2\hat{\Sigma}R_\perp K^{-1}$: the (j,i) coefficient in eigenbases ($j > r \geq i$) is multiplied by $1 - 2\mu_j/\mu_i \in (-1,1)$.*

(iv) Teacher alignment: *let R have columns in $\hat{D}$. If* $\operatorname{range} U_\star \perp \hat{D}$*, the multiset of principal-angle cosines between* $\operatorname{range} W$ *and* $\operatorname{range} U_\star$ *is exactly $\{(1 + \sigma_i(R)^2)^{-1/2} : 1 \leq i \leq r\}$, so $\cos^2_{\min} = (1 + \eta^2/4)^{-1}$ on the symmetric cycle; if $\operatorname{range} U_\star = \hat{D}$ and $\|R\|_{\mathrm{op}} < 1$, then $\cos^2_{\min} = 1$.*

Part (ii) holds on any r-dimensional $\hat{\Sigma}$-invariant subspace S, with floor $\frac{\eta^2}{4}\operatorname{tr}(\hat{\Sigma}|_S)$, for every positive-definite quadratic. For block-isotropic $\hat{\Sigma}$ ($K = \lambda_D I_r$), the cycle condition is $\sigma_i(R_T) \in (0, \eta)$ and the floor is $\lambda_D \eta^2 r/4$. On an exact weight cycle, loss and teacher alignment (when defined) repeat every two steps. Continued alignment change concerns off-cycle trajectories; part (iii) gives only their local transverse linearisation, not a convergence theorem. Numerical checks are in Appendix A.7.

Proof. At the cycle point R_T, write the full-column-rank polar decomposition $\hat{\Sigma}R_T = OP$, with $O^\top O = I_r$ and $P = (R_T^\top \hat{\Sigma}^2 R_T)^{1/2} \succ 0$. Set $K := O^\top \hat{\Sigma} O$ and $P' := \eta K - P$. In part (iii), a prime on a residual denotes its one-step update, and e_i denotes a standard coordinate vector. All matrix remainders there are measured in Frobenius norm with the cycle point fixed.

Proof. (i) $\operatorname{rank}(\hat{\Sigma}R_t) = \operatorname{rank}(R_t) = r$ because $\hat{\Sigma} \succ 0$, so Proposition 2(ii) gives $\|W_{t+1} - W_t\|_F = \eta\sqrt{r}$. Since $\hat{D}$ and $\hat{D}^\perp$ are $\hat{\Sigma}$-invariant and orthogonal, $R_{\hat{D}}^\top \hat{\Sigma} R_\perp = 0$ and the loss splits.

(ii) The columns of $\hat{\Sigma}R_T$ lie in $\hat{D}$ and $P \succ 0$, so O is an orthonormal basis of $\hat{D}$ (here $\dim \hat{D} = r$ is used), $OO^\top$ is the projector onto $\hat{D}$, and invariance gives $\hat{\Sigma}O = OO^\top \hat{\Sigma} O = OK$. Hence $R_{T+1} = R_T - \eta O$ has columns in $\hat{D}$ and $\hat{\Sigma}R_{T+1} = OP - \eta OK = -O(\eta K - P) = -OP'$ with $P' \succ 0$. By uniqueness of the polar factor of a full-column-rank matrix, $\operatorname{Polar}(\hat{\Sigma}R_{T+1}) = -O$, so $R_{T+2} = R_{T+1} + \eta O = R_T$. The hypothesis is symmetric in (P, P') and $(O, -O)$ (note $(-O)^\top \hat{\Sigma}(-O) = K$), so the argument repeats and the residual orbit is 2-periodic. Since $W_t = U_\star + R_t$ with fixed $U_\star$, $W_{t+2} = W_t$ as well. Its phases are distinct because $\|W_{T+1} - W_T\|_F = \eta\|O\|_F = \eta\sqrt{r} > 0$. From $O^\top \hat{\Sigma} R_T = P$ and $O^\top \hat{\Sigma} R_T = O^\top \hat{\Sigma} O O^\top R_T = KO^\top R_T$ we get $R_T = OK^{-1}P$, hence $\hat{\mathcal{L}}(W_T) = \operatorname{tr}(PK^{-1}O^\top \hat{\Sigma} OK^{-1}P) = \operatorname{tr}(PK^{-1}P)$, and likewise $R_{T+1} = -OK^{-1}P'$ with loss $\operatorname{tr}(P'K^{-1}P')$. The map $X \mapsto \operatorname{tr}(XK^{-1}X) = \|K^{-1/2}X\|_F^2$ is strictly convex on symmetric matrices and $P + P' = \eta K$, so the two-step mean is at least its value at $X = \eta K/2$, namely $\frac{\eta^2}{4}\operatorname{tr}(K) = \frac{\eta^2}{4}\sum_{i\leq r}\mu_i$ (O spans the eigenspace of $\mu_1, \ldots, \mu_r$), with equality iff $P = P' = \eta K/2$, i.e. $R_T = \frac{\eta}{2}O$, in which case $R_{T+1} = -\frac{\eta}{2}O = -R_T$. Thus the symmetric weight cycle is $W_T = U_\star + \frac{\eta}{2}O$, $W_{T+1} = U_\star - \frac{\eta}{2}O$: it is centred at the teacher, and its loss is constant at the floor. The singular values $\eta/2$ belong to R_T, not generally to W_T. For $K = \lambda_D I_r$, $P = \lambda_D (R_T^\top R_T)^{1/2}$

and the condition is $\sigma_i(R_T) \in (0, \eta)$. The argument uses only that $\hat{D}$ is an r-dimensional $\hat{\Sigma}$-invariant subspace; for an invariant subspace of dimension larger than r, $\operatorname{range} O$ need not be invariant and the conclusion can fail.

(iii) Write $G = \hat{\Sigma}R = G_{\hat{D}} + G_{\perp}$ with $G_{\hat{D}} = \hat{\Sigma}R_{\hat{D}}$ and $G_{\perp} = \hat{\Sigma}R_{\perp}$; invariance gives $G_{\hat{D}}^{\top}G_{\perp} = 0$, so $G^{\top}G = G_{\hat{D}}^{\top}G_{\hat{D}} + G_{\perp}^{\top}G_{\perp}$. On the phase of the cycle where $R_{\hat{D}} = R_T$, $G_{\hat{D}} = OP$ and $G_{\hat{D}}^{\top}G_{\hat{D}} = P^2 \succ 0$. The map $S \mapsto S^{-1/2}$ is smooth at positive-definite S, so $(G^{\top}G)^{-1/2} = P^{-1} + O(\|R_{\perp}\|_F^2)$ and $\operatorname{Polar}(G) = G(G^{\top}G)^{-1/2}$ gives

$$R'_{\perp} = R_{\perp} - \eta\hat{\Sigma}R_{\perp}P^{-1} + O(\|R_{\perp}\|_F^3), \qquad R'_{\hat{D}} = R_{\hat{D}} - \eta O + O(\|R_{\perp}\|_F^2).$$

On the next phase $G_{\hat{D}} = -OP'$ and the same computation gives the factor P'^{-1}. In the eigenbasis of $\hat{\Sigma}|_{\hat{D}^{\perp}} = \sum_{j>r} \mu_j q_j q_j^{\top}$ the row $q_j^{\top}R_{\perp}$ is therefore multiplied by $I - \eta\mu_j P^{-1}$ and then by $I - \eta\mu_j P'^{-1}$. Both are symmetric with eigenvalues $1 - \eta\mu_j/\pi$, π an eigenvalue of P (resp. P'), which lie in $(-1, 1)$ when $\pi > \eta\mu_j/2$; since $\mu_j \le \mu_{r+1}$, the condition $P, P' \succ \frac{\eta\mu_{r+1}}{2} I$ makes each row multiplier have spectral norm below one. Thus the full linearised complement map contracts in Frobenius norm. On the symmetric cycle $P = P' = \frac{\eta}{2}K$, the per-step map is $R_{\perp} \mapsto R_{\perp} - 2\hat{\Sigma}R_{\perp}K^{-1}$, and with $K = V \operatorname{diag}(\mu_1, \dots, \mu_r)V^{\top}$ the coefficients $c_{ji} = q_j^{\top}R_{\perp}Ve_i$ are multiplied by $1 - 2\mu_j/\mu_i$, which lies in $(-1, 1)$ because $0 < \mu_j < \mu_i$.

(iv) Let R have columns in $\hat{D}$. If $\hat{D} \perp \operatorname{range} U_{\star}$, then $U_{\star}^{\top}R = 0$, so $W^{\top}W = I_r + R^{\top}R$ and $U_{\star}^{\top}W = I_r$. The columns of $W(W^{\top}W)^{-1/2}$ are an orthonormal basis of $\operatorname{range} W$, and the principal-angle cosines are the singular values of $U_{\star}^{\top}W(W^{\top}W)^{-1/2} = (I_r + R^{\top}R)^{-1/2}$, i.e. the multiset $\{(1 + \sigma_i(R)^2)^{-1/2} : 1 \le i \le r\}$. In particular, $\cos^2_{\min} = 1/(1 + \|R\|^2_{\mathrm{op}})$; on the symmetric cycle all $\sigma_i = \eta/2$. If $\operatorname{range} U_{\star} = \hat{D}$, then $R = U_{\star}U_{\star}^{\top}R$ and $W = U_{\star}(I_r + U_{\star}^{\top}R)$ with $\|U_{\star}^{\top}R\|_{\mathrm{op}} = \|R\|_{\mathrm{op}} < 1$, so $\operatorname{range} W = \hat{D}$ and all cosines equal 1. □

Exact cycles and changing alignment. The exact cycles in part (ii) repeat the full weight matrix, hence any fixed function of it, every two steps. In particular, teacher alignment cannot continue improving along either parity subsequence after exact cycle entry. Part (iii) instead describes perturbations off the cycle family, where a nonzero complementary residual can decay while alignment changes. Its local linearisation does not prove convergence from arbitrary initialisation.

The minimum is not the plateau. $\hat{\mathcal{L}} \ge 0$ with equality at $W = U_{\star}$, and gradient descent with step $0 < \ell < 1/\mu_1$ contracts R by $\max_k |1 - 2\ell\mu_k| < 1$ per step. The floor of part (ii) is therefore a property of the polar update: every full-column-rank matrix-Muon 2-cycle in $\hat{D}$ has two-step mean loss at least $\frac{\eta^2}{4}\sum_{i \le r}\mu_i$. For completeness, every such full-rank 2-cycle has the form in part (ii): returning after two steps forces its polar factors to be O and $-O$. Writing $\hat{\Sigma}R_T = OP$ and $\hat{\Sigma}R_{T+1} = -OP'$ with $P, P' \succ 0$, the update and $\hat{\Sigma}O = OK$ give $P + P' = \eta K$. Thus $0 \prec P \prec \eta K$. Rank-deficient 2-cycles use fewer than r directions and can sit lower.

A.6 Loss envelopes: where the feature-learning signal sits in the loss

Write $H_t := \sum_{k \ge 2} \mu_k z_{t,k}^2$ for the loss held by the draining modes. The lower phase means the parity class with the smaller leading-coordinate loss in the limiting scalar cycle; the exact identity below holds for any two times. Lemma 6 gives, for any two times t, T on the lower phase of the cycle, the exact identity

$$\hat{\mathcal{L}}(w_t) - \hat{\mathcal{L}}(w_T) = (H_t - H_T) + \mu_1\big(z_{t,1}^2 - z_{T,1}^2\big). \tag{59}$$

For the frozen-gradient symmetric cycle of Proposition 10(c), one has exactly $\mu_1 a_g^2 = \mu_1\eta^2/4 - g^2/\mu_1$. That identity does not hold instantaneously for a trajectory with changing off-mode content: the sum rule constrains two adjacent amplitudes, not either amplitude separately.

Under the hypotheses of Lemma 9, its phase-convergence estimate instead gives, for $t, T \ge T_1$ of the same parity,

$$\big|\mu_1\big(z_{t,1}^2 - z_{T,1}^2\big)\big| \le C\,\frac{g_t^2 + g_T^2}{\mu_1}, \qquad g_t^2 := \sum_{k \ge 2} \mu_k^2 z_{t,k}^2,$$

where C may depend on the fixed spectral contraction margin. For a single off-mode k, $g_t^2/\mu_1 = (\mu_k/\mu_1)H_t$. Thus slow modes can make the offset small compared with the hidden-loss level, but this estimate does not control the ratio of changes between arbitrarily close times.

For the auxiliary run initialised with $z_0 = 1.5\eta e_1 + 3e_8$, the measured ratio of the lower-envelope loss change to the hidden-loss change is 1.000. For the generic asymmetric showcase it is 0.797–0.798 from $t = 300$ to 6000: in that run the changing cycle offset cancels about 20% of the hidden-loss change. These are configuration-specific observations. The apparent plateau can therefore retain a small, measurable loss signal while alignment improves.

A.7 Numerical consistency checks

Reading Figure 1. Learning toys: nearly flat cycle-averaged loss while teacher alignment improves. (a,b) M2: NGD, $p = 8$, $n = 6000$, $\eta = 0.3$, low-variance teacher. **(c,d)** M3: matrix Muon, $p = 30$, $r = 5$, $n = 5000$, $\lambda_D = 1$, $\lambda_\perp = 10^{-3}$, $\eta = 0.3$, $\operatorname{range} U_\star \subset D^\perp$. Loss panels: adjacent-step bands and means, symmetric-cycle floors dashed; alignment panels: orthogonal-teacher reference $(1 + \eta^2/4)^{-1} = 0.978$ dashed. *Blue*: balanced leading components with nonzero off-mode residuals; the final means match the floors to relative 10^{-4} while alignment rises $0.13 \to 0.97$ (M2) and $0.49 \to 0.978$ (M3). *Orange*: generic starts; final means 1.06 and 1.35 times the floors. These finite-sample runs are numerical illustrations, not verified instances of all theorem hypotheses (details in Appendix A.7; Figure 8 compares optimisers).

We checked the formulas proved above against finite runs: the coordinate reduction and loss identities of Lemma 6, the scalar cycle relations and drain rates of Proposition 10, the alignment bounds of Theorem 7 and Corollary 8, the envelope identity of Section A.6, and the matrix cycle relations of Proposition 11. The coordinate/loss identities, constructed matrix cycles and exact alignment formulas have numerical residuals below 10^{-12}. Approximate comparisons use check-specific tolerances, including 2% for the slow-mode log-drain rates and 5% for the sample-to-population floor comparison; all 50 checks meet their numerical tolerances. These tolerances are not interval-arithmetic certificates. These checks illustrate the proved statements and their approximations rather than supply the proofs, and they describe particular trajectories: they do not establish convergence from arbitrary initialisation. The configurations are as follows.

Configurations. Here Σ_x denotes the population input covariance. In M3, D is its planted r-dimensional high-variance block, with covariance eigenvalues λ_D on D and $\lambda_\perp$ on $D^\perp$. Thus D and the empirical subspace $\hat{D}$ need not coincide. Each e_k is a standard coordinate vector. For M2, $p = 8$, $n = 6000$, $\eta = 0.3$, and $\Sigma_x = \operatorname{diag}(1, 0.6, 0.3, 0.15, 0.07, 0.03, 0.01, 10^{-4})$, with independent Gaussian samples. Case (A) uses $u = e_1$ and $w_0 = -e_1 + 3e_8$; case (B) uses $u = e_7$ and $w_0 = 2e_1 + 3e_8$. Both are run for 20000 steps. The symmetric initialisation check instead uses $z_0 = 1.5\eta e_1 + 3e_8$ and 6001 steps. In these finite samples, alignment of the teacher with a population eigendirection need not imply exact alignment with an empirical eigenvector.

For M3, $p = 30$, $r = 5$, $n = 5000$, $\eta = 0.3$, and $\Sigma_x = P_D + 10^{-3} P_{D^\perp}$. The constructed-cycle checks use an orthonormal basis O of $\hat{D}$, $K = O^\top \hat{\Sigma} O$, and

$$P = \frac{\eta}{2} K^{1/2}(I + \tfrac{1}{2}T)K^{1/2}, \qquad R_0 = OK^{-1}P, \qquad T = T^\top, \quad \|T\|_{\mathrm{op}} = 1.$$

Thus $\eta K/4 \preceq P \preceq 3\eta K/4$, so the exact cycle condition is satisfied. The cycle is followed for 2000 steps. Complement perturbations are of the form $\varepsilon F \Xi$, with F an orthonormal basis of $\hat{D}^\perp$ and Ξ a Gaussian matrix: $\varepsilon = 10^{-6}$ for the one- and two-step checks, and 10^{-3} for the 1200-step comparison. The M3 trajectory checks use a teacher in $D^\perp$, a Gaussian initial matrix projected onto D, and 4000 steps, matching the orange run in Figure 1. For the blue M2 run, instead, $z_0 = (\eta/2)e_1 - e_7 + 3e_8$ in empirical coordinates. The blue M3 run uses $W_0 = U_\star + (\eta/2)U_D + U_{\mathrm{junk}}$, where U_D is an orthonormal basis of the population block D and U_{junk} is an orthonormal r-frame perpendicular to both D and the teacher. Thus this initialisation has a nonzero complement residual; it is not an exact cycle. The plotted blue runs do not establish the conditional NGD margin or nonlinear convergence of the matrix complement. The M1 runs use the configurations of Figure 6. The frozen-gradient scalar check uses $\mu_1 = 1$, $\eta = 0.3$, $g = 0.05$, and 4000 iterations; its two-step multiplier is estimated by a centred difference with increment 10^{-7}.

For Figure 1, the settling time is the first reported step after the last adjacent-step loss mean that differs by more than 5% from the average of the final 200 adjacent-step means. In Figure 1(a,b) the hidden residual contributes about 4% of the floor (symmetric-cycle e-fold ≈ 5152 steps). The final adjacent-step bands of the blue runs are below 1%, and the blue M2 alignment is quoted from step 90. From the settled steps 89 (M2) and 17 (M3) of the orange runs, alignment rises $0.07 \to 0.96$–0.99 and $0.04 \to 0.93$–0.96. The population teacher directions need not be orthogonal to the empirical cycle subspace, so the dashed alignment level is a reference.

A.8 Additional figures

The additional figures use the following notation: λ_k for the eigenvalues of $\hat{\Sigma}$ (our μ_k; "EoS floor $2\lambda_1\eta^2/8$" is our $\mu_1\eta^2/4$), $y_{t,k}$ and $\tilde{w}$ for the residual coordinates and residual (our $z_{t,k}$, $w-u$), v_1 for q_1, V for the planted data block (our D), U_t for the teacher (our $U_\star$), $M = r$, and "V-block floor" for the population value $\lambda_D\eta^2 r/4$.

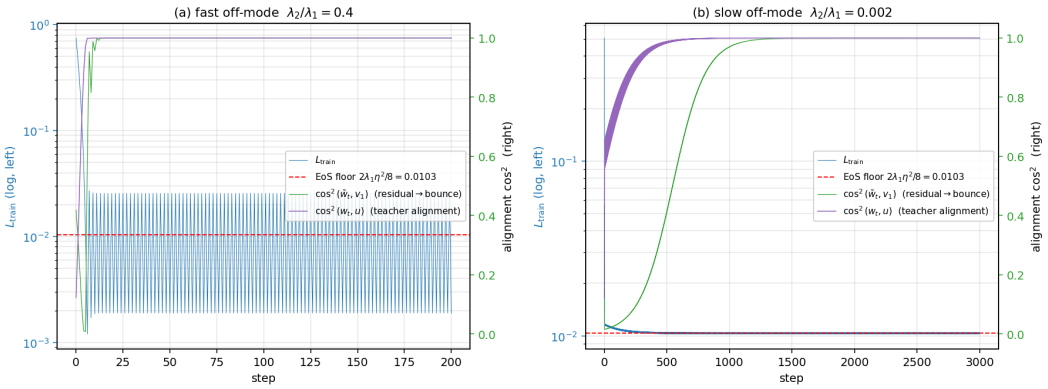


Figure 6: M1: two-dimensional NGD on linear regression. $n = 3000$, $\Sigma_x = \mathrm{diag}(1, \lambda_2)$ with $\lambda_2 = 0.4$ (a) or 0.002 (b), teacher $u = e_1$, $\eta = 0.2$, $w_0 = (0.3, 0.8)$. Loss (log, left axis) and alignments (right axis). (a) The fast off-mode leaves the cycle on the asymmetric member $\{0.157, -0.043\}$, so the plateau is $1.33\times$ the floor, as Proposition 10(a) predicts $(2(a^2 + (\eta - a)^2)/\eta^2 = 1.328)$. (b) The slow, loss-visible off-mode symmetrises the cycle onto $\pm\eta/2$ (plateau/floor $= 1.000$): the loss settles at step 120 and stays on the floor for 2900 further steps while the residual alignment climbs $0.04 \to 1.00$ and the teacher alignment $0.82 \to 1.00$. Here the misalignment sits in a loss-visible mode; the invisible regime is shown with a dedicated hidden mode in M2.

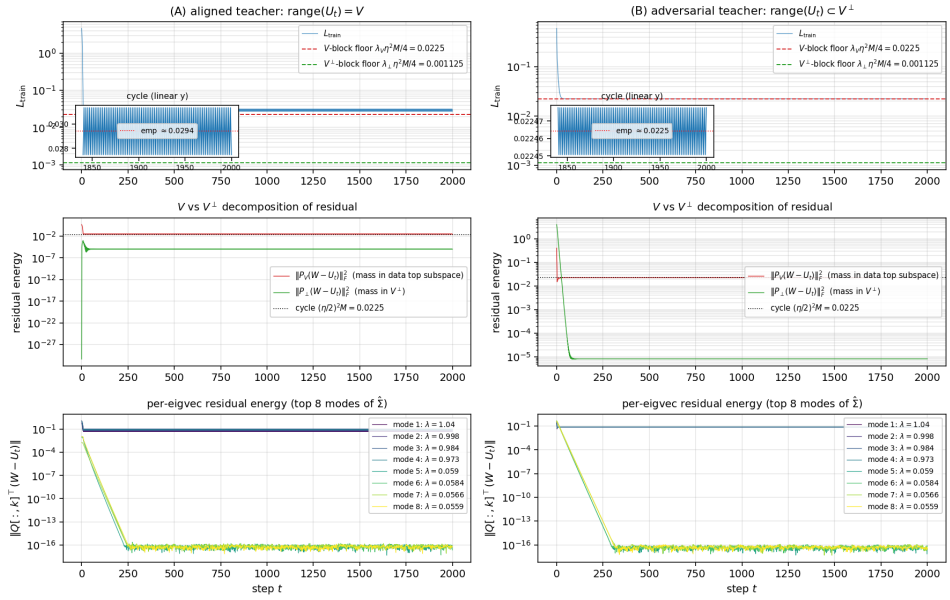


Figure 7: M3: matrix Muon on multi-output regression. $p = 30$, $r = 4$, $n = 4000$, $\eta = 0.15$, $\lambda_D = 1$, $\lambda_\perp = 0.05$, both runs initialised in D; teacher (A) $= D$, (B) $\subset D^\perp$. **Top:** loss with the population floors $\lambda_D\eta^2 r/4 = 0.0225$ and $\lambda_\perp\eta^2 r/4$ (insets: the cycle; (A) sits on an asymmetric member, $1.31\times$, (B) nearly symmetric). **Middle:** residual energy in the population blocks D, $D^\perp$ (the $D^\perp$ part levels off near 10^{-5} only because $\hat{D}$ is slightly rotated from D). **Bottom:** residual energy in the eigenbasis of $\hat{\Sigma}$: the four displayed complement modes ($\mu \approx 0.056$–0.059) decay toward numerical precision; the factor $1 - 2\mu_j/\mu_i \approx 0.89$ is the symmetric-cycle linearised reference from Proposition 11(iii), and asymmetric cycles have the two-phase multipliers stated there. The late teacher alignment is approximately 1 (A) and 0.9944 (B), close to the ideal symmetric-cycle reference $(1 + \eta^2/4)^{-1}$ (the exact formula in Proposition 11(iv) requires empirical-subspace orthogonality), and the step length $\eta\sqrt{r}$ holds to four digits (Proposition 11(i)).

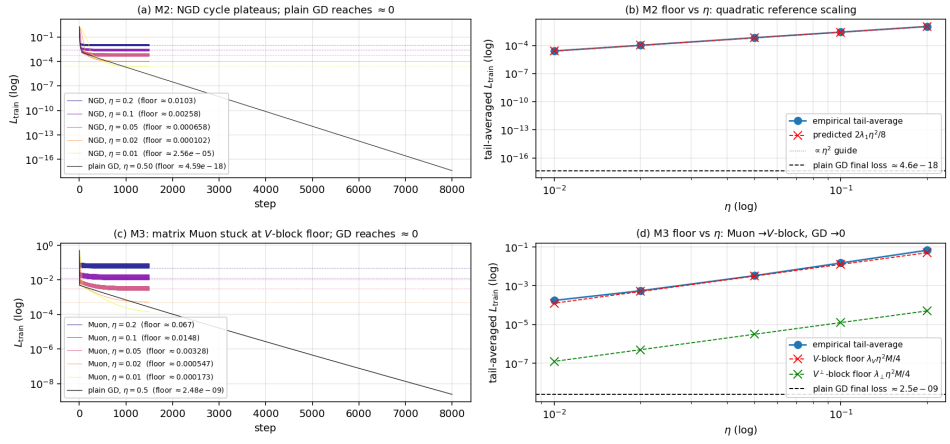


Figure 8: The plateau belongs to the optimiser. (a,b) M2 ($p = 9$, $n = 2000$, $\Sigma_x = \mathrm{diag}(1, 0.6, 0.3, 0.15, 0.07, 0.03, 0.012, 0.005, 0.002)$, teacher e_1): NGD at $\eta \in \{0.2, 0.1, 0.05, 0.02, 0.01\}$ stops at floors within 0–4% of $\mu_1\eta^2/4$; plain GD (step 0.5) reaches 4.6×10^{-18}. (c,d) M3 ($p = 30$, $r = 5$, $n = 2000$, $\lambda_D = 1$, $\lambda_\perp = 10^{-3}$, adversarial teacher): matrix Muon stops at 1.05–$1.39 \times \lambda_D\eta^2 r/4$ (a population reference; both sampling and the asymmetric-cycle excess of Proposition 11(ii) affect the ratio), far above the $D^\perp$-scale value (green); plain GD reaches 2.5×10^{-9}.

B PROOFS AND SUPPORTING DERIVATIONS FOR SECTION 5

This appendix collects the proofs of Proposition 2, Theorem 13, and Proposition 15 (Sections B.1, B.2, B.3), together with the full assumptions and proof of the conditional drift result, Proposition 4, stated in Section 5.3 (Section B.4 below).

Notation for these proofs. We use $W \in \mathbb{R}^{p\times r_s}$, input dimension p, student width r_s, teacher subspace V of dimension r_t, and $M = \min(p, r_s)$. The matrices P_V and $P_\perp = I_p - P_V$ are the orthogonal projectors onto V and $V^\perp$; I_k is the $k \times k$ identity. For vectors, $\|\cdot\| = \|\cdot\|_2$; for matrices, $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{op}}$ are the Frobenius and operator norms. The map $\Phi(W) = W - \eta\,\mathrm{Polar}(\nabla\mathcal{L}(W))$ uses step coefficient $\eta > 0$. Here $\mathcal{L}$ is the loss specified in each statement, and $r^{\mathrm{grad}} = \mathrm{rank}(\nabla\mathcal{L}(W))$. We write $A_{\mathrm{sub}}(W) = \|P_V W\|_F^2/\|W\|_F^2$ for $W \neq 0$. For an empirical student AGOP $G_s(W)$, $\cos^2_{\min}(W)$ denotes the smallest squared cosine of its leading r_t-dimensional eigenspace with V (Eq. (8)–Eq. (9)). Singular values are ordered from largest to smallest. The symbols tr and rank denote trace and matrix rank; $O(k) = \{Q \in \mathbb{R}^{k\times k} : Q^\top Q = I_k\}$, and $\mathrm{St}(p,k) = \{Q \in \mathbb{R}^{p\times k} : Q^\top Q = I_k\}$.

In the first three subsections, $O(\cdot)$ denotes the standard asymptotic notation: $f(\eta) = O(g(\eta))$ as $\eta \to 0$ means $\limsup_{\eta\to 0} |f(\eta)/g(\eta)| < \infty$. Constants in $O(\cdot)$ may depend on p, r_t, r_s, the spectrum of the Hessian along the trajectory, and the teacher activation; they never depend on η itself. Section B.4 instead uses bounds uniform across its stated family and trajectory window, with the learning-rate rule included among the fixed family parameters.

B.1 PROOF OF PROPOSITION 2 (POLAR-UPDATE IDENTITIES)

We prove parts (i), (ii), (iii) of the proposition, using only the orthonormality of the SVD factors and the definition of the polar factor $\mathrm{Polar}(g) := UV_g^\top$ for a compact SVD $g = U\Sigma V_g^\top$. The leakage scalars used in part (iii) are recorded formally for reference:

Definition 12 (leakage scalars)**.** *Let $g \in \mathbb{R}^{p\times r_s}$ have $\mathrm{rank}(g) = r \geq r_t$ and compact SVD $g = U\Sigma V_g^\top$ with $U \in \mathbb{R}^{p\times r}$ orthonormal columns $u_1, \ldots, u_r$. Given the teacher subspace $V \subseteq \mathbb{R}^p$ from Eq. (2),*

$$\varepsilon_V^2 := \sum_{i=1}^{r_t} \|P_\perp u_i\|^2, \qquad \varepsilon_\perp^2 := \sum_{i=r_t+1}^{r} \|P_V u_i\|^2. \tag{60}$$

ε_V^2 measures how much of the squared norm of the first r_t left singular vectors of g falls outside V; $\varepsilon_\perp^2$ measures how much of the squared norm of the remaining left singular vectors falls inside V. Both vanish iff the left singular vectors of g split exactly into a V-block and a $V^\perp$-block.

Part (i): $\|\mathrm{Polar}(g)\|_F^2 = r$. Since $U \in \mathbb{R}^{p\times r}$ has orthonormal columns ($U^\top U = I_r$),

$$\|UV_g^\top\|_F^2 \;=\; \mathrm{tr}\big(V_g\,U^\top U\,V_g^\top\big) \;=\; \mathrm{tr}(V_g V_g^\top) \;=\; \mathrm{tr}(I_r) \;=\; r. \tag{61}$$

Part (ii): $\|\Phi(W) - W\|_F = \eta\sqrt{r^{\mathrm{grad}}}$. By definition, $\Phi(W) - W = -\eta\,\mathrm{Polar}(\nabla\mathcal{L}(W))$, so $\|\Phi(W) - W\|_F = \eta\,\|\mathrm{Polar}(\nabla\mathcal{L}(W))\|_F = \eta\sqrt{r^{\mathrm{grad}}}$ by part (i). Generically along a trajectory $r^{\mathrm{grad}} = M$, giving the stated $\eta\sqrt{M}$.

Part (iii): leakage decomposition. Here $r = M \geq r_t$, as in the full-rank statement of the proposition. Writing $\mathrm{Polar}(g) = UV_g^\top$ and using $V_g^\top V_g = I_r$,

$$\|P_V\,\mathrm{Polar}(g)\|_F^2 \;=\; \mathrm{tr}\big(V_g(P_V U)^\top(P_V U)V_g^\top\big) \;=\; \mathrm{tr}\big((P_V U)^\top(P_V U)\big) \;=\; \sum_{i=1}^{r}\|P_V u_i\|^2, \tag{62}$$

where u_i is the i-th column of U. The $V/V^\perp$ split $\|P_V u_i\|^2 = 1 - \|P_\perp u_i\|^2$ then gives

$$\begin{aligned}\|P_V\,\mathrm{Polar}(g)\|_F^2 &= \sum_{i=1}^{r_t}\big(1 - \|P_\perp u_i\|^2\big) + \sum_{i=r_t+1}^{r}\|P_V u_i\|^2 \\ &= r_t - \varepsilon_V^2 + \varepsilon_\perp^2,\end{aligned} \tag{63}$$

with $\varepsilon_V^2, \varepsilon_\perp^2$ as in Definition 12. The complementary statement $\|P_\perp \operatorname{Polar}(g)\|_F^2 = (M - r_t) + \varepsilon_V^2 - \varepsilon_\perp^2$ follows from $\|\operatorname{Polar}(g)\|_F^2 = \|P_V \operatorname{Polar}(g)\|_F^2 + \|P_\perp \operatorname{Polar}(g)\|_F^2 = M$ (part (i) at $r = M$). The identities are exact under no further smallness assumption on the leakage.

Connection to data. The identities hold exactly for every gradient of full column rank; how close $\|P_V \operatorname{Polar}(g)\|_F^2$ is to r_t depends on the leakage scalars, which we did not measure along the trajectories. The polar-mass check of Appendix C.4 (Table 3) finds the V-mass of the update approximately 2–5% below r_t on three ReLU checkpoints. Figure 3(b) reports a different quantity, the V-mass of the top AGOP eigenvectors.

B.2 Period-2 orbits on the isotropic quadratic

Proposition 2 constrains the *shape* of any period-2 orbit of the polar map but not its existence. On the isotropic quadratic there is an explicit family.

Theorem 13 (period-2 cycle on the isotropic quadratic)**.** *Let $\mathcal{L}(W) = \frac{\lambda}{2}\|W\|_F^2$ on $\mathbb{R}^{p\times r_s}$ with $\lambda > 0$ and $r_s \le p$. The matrix Muon iteration Φ admits a continuum of* symmetric *period-2 orbits $\{W^*, -W^*\}$ of Frobenius radius $\|W^*\|_F = \frac{\eta}{2}\sqrt{M}$. Concretely, $W^* = \frac{\eta}{2} U_0 V_0^\top$ for any orthonormal columns $U_0 \in \mathbb{R}^{p\times r_s}$, $V_0 \in \mathbb{R}^{r_s\times r_s}$.*

Existence on arbitrary $A \succ 0$ holds on every r_s-dimensional A-invariant subspace (Proposition 11(ii)). On the 1-HL ReLU loss we observe only approximately period-2 loss oscillations, while the parameters continue to evolve (Appendix C.4, Test A).

Proof of Theorem 13. We use the symmetric ansatz $W_B = -W_A$ for the period-2 condition $\Phi(W_A) = W_B$:

$$W_A - \eta \operatorname{Polar}(\lambda W_A) \;=\; -W_A \quad\Longleftrightarrow\quad 2W_A \;=\; \eta \operatorname{Polar}(W_A), \tag{64}$$

where we used $\operatorname{Polar}(cX) = \operatorname{Polar}(X)$ for any $c > 0$.

We verify that $W^* := (\eta/2)\, U_0 V_0^\top$, for any orthonormal $U_0 \in \mathbb{R}^{p\times r_s}$ and $V_0 \in \mathbb{R}^{r_s\times r_s}$, satisfies Eq. (64). The compact SVD of W^* is $W^* = (\eta/2)\, U_0 \Sigma V_0^\top$ with $\Sigma = I_{r_s}$ (both U_0, V_0 are orthonormal), so all singular values equal $\eta/2$ and $\operatorname{Polar}(W^*) = U_0 V_0^\top$. Therefore $\eta \operatorname{Polar}(W^*) = \eta\, U_0 V_0^\top = 2W^*$, which is the right-hand side of Eq. (64).

The orbit is genuinely period-2 because a fixed point of Φ would require $\operatorname{Polar}(\lambda W) = 0$, contradicting $\|\operatorname{Polar}(g)\|_F^2 = M > 0$ at full column rank (Proposition 2(i)). The continuum of solutions is parameterised by the product of Stiefel manifolds $(U_0, V_0) \in \operatorname{St}(p, r_s) \times O(r_s)$, modulo the right-action of $O(r_s)$. □

Remark 14 (extension to general $A \succ 0$)**.** *For $\mathcal{L}(W) = \frac{1}{2}\operatorname{tr}(W^\top A W)$ with $A \succ 0$ general, the symmetric ansatz becomes $2W = \eta \operatorname{Polar}(AW)$. For arbitrary $A \succ 0$ the symmetric ansatz is solved exactly by $W^* = \frac{\eta}{2} O$ whenever the columns of O are an orthonormal basis of an r_s-dimensional A-invariant subspace, since then $AW^* = \frac{\eta}{2} O(O^\top A O)$ with $O^\top A O \succ 0$ and $\operatorname{Polar}(AW^*) = O$ (Proposition 11(ii) gives the whole family of such orbits); stability for general A and orbits outside invariant subspaces remain open. In the network experiments only the loss has a period-2 signature (Test A, Appendix C.4); the weights keep moving, so they do not sit on an orbit of this kind.*

B.3 Block-orthogonal equivariance on a block-isotropic quadratic

Proposition 15 (block-orthogonal equivariance)**.** *Let $\mathcal{L}(W) = \frac{1}{2}\operatorname{tr}(W^\top A W)$ with $A = \operatorname{diag}(\lambda I_{r_t}, \mu I_{p-r_t})$ for some $\lambda, \mu > 0$. The matrix Muon iteration map Φ defined in Eq. (5) commutes with the block-orthogonal group $G = \text{blk-diag}(Q, R) \in O(r_t) \times O(p - r_t)$ acting on W by left multiplication: $\Phi(GW) = G\,\Phi(W)$ for every W at which $\operatorname{Polar}(AW)$ is defined. Consequently, if the law of W_0 is invariant under $O(r_t) \times O(p - r_t)$, the law of W_t is invariant under the same group action for every $t \ge 0$.*

Proof of Proposition 15. Let $G \in O(r_t) \times O(p - r_t) \subset O(p)$, written in block form as $G = \text{blk-diag}(Q, R)$ with $Q \in O(r_t)$, $R \in O(p - r_t)$. Since $A = \operatorname{diag}(\lambda I_{r_t}, \mu I_{p-r_t})$ has the same block

structure, $G^\top AG = A$ (equivalently $AG = GA$). For any $W \in \mathbb{R}^{p\times r_s}$,

$$\nabla\mathcal{L}(GW) \;=\; A\,GW \;=\; G\,AW \;=\; G\,\nabla\mathcal{L}(W). \tag{65}$$

The polar factor is left-equivariant under orthogonal transformations: for any $G \in O(p)$ and full-column-rank $X \in \mathbb{R}^{p\times r_s}$, the compact SVD of GX is $GX = (GU)\Sigma V^\top$, so $\mathrm{Polar}(GX) = (GU)V^\top = G\,\mathrm{Polar}(X)$. Combining the two equivariances gives

$$\Phi(GW) \;=\; GW - \eta\,\mathrm{Polar}(AGW) \;=\; GW - \eta\,G\,\mathrm{Polar}(AW) \;=\; G\,\Phi(W), \tag{66}$$

which is the equivariance claim. Iterating, $\Phi^t(GW_0) = G\,\Phi^t(W_0)$ for every t. If the law of W_0 is $O(r_t)\times O(p-r_t)$-invariant, the law of $W_t = \Phi^t(W_0)$ is invariant under the same group action. □

Caveat. Block-orthogonal invariance does not pin down the relative energy of W_t between the V- and $V^\perp$-blocks: an $O(r_t)\times O(p-r_t)$-invariant distribution can place arbitrary relative mass in the two blocks. In particular, $\mathbb{E}_{W_0}[A_{\mathrm{sub}}(W_t)] = r_t/p$, where the expectation is over initialisation, does not follow from the symmetry alone. Block-orthogonal equivariance neither establishes nor excludes a preference for the teacher subspace under a fixed Hessian. Identifying the cause of the observed preference in the 1-HL ReLU network requires an additional dynamical argument; the proposition does not attribute it to changing activation patterns.

B.4 Assumptions and proof of conditional secular drift

This subsection proves Proposition 4. Write $M = \min(p, r_s)$, $q = r_t/M$, $A_t = \cos^2_{\min}(W_t)$, $e_k = 1 - A_{2k}$ and $v_k = (A_{2k+2} - A_{2k})/2$. Fix the remaining family parameters θ, including the learning-rate rule. Here k indexes pairs of optimisation steps, $\asymp_\theta$ means upper and lower bounds by positive constants depending only on θ, and $\lesssim_\theta$ has the corresponding one-sided meaning. All matrix norms in this subsection are operator norms, including those in remainder bounds. Unadorned constants C, K in bounds are finite and uniform over the stated family and window, and may change between estimates; $C(\theta)$ denotes the fixed leading prefactor specified below. The base assumptions yield the two-sided bound Eq. (14). Under the additional stationary slow-mode conditions specified below, the sharper conclusion is

$$v_k = C(\theta)q^\alpha e_k(1 + o(1)), \tag{67}$$

with explicit uniform control

$$\left|\frac{v_k}{C(\theta)q^\alpha e_k} - 1\right| \le K_\theta\left(q^\nu + \sqrt{e_k} + e^{-\omega k q^\alpha}\right). \tag{68}$$

The exponent and prefactor depend on the specified family, including its learning-rate convention. These statements concern same-parity alignment drift; an exact repeating weight orbit is not assumed.

Feature coordinates and residence. Assume that, on a trajectory window $W_0, \ldots, W_{2N}$, the leading rank-r_t student AGOP subspace is separated from the remaining spectrum. Near the teacher subspace it has a graph representation

$$\mathcal{S}(W_{2k}) = \mathrm{range}(U + U_\perp Z_k),$$

All two-step conclusions below apply for $0 \le k < N$. Here $U \in \mathbb{R}^{p\times r_t}$ and $U_\perp \in \mathbb{R}^{p\times(p-r_t)}$ are orthonormal frames for the teacher subspace and its complement, so $Z_k \in \mathbb{R}^{(p-r_t)\times r_t}$. For the paper's worst-principal-angle metric,

$$A_{2k} = \frac{1}{1 + \|Z_k\|^2_{\mathrm{op}}}, \qquad e_k = \frac{\|Z_k\|^2_{\mathrm{op}}}{1 + \|Z_k\|^2_{\mathrm{op}}}. \tag{69}$$

Orthogonal changes of these frames preserve the formula. Residence in the stated window and any near period-two loss condition are separate assumptions; the alignment calculation does not establish them.

Finite-order transverse dynamics. Put $z = q^\nu$, with a prescribed $\nu > 0$. For an integer $j \geq 1$, suppose that $\varepsilon = z^j$ and

$$Z_{k+1} = Q_k\{Z_k - \varepsilon(L_k Z_k + Z_k R_k) + E_k\}S_k^\top, \tag{70}$$

where $Q_k, L_k \in \mathbb{R}^{(p-r_t)\times(p-r_t)}$, $S_k, R_k \in \mathbb{R}^{r_t \times r_t}$, and E_k has the same shape as Z_k. The matrices Q_k, S_k are orthogonal, L_k, R_k are symmetric positive semidefinite, and, uniformly across the family and window,

$$\lambda_{\min}(L_k) + \lambda_{\min}(R_k) \geq a > 0, \tag{71}$$

$$\|L_k\|_{\mathrm{op}} + \|R_k\|_{\mathrm{op}} \leq b < \infty, \tag{72}$$

$$\|E_k\|_{\mathrm{op}} \leq K\varepsilon\big(z\|Z_k\|_{\mathrm{op}} + \|Z_k\|_{\mathrm{op}}^2\big). \tag{73}$$

These are sufficient conditions for dissipation in the operator norm, which is the norm relevant to the worst principal angle. Positive definiteness of an arbitrary generator in Frobenius norm alone does not imply them.

Proof of the two-sided law. Write $x = \|Z_k\|_{\mathrm{op}}$ and $x_+ = \|Z_{k+1}\|_{\mathrm{op}}$. The linear term can be replaced by $(I - \varepsilon L_k)Z_k(I - \varepsilon R_k)$, with an additional error at most $b^2\varepsilon^2 x$. This is absorbed in Eq. (73), since $\varepsilon = z^j \leq z$ for $z \leq 1$. For positive definite P, T,

$$\sigma_{\min}(P)\sigma_{\min}(T)\|Z\|_{\mathrm{op}} \leq \|PZT\|_{\mathrm{op}} \leq \|P\|_{\mathrm{op}}\|T\|_{\mathrm{op}}\|Z\|_{\mathrm{op}}.$$

It follows that

$$[1 - \varepsilon b - C\varepsilon(z + x)]x \leq x_+ \leq [1 - \varepsilon a + C\varepsilon(z + x)]x.$$

The exact graph identity gives

$$\frac{v_k}{\varepsilon e_k} = \frac{x^2 - x_+^2}{2\varepsilon x^2(1 + x_+^2)}.$$

Hence

$$a - C(z + \sqrt{e_k}) \leq \frac{v_k}{q^{\nu j} e_k} \leq b + C(z + \sqrt{e_k}). \tag{74}$$

For sufficiently small q, e_k, this proves Eq. (14) with $\alpha = \nu j$. The upper bound on x_+ also keeps the graph error small. □

Obtaining the exponent without analyticity. The expansion above can be obtained from a finite Taylor expansion of the entire two-step graph map in moving coordinates. Existence of such a graph map is an additional closure hypothesis: the AGOP subspace alone need not determine its next iterate. Other state coordinates may be included as parameters if the stated estimates hold uniformly in them. Derivative bounds below use norms induced by the operator norm of graph matrices, not dimension-dependent conversions from the Frobenius norm. When dimensions vary, this means either a common reduced coordinate space or a smooth parameter extension on each fixed-dimensional graph space, evaluated at its actual $z = q^\nu$; all stated derivative and remainder bounds must then be uniform across those extensions. It is not a derivative of matrix dimensions. Specifically, suppose $F_z(0) = 0$, $F_0(Z) = Z$, and $\partial_z^\ell F_z(Z)|_{z=0} = 0$ for $1 \leq \ell < j$, throughout the local graph neighborhood. If

$$\frac{1}{j!}\partial_z^j F_z(Z)\big|_{z=0} = -LZ - ZR + O(\|Z\|_{\mathrm{op}}^2)$$

and the mixed derivative $D_Z\partial_z^{j+1}F$ is uniformly bounded, Taylor's theorem gives Eq. (70)–Eq. (73). Only finite smoothness and a nonzero dissipative term are used. The integer j is the first active damping order; the rank-ratio exponent is $\alpha = \nu j$.

Checking these derivatives only for the linearization at $Z = 0$ is weaker. Ordinary uniform C^2 control then gives an unscaled $O(\|Z\|_{\mathrm{op}}^2)$ nonlinear remainder. Writing $e = e_k$ and $v = v_k$, in that case the same proof gives

$$a - C\left(z + \frac{\sqrt{e}}{q^\alpha}\right) \leq \frac{v}{q^\alpha e} \leq b + C\left(z + \frac{\sqrt{e}}{q^\alpha}\right).$$

Thus a small fixed relative-error bound holds only in a shrinking neighborhood $e \lesssim q^{2\alpha}$. A uniform small-error neighborhood requires the scaled nonlinear control above or another condition that supplies it.

One sufficient set of slow-mode conditions. For the sharper result, write $Y_k = O_k^\top Z_k T_k$ for orthogonal changes of the left and right frames, so $\|Y_k\|_{\mathrm{op}} = \|Z_k\|_{\mathrm{op}}$. Suppose that in these coordinates the exact transverse dynamics are stationary and have the form

$$Y_{k+1} = (I - \varepsilon L(q))Y_k(I - \varepsilon R(q)) + N_q(Y_k), \qquad \|N_q(Y)\|_{\mathrm{op}} \le K\varepsilon\|Y\|_{\mathrm{op}}^2. \tag{75}$$

Assume the preceding bounds on L, R. Write their smallest eigenvalues as $\lambda_0(q), \rho_0(q)$, and their corresponding orthogonal projectors as $P_L(q), P_R(q)$. Assume that these slow eigenspaces are separated from the other eigenvalues by a uniform gap $\gamma > 0$, and

$$\lambda_0(q) + \rho_0(q) = C(\theta) + O(q^\nu), \qquad C(\theta) > 0.$$

Finally, for a fixed $\delta > 0$, assume

$$\|P_L Y_0 P_R\|_{\mathrm{op}} \ge \delta\|Y_0\|_{\mathrm{op}},$$

with $\|Y_0\|_{\mathrm{op}}$ sufficiently small depending on δ. The quantitative projection condition prevents nonlinear forcing from canceling the slow component. These conditions are additional to the finite-order expansion.

Proof of the shared-prefactor law. Set $x_0 = \|Y_0\|_{\mathrm{op}}$, $\mathcal{P}Y = P_L Y P_R$, and

$$\mu_0 = (1 - \varepsilon\lambda_0)(1 - \varepsilon\rho_0) = 1 - C(\theta)\varepsilon + O(\varepsilon q^\nu).$$

The linear map equals $\mu_0 I$ on the slow subspace. Its complement splits into the three blocks $(I - P_L)YP_R$, $P_L Y(I - P_R)$, and $(I - P_L)Y(I - P_R)$. The n-step linear action on the complement has norm at most $3\mu_0^n e^{-c\gamma\varepsilon n}$; this estimate is independent of matrix dimensions.

For sufficiently small x_0, global contraction gives $\|Y_k\|_{\mathrm{op}} \le x_0 r^k$, with $r = \mu_0 + K\varepsilon x_0 < 1$ and $r^2/\mu_0 \le 1 - c\varepsilon$. For every k in the window, write

$$D_k = \mathcal{P}Y_0 + \sum_{n=0}^{k-1} \mu_0^{-n-1}\mathcal{P}N_q(Y_n), \qquad \mathcal{P}Y_k = \mu_0^k D_k.$$

Each nonlinear term carries ε, which cancels the $O(\varepsilon^{-1})$ geometric sum, so that $\|D_k - \mathcal{P}Y_0\|_{\mathrm{op}} \le Cx_0^2$, uniformly in k. Choosing x_0 small relative to δ gives $\|D_k\|_{\mathrm{op}} \ge \delta x_0/2$. The same sums, projected onto the complementary blocks, give

$$\|(I - \mathcal{P})Y_k\|_{\mathrm{op}} \le Cx_0\mu_0^k e^{-\omega\varepsilon k}, \qquad \|Y_k\|_{\mathrm{op}} \ge c_\delta x_0 \mu_0^k.$$

These estimates use only iterates within the stated window.

Let $\mathcal{T}_q Y = (I - \varepsilon L)Y(I - \varepsilon R)$. The operator $\mathcal{T}_q - \mu_0 I$ vanishes on the slow subspace and has operator-norm-to-operator-norm bound $O(\varepsilon)$. Consequently

$$\left|\|Y_{k+1}\|_{\mathrm{op}} - \mu_0\|Y_k\|_{\mathrm{op}}\right| \le C\varepsilon\|Y_k\|_{\mathrm{op}}\left(e^{-\omega\varepsilon k} + \|Y_k\|_{\mathrm{op}}\right).$$

This extra factor of ε is essential when converting mode selection to a normalized drift rate. Substitution in the exact graph identity, together with

$$\frac{1 - \mu_0^2}{2} = C(\theta)\varepsilon + O(\varepsilon q^\nu),$$

proves Eq. (67). □

Interpretation and limits. A fixed-factor error reduction takes $\Theta_\theta(q^{-\alpha})$ optimization steps, provided the plateau window lasts long enough. The shared-prefactor relative error vanishes when $q \to 0$, $e_k \to 0$, and $kq^\alpha \to \infty$. The slow-mode hypotheses need not follow from a leading Taylor expansion: they must hold for the actual linearization. Likewise, arbitrary perturbations of a multiple slow eigenvalue need not preserve its gap.

Finite smoothness alone is insufficient: a positive damping factor $\exp(-1/q^2)$, extended by zero at $q = 0$, is smooth but has no finite nonzero power-law term. A finite-order nondegeneracy condition is therefore necessary for this route. If contraction remains nonzero as $q \to 0$, an exponent-zero variant uses the exact discrete coefficient $(1 - \mu(0)^2)/2$. None of these statements identifies the exponent for the full empirical Muon family without checking its transverse dynamics.

C ADDITIONAL EXPERIMENTS AND THEORY-VS-EXPERIMENT ALIGNMENT

This appendix collects the multi-activation extension of the experiments in Section 5.1 (the GELU and SiLU teacher grids) and the per-config quantitative detail for each of the seven theory tests on the long-run ReLU checkpoints. The tables in Section C.4 are computed from the saved long-run ReLU checkpoints. The smooth-teacher qualitative story is carried by Figures 10 and 11; we do not include quantitative tables for those grids in this version.

Notation and diagnostics. Here p, r_t, r_s, n, η denote input dimension, teacher rank, student width, training-set size and the polar-update coefficient; $M = \min(p, r_s)$. The symbol σ_t in table headers names the teacher activation (the subscript means "teacher", not optimisation time). Write $V = \mathrm{span}(U)$, $P_V = UU^\top$, $P_\perp = I_p - P_V$, and $L_t = L_{\mathrm{train}}(W_t)$. For the empirical input AGOP $G_s(W) = n^{-1}\sum_{i=1}^n \nabla_x f_W(x_i)\nabla_x f_W(x_i)^\top$, let $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ and $v_1, \ldots, v_p$ be its ordered eigenvalues and orthonormal eigenvectors. The two alignment summaries are the mean and minimum squared cosines of the principal angles between $\mathrm{span}(v_1, \ldots, v_{r_t})$ and V (Eq. (9)). We distinguish the direction-only mass $\sum_{j \leq r_t} \|P_V v_j\|_2^2$ from the eigenvalue-weighted fraction

$$\mathrm{mass}_V(G_s) := \frac{\mathrm{tr}(P_V G_s)}{\mathrm{tr}(G_s)} = \frac{\sum_j \lambda_j \|P_V v_j\|_2^2}{\sum_j \lambda_j}, \qquad A_{\mathrm{sub}}(W) := \frac{\|P_V W\|_F^2}{\|W\|_F^2}.$$

The first fraction requires $\mathrm{tr}(G_s) > 0$ and the second $W \neq 0$. The head-fit losses $L_{V,\mathrm{opt}}$ and $L_{\text{AGOP-opt}}$ refit the head after projecting W onto V or the student's selected AGOP subspace, respectively, as in Eq. (11). Empirical means, standard deviations and correlations below are taken over the stated training-step window; aggregation over seeds is specified separately.

C.1 PLATEAU-SEARCH GRID: RELU, GELU, AND SILU TEACHERS

We sweep (p, r_t, r_s, η) at fixed seed 5, full batch, $n = 15{,}000$, and 8,000–10,000 training steps. The ReLU sweep covers $p \in \{100, 150, 200\}$, $r_t \in \{4, 6, 8, 10, 12\}$, $r_s \in \{30, 50, 80, 100\}$, $\eta \in \{1.0, 1.25, 1.5, 2.0, 2.5\}$; the smooth-teacher sweep adds large-r_t amplification configs ($r_t \in \{12, 16, 20\}$) at the activation's EoS learning rate ($\eta \approx 0.7$ for GELU, $\eta \approx 0.85$ for SiLU). The cleanest *rising-loss-with-learning* regimes come from the smooth-teacher amplification configs: SiLU $r_t = 20$, $\eta = 0.85$ exhibits a loss that climbs from 1.92 to 2.06 while $\cos^2_{\min}$ grows by $+0.43$ over the second half of training; GELU $r_t = 16$, $\eta = 0.85$ shows the analogous picture on a smaller amplitude. The cleanest-by-score configs from each sweep populate Figures 9, 10, and 11. Table 2 lists all 33 configurations with their outcomes.

C.2 PER-TEACHER TOP-6 CONFIGURATIONS AND THE FULL RELU 10-CONFIG GRID

The effective-rank curves count, among the 12 recorded leading AGOP eigenvalues, those at least $0.05\lambda_1$ or at least $0.5\lambda_2$, respectively. The latter threshold also defines the selected rank $\tilde{r}$ used by the head-refitting diagnostic (over the spectrum used for that diagnostic).

Figure 9 shows six ReLU configurations and Figures 10–11 six SiLU and six GELU configurations, selected on the seed-5 sweep with the scores of Section C.3 and shown with all seeds; the selection and the column order are those of the seed-5 ranking. Figure 12 reports the underlying full 10-config ReLU plateau-search grid from which Figure 9 selects.

Reading Figure 2. Loss plateau with continued feature learning in the 1-HL student (M4). One configuration per ReLU, SiLU and GELU teacher activation; matrix Muon, 8,000 steps. **Top:** training loss (per-step range, running minimum and mean), test loss and oracles $L_{V,\mathrm{opt}}$, $L_{\text{AGOP-opt}}$. **Bottom:** AGOP alignment $\overline{\cos^2}$, $\cos^2_{\min}$ rises during the period-2 loss plateau. Medians over 53 seeds; shaded interquartile range. Figure 13 shows the same columns with all four rows.

C.3 CONFIGURATION SELECTION

The configurations shown in Figures 9–13 were chosen before the multi-seed runs, on the seed-5 sweep described at the start of this appendix, with the following per-trajectory scores; the multi-seed figures keep those choices and their order. The original seed-5 figures span 10,000 steps; the

Table 2: **All** 33 **one-hidden-layer teacher configurations** ($p = 100$, $n = 15{,}000$, 5 seeds each: $5, 11, 23, 37, 41$; medians over seeds). $1 - \rho_2$: second-half period-2 signature (values below 10^{-10} shown as 10^{-10}); $L/L_{V,\mathrm{opt}}$: late cycle-mean training loss over the head oracle; $\Delta \cos^2_{\min}$: change from the midpoint to the end of training; final $\cos^2_{\min}$ at the last step. Outcome, from the medians: *not recovered* if final $\cos^2_{\min} < 0.2$; otherwise *aligned before plateau* if $\Delta \cos^2_{\min} < 0.005$ and final $\cos^2_{\min} \geq 0.95$; *no late gain* if $\Delta \cos^2_{\min} < 0.005$ otherwise; *late gain* if $\Delta \cos^2_{\min} \geq 0.005$.

Teacher	r_t	r_s	η	$1 - \rho_2$	$L/L_{V,\mathrm{opt}}$	$\Delta \cos^2_{\min}$	final $\cos^2_{\min}$	Outcome
ReLU	4	50	2	4.6×10^{-7}	3045	-0.028	0.670	no late gain
ReLU	6	50	1.5	9.2×10^{-9}	173	$+0.016$	0.955	late gain
ReLU	8	30	1.5	1.0×10^{-10}	16.2	$+0.532$	0.831	late gain
ReLU	8	50	1	1.0×10^{-10}	13.3	$+0.289$	0.857	late gain
ReLU	8	50	1.5	1.2×10^{-10}	40.9	$+0.189$	0.907	late gain
ReLU	12	50	1.5	1.0×10^{-10}	23.1	$+0.077$	0.820	late gain
ReLU	16	50	1.5	1.9×10^{-10}	11.3	$+0.079$	0.096	not recovered
ReLU	16	50	2	1.0×10^{-10}	26.7	$+0.093$	0.104	not recovered
ReLU	20	50	2	1.8×10^{-10}	16.1	$+0.052$	0.065	not recovered
GELU	4	50	0.7	6.4×10^{-6}	7.8	$+0.001$	0.995	aligned before plateau
GELU	6	50	0.7	1.3×10^{-7}	7.5	$+0.012$	0.990	late gain
GELU	8	30	0.7	2.6×10^{-9}	3.0	$+0.015$	0.977	late gain
GELU	8	50	0.5	5.5×10^{-9}	4.2	$+0.038$	0.967	late gain
GELU	8	50	0.7	1.7×10^{-9}	6.6	$+0.066$	0.960	late gain
GELU	12	50	0.6	3.9×10^{-10}	3.3	$+0.447$	0.758	late gain
GELU	12	50	0.7	2.3×10^{-10}	4.1	$+0.242$	0.823	late gain
GELU	12	50	0.85	2.1×10^{-10}	4.8	$+0.069$	0.886	late gain
GELU	12	80	0.7	1.8×10^{-10}	8.4	$+0.405$	0.557	late gain
GELU	16	50	0.7	1.7×10^{-10}	3.0	$+0.100$	0.809	late gain
GELU	16	50	0.85	1.5×10^{-10}	4.0	$+0.359$	0.517	late gain
GELU	20	50	0.7	2.4×10^{-10}	1.9	$+0.242$	0.556	late gain
SiLU	4	50	0.85	4.9×10^{-6}	8.4	-0.000	0.989	aligned before plateau
SiLU	6	50	0.85	1.8×10^{-6}	5.6	$+0.004$	0.987	aligned before plateau
SiLU	8	30	0.85	7.2×10^{-9}	2.4	$+0.017$	0.977	late gain
SiLU	8	50	0.6	1.1×10^{-8}	3.2	$+0.015$	0.977	late gain
SiLU	8	50	0.85	1.3×10^{-8}	5.6	$+0.016$	0.975	late gain
SiLU	12	50	0.7	5.8×10^{-10}	2.7	$+0.279$	0.894	late gain
SiLU	12	50	0.85	6.1×10^{-10}	4.1	$+0.312$	0.889	late gain
SiLU	12	50	1	6.6×10^{-10}	5.2	$+0.176$	0.916	late gain
SiLU	12	80	0.85	5.3×10^{-10}	11.5	$+0.211$	0.905	late gain
SiLU	16	50	0.85	1.9×10^{-10}	2.6	$+0.277$	0.845	late gain
SiLU	16	50	1	1.3×10^{-10}	3.9	$+0.376$	0.793	late gain
SiLU	20	50	0.85	1.0×10^{-10}	3.3	$+0.384$	0.415	late gain

retrospective rankings below use the 8,000-step trajectories of the multi-seed replication. For a run of T steps with per-step loss $L_1, \ldots, L_T$, let $\bar{L}$ be its running mean over windows of $\max(3, \lfloor T/30 \rfloor)$ steps, $s = \lfloor T/2 \rfloor$, L_{init} the mean of the first five losses and L_p the mean of $\bar{L}$ over steps s to T. Here T counts training steps, whereas n counts training examples. The slice $[s : T]$ restricts a sequence to that window. Both std and sd denote its empirical standard deviation. With

$$\mathrm{flat} = \max\Big\{0,\, 1 - \frac{\mathrm{std}(\bar{L}_{[s:T]})}{L_p}\Big\}, \qquad \mathrm{grow} = \max\{0,\, \cos^2_{\min}(W_T) - \cos^2_{\min}(W_s)\},$$

$$\mathrm{drop} = \min\Big\{1, \max\Big\{0, \log_{10} \frac{L_{\mathrm{init}}}{L_p}\Big\}\Big\},$$

the ReLU score is $\min(\mathrm{flat}, 1) \cdot \min(\mathrm{grow}, 1) \cdot \mathrm{drop}$: a flat second half, alignment growth after the midpoint, and loss reduction from initialisation, capped at one decade. For SiLU and GELU it is multiplied by $e^{-|\bar{k} - r_t|}/(1 + \mathrm{sd}(k))$, where k_t is the AGOP effective rank at the 50% threshold (the number of eigenvalues at least half the second largest) over the second half, with mean $\bar{k}$ and standard deviation $\mathrm{sd}(k)$. The scores reward a flat loss, alignment growth and some loss reduction; they do not involve the head-fit oracle. Ranked instead by the median score over all seeds, the six ReLU configurations are the same, in the order (e), (f), (b), (a), (g), (j) rather than the plotted (e), (b), (f), (g), (a), (j); among the six SiLU and the six GELU configurations shown, the configurations of Figure 13 rank third (first are $r_t = 20$, $\eta = 0.85$ for SiLU and $r_t = 10$, $\eta = 0.7$ for GELU).

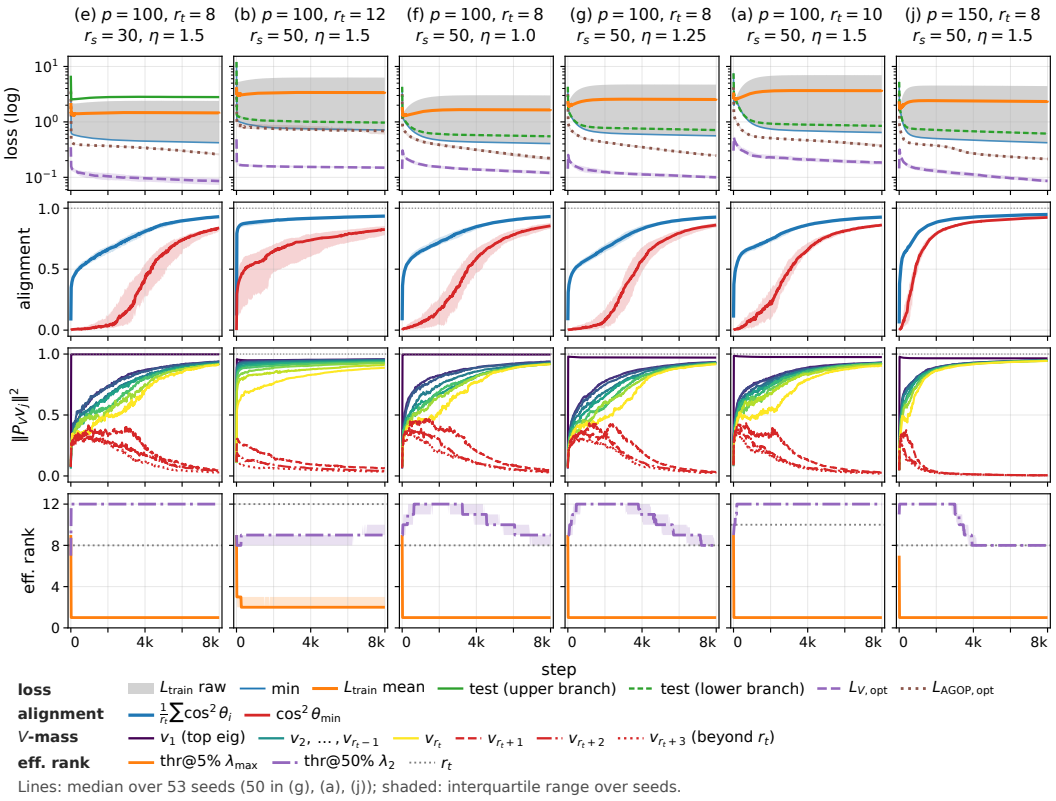


Figure 9: Top-6 ReLU configurations from the 10-configuration plateau-search grid: features (P1) and (P2). Columns are configurations, lettered as in Figure 12; lines are medians over 53 seeds (50 for (g), (a), (j)) and shading is the interquartile range. **Row 1** training loss (per-step range, running minimum, running mean), test loss and the oracles $L_{V,\text{opt}}$, $L_{\text{AGOP-opt}}$ (Section 3). Test loss and oracles are logged every 20 steps, so they sit on one branch of the period-2 loss oscillation, the same for every seed; the legend names the branch. Over the last 30% of training the cycle-mean training loss is 14–26 $\times$ $L_{V,\text{opt}}$ and 5–11 $\times$ $L_{\text{AGOP-opt}}$. **Row 2** AGOP direction-only alignment $\overline{\cos^2}$, $\cos^2_{\min}$ (verifies (P2)). **Row 3** per-eigenvector V-mass $\|P_V v_j\|^2$ for the top $r_t + 3$ AGOP eigenvectors. **Row 4** AGOP effective rank from the top 12 eigenvalues; dotted line at r_t.

C.4 Theory–experiment overlay and test-by-test detail

Figure 3 overlays theory and experiment for the 15 multi-activation configurations; the tests below give the per-configuration evidence. All use $p = 100$ and $r_s = 50$. ReLU (blue): $(r_t, \eta) = (4, 2)$ circle, $(6, 1.5)$ down-triangle, $(8, 1.5)$ square, $(12, 1.5)$ diamond, $(16, 2)$ star. GELU, $\eta = 0.7$ (purple): $r_t = 4, 8, 12, 16$ as square, down-triangle, diamond, star; GELU $r_t = 16$, $\eta = 0.85$: magenta plus. SiLU, $\eta = 0.85$ (brown): $r_t = 4, 8, 12, 16$ with the same shapes; SiLU $r_t = 20$, $\eta = 0.85$: olive plus.

Reading Figure 3. Theory–experiment overlay: 15 ReLU, GELU and SiLU teacher configurations. Medians over 53 seeds; interquartile ranges as error bars or shading. **(a)** Second-half per-step loss period-2 signature $1 - \rho_2$: at most about 6.3×10^{-6}, with values below 10^{-10} floored for display (supports (P1)). **(b)** Final top-r_t AGOP V-mass $\sum_{j \le r_t} \|P_V v_j\|^2$ against its maximum r_t (AGOP alignment; this is not the gradient leakage of Proposition 2(iii)): within 7% for $r_t \le 12$, except ReLU r_t=4, η=2 (10%); 6–28% below for $r_t \ge 16$. **(c)** Plateau loss (cycle mean, last 30% of steps) vs. head oracle $L_{V,\text{opt}}$: 23–3000$\times$ above diagonal for ReLU, 2.7–8.4$\times$ for GELU and SiLU. **(d)** Late-half $\cos^2_{\min}$ secant vs. r_t/M: log-log slope 1.8 across the 12 configurations with late-half gain > 0.005; three saturated r_t=4 configurations are open markers. This descriptive raw-gain fit, not the

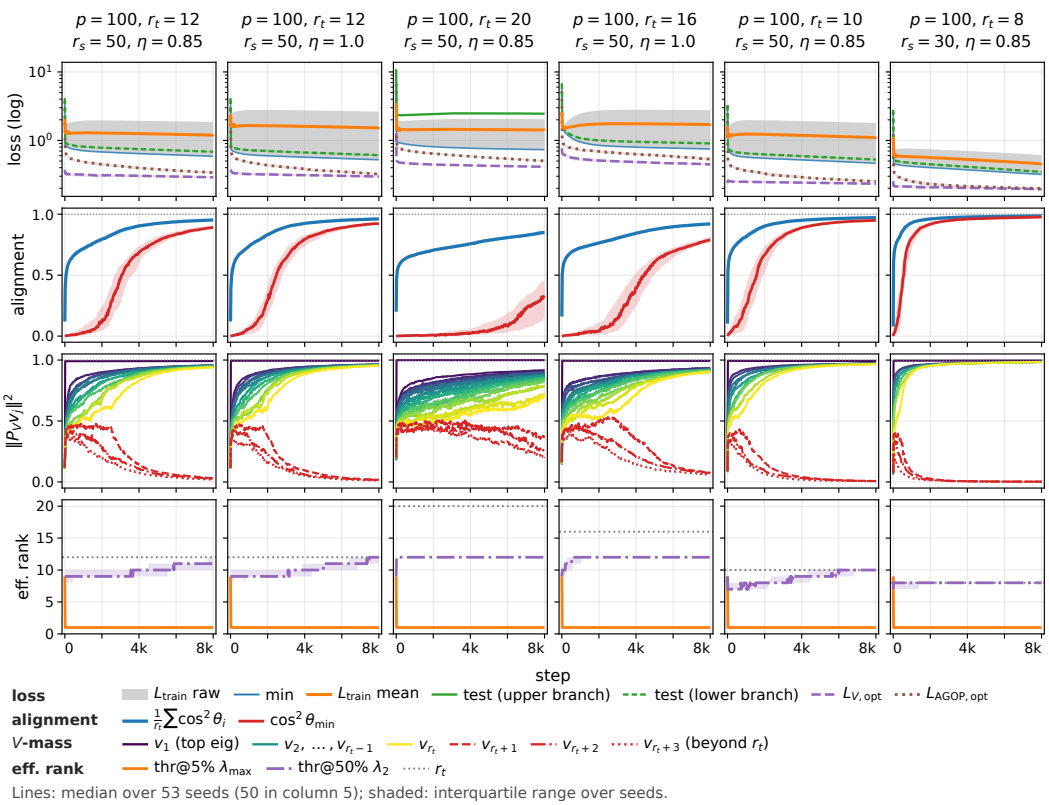


Figure 10: Six SiLU configurations, selected on seed 5 with the smooth-teacher score of Section C.3. Same four-row layout and statistics as Figure 9: medians over 53 seeds (50 in column 5), interquartile range shaded.

normalised rate v_k/e_k, does not test Proposition 4. **(e)** Off-V AGOP-mass proxy $(1 - \text{mass}_V) \cdot M$ vs. $\eta^2(M - r_t)$, with a unit-slope reference line whose prefactor is the median ratio (approximately 0.09). The slope is fixed, not fitted. This comparison does not measure the fixed-head loss residual in Claim 3. **(f)** $\cos^2_{\min}$ (solid), $\overline{\cos^2}$ (dashed): four selected trajectories (illustrates (P2)).

Test A: period-2 loss signature. Over the second half of training, the seed median of $1 - \rho_2$, with $\rho_2 := -\text{Corr}(\Delta L_t, \Delta L_{t+1})$ and $\Delta L_t := L_{t+1} - L_t$, is at most about 6.3×10^{-6} in all 15 configurations of Figure 3(a); values below 10^{-10} are floored for display. The loss increments therefore alternate almost perfectly; this does not establish an exact two-cycle of the weights (Section 5.1).

Test B: V/V$^\perp$ polar mass split (Proposition 2). Computed by taking the polar update on the saved final W and splitting into $\|P_V \,\text{Polar}(g)\|_F^2$ and $\|P_\perp \,\text{Polar}(g)\|_F^2$, where g is the gradient reconstructed as specified in the table caption.

Test C: plateau height vs head-fit limit (Claim 3). For a trajectory of T steps, measure $L_{\text{plateau}} := \text{mean}(L_{[\lfloor 0.7T \rfloor : T]})$ and compare to the head-fit oracle $L_{V,\text{opt}}$, which optimises the head amplitudes after projecting W onto V.

Test D: drift growth rate. Measure the empirical growth rate of $\cos^2_{\min}$ over the second half of training, and compare to the naive scale $\eta\, r_t/M$ (per step). In Table 5, the measured rate is the finite secant $[\cos^2_{\min}(W_{t_b}) - \cos^2_{\min}(W_{t_a})]/(t_b - t_a)$ from the first logged checkpoint in the second half, t_a, to the final checkpoint t_b, rather than a continuous-time derivative.

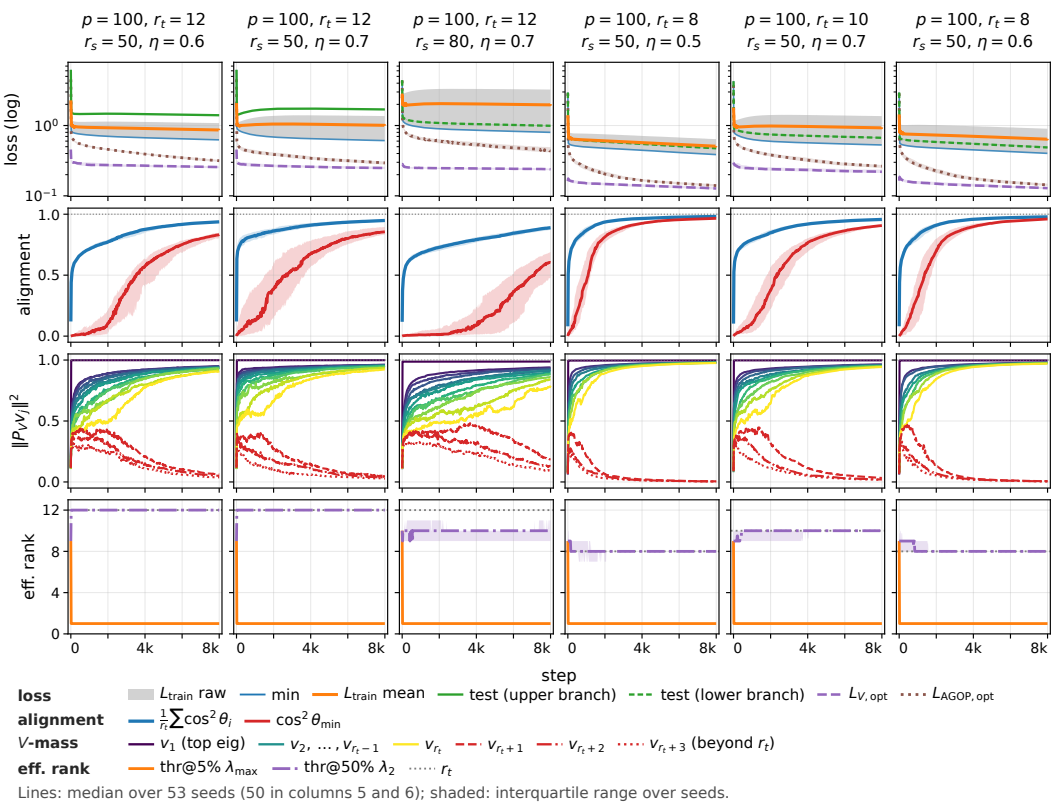


Figure 11: Six GELU configurations, selected on seed 5 with the smooth-teacher score of Section C.3. Same four-row layout and statistics as Figure 9: medians over 53 seeds (50 in columns 5 and 6), interquartile range shaded.

Table 3: Polar mass split on three long-run ReLU checkpoints. The total equals M to numerical precision (Proposition 2(i)); the V-mass is approximately 2–5% below r_t. The gradient is recomputed from the saved weights with a unit output head, and the leakage scalars of Proposition 2(iii) were not measured separately.

	V-mass		$V^\perp$-mass		Total mass	
Config (σ_t, r_t, r_s)	Predicted	Measured	Predicted	Measured	M	Measured
ReLU, $r_t = 4$, $r_s = 50$	4	3.79	46	46.21	50	50.00
ReLU, $r_t = 8$, $r_s = 50$	8	7.86	42	42.14	50	50.00
ReLU, $r_t = 12$, $r_s = 50$	12	11.76	38	38.24	50	50.00

Test E: $L_{\text{AGOP-opt}}$ below the training loss. On the long-run ReLU configs in Table 6, $L_{\text{AGOP-opt}}$ lies well below L_{train} over most of the plateau, though not at every checkpoint; on the six cleanest ReLU configurations of Figure 9 the seed-median ratio of the cycle-mean training loss to $L_{\text{AGOP-opt}}$ over the last 30% of training is about 5–11, and the same ordering is visible in the loss panels of Figures 10 and 11. The diagnostic uses no teacher information. Because it projects W before refitting the head, it shows that the learned directions support a better predictor, not that the head alone limits the loss.

Test F: weight mass versus AGOP mass. Across the long-run ReLU configs, top-r_t AGOP eigvecs align with V ($\overline{\cos^2} \in [0.85, 0.96]$) and the AGOP eigenvalue mass concentrates ($\text{mass}_V \in [0.83, 0.90]$), but the weight Frobenius mass fraction A_{sub} stays in $[0.29, 0.50]$. If the weight Gram matrix $WW^\top$ were proportional to the AGOP, the simplest form of the Neural Feature Ansatz, the

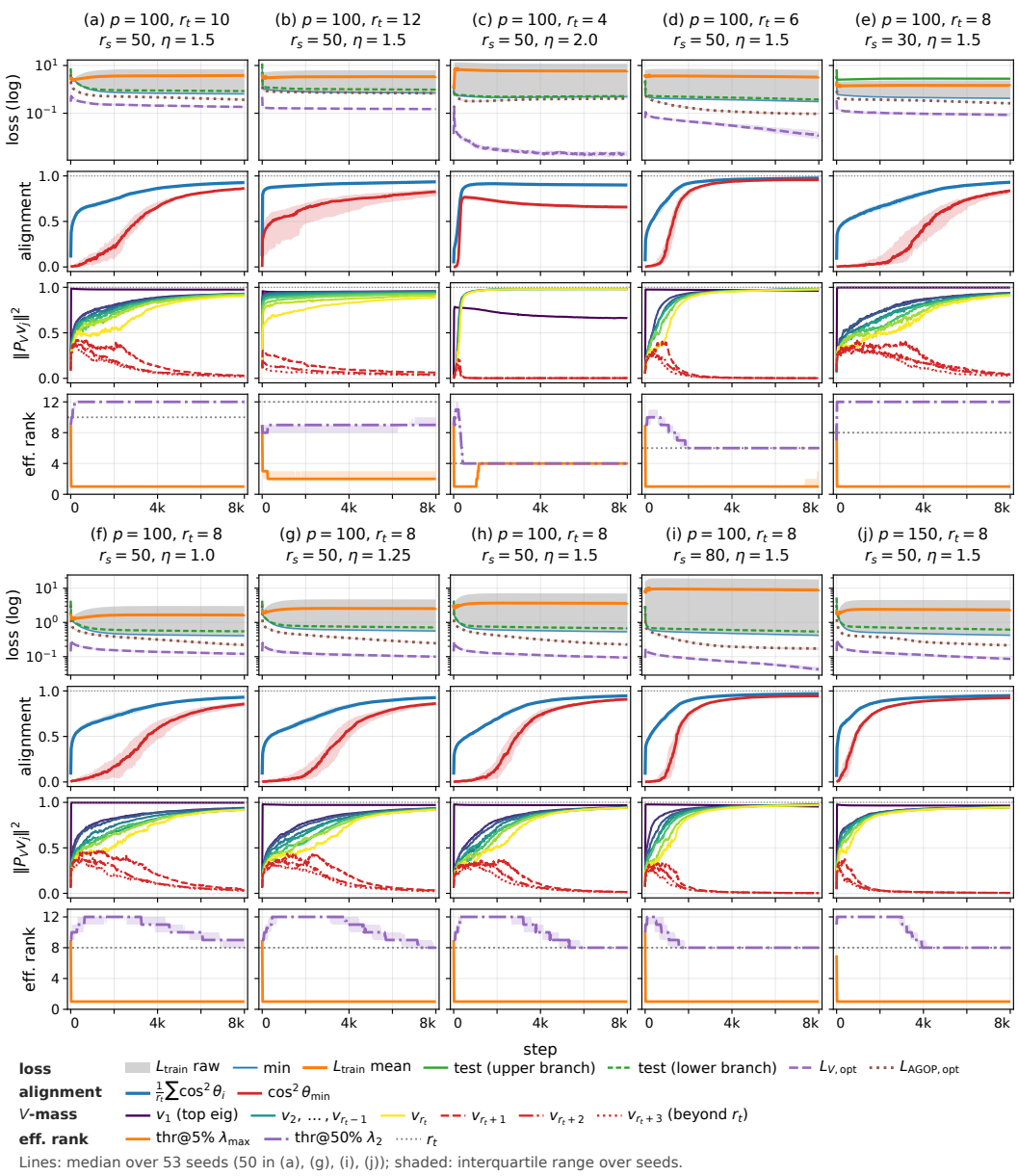


Figure 12: Full 10-configuration plateau-search grid for the ReLU teacher (superset of Figure 9). Each column is one (p, r_t, r_s, η) configuration; rows and statistics as in Figure 9 (53 seeds; 50 in (a), (g), (i), (j)). Configurations span $r_t \in \{4, 6, 8, 10, 12\}$, $r_s \in \{30, 50, 80\}$, $p \in \{100, 150\}$ and $\eta \in \{1.0, 1.25, 1.5, 2.0\}$.

two mass fractions would be equal; they differ by a large margin. Power-law variants of the ansatz (e.g. with the square root of the AGOP) are not tested here.

Test G: comparison with the naive per-step scale. The directly measured statistic is the late-window alignment secant of Test D. Its ratio to $\eta r_t / M$ lies between 1.10×10^{-4} and 9.16×10^{-4} on the four configurations in Table 5. Thus this scale substantially exceeds the observed alignment drift. This comparison does not measure a one-step weight-update signal-to-noise ratio or verify the conditional drift hypotheses.

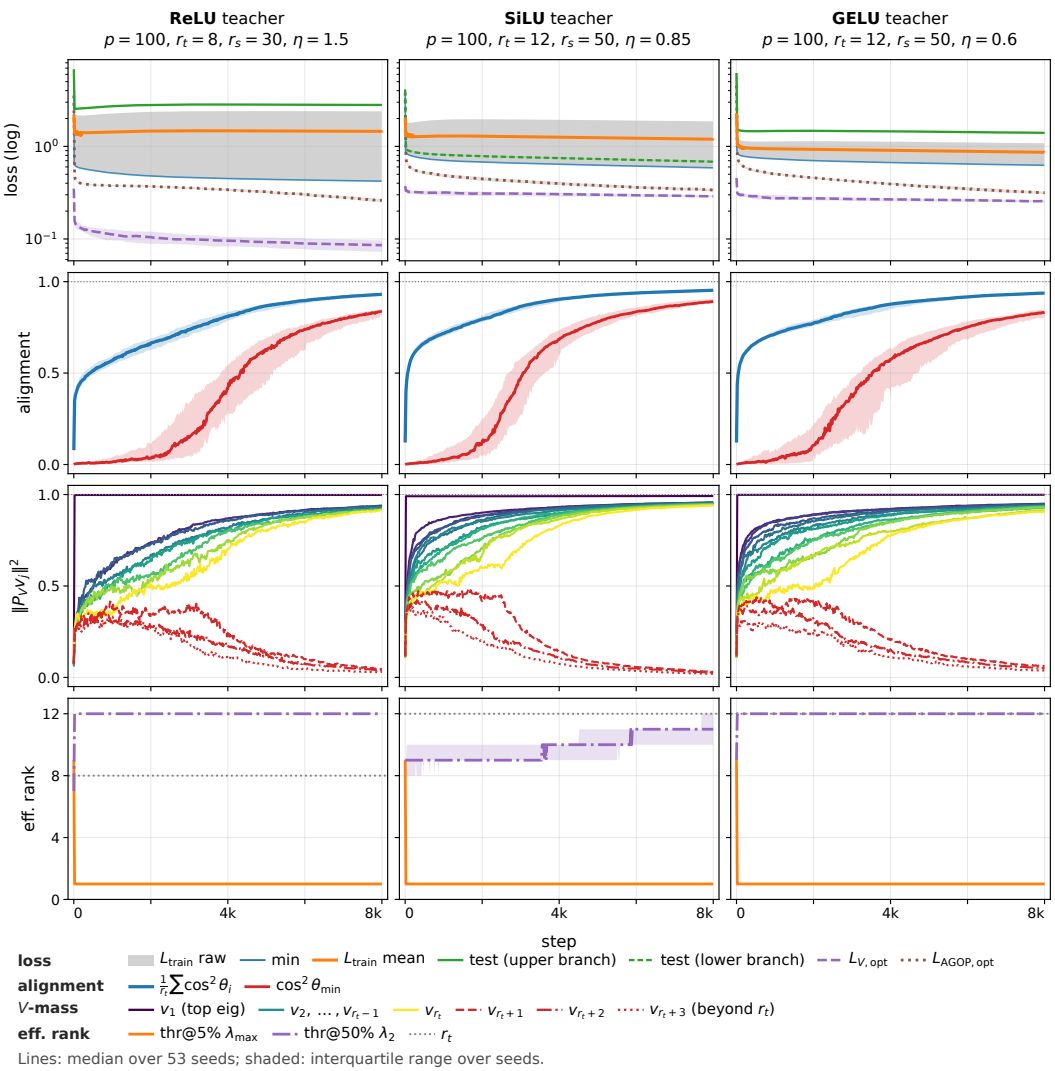


Figure 13: Cleanest configuration per teacher activation: ReLU, SiLU, GELU. The four-row diagnostic (loss decomposition, AGOP direction-only alignment, per-eigvec V-mass, AGOP effective rank) reproduces features (P1) and (P2) for all three teacher activations. Per-activation configurations are the seed-5 top choices of Section C.3, kept for the multi-seed runs (with all seeds, the SiLU and GELU choices rank third among the six shown). Per-teacher 10-config grids and 6-config rankings are in Figures 10–12 above. Lines are medians over 53 seeds; shading is the interquartile range.

Table 4: Plateau height vs head-fit oracle. The plateau loss sits $10\times$–$36\times$ above $L_{V,\mathrm{opt}}$, the loss after projecting W onto V and refitting the head: the representation supports a much better predictor than the trained network realises (Test E and Claim 3).

$(\sigma_t, p, r_t, r_s, \eta)$	L_{plateau}	$L_{V,\mathrm{opt}}$	ratio $L_{\mathrm{plateau}}/L_{V,\mathrm{opt}}$
ReLU, $(100, 8, 50, 1.0)$	1.70	0.18	9.6
ReLU, $(100, 8, 50, 1.5)$	3.74	0.10	36.0
ReLU, $(100, 12, 50, 1.5)$	3.48	0.16	21.5

Table 5: Drift rate scaling. The measured per-step rate is $O(10^{-4})$ across configs, while the comparison scale $\eta r_t/M$ is $O(10^{-1})$; the table reports their ratio (Test G). This raw rate is not the normalised rate of Proposition 4, whose conditions we have not verified here. The table covers $r_t \in \{6, 8, 10, 12\}$ at fixed $\eta = 1.5$.

$(\sigma_t, p, r_t, r_s, \eta)$	predicted $\eta r_t/M$	measured secant	ratio
ReLU, $(100, 6, 50, 1.5)$	0.18	2.0×10^{-5}	$+1.13 \times 10^{-4}$
ReLU, $(100, 8, 50, 1.5)$	0.24	2.2×10^{-4}	$+9.16 \times 10^{-4}$
ReLU, $(100, 10, 50, 1.5)$	0.30	2.6×10^{-4}	$+8.62 \times 10^{-4}$
ReLU, $(100, 12, 50, 1.5)$	0.36	4.0×10^{-5}	$+1.10 \times 10^{-4}$

Table 6: Weight mass versus AGOP mass. Direction-only alignment $\overline{\cos^2}$ lies in $[0.85, 0.96]$ and eigenvalue-weighted AGOP mass-in-V in $[0.83, 0.90]$; the weight Frobenius mass-in-V A_{sub} caps at ≈ 0.50 across $r_t \in \{6, 8, 10, 12\}$ at fixed $\eta = 1.5$.

$(\sigma_t, p, r_t, r_s, \eta)$	$\overline{\cos^2}$	$\text{mass}_V(G_s)$	A_{sub}
ReLU, $(100, 6, 50, 1.5)$	0.96	0.90	0.35
ReLU, $(100, 8, 50, 1.5)$	0.86	0.85	0.29
ReLU, $(100, 10, 50, 1.5)$	0.85	0.83	0.29
ReLU, $(100, 12, 50, 1.5)$	0.86	0.84	0.50

C.5 How theory and experiment align: a discussion

Combining the long-run ReLU tables in Section C.4 with the top-6 figures for ReLU, SiLU, and GELU (Figures 9, 10, 11) lets us distinguish exact identities from empirical comparisons. The qualitative loss/alignment separation occurs across the three teacher activations; the polar-mass and weight-versus-AGOP-mass tables are restricted to ReLU.

Exact identities (machine precision). The polar Frobenius identity (Proposition 2) holds to machine precision in every run. It is an algebraic fact about the polar map, independent of the activation, the data distribution and the learning rate. The period-2 loss signature (Test A) is an empirical observation, not an algebraic fact.

Structural identities. The polar Frobenius identity and the $V/V^{\perp}$ decomposition are exact (Proposition 2). The measured V-mass of the polar update (Test B, Table 3) is approximately 2–5% below r_t on three ReLU checkpoints; since the leakage scalars were not measured, we do not attribute this gap quantitatively. The AGOP V-mass of Figure 3(b) is a separate, dynamical quantity.

Drift rate and plateau height. The measured drift rate (Test D) is about 10^{-3} of $\eta r_t/M$ per step (Table 5; see Test G for the range of ratios); Proposition 4 gives sufficient conditions for a power-law normalised rate, which we have not verified along these trajectories. The plateau height (Test C, Claim 3) sits 10–36$\times$ above the head-fit oracle $L_{V,\text{opt}}$; the off-V prefactor at high η remains an open quantitative target.

Diagnostic gap. Over the last 30% of training the seed-median ratio of the cycle-mean training loss to $L_{\text{AGOP-opt}}$ is about 5–11 on the six cleanest ReLU configurations (Test E): the learned directions support a better predictor than the trained network realises. The mass gap of Test F is similarly clear: $\overline{\cos^2}$ saturates above 0.85 while A_{sub} caps in the range $[0.29, 0.50]$ across the long-run ReLU configs in Table 6. The per-eigenvector V-mass rows of Figures 10 and 11 show AGOP directional alignment for the smooth teachers; they do not measure weight-mass concentration.

Comparison across teacher activations. Smooth teachers (GELU, SiLU) exhibit the qualitative loss/alignment separation at tested learning rates roughly half those of the ReLU examples (GELU: $\eta \approx 0.7$; SiLU: $\eta \approx 0.85$; ReLU: $\eta \approx 1.5$). Training loss *rising* during alignment growth is especially visible for SiLU $r_t = 20$ (Figure 10). The polar-mass and weight-versus-AGOP-mass tables here are restricted to ReLU and do not establish the same quantitative comparisons for the smooth teachers.

D DEEP GAUSSIAN TEACHER–STUDENT EXPERIMENTS

This appendix specifies the deep-network experiments: data, teacher, student, optimiser, metrics, the scoring protocol, the configurations, and the depth results.

Notation. Here p is the input dimension, r_t the teacher-subspace dimension, m the hidden width, L the number of hidden layers and T the number of training steps. We write $W = (W_1, \dots, W_L)$ for all hidden matrices, reserving L_t for the scalar training loss at step t. The training, test and AGOP-subset sizes are n, n_{test} and n_m, respectively. The norms $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_F$ denote vector ℓ_1, Euclidean and matrix Frobenius norms; I_p is the identity matrix. The learning rates η_1 and η_{deep} are coefficients of the polar factor, not the Frobenius lengths of the resulting updates.

D.1 DATA, TEACHER AND STUDENT

Data and teacher. Inputs are $x \sim \mathcal{N}(0, I_p)$ and labels are the noise-free outputs of a fixed teacher. Two seeds fix all randomness: the *problem seed* s fixes the teacher, the training and test inputs and the AGOP subset (generators seeded with $10000 + s$, $20000 + s$, $30000 + s$ and $50000 + s$), and the *initialisation seed* s' fixes the student's initial weights (generator seeded with $40000 + s'$). The teacher subspace is $V = \operatorname{span}(U)$, where $U \in \mathbb{R}^{p \times r_t}$ has orthonormal columns (QR factor of a Gaussian matrix). The default teacher has two layers,

$$f^\star(x) \;=\; \sum_{j=1}^{2r_t} \Big[\mathrm{ReLU}\big(\mathrm{ReLU}(x^\top U)\,[Q,\,-Q]\big) \Big]_j \;=\; \big\| Q^\top \mathrm{ReLU}(U^\top x) \big\|_1, \tag{76}$$

with $Q \in \mathbb{R}^{r_t \times r_t}$ a random orthogonal matrix, so its input-relevant subspace is again V. The labels are rescaled by one constant per problem so that their root mean square on the training inputs equals that of the one-layer teacher $f_1^\star(x) = \sum_{j \le r_t} \mathrm{ReLU}(u_j^\top x)$, which one breadth arm uses instead.

Student. The student is a residual ReLU network of depth L and width m, without biases:

$$h_1 = \mathrm{ReLU}(x^\top W_1), \qquad h_\ell = h_{\ell-1} + \alpha\,\mathrm{ReLU}(h_{\ell-1} W_\ell) \quad (\ell = 2, \dots, L), \qquad f_W(x) = h_L\, a, \tag{77}$$

with $W_1 \in \mathbb{R}^{p \times m}$, $W_\ell \in \mathbb{R}^{m \times m}$ and residual scale $\alpha = 1/(L-1)$ unless stated otherwise. The head $a = (c_a/\sqrt{p})\,\mathbf{1}_m$ is frozen. At initialisation W_1 has i.i.d. $\mathcal{N}(0, 1/p)$ entries and each W_ℓ, $\ell \ge 2$, i.i.d. $\mathcal{N}(0, 1/m)$ entries, all multiplied by a gain g. One arm, used in Appendix E, replaces the residual network by a plain MLP, $h_\ell = \mathrm{ReLU}(h_{\ell-1} W_\ell)$, with He initialisation ($g = \sqrt{2}$), $\eta_1 = 0.6$ and $\eta_{\text{deep}} = 0.02$.

Optimiser. Every hidden matrix is trained by the exact matrix-Muon step on the full-batch loss $\mathcal{L}(W) = \frac{1}{n} \sum_{i \le n} \big(f_W(x_i) - y_i\big)^2$,

$$W_1 \leftarrow W_1 - \eta_1\, \mathrm{Polar}\big(\nabla_{W_1} \mathcal{L}\big), \qquad W_\ell \leftarrow W_\ell - \eta_{\text{deep}}\, \mathrm{Polar}\big(\nabla_{W_\ell} \mathcal{L}\big) \quad (\ell \ge 2), \tag{78}$$

where $\mathrm{Polar}(G) = U_G V_G^\top$ is formed from the thin SVD $G = U_G \Sigma V_G^\top$. There is no momentum, Newton–Schulz iteration, weight decay or learning-rate schedule; arithmetic is float64. Runs last $T = 1000$ steps.

D.2 METRICS AND SCORING PROTOCOL

Reading Figure 4. Deep residual students under exact-polar Muon: cycling loss, rising alignment. Columns: Table 8's depth arms and D8 on fresh seeds. **Row 1:** training loss: per-step range between period-2 branches (grey), running minimum, the lower branch (blue), 20-step running mean (orange). **Row 2:** top-r_t input-AGOP eigenspace alignment with the teacher subspace, $\overline{\cos^2}$ and $\cos^2_{\min}$, averaged over each two-step checkpoint pair. Median lines with interquartile shading over 100 runs per column (20 problems $\times$ 5 initialisations). The dotted line starts the scoring window. Across columns, the median per-run cycle-mean loss change is at most about 2.6%, with falling lower branch and rising alignment measures. Table 9 instead reports medians over problems of the per-problem medians; their largest cycle-mean loss change is approximately 2.62%.

Metrics. The end-to-end input AGOP $G_s = \frac{1}{n_m}\sum_{i\le n_m} \nabla_x f_W(x_i)\nabla_x f_W(x_i)^\top$ is computed on a fixed random subset of n_m training inputs; its top-r_t eigenspace is compared with V through principal angles, giving the mean alignment $\overline{\cos^2}$ and the weakest-direction alignment $\cos^2_{\min}$ of Eq. (9). Explicitly, if B has orthonormal columns spanning the top-r_t eigenspace of G_s, these are $r_t^{-1}\|U^\top B\|_F^2$ and $\sigma_{\min}(U^\top B)^2$, where $\sigma_{\min}$ is the smallest singular value. Both are evaluated at pairs of consecutive steps $(t, t+1)$, every 100 steps, and at $(999, 1000)$. The loss cycle is summarised by $\rho_2 = -\mathrm{Corr}(\Delta L_t, \Delta L_{t+1})$ with $\Delta L_t = L_{t+1} - L_t$, by the two-step ratio $R_2 = \mathrm{median}_t\, |L_{t+2} - L_t|/|L_{t+1} - L_t|$, and by the cycle mean $\bar{L}_t = \frac{1}{2}(L_t + L_{t+1})$. The held-out refit loss replaces the frozen head by a least-squares head on the last hidden layer h_L, fitted on the n_m AGOP inputs (ridge $10^{-8}\,\mathrm{tr}(H^\top H)/m$), where $H \in \mathbb{R}^{n_m \times m}$ has rows $h_L(x_i)$, and evaluates it on the n_{test} test inputs.

Window and parity. All statistics use the window from step 200 to step 1000. Because the loss alternates between two branches, the gain of a metric is its change between the window ends computed on each parity, even steps $200 \to 1000$ and odd steps $201 \to 999$, and averaged over the two.

Strict run test. A run passes if, on the window, all of the following hold: (i) $\rho_2 \ge 0.99$; (ii) $R_2 \le 0.01$; (iii) the relative change of $\bar{L}_t$ between the window ends, divided by the window length in hundreds of steps, is at most 0.02 in absolute value; (iv) $\overline{\cos^2}$ gains at least 0.05; (v) $\cos^2_{\min}$ gains a positive amount; (vi) both gains are positive on each parity separately.

Aggregation. The problem seed, not the run, is the unit of evidence. In the 20×5 campaigns a problem succeeds if at least 4 of its 5 initialisations pass, and an arm is robust if at least 15 of its 20 problems succeed. The 12×4 campaigns of Appendices E and F use 3 of 4 and 9 of 12.

D.3 Configurations and seeds

Table 7 lists the settings shared by every run and Table 8 the depth arms. Every campaign has its own block of problem seeds and initialisation seeds, and the blocks are pairwise disjoint: 20 problems $\times$ 5 initialisations for D6–D12, for D16 and for the fresh-D8 block, and 12 problems $\times$ 4 initialisations for the runs of Appendices E and F.

Table 7: Settings shared by every deep run unless an arm overrides them.

input dimension p	64	teacher rank r_t	4
student width m	32	teacher depth	2 Eq. (76)
training inputs n	4096	test inputs n_{test}	8192
AGOP inputs n_m	2048	steps T	1000
initial gain g	1	head scale c_a	1
residual scale α	$1/(L-1)$	architecture	residual Eq. (77)

Table 8: Depth arms. The fresh-D8 block reuses the D8 configuration with new seeds.

arm	L	η_1	η_{deep}	c_a	α
D6	6	0.66	3.125×10^{-3}	0.75	1/5
D8	8	0.66	7.8125×10^{-4}	1	1/7
D10	10	0.70	3.90625×10^{-4}	1	1/9
D12	12	0.70	1.953125×10^{-4}	1	1/11
D16	16	0.74	1.953125×10^{-4}	1	1/15

D.4 Depth results

Figure 4 shows the trajectories and Table 9 the scores.

Scope of these runs. The step sizes are on two time scales ($\eta_1 \approx 0.7$, $\eta_{\mathrm{deep}} \le 3 \times 10^{-3}$) and the residual branches are scaled by $1/(L-1)$, so we describe these runs as feature learning in a

Table 9: Depth campaigns under the strict protocol (Appendix D.2): runs passing, problems succeeding, and, as medians over problems of the per-problem median over initialisations, the window gains of $\overline{\cos^2}$ and $\cos^2_{\min}$ and the relative change over the window of the cycle-mean training loss and of the held-out refit loss.

arm	strict runs	problems	$\overline{\cos^2}$ gain	$\cos^2_{\min}$ gain	cycle mean	held-out refit
D6	94/100	20/20	$+0.120$	$+0.384$	$+2.62\%$	-6.15%
D8	97/100	20/20	$+0.206$	$+0.515$	$+0.62\%$	-6.88%
D10	96/100	19/20	$+0.200$	$+0.441$	$+1.10\%$	-5.38%
D12	95/100	20/20	$+0.193$	$+0.436$	$+1.55\%$	-5.20%
D16	96/100	19/20	$+0.187$	$+0.418$	$+1.75\%$	-4.60%
Fresh D8	91/100	20/20	$+0.153$	$+0.396$	$+1.24\%$	-6.58%

deep trainable network, not as every layer learning the teacher directions. A period-2 loss is not a period-2 parameter orbit: the two loss branches keep drifting inside the window while the cycle mean stays nearly flat. The polar step uses every column of the thin SVD, and deep-layer gradients can be numerically rank-deficient; runs reproduce exactly in the pinned software stack. The data are synthetic Gaussian, training is full batch with the exact polar factor, no momentum, a frozen head and fixed step sizes.

E FREEZING THE DEEP LAYERS SWITCHES THE PLATEAU ON

In the deep runs of Appendix D the deeper layers take steps 200 to 3,800 times smaller than the first layer (Table 8). This appendix turns that observation into an experiment. Keeping everything else fixed, we change how much of the network between W_1 and the head may train, and ask when the plateau with continued feature learning appears:

- with the deep layers frozen or training slowly, the cycle-mean loss stays flat while the alignment keeps rising, and the learned alignment sits in W_1;
- with the deep layers training at a rate comparable to the first layer, the alignment still rises, but the loss decreases through the window and part of the alignment moves into the deep layers;
- freezing the deep layers in the middle of a run makes a decreasing loss flat, and unfreezing them makes a flat loss decrease, while the alignment keeps rising in both cases;
- a single trainable layer directly after W_1 is enough to remove the plateau, while a single trainable layer next to the head is not.

E.1 SETUP

All runs use the problems, student, head, optimiser, metrics and scoring protocol of Appendix D and one 12×4 seed block (12 problems $\times$ 4 initialisations, 48 runs per arm); only the per-layer step sizes change. There are 1,872 runs, none of which diverged.

Notation. The first hidden matrix is W_1 and the deeper matrices are $W_2, \dots, W_L$; η_1 and η_{deep} multiply their polar updates. The width is m, the total number of training steps is T, and $\bar{L}_t = (L_t + L_{t+1})/2$ is the cycle mean of the training loss L_t. For a teacher basis U and a leading input-AGOP basis B, both with orthonormal columns, $\overline{\cos^2} = r_t^{-1}\|U^\top B\|_F^2$ and $\cos^2_{\min} = \sigma_{\min}(U^\top B)^2$. Writing $\widehat{\mathbb{P}}_n$ for the uniform distribution on the training pairs, the "offset" is the square of the mean residual, $\left(\mathbb{E}_{(x,y)\sim\widehat{\mathbb{P}}_n}[f_W(x) - y]\right)^2$; it is distinct from the MSE $L_t = \mathbb{E}_{(x,y)\sim\widehat{\mathbb{P}}_n}[(f_{W_t}(x) - y)^2]$.

Step-size ladder. At depth 8 we keep $\eta_1 = 0.66$ and set $\eta_{\text{deep}} \in \{0,\ 7.8 \times 10^{-4},\ 3.1 \times 10^{-3},\ 0.0125,\ 0.02,\ 0.05,\ 0.2,\ 0.66\}$, from frozen deep layers through the D8 value of Table 8 ($\eta_1/845$) to the same step for every layer. The same ladder is run for the plain network of Appendix D.1 ($\eta_1 = 0.6$, He initialisation). At depths 6 and 16 we run four points: frozen, the value of Table 8, $\eta_1/50$ and η_1.

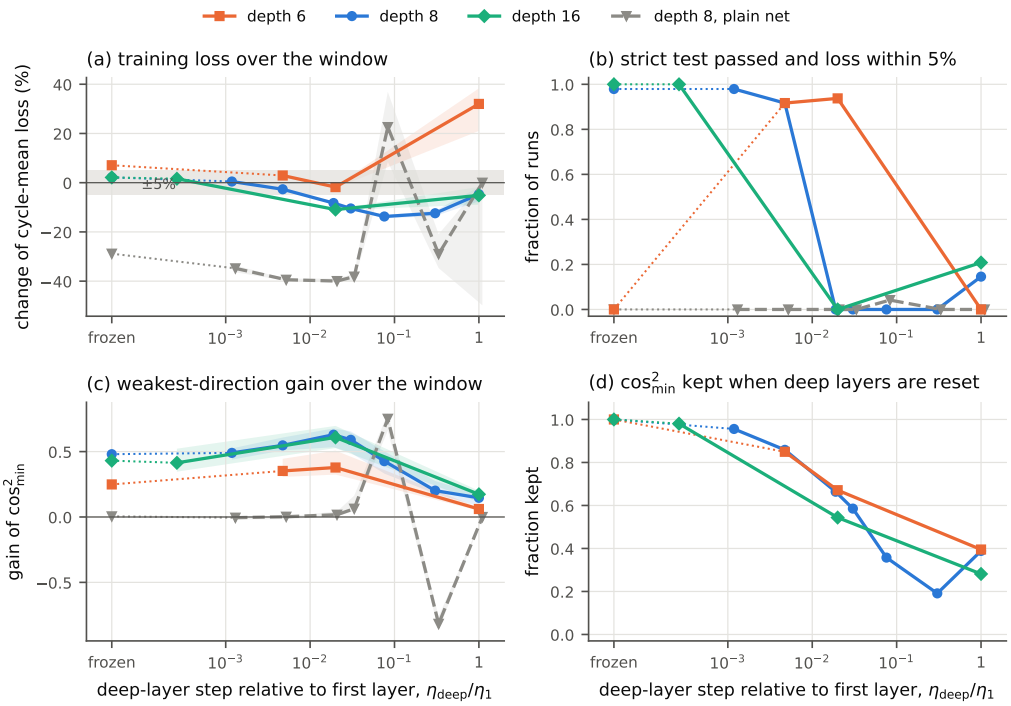


Figure 14: Slowing the deep layers makes the loss flat; training them makes it fall. Residual students of depth 6, 8 and 16 and the depth-8 plain network, against the deep-layer step relative to the first layer's. **(a)** Change of the cycle-mean training loss over the window (grey band: $\pm 5\%$). **(b)** Fraction of runs that pass the strict test with the cycle mean within 5%. **(c)** Window gain of $\cos^2_{\min}$. **(d)** Fraction of the final $\cos^2_{\min}$ that remains when the deep layers are reset to their initialisation. Medians over problems of the per-problem median over initialisations; shading is the interquartile range over problems; the frozen point is joined to the others by a dotted segment.

Partial freezing. At depth 8 the first layer and a block of j of the seven deep layers train with $\eta_1 = 0.66$ and the other deep layers are frozen, for $j \in \{1, 2, 4, 6\}$, with the block either next to W_1 ($W_2, \ldots, W_{j+1}$) or next to the head ($W_{9-j}, \ldots, W_8$). The end points $j = 0$ and $j = 7$ are the ladder arms with $\eta_{\text{deep}} = 0$ and $\eta_{\text{deep}} = \eta_1$.

Switching during a run. At depth 8 with $T = 1500$, the deep layers train with η_{on} until step 500 and are frozen afterwards, or are frozen until step 500 and train with η_{on} afterwards, for $\eta_{\text{on}} \in \{0.0125, 0.66\}$. The controls keep the deep layers frozen, or training with η_{on}, for all 1500 steps. The first layer always trains with $\eta_1 = 0.66$. The switching runs follow the same protocol step for step (with a constant rate they reproduce the ladder runs' losses exactly) and score two windows with the strict test: steps $200 \to 500$ before the switch and $700 \to 1500$ after it.

Flatness and localisation. The strict test of Appendix D.2 allows the cycle-mean loss to change by up to 2% per hundred steps, about $\pm 16\%$ over the window. We therefore also count runs that pass it with the cycle mean changing by at most 5%. To see where the learned alignment lives, we recompute $\cos^2_{\min}$ at the end of the window with the deep layers reset to their initial values and report the fraction of the trained network's $\cos^2_{\min}$ that remains.

E.2 Results

Step size. Figure 14 gives the ladders. At depth 8, with the deep layers frozen, at the D8 step ($\eta_1/845$) or at $\eta_1/211$, 142 of 144 runs pass the strict test and 138 keep the cycle mean within 5%; the weakest direction gains $+0.48$ to $+0.55$, and resetting the deep layers keeps 86–100% of the final $\cos^2_{\min}$, so the alignment sits in W_1. From $\eta_1/53$ upwards the alignment still rises (gains $+0.20$ to

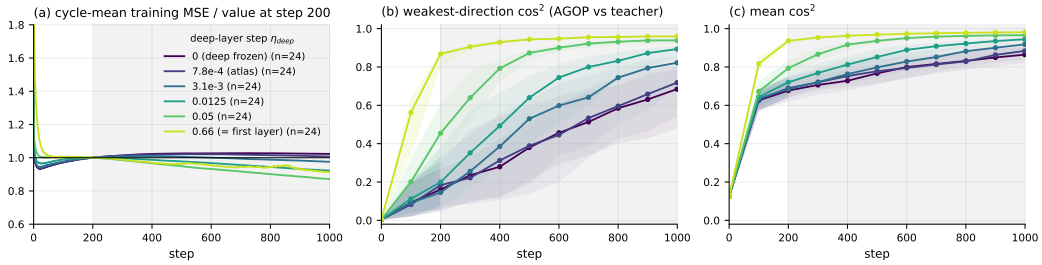


Figure 15: The deep-layer step ladder at depth 8 over time. Residual students, first-layer step 0.66 fixed, deep-layer step from 0 (frozen) to 0.66 (the same step for every layer); medians over 24 runs (12 problems × 2 initialisations), interquartile range shaded; grey: the scored window. **(a)** Cycle-mean training loss divided by its value at step 200. **(b)** $\cos^2_{\min}$. **(c)** $\overline{\cos^2}$. With the deep layers frozen or slow (up to $\eta_1/211$) the loss stays flat while the weakest direction aligns; from $\eta_1/53$ on the loss falls through the window and the alignment forms earlier.

+0.63), but the cycle mean falls by 8–14% over the window, no run stays within 5%, the held-out refit loss falls by 22–43%, and the deep layers carry a growing part of the alignment (only 19–66% remains when they are reset). The strict test still passes most of these runs because of its ±16% tolerance. With the same step for every layer most of the alignment forms before the window, and 19 of 48 runs pass. Depth 16 behaves like depth 8: frozen or at the D16 step, all 96 runs pass and stay within 5%; at $\eta_1/50$ none stays within 5% and the cycle mean falls by 10.9%. At depth 6 the cycle mean drifts up by 7% when the deep layers are frozen entirely, mostly through the offset part $\left(\mathbb{E}_{\widehat{\mathbb{P}}_n}[f_W(x) - y]\right)^2$, and the loss is flattest at small non-zero steps ($\eta_1/211$ and $\eta_1/50$: 89 of 96 runs within 5%); the same step for every layer makes the depth-6 loss rise by 32%. The plain network fails the strict test at every step, including with its deep layers frozen, where the cycle mean falls by 29%. Figure 15 shows the depth-8 trajectories.

Reading Figure 5. The main-text figure shows depth-8 students trained for $T = 1500$ steps, with the deep layers switched at step 500 (dotted line). The first layer always trains with $\eta_1 = 0.66$. Colour gives the current state of the deep layers: blue is frozen and orange is training. Dashed curves change state at step 500 and are drawn only after the switch, since before it they coincide with the control they start from. Panels (a,b) use deep-layer step 0.0125 while training, and (c,d) use 0.66. Panels (a,c) show cycle-mean training loss divided by its last value before the switch; (b,d) show $\cos^2_{\min}$. Curves are medians over 48 runs; the shaded bands mark the two scoring windows, not uncertainty intervals.

Switching. Figure 5 switches the deep layers during a run. With $\eta_{\text{on}} = 0.0125$ the loss responds at once. Freezing the deep layers at step 500 stops the decrease: over steps 700–1500 the cycle mean changes by −3.0% and stays within 5% in all 48 runs, against −9.7% and none of 48 in the control that keeps training them, while $\cos^2_{\min}$ rises by the same amount as in the control (+0.12). Unfreezing them at step 500 does the opposite: the flat loss starts to fall (−10.8%, no run within 5%) while the alignment keeps rising (+0.18). Only 19 of the 48 runs frozen at step 500 pass the strict test, because by then the features are nearly aligned and $\overline{\cos^2}$ gains less than the test's threshold of 0.05 (+0.04). With $\eta_{\text{on}} = \eta_1$ the features are aligned before the switch ($\cos^2_{\min} \approx 0.95$). Freezing then removes the part of the loss oscillation driven by the deep layers' own steps, which lowers the loss by about 15% within 100 steps, after which the loss stays flat (40 of 48 runs within 5%) with no further alignment gain; unfreezing lowers the loss by about 20% within 100 steps while $\cos^2_{\min}$ jumps from about 0.45 to 0.85.

Partial freezing. Which deep layers train matters more than how many (Figure 16). Training only the deep layer next to the head, with the first layer's step, keeps the plateau: 47 of 48 runs pass the strict test, 45 stay within 5%, and $\cos^2_{\min}$ gains +0.51. Training only the deep layer next to W_1 removes it: the cycle mean falls by 21% and 1 of 48 runs stays within 5%, even though the alignment

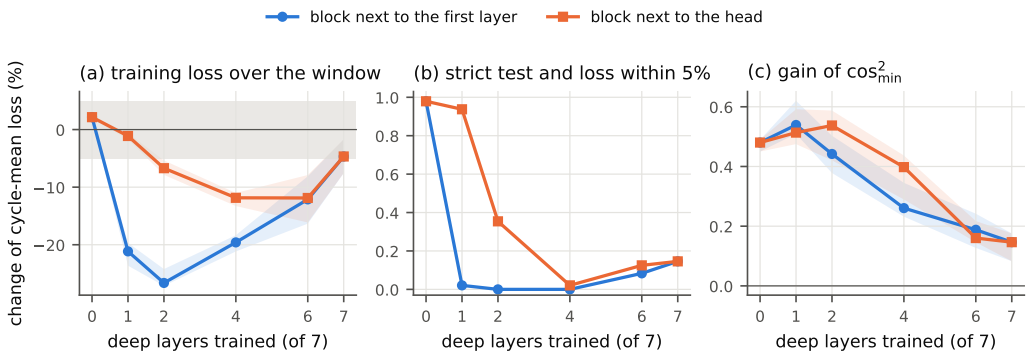


Figure 16: Which deep layers may train. Depth 8; the first layer and j deep layers train with $\eta_1 = 0.66$ and the others are frozen; the trained block sits next to W_1 (blue) or next to the head (orange). **(a)** Window change of the cycle-mean loss (grey band: $\pm 5\%$). **(b)** Fraction of runs passing the strict test with the loss within 5%. **(c)** Window gain of $\cos^2_{\min}$. Medians over problems; shading is the interquartile range.

still sits mostly in W_1 (90% remains when the deep layers are reset). Larger trained blocks remove the plateau from either end: for $j \geq 4$ at most 7 of 48 runs stay within 5%.

E.3 Mechanism

With the deep layers frozen the student is $x \mapsto g\big(\mathrm{ReLU}(W_1^\top x)\big)$ for a fixed map g (the residual stack at its initialisation followed by the frozen head), and only W_1 trains, with the large polar step. This is the one-hidden-layer setting of the main text with a fixed nonlinear readout in place of a fixed linear one, and it behaves the same way. The loss locks into the period-2 cycle, whose high branch is 83–92% offset (a shift of the mean prediction); the cycle mean stays flat; and the polar step keeps rotating the dominant input directions of W_1 towards the teacher subspace. When the deep layers also train, they keep fitting the targets with whatever features W_1 currently provides, so the loss decreases, and they come to carry part of the alignment (Figure 14d). The partial-freezing arms place the effect: the layer directly after W_1 is enough to turn the plateau into a steady decrease, even while the alignment stays in W_1, and the layer next to the head is not. The plateau with continued feature learning is therefore a property of the first layer trained against a fixed, or slowly moving, downstream map; freezing the downstream layers isolates it, and training them, starting with the layer right after W_1, hides it.

Scope. All of this is for residual MLP students of a Gaussian teacher, trained full batch with the exact polar step and a frozen head. At depth 6 the flattest loss needs slow but non-zero deep steps rather than frozen ones. The plain network does not show the effect at the first-layer step of the plain arm, even with its deep layers frozen; we did not tune that step for it. We have not established the effect for other architectures or data.

F A freezing principle for the plateau with feature learning

Appendix E showed that, in deep residual students, freezing the layers after W_1 switches the plateau with continued feature learning on and training them switches it off. Here we state the observation as a working principle and test it in three more ways: at other depths and widths, after other training schemes, and under Muon as it is used in practice (Newton–Schulz and momentum).

Notation. For the deep-network experiments, W_1 is the first hidden matrix, L is depth, m is width, η_1 is the coefficient of the first-layer polar update, and T_0 is the number of pretraining steps. The fixed downstream map g includes all later layers and the head. The teacher basis U and the leading input-AGOP basis B have orthonormal columns; the reported alignment metrics are $\overline{\cos^2} = r_t^{-1}\|U^\top B\|_F^2$ and $\cos^2_{\min} = \sigma_{\min}(U^\top B)^2$. Cycle means, scoring windows and held-out head refits are as defined

in Appendix D.2. The small-head experiment in Section F.3 uses the separate mean-field notation stated there.

Working principle. Write the student as $x \mapsto g\big(\mathrm{ReLU}(W_1^\top x)\big)$, where g collects every layer after the first and the frozen head.

(i) *Isolation.* With g held fixed, normalised (polar) steps on W_1 keep improving the input-AGOP alignment with the teacher subspace, while the training loss stays on the floor set by the W_1 step: the period-2 cycle of Appendix D, whose cycle mean stays flat.

(ii) *Conversion.* Letting g train, with a step of about $\eta_1/50$ or more, turns that feature progress into loss progress: the loss falls, and g comes to carry part of the alignment. The layer directly after W_1 matters most (Figure 16).

The plateau is thus the loss floor of the normalised W_1 step, which better features do not lower while g is fixed; training g, or training W_1 alone by gradient descent or Adam, lowers the loss. The plateau is not needed for the feature learning itself (Section F.2).

F.1 Depth and width

The depth-8 and depth-16 ladders of Appendix E extend to depths 4 and 24 (24 runs per arm, first-layer steps 0.66 and 0.74). With the deep layers frozen, 24 of 24 and 23 of 24 runs pass the strict test with the cycle mean within 5% (+2% in both). Over the window $\cos^2_{\min}$ rises from 0.18 to 0.71 at depth 4 and from 0.22 to 0.60 at depth 24; with the deep layers frozen, this change is due to W_1 alone. With the deep layers at $\eta_1/50$ the alignment still rises ($0.26 \to 0.89$ and $0.26 \to 0.88$), but the cycle mean falls by 7% and 10%, 2 and 0 of 24 runs stay within 5%, and the held-out refit loss falls by 23% and 26% against 6% and 4% with the deep layers frozen. Depth 6 remains the exception noted in Appendix E, with the flattest loss at small non-zero deep steps.

Width is less forgiving at the depth-8 first-layer step, which we did not retune. At width 64 and $\eta_1 = 0.66$ the frozen student still learns ($\cos^2_{\min}$ $0.04 \to 0.87$, 24 of 24 strict), but the cycle mean falls by 10%, so no run stays within 5%. At width 64 with $\eta_1 = 1$, and at width 128 with $\eta_1 = 0.66$, the weakest teacher direction is never learned ($\cos^2_{\min} \le 0.01$). The plain (non-residual) network fails at every first-layer step from 0.05 to 1, with the cycle mean falling by 27–41%.

F.2 Training only W_1 after other schemes

We train the whole depth-8 student with a scheme S for T_0 steps and then train only W_1, with the deep layers frozen, for 1000 steps (Figure 17). S is GD or Adam at a fast or a slow rate (tuned on two development problems), or Muon with one step for every layer or with the two rates of Appendix D; $T_0 \in \{100, 300, 1000\}$.

The features. After 100 or 300 steps of slow GD or Adam, $\cos^2_{\min}$ is 0.00–0.21 at the switch. Two hundred steps later, 15 of the 16 combinations of these four starts with the four W_1 optimisers (polar 0.66, polar 0.2, GD, Adam) have reached 0.93–1.00; the exception is polar 0.66 after 100 steps of Adam (0.45, and 0.73 at the end of the window). Continuing S on the whole network reaches only 0.02–0.71 in the same time, and it takes about 700–1000 steps to catch up. Training W_1 against a fixed downstream map is therefore a fast feature learner in its own right.

The loss. The optimiser on W_1 decides what the loss does. A polar step of 0.66 moves the loss to the floor set by the step and holds it there: after slow GD or Adam the cycle mean first rises by a factor 1.3–2.3 (after 100 steps of Adam it falls by about 20% instead) and then changes by +1% to +11% over the window; after fast GD or Adam, when the loss is already small, it rises by a factor 7 to 2,000. GD and Adam on W_1 keep lowering the loss (21–30% over the window after 100 or 300 steps of slow GD or Adam; W_1-only Adam after 1000 steps lowers it by 10–11%).

When the plateau carries learning. A long plateau during which the features keep improving appears only when the deep layers are still close to their initialisation: after Muon at the two rates of Appendix D for $T_0 = 100$ or 300 (24 and 23 of 24 runs strict and within 5%, $\cos^2_{\min}$ $0.30 \to 0.66$

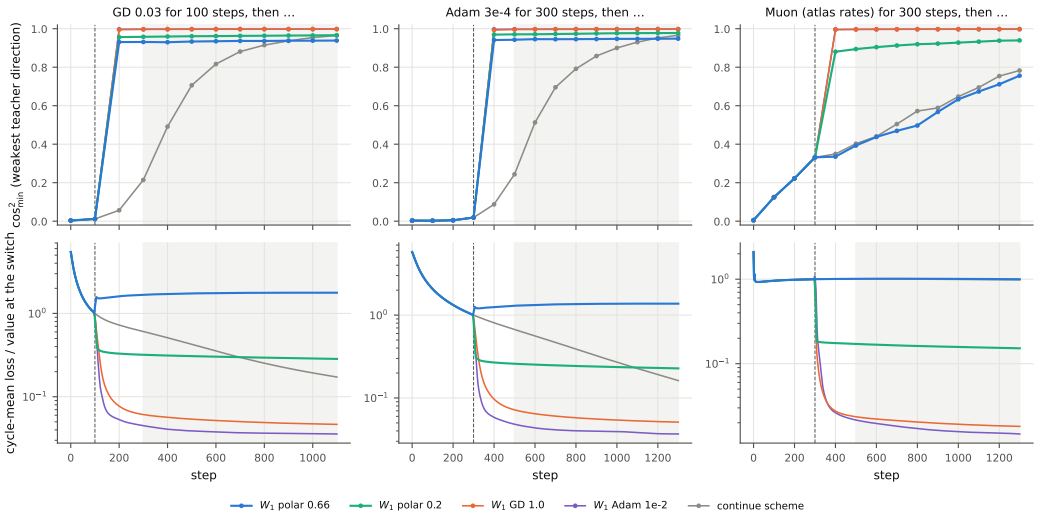


Figure 17: Training only W_1 after partial pretraining. The depth-8 residual student is trained with GD (0.03, 100 steps), Adam (3×10^{-4}, 300 steps) or Muon at the rates of Appendix D (300 steps). At the switch (dashed) the deep layers are frozen and only W_1 trains, with a polar step of 0.66 or 0.2, GD or Adam, or the scheme continues on every layer. **Top** $\cos^2_{\min}$. **Bottom** cycle-mean training loss divided by its value at the switch. Medians over problems of the per-problem median over initialisations; grey: the scored window.

and $0.39 \to 0.76$ over the window), or with the deep layers frozen from initialisation (23 of 24, $0.17 \to 0.60$). After 100 steps of GD or Adam the learning is finished before the scored window begins; the loss is then flat, but there is little left to learn. The one diverging branch is W_1-only GD after 1000 steps of Muon with one step for every layer (12 of 24 runs).

F.3 Newton–Schulz and momentum: two regimes

The convergent comparison uses $X_0 = G/(\|G\|_F + 10^{-7})$ and eight iterations of the quintic map

$$X_{j+1} = 2X_j - \frac{3}{2}(X_j X_j^\top)X_j + \frac{1}{2}(X_j X_j^\top)^2 X_j. \tag{79}$$

Tall matrices are transposed before iteration, and the result is transposed back. This specifies the coefficients $(2, -1.5, 0.5)$ of the convergent comparison, distinct from Muon's quintic approximation.

Whether the plateau survives Muon's approximations depends on what holds the loss on its floor.

Small-head notation. For this experiment only, d is the input dimension, r the teacher rank, $D = d - r$, and m the number of hidden units. The balanced ReLU teacher and student are those of Appendix G.1.1: the fixed head coefficient is α/m, the intercept is refitted, and $\kappa = \sigma_0/(\bar{\eta}\sqrt{d})$ for initial columns $N(0, \sigma_0^2 I_d/d)$ and polar step $\eta = \bar{\eta}\sqrt{m}$. Here α denotes head scale, not the residual-branch scale of the deep network. For orthonormal teacher directions $u_1, \dots, u_r$, let $\mathcal{U} = \mathrm{span}(u_1, \dots, u_r)$ be their subspace, with projector P_U. If P_r projects onto the leading r AGOP eigenvectors, $A_r = r^{-1}\,\mathrm{tr}(P_U P_r)$. Write $\bar{u} = r^{-1/2}\sum_{i=1}^r u_i$; λ_c and $\lambda_\perp$ are the mean-field AGOP eigenvalues on the contrast space $\mathcal{U} \cap \bar{u}^\perp$ and the orthogonal complement $\mathcal{U}^\perp$, respectively (the common head factor cancels in their ratio). Neuron teacher energy is $\|P_U w_j\|_2^2/\|w_j\|_2^2$; its median is over neurons. In the momentum experiments, β is the exponential-moving-average coefficient and "Nesterov" specifies the look-ahead buffer combination.

Small fixed head: the plateau survives. In the one-hidden-layer student with a small fixed head (mean-field scale, a balanced two-direction ReLU teacher, population loss; $d \in \{64, 402\}$, 3 seeds each), the loss plateau is produced by the head scale, and the orthogonaliser mainly rescales how fast each direction moves. Muon's Newton–Schulz quintic keeps the plateau with feature learning at the

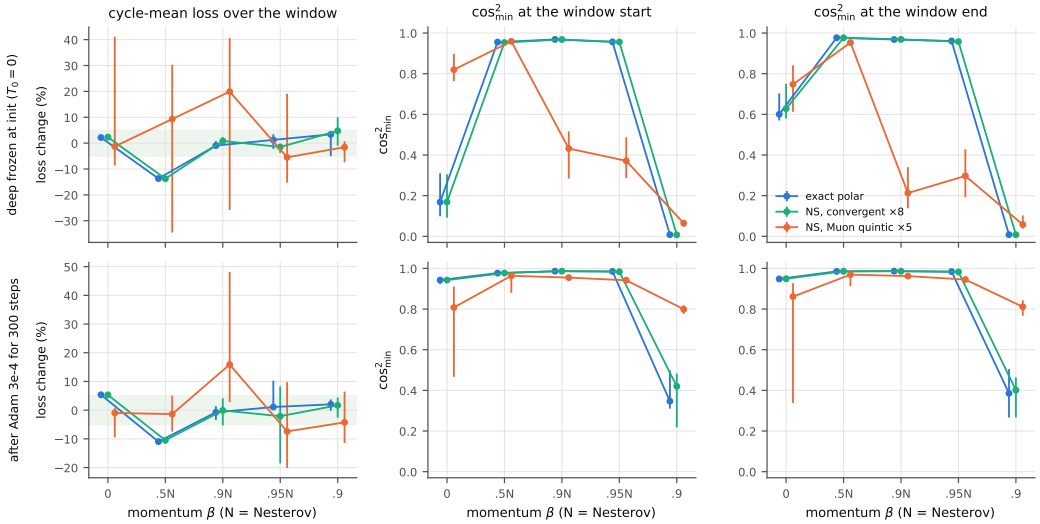


Figure 18: Muon as used in practice, depth 8. Only W_1 trains (step 0.66, deep layers frozen), from initialisation (top) or after 300 steps of Adam (bottom). Colours: the exact polar factor, a convergent quintic iteration (8 steps; Eq. (79)) and Muon's quintic (5 steps); x-axis: momentum (EMA buffer as in Muon; N: Nesterov). **Left** cycle-mean loss change over the window. **Middle, right** $\cos^2_{\min}$ at the start and at the end of the window. Medians over problems; bars: interquartile range.

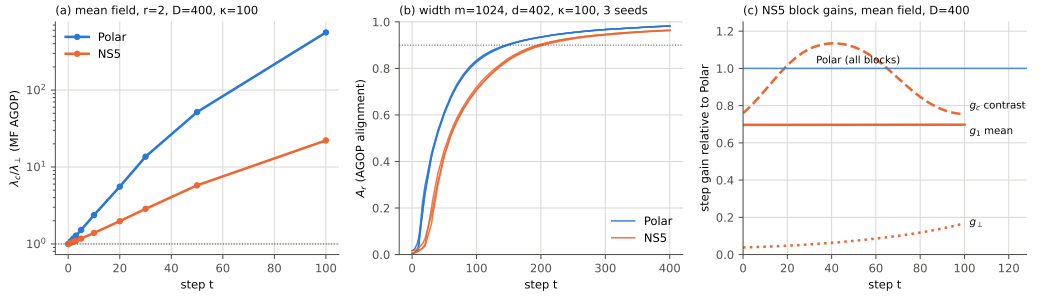


Figure 19: Newton–Schulz (NS5) against exact Polar in the small-head setting ($r = 2$, $\kappa = 100$). **(a)** Mean-field AGOP ratio $\lambda_c/\lambda_\perp$ at $D = 400$ (computed): both optimisers put the teacher subspace on top from the first step, and the gap grows more slowly under NS5. **(b)** AGOP alignment A_r at $d = 402$, $m = 1024$, $\alpha = 10^{-3}$, 3 seeds. **(c)** NS5 step gains on the mean, contrast and $\perp$ blocks relative to Polar, which has gain 1 on every block.

level of the AGOP and the teacher subspace: the AGOP aligns with the teacher subspace ($A_r \geq 0.9$) within 120–280 steps for the exact polar factor, for the quintic with 2 or 5 steps and for 7 steps of a convergent quintic. The speed of the loss is set by the iteration's gain on the mean direction (0.70 for an odd and 1.1 for an even number of Muon steps), and the specialisation of individual neurons by its gain on the remaining directions, so that more steps are needed at larger width. In the population, momentum leaves the alignment and the plateau unchanged and delays specialisation slightly; with minibatches it shortens the plateau and speeds up specialisation. Figures 19, 20 and 21 show these results.

Large polar step: the plateau needs an accurate orthogonaliser. In the deep student the floor is the period-2 loss oscillation of the large polar step on W_1, and anything that changes that step changes the oscillation (Figure 18). From frozen deep layers, a convergent Newton–Schulz iteration, which sits within 2×10^{-9} of the exact polar factor, reproduces it: 23 of 24 runs strict and within 5% for both, $\cos^2_{\min}$ $0.17 \to 0.60$ (exact) and $0.17 \to 0.63$ (Newton–Schulz). Muon's quintic stays

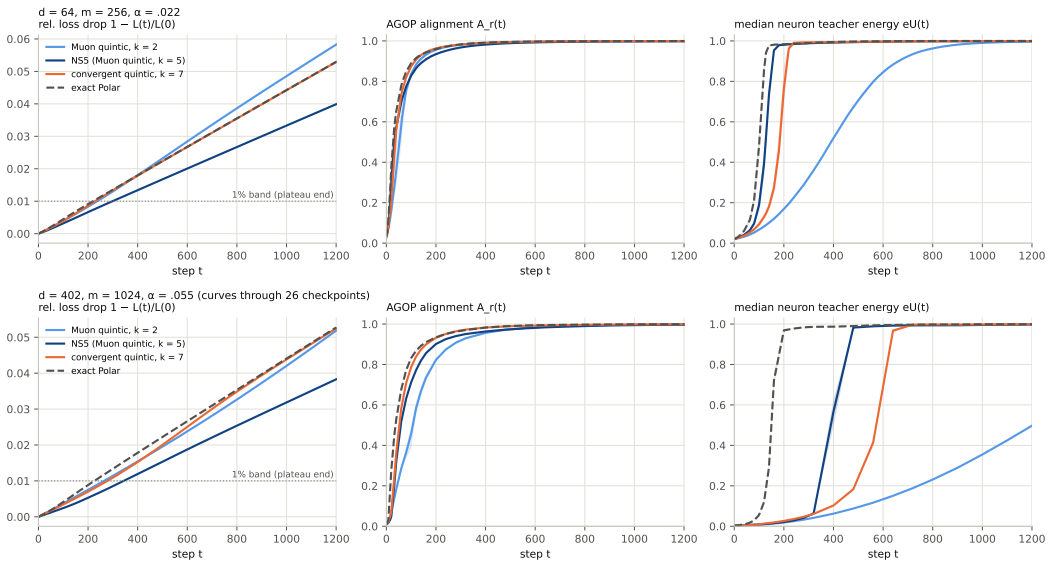


Figure 20: Training trajectories for four orthogonalisers (small-head setting, population loss, no momentum; $d = 64$ and $d = 402$): Muon's quintic with 2 and 5 steps, 7 steps of a convergent quintic, and exact Polar. Shown are $1 - L(t)/L(0)$, the AGOP alignment $A_r(t)$ and the median neuron teacher energy; lines are means over 3 seeds and bands the seed range. The AGOP aligns early ($A_r \geq 0.9$ by step 120–280), while the neurons specialise at very different times.

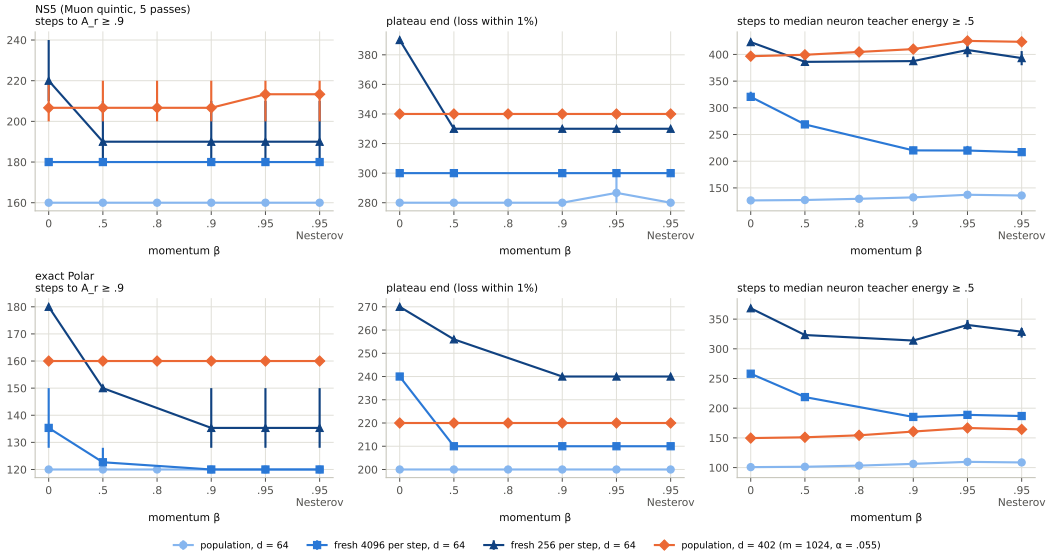


Figure 21: Momentum for NS5 and exact Polar (small-head setting; 3 seeds, bars show the seed range). Rows: NS5 and Polar. Columns: time to $A_r \geq 0.9$, end of the loss plateau, and time to a median neuron teacher energy of 0.5, against the momentum $\beta \in \{0, 0.5, 0.8, 0.9, 0.95\}$ and 0.95 with Nesterov; population loss at $d = 64$ and $d = 402$, and fresh minibatches of 4096 or 256 samples at $d = 64$. In the population, momentum leaves the alignment and the plateau unchanged and delays specialisation slightly; with minibatch noise it shortens the plateau and speeds up specialisation.

0.18–0.21 from the polar factor, breaks the clean cycle, and no run passes. Nesterov momentum (0.5–0.95) learns the features before the plateau forms: $\cos^2_{\min}$ is already 0.95–0.97 at the start of the window, after which the loss either keeps falling (-14% at 0.5) or stays flat with nothing left to learn (17–20 of 24 runs flat at 0.9 and 0.95). Plain EMA momentum at 0.9 never aligns the

weakest direction (0.01), and Muon's own update, the quintic with Nesterov momentum 0.95, stalls ($0.37 \to 0.30$).

The two regimes agree with the principle: in both, the features are learned by the orthogonalised first-layer step against a fixed readout. They differ in what keeps the loss flat, and that determines how much approximation the plateau tolerates.

F.4 Other architectures and scope

We also ran a transformer on sequences of Gaussian tokens with a planted rank-4 teacher, and two vision transformers and a CNN on the 8×8 optical-digit images of Alpaydin & Kaynak (1998) (Table 10). For the transformer, alignment is measured against the planted rank-4 teacher subspace; for the digit-image models, against a fixed rank-8 input-AGOP reference subspace learned beforehand. For those digit models, the principal-angle definitions above use the rank-8 reference basis in place of U and $r_t = 8$. After removing LayerNorm (LN) and, when $T_0 > 0$, partial AdamW pretraining, only the first layer is trained with exact polar steps; one setting per model was chosen on two development seeds and run on 10 fresh initialisations. The loss sits on a clean period-2 cycle with a flat mean in 8–10 of 10 runs for the transformer, the 4-patch vision transformer and the CNN, and the centred error is flat as well. Only the transformer keeps learning on that cycle, and only weakly (mean $\cos^2$ $+0.009$ over the window, 2 of 10 runs pass the joint criterion defined below); Muon's quintic on the first layer removes the clean cycle in all three models and Nesterov momentum removes it in the transformer, and training every matrix removes the plateau. We therefore claim the principle for residual MLP students of a Gaussian teacher (depths 4 to 24), for the one-hidden-layer small-head setting, and in weak form for the transformer, and not for plain MLPs, for wider MLPs at an untuned first-layer step, or for the vision transformers and the CNN.

Table 10: Settings of the architecture runs. Every run minimises the MSE through a frozen head with orthonormal rows, full batch. Stage 1 trains all non-head parameters with AdamW (learning rate 10^{-3}, zero weight decay) for T_0 steps and is omitted when $T_0 = 0$. Stage 2 trains only the first-layer (input-embedding) weight by $W \leftarrow W - \eta\,\mathrm{Polar}(G)$, where $G = \nabla_W L_{\mathrm{train}}$, for 1600 steps, scored on steps 200–1600. The update coefficient is $\eta = \delta\|W_{\mathrm{start}}\|_F/\sqrt{M}$, where W_{start} is the stage-2 initial weight and $M = \min(n_{\mathrm{rows}}, n_{\mathrm{cols}})$ after flattening any convolutional kernel to a matrix: $M = 32, 16, 64, 9$ for the four rows below, respectively. Transformer data: sequences of 8 Gaussian tokens in $\mathbb{R}^{32}$; the teacher applies one attention layer and a ReLU layer of width 16 to the tokens' projections onto a rank-4 subspace, averages over tokens and outputs 4 standardised targets; 1024 training sequences, and ten fresh planted problems in the confirmation runs. Digit data: 500 training and 500 test images, one-hot targets.

model	data	architecture	LN removed	T_0	δ
transformer	Gaussian tokens	1 block, width 64, 4 heads	output	0	1.0
ViT, four 4×4 patches	digits	1 block, width 64, 4 heads	all	500	0.03
ViT, one 8×8 patch	digits	1 block, width 64, 4 heads	all	500	1.0
CNN	digits	3 conv. layers, channels 16/32/64	output	200	0.1

Architecture-specific success criterion. These runs use a separate joint test from the strict residual-network test of Appendix D.2. Time is measured from the start of stage 2, and the scoring window is $s = 200$ to $e = 1600$, inclusive. For a model with K outputs and training residuals $r_i(t) = f_t(x_i) - y_i \in \mathbb{R}^K$, let $L_t = (nK)^{-1}\sum_i \|r_i(t)\|_2^2$ and $E_t = L_t - K^{-1}\|n^{-1}\sum_i r_i(t)\|_2^2$ be the training MSE and centred training error. Define $d_t = L_{t+1} - L_t$ and the adjacent-pair means $C_t = (L_t + L_{t+1})/2$ and $H_t = (E_t + E_{t+1})/2$, for $s \le t < e$. The loss-cycle signature requires

$$\rho_2 = -\operatorname{Corr}_{t=s}^{e-2}(d_t, d_{t+1}) \ge 0.99, \qquad R_2 = \operatorname{median}_{t=s}^{e-2} \frac{|L_{t+2} - L_t|}{\max\{|d_t|, \epsilon\}} \le 0.01, \tag{80}$$

and a normalized phase gap

$$\frac{\operatorname{median}_{s\le t<e}|d_t|}{\max\{\operatorname{median}_{s\le t<e} C_t, 10^{-30}\}} \ge 0.01, \qquad \epsilon = 10^{-15}\max\{\operatorname{median}_{s\le t<e}|d_t|, 1\}. \tag{81}$$

We set $\rho_2 = 0$ if either increment sequence has zero variance. Flat cycle-mean loss and flat centred error mean, respectively,

$$\left|\frac{C_{e-1} - C_s}{\max\{C_s, 10^{-30}\}}\right| \le 0.05, \qquad \left|\frac{H_{e-1} - H_s}{\max\{H_s, 10^{-30}\}}\right| \le 0.15. \tag{82}$$

These are endpoint changes over the complete window, with no division by its duration. Let B_t have orthonormal columns spanning the leading rank-r_t input-AGOP eigenspace, and let U be the orthonormal planted teacher basis for the transformer ($r_t = 4$) or the fixed reference basis for the digit models ($r_t = 8$). The alignment measures are $a_t = \|U^\top B_t\|_F^2 / r_t$ and $b_t = \sigma_{\min}(U^\top B_t)^2$. For either $q = a$ or $q = b$, define the even and odd gains by $g_q^{\text{even}} = q_e - q_s$ and $g_q^{\text{odd}} = q_{e-1} - q_{s+1}$, with $g_q = (g_q^{\text{even}} + g_q^{\text{odd}})/2$. A run passes all criteria if it has the loss-cycle signature, both flatness bounds, $g_a \ge 0.05$, $g_b > 0$, and $g_a^{\text{even}}, g_a^{\text{odd}}, g_b^{\text{even}}, g_b^{\text{odd}} > 0$.

Seeds and fixed reference. The development initialisation seeds are $1, 2$ for the transformer and $26090971, 5807101$ for the digit models; confirmation uses 9350101–9350110, paired with transformer problem seeds 12–21. Digit splits use seeds 26090961, 26090962 and 26090963 for the test split, the 128-image diagnostic split and the training subset, respectively. The fixed digit reference is stored as a 64×8 basis in `ref_basis_v4.b64`; the decoded basis has SHA-256 prefix `cefa0672cd516a63` and is reused unchanged across runs.

G THE MEAN-FIELD THEOREM: COMPLETE PROOF

Notation. This appendix is self-contained and uses the notation of the proof: the input dimension is d (p in the main text), the width is m (r_s), the teacher subspace is U (V in the main text), and V denotes the target variance $\mathrm{Var}(Y)$ (σ_Y^2 in Section 5.4). Theorem 5 is Theorem 16 below. For a vector z, $|z|$ is its Euclidean norm; for a deterministic matrix A, $\|A\|$ is its operator norm. Expectations are written $\mathbb{E}[\,\cdot\,]$: a subscript X denotes the Gaussian input, and a subscript ν or $w \sim \nu$ denotes the particle law; omitted subscripts use the locally specified law. Repeated particles $w, w' \sim \nu$ are independent unless stated otherwise. For a random scalar or vector Z, $\|Z\|_{L^2(\nu)} = (\mathbb{E}_\nu[|Z|^2])^{1/2}$; the reduced dynamics abbreviates this to $\|Z\|$. We write $\mathbf{1}\{B\}$ for the indicator of an event B, $a_+ := \max(a, 0)$ and $a_- := \max(-a, 0)$. Both $\log$ and $\ln$ denote the natural logarithm. The Euler gamma function is $\Gamma(a) := \int_0^\infty t^{a-1}e^{-t}\,dt$ for $a > 0$.

Computer-assisted step. Region N uses parameter choices obtained by computational search and recorded here as exact rational numbers. These choices are verified against the analytically derived certificate inequalities using outward-rounded interval arithmetic (Section G.6.3). The exact inputs and enclosure formulas below specify a finite calculation that can be checked independently of the search procedure.

G.1 MAIN RESULT

We consider a two-layer ReLU network trained by the literal Polar (Muon) update from an ordinary Gaussian initialization, in the mean-field limit, on a balanced ReLU teacher with $r \geq 2$ orthonormal directions. On a single trajectory the loss stays flat to within 1% over a window of $\Theta(\kappa)$ updates. Inside that window, the top-r eigenspace of the network's average gradient outer product (AGOP) becomes exactly the teacher subspace, whereas initially the AGOP is isotropic. Later, the loss drops by a fixed fraction of the target variance. Section G.1 states the result (Theorem 16); Sections G.2–G.7 contain the complete proof.

G.1.1 SETTING

Data and teacher. $X \sim N(0, I_d)$. Let $u_1, \ldots, u_r \in \mathbb{R}^d$ be orthonormal ($r \geq 2$), $U := \mathrm{span}\{u_i\}$, P_U its orthogonal projector and $D := d - r$. Without loss of generality $u_i = e_i$. The teacher is the balanced ReLU sum

$$Y = \gamma \sum_{i=1}^{r} \mathrm{ReLU}(u_i^\top X), \qquad \gamma := \sqrt{\frac{2\pi}{r(r + \pi - 1)}}, \tag{83}$$

normalized so that $\mathbb{E}[Y^2] = 1$. Its variance is

$$V := \mathrm{Var}\, Y = \gamma^2 r \mathsf{k}_1 = (\pi - 1)/(r + \pi - 1), \tag{84}$$

where $\mathsf{k}_1 := (\pi - 1)/(2\pi)$.

We write $\bar{e} := r^{-1/2} \sum_i u_i$ for the mean teacher direction, $P_1 := \bar{e}\bar{e}^\top$, $P_c := P_U - P_1$ (the *contrast* projector) and $c_0 := \gamma/(2\pi)$. For unit vectors a, b with $a^\top b = \rho$,

$$\mathrm{Cov}\left(\mathrm{ReLU}(a^\top X), \mathrm{ReLU}(b^\top X)\right) = \mathsf{k}(\rho) := \frac{\sqrt{1 - \rho^2} + \rho(\pi - \arccos\rho) - 1}{2\pi}, \tag{85}$$
$$\mathsf{k}(1) = \mathsf{k}_1.$$

The kernel in Eq. (85) will express the population loss in terms of particle directions.

Student and loss. For $W = [w_1, \ldots, w_m] \in \mathbb{R}^{d \times m}$ the student is

$$f_W(x) = b + \frac{\alpha}{m} \sum_{j=1}^{m} \mathrm{ReLU}(w_j^\top x),$$

with a fixed head α/m ($\alpha > 0$) and no hidden biases. The output intercept is refitted at every step:

$$b = b(W) := \mathbb{E}[Y - (\alpha/m) \sum_j \mathrm{ReLU}(w_j^\top X)].$$

Thus the residual $R := Y - f_W(X)$ is centred. The population loss is $L(W) := \mathbb{E}[(Y - f_W(X))^2]$, and $\hat{w} := w/|w|$.

Force and the Polar step. The force is

$$F(W) := \mathbb{E}\big[R\,X\,\mathbf{1}\{W^\top X > 0\}^\top\big] \in \mathbb{R}^{d\times m}, \qquad F_j = \mathbb{E}\big[R\,X\,\mathbf{1}\{w_j^\top X > 0\}\big], \tag{86}$$

so that $\nabla_W L = -(2\alpha/m)F$. For a compact singular value decomposition $F = A\Sigma B^\top$, $\mathrm{Polar}(F) := AB^\top$, with $\mathrm{Polar}(0) := 0$. The hidden weights move by

$$W(t+1) = W(t) + \eta\ \mathrm{Polar}\big(F(W(t))\big), \qquad t = 0, 1, 2, \dots \tag{87}$$

For a unit vector u and a unit vector $\hat{w}$ at angle θ, the two Gaussian identities are

$$\begin{aligned} \mathbb{E}[\mathrm{ReLU}(u^\top X)X\mathbf{1}\{\hat{w}^\top X > 0\}] &= \frac{\pi-\theta}{2\pi}u + \frac{\sin\theta}{2\pi}\hat{w}, \\ \mathbb{E}[X\mathbf{1}\{\hat{w}^\top X > 0\}] &= \hat{w}/\sqrt{2\pi}. \end{aligned} \tag{88}$$

With $\rho_i := u_i^\top \hat{w}$ and $\theta_i := \arccos\rho_i$, centring the teacher in Eq. (88) gives its force on a unit with direction $\hat{w}$:

$$\begin{aligned} g(\hat{w}) &= c_0 \sum_{i=1}^{r} \Big[(\pi - \theta_i)\,u_i - (1 - \sin\theta_i)\,\hat{w}\Big] \\ &= c_0 \sum_{i=1}^{r} \Big[\big(\tfrac{\pi}{2} + \arcsin\rho_i\big)u_i - \big(1 - \sqrt{1-\rho_i^2}\big)\hat{w}\Big]. \end{aligned} \tag{89}$$

With θ_{jk} the angle between w_j and w_k, the full force is

$$F_j = g(\hat{w}_j) - \frac{\alpha}{2\pi m}\sum_{k=1}^{m}|w_k|\Big[(\pi - \theta_{jk})\hat{w}_k - (1 - \sin\theta_{jk})\hat{w}_j\Big]. \tag{90}$$

The second term is the *self-force*.

AGOP and alignment. The AGOP of the network is

$$M(W) := \mathbb{E}\big[\nabla_x f_W(X)\nabla_x f_W(X)^\top\big] = \Big(\frac{\alpha}{m}\Big)^2 WHW^\top, \qquad H_{jk} := \frac{\pi - \theta_{jk}}{2\pi}.$$

A *leading-r selector* is any orthogonal projector P_r onto an r-dimensional span of eigenvectors for the r largest eigenvalues of M. The alignment with the teacher subspace is $A_r(W) := \frac{1}{r}\,\mathrm{tr}(P_U P_r)$. Every statement about A_r below holds for every leading selector; this follows once we prove the strict gap $\lambda_r > \lambda_{r+1}$, which makes P_r unique.

Initialization and mean-field scaling. The columns $w_j(0)$ are IID $N(0, \sigma_0^2 I_d/d)$. We take $\eta = \bar{\eta}\sqrt{m}$ with $\bar{\eta}$ fixed and use the free scale parameter

$$\kappa := \frac{\sigma_0}{\bar{\eta}\sqrt{d}} = \frac{\sigma_0}{\eta}\sqrt{\frac{m}{d}}. \tag{91}$$

For example, $\sigma_0 = 1$, $\eta = .02$ and $m = 16d$ give $\kappa = 200$. Ordinary initialization means $\kappa \gg 1$.

Mean-field dynamics. A particle law ν on $\mathbb{R}^d$ acts on each of its particles through the force $F(w;\nu)$. This is Eq. (90) with the average over k replaced by $\mathbb{E}_{w'\sim\nu}[\cdot]$ (Section G.2.1). Put $\Sigma_\nu := \mathbb{E}_{w\sim\nu}[F(w;\nu)F(w;\nu)^\top]$. The mean-field (MF) Polar dynamics moves every particle by

$$w \mapsto w + \bar{\eta}\,\Sigma_{\nu_t}^{\dagger 1/2} F(w;\nu_t), \qquad \nu_0 = N(0, \sigma_0^2 I_d/d), \tag{92}$$

where $\Sigma^{\dagger 1/2} := (\Sigma^\dagger)^{1/2}$ is the square root of the Moore–Penrose inverse. When $\Sigma_\nu \succ 0$ it equals $\Sigma_\nu^{-1/2}$. The MF loss $L(\nu)$ and the MF AGOP $\alpha^2\mathcal{M}(\nu)$, with $\mathcal{M}(\nu) := \mathbb{E}_{w,w'\sim\nu}[H(w,w')\,ww'^\top]$, are defined in the same way.

At every width m the update Eq. (87) is exactly the map Eq. (92) applied to the empirical measure of the columns; the factors $\sqrt{m}$ cancel, and no condition such as $m \geq d$ is needed (Lemma 17). The width-m loss and AGOP are the MF ones evaluated at that measure.

On any horizon along which $\Sigma_{\nu_t} \succ 0$ (in the theorem below, $[0, t_F]$), the width-m dynamics converges to the MF one as $m \to \infty$ at fixed $\bar{\eta}$ (Proposition 29(b)). Theorem 16 is a statement about the MF dynamics itself.

G.1.2 MAIN THEOREM

Theorem 16 (Plateau, then feature learning, then loss drop)**.** *Let $r \geq 2$, $d \geq 64r$ and $D = d - r \geq 400$, and let $\tau_F \in (0, .101)$, $\kappa_{\min} < \infty$ and $\delta > 0$ be the numbers of Theorem 52; they depend only on (r, D) and are described below. Let $\kappa \geq \kappa_{\min}$, $t_F := \lfloor \tau_F \kappa \rfloor$ (so $t_F \geq 2$), let $t_P \geq t_F$ be any integer, and put*

$$\alpha^*(t_P) := \frac{\bar{y}\,\sqrt{V}}{\bar{\eta}\,(\kappa + t_P)}, \qquad \bar{y} = .0084587115\ldots, \tag{93}$$

where $\bar{y}$ is the positive root of $1.17933\,\bar{y} + .34085\,\bar{y}^2 = .01$. There is a number $a_0 = a_0(r, d, \kappa) > 0$ with the following properties. Consider the MF trajectory Eq. (92) started from $\nu_0 = N(0, \sigma_0^2 I_d/d)$, $\sigma_0 = \kappa\bar{\eta}\sqrt{d}$, with head scale $\alpha > 0$.

(P) Plateau. *If $\alpha \leq \alpha^*(t_P)$, then $L(t) \leq 1.01\,L(s)$ and $L(t) \geq .99\,V$ for all $0 \leq s, t \leq t_P$.*

(F) Feature learning. *$\mathcal{M}(\nu_0)$ is a multiple of I_d. If $\alpha\bar{\eta} \leq a_0$, then $\Sigma_{\nu_t} \succ 0$ for $t \leq t_F$, and*

$$\begin{aligned} \mathcal{M}(\nu_{t_F}) &= \lambda_1 P_1 + \lambda_c P_c + \lambda_\perp (I_d - P_U), \\ \min(\lambda_1, \lambda_c) &\geq \left(1 + \tfrac{\delta}{2}\right) \lambda_\perp. \end{aligned} \tag{94}$$

By Eq. (94), at t_F the top-r eigenspace of the MF AGOP is exactly U, $A_r = 1$ for every leading selector, and $\lambda_r \geq (1 + \delta/2)\,\lambda_{r+1}$.

(R) Later loss drop. *If $\alpha \leq \alpha^*(t_P)$, then $L(0) - L(t_2) \geq .03\,V$ at $t_2 := \lceil .05086\,\sqrt{V}/(\alpha\bar{\eta}) \rceil$, that is, once the clock $\alpha\bar{\eta}t/\sqrt{V}$ has reached .05086. At $\alpha = \alpha^*(t_P)$ this gives $t_2 \leq 6.013\,(\kappa + t_P) + 1$; for instance $t_2 \leq 6.62\,\kappa + 1$ when $t_P \leq .1007\,\kappa$.*

In particular, with $\alpha^*(t_P)$ from Eq. (93), (P), (F) and (R) hold together for every $\alpha \in \left(0, \min\{\alpha^*(t_P), a_0/\bar{\eta}\}\right]$. The loss comparisons in (P) and (R) use the variance Eq. (84).

The constants. Two regions cover every (r, D) with $D \geq \max(400, 63r)$; the values are those of Theorem 52.

- *Region N*: $2 \leq r \leq 80$ and $D \leq D_A(r)$, with $D_A(r)$ the right end of the last D-interval of r in Table 12. The constants are read, row by row, from the 80 certificates of Table 12. They satisfy $\tau_F \in [.013740, .088861)$, $\kappa_{\min} \in [91, 777]$ and $1 + \delta \geq 1.006223$ (≥ 1.012711 for $r \geq 3$, ≥ 1.079513 for $r \geq 10$).
- *Region A*: all other D, which includes every $D \geq 63r$ when $r \geq 81$. The constants are in closed form (Lemma 64): $\tau_F \leq .100667\sqrt{r^2 - 1}/M$ and $\kappa_{\min} = 4/\tau_F$, and $1 + \delta$ is the smaller of two explicit bounds, $\lambda_c/\lambda_\perp \geq \exp(.0176\sqrt{r^2 - 1}/M)$ and a bound $\lambda_1/\lambda_\perp > 1$, at t_F for the $\alpha = 0$ dynamics. Here $M = M(r, D) \approx r + 2\sqrt{r \ln D} + 2 \ln D$.

For example, at $r = 2$, $400 \leq D \leq 14\,930$ and $\kappa = 1000$: $t_F = 13$; and with $t_P = t_F$ and $\alpha = \alpha^*(t_P)$, the loss has dropped by $.03V$ at $t_2 = 6\,091$.

G.1.3 REMARKS

(i) *The finite computational step.* (P) is proved in closed form, and (R) uses an explicit logarithm integral and finite rational endpoint bounds. In (F), region N is covered by 80 explicit inequality certificates verified with outward interval arithmetic (Section G.6.3). Every certificate input and enclosure rule is included in this proof.

(ii) *The head scale in (F).* (F) is first proved for the $\alpha = 0$ dynamics (Theorem 52) and then transferred to small $\alpha > 0$ by continuity of the MF map in the head scale (Proposition 29(a)); the threshold a_0 is not explicit.

(iii) *The window is long in steps, not in movement.* Each MF step moves the particles by at most $\bar{\eta}\sqrt{d}$ in RMS (Lemma 31(a)), so over $[0, t_F]$ the RMS displacement is at most $\bar{\eta}\sqrt{d}\,t_F \leq \tau_F\sigma_0 < .101\,\sigma_0$. The count t_F is large because the step is small compared with the initialization.

(iv) *The initial alignment.* The initial MF AGOP is isotropic, so its top-r space is not determined. A finite-width initialization comparison additionally requires a specified leading selector when eigenvalues tie; no selector-independent, all-width Haar assertion is used in this theorem.

(v) *Scope.* The data are population Gaussian; the teacher is a balanced sum of ReLUs on orthonormal directions; the head is fixed, common to all units and small; the intercept is refitted exactly. (F) is proved at the single time t_F. The certified margin $1+\delta$ is small, whereas the observed one is much larger. The width-m dynamics converges to the MF one on $[0, t_F+1]$ (Proposition 29(b)); no rate is claimed.

G.1.4 Structure of the proof

Section G.2 sets up the MF map and reduces (F) to two scalar inequalities; Section G.3 proves (P) and (R); Sections G.4–G.6 prove (F) for the head-free dynamics; Section G.7 assembles the proof and transfers (F) to small heads. Every section opens with a short statement of what it proves and where it is used, and Table 11 lists where each part of Theorem 16 is proved.

Table 11: Where the parts of Theorem 16 are proved.

part of Theorem 16	proved in	main ingredients
trajectory	Section G.2	Lemmas 17, 18, 20; Proposition 21
(P)	Theorem 42	Lemmas 31, 36, 39
(F) at $\alpha = 0$	Theorem 52	Propositions 24, 27, 58; Lemmas 49, 50, 51, 61, 62, 64; Proposition 65
(F) at small $\alpha > 0$	Section G.7, Step 3	Lemma 28; Proposition 29(a)
(F) at $t = 0$	Section G.7, Step 4	Proposition 27(4)
(R)	Corollary 45	Theorem 44; Lemmas 31, 39

G.2 The mean-field map

What this section proves. The width-m Polar update is exactly a map Φ on particle laws applied to the empirical measure of the columns (Lemma 17), and up to a dilation the trajectory depends on the head α and the step $\bar{\eta}$ only through $\alpha\bar{\eta}$ (Lemma 18). Along the trajectory the laws are absolutely continuous and invariant under $S_r \times O(D)$ (Sections G.2.2–G.2.3). At $\alpha = 0$ a particle reduces to a point $(\omega, \zeta) \in \mathbb{R}^r \times \mathbb{R}$ with explicit updates (Section G.2.4), and the MF AGOP has a three-block form whose top-r space is U as soon as two scalar inequalities hold (Proposition 27). Finally the trajectory is continuous in the head scale at $\alpha = 0$ and in the width (Section G.2.6).

Where it is used. Everywhere: Sections G.2.1–G.2.3 underlie Sections G.3–G.6; Section G.2.4 and Section G.2.5 are the language of the certificate; Section G.2.6 transfers (F) from $\alpha = 0$ to small $\alpha > 0$ in Section G.7.

G.2.1 Definitions and the exact identity

$\mathcal{P}_p(\mathbb{R}^d)$ is the set of Borel probability measures with finite p-th moment, and W_p is the p-Wasserstein distance ($W_1 \le W_2$). For $W = [w_1, \dots, w_m]$ put $\mu_W := m^{-1}\sum_j \delta_{w_j}$, where δ_w is the point mass at w. For a measurable map T, its pushforward is $(T_{\#}\mu)(B) := \mu(T^{-1}(B))$ for Borel sets B. For a unit vector $\hat{w}$ and $w' \in \mathbb{R}^d$ at angle $\theta = \theta(\hat{w}, w')$, let

$$\Upsilon(\hat{w}, w') := \frac{1}{2\pi}\Big[(\pi - \theta)\, w' - (1 - \sin\theta)|w'|\,\hat{w}\Big], \qquad \Upsilon(\hat{w}, 0) := 0.$$

For $\mu \in \mathcal{P}_1(\mathbb{R}^d)$ and $w \neq 0$ define

$$S(\hat{w}; \mu) := \int \Upsilon(\hat{w}, w')\, d\mu(w'), \qquad F(w; \mu) := g(\hat{w}) - \alpha S(\hat{w}; \mu), \qquad F(0; \mu) := 0,$$

$$\Sigma(\mu) := \int F(w;\mu)F(w;\mu)^\top d\mu(w),$$

$$T_\mu(w) := w + \bar\eta\,\Sigma(\mu)^{\dagger 1/2}F(w;\mu),$$

$$\Phi(\mu) := (T_\mu)_\# \mu,$$

where $A^{\dagger 1/2} := (A^\dagger)^{1/2}$ and $\dagger$ is the Moore–Penrose inverse. We also write Σ_μ for $\Sigma(\mu)$. The MF trajectory is $\nu_0 := N(0, \sigma_0^2 I_d/d)$ and $\nu_{t+1} := \Phi(\nu_t)$; this is Eq. (92). When the dependence on the head scale matters we write F_α, Σ_α, $T_{\mu,\alpha}$, Φ_α and $\nu_\alpha(t)$, and for the step size, $\Phi_{\alpha,\bar\eta}$. With $H(w,w') := (\pi - \theta(\hat w, \hat w'))/(2\pi)$ and the kernel k of Section G.1.1, the MF loss and the MF AGOP (up to the factor α^2) are

$$\mathcal{L}(\mu) := V - 2\alpha\gamma \int |w| \sum_{i=1}^r \mathsf{k}(u_i^\top \hat w)\, d\mu + \alpha^2 \iint |w||w'|\mathsf{k}(\cos\theta(\hat w, \hat w'))\, d\mu\, d\mu,$$

$$\mathcal{M}(\mu) := \iint H(w,w')\, ww'^\top d\mu\, d\mu,$$

every integrand being 0 when $w = 0$ or $w' = 0$. We write $L(\mu)$ for $\mathcal{L}(\mu)$ and $L(t)$ for $\mathcal{L}(\nu_t)$. Equivalently $F(w;\mu) = \mathbb{E}_X[R_\mu(X)X\mathbf{1}\{w^\top X > 0\}]$, where $R_\mu := Y - \mathbb{E}[Y] - (g_\mu - \mathbb{E}[g_\mu])$ is the centred residual of the MF network output $g_\mu(x) := \alpha \int \mathrm{ReLU}(w'^\top x)\, d\mu(w')$ (Lemma 17(a) below). Note that $F(w;\mu)$ depends on $w \neq 0$ only through $\hat w$.

Lemma 17 (Exact identity)**.** *For every $m \geq 1$, $\alpha \geq 0$ and $W \in \mathbb{R}^{d\times m}$ (zero columns allowed), with $\eta = \bar\eta\sqrt{m}$ and $W^+ := W + \eta\,\mathrm{Polar}(F(W))$:*

(a) column j of the force Eq. (86) is $F(w_j;\mu_W)$, the self-term $k = j$ included, so $F(W)F(W)^\top = m\Sigma(\mu_W)$;

(b) $\mathrm{Polar}(F) = (FF^\top)^{\dagger 1/2}F$ for every real matrix F;

(c) $w_j^+ = T_{\mu_W}(w_j)$ for every j, hence $\mu_{W^+} = \Phi(\mu_W)$;

(d) $L(W) = \mathcal{L}(\mu_W)$ and $M(W) = \alpha^2\mathcal{M}(\mu_W)$.

Consequently $\mu_{W(t)} = \Phi^t(\mu_{W(0)})$ along Eq. (87), for every m.

Proof. Part (a): Identify the force. If $w_j = 0$ the gate vanishes and $F_j = 0 = F(0;\mu_W)$. Let $w_j \neq 0$. With the refitted intercept,

$$R = (Y - \mathbb{E}[Y]) - \frac{\alpha}{m}\sum_k (\mathrm{ReLU}(w_k^\top X) - \mathbb{E}[\mathrm{ReLU}(w_k^\top X)]).$$

By the exact formulas of Section G.1.1, the teacher contribution is

$$\begin{aligned}&\mathbb{E}[YX\mathbf{1}\{\hat w^\top X > 0\}] - \mathbb{E}[Y]\,\mathbb{E}[X\mathbf{1}\{\hat w^\top X > 0\}]\\ &\quad = \gamma\sum_i \Big[\frac{\pi - \theta_i}{2\pi}u_i + \frac{\sin\theta_i}{2\pi}\hat w\Big] - \frac{\gamma r}{2\pi}\hat w = g(\hat w),\end{aligned}$$

using $\mathbb{E}[Y] = \gamma r/\sqrt{2\pi}$.

For $w' \neq 0$ at angle θ to $\hat w$, the same formulas and homogeneity give

$$\mathbb{E}[\mathrm{ReLU}(w'^\top X)X\mathbf{1}\{\hat w^\top X > 0\}] = \frac{\pi - \theta}{2\pi}w' + \frac{\sin\theta}{2\pi}|w'|\hat w,$$

$$\mathbb{E}[\mathrm{ReLU}(w'^\top X)]\,\mathbb{E}[X\mathbf{1}\{\hat w^\top X > 0\}] = \frac{|w'|}{2\pi}\hat w.$$

Their difference is $\Upsilon(\hat w, w')$, and a zero column contributes $0 = \Upsilon(\hat w, 0)$. Therefore

$$\begin{aligned}F_j &= g(\hat w_j) - \frac{\alpha}{m}\sum_k \Upsilon(\hat w_j, w_k) = F(w_j;\mu_W),\\ FF^\top &= \sum_j F_j F_j^\top = m\Sigma(\mu_W).\end{aligned} \tag{95}$$

The same computation with μ in place of μ_W gives $F(w;\mu) = \mathbb{E}[R_\mu X\mathbf{1}\{w^\top X > 0\}]$.

Part (b): Express the polar factor. For a compact SVD $F = A\Sigma_F B^\top$,

$$(FF^\top)^{\dagger 1/2}F = A\Sigma_F^{-1}A^\top A\Sigma_F B^\top = AB^\top. \tag{96}$$

Both sides vanish if $F = 0$.

Part (c): Cancel the width factors. By Eq. (95) and Eq. (96),

$$\mathrm{Polar}(F) = (m\Sigma(\mu_W))^{\dagger 1/2}F = m^{-1/2}\Sigma(\mu_W)^{\dagger 1/2}F.$$

Thus column j of $\eta\,\mathrm{Polar}(F)$ is $\bar\eta\Sigma(\mu_W)^{\dagger 1/2}F(w_j;\mu_W)$: the factors $\sqrt{m}$ cancel exactly, and $\mu_{W^+} = (T_{\mu_W})_\#\mu_W$.

Part (d): Identify the loss and AGOP. The refitted loss is $L(W) = \mathrm{Var}(Y - \frac{\alpha}{m}\sum_j \mathrm{ReLU}(w_j^\top X))$, with covariances

$$\mathrm{Cov}(\mathrm{ReLU}(u_i^\top X), \mathrm{ReLU}(w_j^\top X)) = |w_j|\mathsf{k}(u_i^\top \hat w_j),$$

$$\mathrm{Cov}(\mathrm{ReLU}(w_j^\top X), \mathrm{ReLU}(w_k^\top X)) = |w_j||w_k|\mathsf{k}(\cos\theta_{jk}).$$

This is $\mathcal{L}(\mu_W)$. For the AGOP,

$$\nabla_x f_W = \frac{\alpha}{m}\sum_j \mathbf{1}\{w_j^\top x > 0\}w_j, \qquad \mathbb{E}[\mathbf{1}\{w_j^\top X > 0\}\mathbf{1}\{w_k^\top X > 0\}] = H_{jk}.$$

Hence

$$M(W) = (\frac{\alpha}{m})^2\sum_{j,k} H_{jk}w_j w_k^\top = \alpha^2\mathcal{M}(\mu_W).$$

Since $H(w,w) = \frac{1}{2}$, the diagonal terms need no separate treatment. □

Lemma 18 (Scaling). *For $c > 0$ let $D_c(w) := cw$. For every $\mu \in \mathcal{P}_1$, $\alpha \geq 0$, $\bar\eta > 0$ and $w \neq 0$,*

$$F_\alpha\big(cw; (D_c)_\#\mu\big) = F_{c\alpha}(w;\mu),$$

$$\Sigma_\alpha\big((D_c)_\#\mu\big) = \Sigma_{c\alpha}(\mu),$$

$$\Phi_{\alpha,\bar\eta}\big((D_c)_\#\mu\big) = (D_c)_\#\Phi_{c\alpha,\bar\eta/c}(\mu),$$

and $\mathcal{M}((D_c)_\#\mu) = c^2\mathcal{M}(\mu)$. Consequently, with $\kappa = \sigma_0/(\bar\eta\sqrt{d})$, the MF trajectory with head α and step $\bar\eta$ from $N(0,\sigma_0^2 I_d/d)$ is the image under $D_{\bar\eta}$ of the trajectory with head $\alpha\bar\eta$ and step 1 from $N(0,\kappa^2 I_d)$. The two trajectories have the same whitening matrices Σ, and their MF AGOPs $\mathcal{M}$ differ by the factor $\bar\eta^2$, which changes no eigenspace and no eigenvalue ratio.

Proof. For $c > 0$, $\Upsilon(\hat w, cw') = c\,\Upsilon(\hat w, w')$ (the angle is unchanged), so $S(\hat w; (D_c)_\#\mu) = c\,S(\hat w;\mu)$; and cw has the direction $\hat w$. This gives the identity for F, and the one for Σ follows by the change of variables $w = cw''$.

For the particle map, the same identities give

$$T_{(D_c)_\#\mu}(cw) = cw + \bar\eta\,\Sigma_{c\alpha}(\mu)^{\dagger 1/2}F_{c\alpha}(w;\mu) = c\,T'(w), \tag{97}$$

where T' is the map of μ with head $c\alpha$ and step $\bar\eta/c$. Pushing μ forward in Eq. (97) gives the identity for Φ. Since $H(cw,cw') = H(w,w')$, $\mathcal{M}((D_c)_\#\mu) = c^2\mathcal{M}(\mu)$.

For the consequence take $c = \bar\eta$:

$$N(0,\sigma_0^2 I_d/d) = (D_{\bar\eta})_\# N(0,\kappa^2 I_d),$$

and induct on t. □

G.2.2 Regularity

Lemma 19 (Regularity of the force). *(a) g is continuous on S^{d-1} and C^∞ on $S^{d-1} \setminus \{\pm u_1,\dots,\pm u_r\}$.*

(b) For every unit $\hat w$ and all w_1', w_2': $|\Upsilon(\hat w, w_1') - \Upsilon(\hat w, w_2')| \leq 2\sqrt{d}|w_1' - w_2'|$ and $|\Upsilon(\hat w, w')| \leq |w'|$. Hence $\sup_{\hat w}|S(\hat w;\mu) - S(\hat w;\mu')| \leq 2\sqrt{d}W_1(\mu,\mu') \leq 2\sqrt{d}W_2(\mu,\mu')$.

(c) For $\mu \in \mathcal{P}_1$, $w \mapsto S(\hat w;\mu)$ is C^1 on $\mathbb{R}^d \setminus \{0\}$. Consequently $w \mapsto F(w;\mu)$ is continuous on $\mathbb{R}^d \setminus \{0\}$ and C^1 on the open set $\{w \neq 0 : \hat w \notin \{\pm u_i\}\}$, whose complement is null.

Proof. Part (a): Regularity of the teacher force. By Eq. (89), $g(\hat{w})$ is built from $\hat{w}$ and the functions $\arcsin \rho_i$ and $\sqrt{1-\rho_i^2}$ of $\rho_i = u_i^\top \hat{w}$; these are continuous on $[-1,1]$ and C^∞ on $(-1,1)$, and $|\rho_i| = 1$ iff $\hat{w} = \pm u_i$.

Part (b): Dependence on the particle law. By the proof of Lemma 17(a), $\Upsilon(\hat{w}, w') = \mathbb{E}[\phi_{w'}(X) X \mathbf{1}\{\hat{w}^\top X > 0\}]$ with $\phi_{w'} := \mathrm{ReLU}(w'^\top X) - \mathbb{E}[\mathrm{ReLU}(w'^\top X)]$ (also for $w' = 0$). Put $\psi := \mathrm{ReLU}(w_1'^\top X) - \mathrm{ReLU}(w_2'^\top X)$. As ReLU is 1-Lipschitz, $(\mathbb{E}[\psi^2])^{1/2} \le |w_1' - w_2'|$ and $|\mathbb{E}[\psi]| \le |w_1' - w_2'|$. With $\mathbf{1}\{\cdot\} := \mathbf{1}\{\hat{w}^\top X > 0\}$,

$$\Upsilon(\hat{w}, w_1') - \Upsilon(\hat{w}, w_2') = \mathbb{E}[\psi X \mathbf{1}\{\cdot\}] - \mathbb{E}[\psi]\,\mathbb{E}[X \mathbf{1}\{\cdot\}].$$

The two terms satisfy

$$|\mathbb{E}[\psi X \mathbf{1}\{\cdot\}]| \le (\mathbb{E}[\psi^2])^{1/2} (\mathbb{E}[|X|^2])^{1/2} \le \sqrt{d}\,|w_1' - w_2'|,$$

$$|\mathbb{E}[\psi]|\,|\mathbb{E}[X \mathbf{1}\{\cdot\}]| = |\mathbb{E}[\psi]|/\sqrt{2\pi} \le |w_1' - w_2'|.$$

Also $|\Upsilon(\hat{w}, w')| \le \frac{\pi+1}{2\pi}|w'| \le |w'|$ from the definition. For the last claim integrate against an optimal W_1 coupling.

Part (c): Differentiate the gated expectation. We first claim that if ϕ is continuous with $|\phi(x)| \le C(1+|x|)$, then $h(w) := \mathbb{E}[\phi(X) X \mathbf{1}\{w^\top X > 0\}]$ is C^1 on $\mathbb{R}^d \setminus \{0\}$, with

$$\begin{aligned} \partial_v h(w) &= |w|^{-1} \int_{\hat{w}^\perp} (v^\top y) \phi(y) y\, \gamma_d(y)\, d\sigma(y), \\ |\partial_v h(w)| &\le C_d C |v|/|w|. \end{aligned} \tag{98}$$

Here σ is Lebesgue measure on $\hat{w}^\perp$, and γ_d is the $N(0, I_d)$ density.

To prove the claim, fix a smooth nondecreasing ϑ with $\vartheta = 0$ on $(-\infty, 0]$ and $\vartheta = 1$ on $[1, \infty)$, and put $h_\epsilon(w) := \mathbb{E}[\phi(X) X \vartheta(w^\top X/\epsilon)]$. Differentiating under the expectation gives

$$\partial_v h_\epsilon(w) = \int_{\mathbb{R}} \epsilon^{-1} \vartheta'(|w| z/\epsilon) G_w(z)\, dz,$$

$$G_w(z) := \int_{\hat{w}^\perp} \phi(x) x (v^\top x) \gamma_d(x)\, d\sigma(y), \qquad x = z\hat{w} + y.$$

The function $G_w(z)$ is jointly continuous in (z, w) (dominated convergence), and the weights $\epsilon^{-1}\vartheta'(|w|z/\epsilon)dz$ have mass $1/|w|$ on $z \in [0, \epsilon/|w|]$. So $\partial_v h_\epsilon \to G_w(0)/|w|$ and $h_\epsilon \to h$ (the difference is at most $\mathbb{E}[|\phi(X)||X| \mathbf{1}\{0 < w^\top X < \epsilon\}]$), uniformly on compact subsets of $\mathbb{R}^d \setminus \{0\}$. Hence $h \in C^1$ with the derivative in Eq. (98).

Apply Eq. (98) with $\phi = \phi_{w'}$ ($C = 2|w'|$): $\partial_v \Upsilon(w/|w|, w')$ is continuous in $w \ne 0$ and bounded by $2C_d |w'||v|/|w|$, so for $\mu \in \mathcal{P}_1$ dominated convergence allows differentiation under $\int d\mu(w')$. With (a) this gives the statements about F; the excluded set is a finite union of rays. □

Lemma 20 (Absolute continuity along the MF trajectory)**.** *Let $\mu \ll \mathrm{Leb}$ and $T(w) = w + b(\hat{w})$ ($w \ne 0$), $T(0) = 0$, with b bounded and of class C^1 on an open $S' \subseteq S^{d-1}$ of full measure. Then $T_\# \mu \ll \mathrm{Leb}$. Consequently, for every $\alpha \ge 0$, every $\nu_\alpha(t)$ is absolutely continuous and lies in $\mathcal{P}_2$; in particular $\nu_\alpha(t)(\{0\}) = 0$.*

Proof. Step 1 (Locate the singular set). On the open set $\Omega := \{w \ne 0 : \hat{w} \in S'\}$ (complement null),

$$DT(w) = I + B(\hat{w})/|w|, \qquad B(\hat{w}) := Db(\hat{w})(I - \hat{w}\hat{w}^\top). \tag{99}$$

For fixed $\hat{w}$, $s \mapsto \det(I + sB(\hat{w}))$ is a polynomial equal to 1 at $s = 0$, so it has finitely many zeros; by polar coordinates and Fubini applied to Eq. (99), $Z := \{w \in \Omega : \det DT = 0\}$ is null (and closed in Ω).

Step 2 (Preserve null sets). On $\Omega \setminus Z$, T is a local C^1-diffeomorphism; cover it by countably many open sets U_n on which T is a diffeomorphism onto its image (Lindelöf). For a null set N, each $T^{-1}(N) \cap U_n$ is null. Hence

$$T^{-1}(N) \subseteq (\Omega^c \cup Z) \cup \bigcup_n (T^{-1}(N) \cap U_n)$$

is null and $T_\# \mu(N) = 0$.

Step 3 (Apply the argument along the trajectory). Induct on t from $\nu_0 \ll \mathrm{Leb}$: $T_{\nu_t}(w) = w + b(\hat{w})$ with $b = \bar{\eta}\Sigma_{\nu_t}^{\dagger 1/2} F(\cdot;\nu_t)$, which is bounded and C^1 on $S^{d-1}\setminus\{\pm u_i\}$ by Lemma 19. Second moments stay finite because b is bounded. No invertibility of Σ is needed. □

G.2.3 SYMMETRY

Coordinates. From now on $u_i = e_i$, so $w = (w_U, w_\perp) \in \mathbb{R}^r \times \mathbb{R}^D$, and $P_\perp := I_d - P_U$. We identify vectors of U with their coordinate vectors in $\mathbb{R}^r$; then $\bar{e} = \mathbf{1}/\sqrt{r}$ and $P_c = I_r - \bar{e}\bar{e}^\top$ on U. For a random variable Z (scalar or vector) under a particle law ν, $\|Z\| := (\mathbb{E}[|Z|^2])^{1/2}$ is its $L^2(\nu)$ norm.

The symmetry group. $\mathcal{G} \subset O(d)$ is the group generated by the permutation matrices of the r coordinates of U (acting as the identity on $U^\perp$) and by $O(D)$ acting on $U^\perp$ (as the identity on U). Thus $\mathcal{G} \cong S_r \times O(D)$. Every $Q \in \mathcal{G}$ commutes with P_1, P_c and $P_\perp$. The law of X is $\mathcal{G}$-invariant and the teacher is $\mathcal{G}$-invariant: $Y(QX) = Y(X)$.

Proposition 21 (Symmetry and block form of Σ_ν)**.** *Let $\alpha \geq 0$ and let $F(\cdot;\nu)$, Σ_ν and T_ν be as in Section G.2.1.*

(a) $g(Q\hat{w}) = Q\,g(\hat{w})$ for every $Q \in \mathcal{G}$ and every unit vector $\hat{w}$. If ν is $\mathcal{G}$-invariant, then $F(Qw;\nu) = QF(w;\nu)$ and $Q\Sigma_\nu Q^\top = \Sigma_\nu$ for every $Q \in \mathcal{G}$.

(b) Let ν be $\mathcal{G}$-invariant with $\nu(\{0\}) = 0$. Then, with $F = F(w;\nu)$ and $w \sim \nu$,

$$\Sigma_\nu = a_1 P_1 + a_c P_c + a_\perp P_\perp, \qquad a_1 := \mathbb{E}[(\bar{e}^\top F)^2], \quad a_c := \frac{\mathbb{E}[|P_c F|^2]}{r-1}, \quad a_\perp := \frac{\mathbb{E}[|P_\perp F|^2]}{D}.$$

(c) If ν is $\mathcal{G}$-invariant, then $T_\nu(Qw) = QT_\nu(w)$ for all $Q \in \mathcal{G}$ and w, and $(T_\nu)_\#\nu$ is $\mathcal{G}$-invariant. Consequently every law ν_t of the MF trajectory Eq. (92) is $\mathcal{G}$-invariant, for every $\alpha \geq 0$.

Proof. Part (a): Equivariance of the force. Write $g(\hat{w}) = c_0\big[\sum_i(\pi - \theta_i)u_i - s\,\hat{w}\big]$ with $\rho_i = u_i^\top \hat{w}$, $\theta_i = \arccos\rho_i$ and $s := \sum_i(1 - \sin\theta_i)$ Eq. (89). Let Q be a permutation of the coordinates of U, say $Qu_i = u_{\pi(i)}$. Then $u_{\pi(i)}^\top Q\hat{w} = u_i^\top \hat{w}$, so the angles are permuted, s is unchanged, and $g(Q\hat{w}) = c_0[\sum_i(\pi - \theta_i)u_{\pi(i)} - sQ\hat{w}] = Qg(\hat{w})$. Let $Q = I_r \oplus O$ with $O \in O(D)$. Then all ρ_i and s are unchanged, $Qu_i = u_i$, and again $g(Q\hat{w}) = Qg(\hat{w})$. These elements generate $\mathcal{G}$.

For the self-force, $\Upsilon(Q\hat{w}, Qw') = Q\Upsilon(\hat{w}, w')$ because Q preserves angles and norms, so by invariance of ν,

$$S(Q\hat{w};\nu) = \int \Upsilon(Q\hat{w}, Qw'')\,d\nu(w'') = QS(\hat{w};\nu).$$

Hence $F(Qw;\nu) = QF(w;\nu)$, and invariance of ν gives

$$Q\Sigma_\nu Q^\top = \mathbb{E}[F(Qw;\nu)F(Qw;\nu)^\top] = \Sigma_\nu. \tag{100}$$

Part (b): Identify the three blocks. By Eq. (100), Σ_ν commutes with $\mathcal{G}$.

- *Cross block.* $R := I_r \oplus (-I_D) \in \mathcal{G}$ gives $P_U\Sigma_\nu P_\perp = P_U R\Sigma_\nu R P_\perp = -P_U\Sigma_\nu P_\perp$, hence $P_U\Sigma_\nu P_\perp = 0$.
- *$\perp$ block.* Its restriction $S_\perp$ to $U^\perp$ commutes with every reflection $I_D - 2ee^\top$ ($|e| = 1$), hence with $ee^\top$. So every unit vector e is an eigenvector of $S_\perp$, and $S_\perp = a_\perp I_D$ with $a_\perp = \mathrm{tr}\,S_\perp/D$.
- *U block.* Its restriction S_U commutes with every permutation matrix, so $(S_U)_{ij} = (S_U)_{\pi(i)\pi(j)}$ for all $\pi \in S_r$. Hence all diagonal entries are equal, say to x, and (since $r \geq 2$) all off-diagonal entries are equal, say to y. So $S_U = (x-y)P_c + (x - y + ry)P_1$, and $a_1 = \bar{e}^\top S_U \bar{e}$, $a_c = \mathrm{tr}(P_c S_U)/(r-1)$.

Part (c): Propagate the symmetry. By (a), Σ_ν commutes with every $Q \in \mathcal{G}$, hence so do $\Sigma_\nu^\dagger$ and $\Sigma_\nu^{\dagger 1/2}$ (functions of Σ_ν). Thus for $w \neq 0$,

$$T_\nu(Qw) = Qw + \bar{\eta}\Sigma_\nu^{\dagger 1/2} QF(w;\nu) = QT_\nu(w),$$

and $T_\nu(0) = 0$. For a Borel set A, equivariance gives $T_\nu^{-1}(QA) = QT_\nu^{-1}(A)$, hence

$$((T_\nu)_\#\nu)(QA) = \nu(QT_\nu^{-1}(A)) = ((T_\nu)_\#\nu)(A).$$

Since ν_0 is $O(d)$-invariant and $\nu_t(\{0\}) = 0$ for every t (Lemma 20), induction on t concludes. □

G.2.4 The reduced dynamics at $\alpha = 0$

Conventions for Sections G.2.4–G.2.5 and Sections G.4–G.6.

- $\alpha = 0$. With the refitted intercept and $\alpha = 0$ the residual is $Y - \mathbb{E}[Y]$, and the force on a particle with direction $\hat{w}$ is the teacher force $g(\hat{w})$ of Eq. (89). Small $\alpha > 0$ is reached by continuity (Section G.2.6 and Section G.7).
- *Units* $\bar{\eta} = 1$. By Lemma 18 (at $\alpha = 0$ the head plays no role), the trajectory with step $\bar{\eta}$ is the image under $w \mapsto \bar{\eta}w$ of the trajectory with step 1 from $N(0, \kappa^2 I_d)$, κ as in Eq. (91). We work with the latter, that is, we write w for $w/\bar{\eta}$; then $\nu_0 = N(0, \kappa^2 I_d)$.

Definition 22 (MF dynamics at $\alpha = 0$)**.** *For a law ν on $\mathbb{R}^d$ with $\nu(\{0\}) = 0$ put $\Sigma_\nu := \mathbb{E}_{w\sim\nu}[g(\hat{w})g(\hat{w})^\top]$ and $T_\nu(w) := w + \Sigma_\nu^{\dagger 1/2} g(\hat{w})$ ($w \neq 0$), $T_\nu(0) := 0$. The $\alpha = 0$ trajectory is $\nu_0 = N(0, \kappa^2 I_d)$, $\nu_{t+1} := (T_{\nu_t})_\#\nu_t$. This is the map Φ of Section G.2.1 at $\alpha = 0$ in units $\bar{\eta} = 1$. By Lemma 26 below, $\Sigma_{\nu_t} \succ 0$ for every t, so $\Sigma_{\nu_t}^{\dagger 1/2} = \Sigma_{\nu_t}^{-1/2}$. The object of (F) is*

$$\mathcal{M}(\nu) := \mathbb{E}_{w,w'}[H(w, w')\, ww'^\top], \qquad H(w, w') := \frac{\pi - \angle(w, w')}{2\pi},$$

with $w, w' \sim \nu$ independent. At $\alpha = 0$, $\mathcal{M}$ is a surrogate: the actual network AGOP is zero. For $\alpha > 0$, the actual AGOP is $\alpha^2\bar{\eta}^2\mathcal{M}$, evaluated at the corresponding positive-head law in unit-step coordinates; the positive factor preserves eigenspaces and eigenvalue ratios.

Definition 23 (Reduced state)**.** *Let ν be $\mathcal{G}$-invariant with $\nu(\{0\}) = 0$. A* reduced representation *of $w \sim \nu$ is a triple $(\omega, \zeta, n) \in \mathbb{R}^r \times \mathbb{R} \times S^{D-1}$ such that*

$$w_U = \omega, \qquad w_\perp = \sqrt{D}\,\zeta\, n,$$

with n uniform on S^{D-1} and independent of (ω, ζ). The sign of ζ is part of the state. The reduced law *is the law of (ω, ζ). For each particle we write*

$$|w| = (|\omega|^2 + D\zeta^2)^{1/2}, \quad \rho := \omega/|w| = \hat{w}_U, \quad \rho_\infty := \max_i |\rho_i|, \quad \omega_1 := \bar{e}^\top\omega, \quad \xi := P_c\omega,$$

$$s := s(\rho) = \sum_{i=1}^r \big(1 - \sqrt{1 - \rho_i^2}\big), \qquad b := s/|w|,$$

$$f(\rho) := \sum_{i=1}^r \Big(\frac{\pi}{2} + \arcsin\rho_i\Big) - s\sum_{i=1}^r \rho_i, \qquad h(\rho) := P_c\big(\arcsin\rho - s\rho\big),$$

where arcsin *acts coordinatewise. The* normalizations *of ν are the $L^2(\nu)$ quantities*

$$\|f\|, \qquad \bar{a} := \frac{\mathbb{E}[|h|^2]}{r-1}, \qquad N := \|b\zeta\| = \Big(\frac{\mathbb{E}[s^2(1 - |\rho|^2)]}{D}\Big)^{1/2},$$

and its contrast scale *is $\sigma_c := (\mathbb{E}[|\xi|^2]/(r-1))^{1/2}$.*

The identity $N = \|b\zeta\| = (\mathbb{E}[s^2(1 - |\rho|^2)]/D)^{1/2}$ holds because $|w|^2(1 - |\rho|^2) = D\zeta^2$.

Proposition 24 (Exact reduced dynamics)**.** *Let ν be $\mathcal{G}$-invariant with $\nu(\{0\}) = 0$. A reduced representation exists: take $\zeta := |w_\perp|/\sqrt{D}$ and $n := w_\perp/|w_\perp|$ (an independent uniform vector if $w_\perp = 0$). The existence claim uses only the invariance of ν; in (a)–(c), $\alpha = 0$.*

(a) Per particle, $g_U = c_0\big(\frac{\pi}{2} + \arcsin\rho_i - s\rho_i\big)_{i\le r}$, $g_\perp = -c_0 s\,\hat{w}_\perp$, and $s \in [|\rho|^2/2, |\rho|^2]$. Moreover $\bar{e}^\top g = c_0 f/\sqrt{r}$, $P_c g = c_0 h$ and $g_\perp = -c_0\sqrt{D}\, b\zeta\, n$. Hence $a_1 = c_0^2\|f\|^2/r$, $a_c = c_0^2\bar{a}$, $a_\perp = c_0^2 N^2$, and $\Sigma_\nu \succ 0$ iff $\bar{a} > 0$ and $N > 0$.

(b) Let $\Sigma_\nu \succ 0$. Then $T_\nu(w) = (\omega^+, \sqrt{D}\,\zeta^+ n)$ with the same *n, where*

$$\omega_1^+ = \omega_1 + \delta\omega_1, \;\; \delta\omega_1 := \frac{f(\rho)}{\|f\|}; \qquad \xi^+ = \xi + \delta\xi, \;\; \delta\xi := \frac{h(\rho)}{\sqrt{\bar{a}}};$$

$$\zeta^+ = \zeta + \delta\zeta = \zeta\Big(1 - \frac{b}{N}\Big), \qquad \delta\zeta := -\frac{b\zeta}{N}.$$

The constant c_0 cancels. Consequently (ω^+, ζ^+, n) is a reduced representation of $(T_\nu)_\#\nu$. The reduced law evolves autonomously, by a map that depends on the law only through $\|f\|$, $\bar{a}$ and N.

(c) (Whitening.) $\mathbb{E}[\delta\omega\,\delta\omega^\top] = I_r$, i.e. $\mathbb{E}[\delta\omega_1^2] = 1$, $\mathbb{E}[|\delta\xi|^2] = r - 1$ and $\mathbb{E}[\delta\omega_1\delta\xi] = 0$. Also $\mathbb{E}[(\delta\zeta)^2] = 1$.

Proof. First establish the reduced representation. Let $Q = I_r \oplus O$ with $O \in O(D)$. Invariance of ν gives $(\omega, \zeta, On) \overset{d}{=} (\omega, \zeta, n)$ on $\{w_\perp \neq 0\}$; on $\{w_\perp = 0\}$ this holds because the auxiliary vector is uniform and independent. So the conditional law of n given (ω, ζ) is $O(D)$-invariant, hence uniform.

Part (a): Resolve the force into its blocks. The coordinate formula for g_U is Eq. (89) with $\hat{w}_U = \rho$, and $g_\perp = -c_0 s\hat{w}_\perp$. For $x \in [-1, 1]$, $1 - \sqrt{1 - x^2} = x^2/(1 + \sqrt{1 - x^2}) \in [x^2/2, x^2]$; summing gives the bounds on s. Next,

$$\bar{e}^\top g = c_0 r^{-1/2} \sum_i \Big(\frac{\pi}{2} + \arcsin\rho_i - s\rho_i\Big) = c_0 f/\sqrt{r},$$

and $P_c g = c_0 h$ because P_c annihilates the constant vector $\frac{\pi}{2}\mathbf{1}$. Also $\hat{w}_\perp = \sqrt{D}\zeta n/|w|$, so $g_\perp = -c_0\sqrt{D}(s/|w|)\zeta n = -c_0\sqrt{D}b\zeta n$. The formulas for $a_1, a_c, a_\perp$ follow from Proposition 21(b) and $|g_\perp|^2 = c_0^2 D b^2\zeta^2$. Finally $a_1 > 0$ always, because $f > 0$ (Lemma 25(i)).

Part (b): Read off the reduced update. By Proposition 21(b),

$$T_\nu(w) = w + a_1^{-1/2}(\bar{e}^\top g)\bar{e} + a_c^{-1/2} P_c g + a_\perp^{-1/2} g_\perp.$$

With (a),

$$\delta\omega_1 = a_1^{-1/2} c_0 f/\sqrt{r} = f/\|f\|, \qquad \delta\xi = a_c^{-1/2} c_0 h = h/\sqrt{\bar{a}}.$$

The $\perp$ part is

$$\sqrt{D}\zeta n - a_\perp^{-1/2} c_0\sqrt{D}b\zeta n = \sqrt{D}\, n\, \zeta(1 - b/N).$$

The new triple has the same n, which is still uniform and independent of (ω^+, ζ^+), because the latter is a function of (ω, ζ) and of the three normalizations.

Part (c): Verify whitening. The scalar normalizations give

$$\mathbb{E}[\delta\omega_1^2] = \mathbb{E}[f^2]/\|f\|^2 = 1, \qquad \mathbb{E}[|\delta\xi|^2] = \mathbb{E}[|h|^2]/\bar{a} = r - 1.$$

Next, $\mathbb{E}[(\bar{e}^\top g)P_c g] = P_c\Sigma_\nu\bar{e} = a_1 P_c\bar{e} = 0$, so $\mathbb{E}[\delta\omega_1\delta\xi] = 0$. Also $\mathbb{E}[\delta\xi\delta\xi^\top] = P_c\Sigma_\nu P_c/a_c = P_c$, so

$$\mathbb{E}[\delta\omega\delta\omega^\top] = P_1 + P_c = I_r.$$

Finally $\mathbb{E}[(\delta\zeta)^2] = \mathbb{E}[b^2\zeta^2]/N^2 = 1$.

□

At $\alpha = 0$ the reduction applies from $t = 0$; the initial reduced law is given in Section G.4.1.

Lemma 25 (Per-particle signs)**.** *Let $B_r := \{\rho \in \mathbb{R}^r : |\rho| \leq 1\}$, $f_{\min} := \min_{B_r} f$, $f_{\max} := \max_{B_r} f$ and $c_1 := f_{\min}/f_{\max}$. For every $\rho \in B_r$:*

(i) $\frac{\pi}{2}(r - \sqrt{r}) \leq f_{\min} \leq f(\rho) \leq f_{\max} \leq \frac{\pi}{2}(r + \sqrt{r})$. Hence $f_{\min} > 0$, $c_1 \geq (\sqrt{r} - 1)/(\sqrt{r} + 1)$, and in every step of Proposition 24(b), $\delta\omega_1 = f/\|f\| \in [c_1, 1/c_1]$ for every particle.

(ii) With $\bar\rho := r^{-1}\sum_i \rho_i$, for every particle

$$\xi \cdot P_c g_U = c_0|w|\Big[\sum_{i=1}^{r}(\rho_i-\bar\rho)(\arcsin\rho_i-\rho_i)+(1-s)|P_c\rho|^2\Big] \;\ge\; c_0(1-s)\frac{|\xi|^2}{|w|} \;\ge 0.$$

Equivalently $\xi\cdot h \ge (1-s)|\xi|^2/|w| \ge 0$.

(iii) $0 \le s \le |\rho|^2 \le 1$, *hence* $b \in [0, 1/|w|]$.

Proof. Part (i): Bounds on the mean force. The function f is continuous on the compact set B_r, so $f_{\min}$ and $f_{\max}$ exist. Fix $s \in [0,1]$ and put $\psi_s(x) := \arcsin x - sx$, an odd function on $[-1,1]$. For $x \in [0,1]$, $x \le \arcsin x \le \frac{\pi}{2}x$: the first inequality holds because $\arcsin x - x$ is nondecreasing and vanishes at 0, the second by convexity of arcsin on $[0,1]$. Hence $0 \le (1-s)x \le \psi_s(x) \le \frac{\pi}{2}x$ there, and by oddness $|\psi_s(x)| \le \frac{\pi}{2}|x|$ on $[-1,1]$. Since $f(\rho) = \frac{r\pi}{2} + \sum_i \psi_{s(\rho)}(\rho_i)$ and $s(\rho) \in [0,1]$ by (iii),

$$\Big|f(\rho) - \frac{r\pi}{2}\Big| \le \frac{\pi}{2}\sum_i|\rho_i| \le \frac{\pi}{2}\sqrt{r}|\rho| \le \frac{\pi}{2}\sqrt{r}.$$

This gives the bounds, $f_{\min} \ge \frac{\pi}{2}(r-\sqrt{r}) > 0$ for $r \ge 2$, and $c_1 \ge (\sqrt{r}-1)/(\sqrt{r}+1)$. Since $f_{\min} \le \|f\| \le f_{\max}$, we get $\delta\omega_1 \in [f_{\min}/f_{\max}, f_{\max}/f_{\min}] = [c_1, 1/c_1]$.

Part (ii): Sign of the contrast force. We have $\xi = |w|P_c\rho$ and $P_c g_U = c_0 h = c_0 P_c(\arcsin\rho - s\rho)$. Since P_c is an orthogonal projector,

$$\xi\cdot P_c g_U = c_0|w|\big[P_c\rho\cdot\arcsin\rho - s|P_c\rho|^2\big], \qquad P_c\rho\cdot\arcsin\rho = \sum_i(\rho_i-\bar\rho)(\arcsin\rho_i-\rho_i)+|P_c\rho|^2.$$

This is the identity. The sum is nonnegative by Chebyshev's sum inequality: for any reals x_i, y_i with $\bar x := r^{-1}\sum_i x_i$,

$$\sum_i(x_i-\bar x)y_i = \frac{1}{2r}\sum_{i,j}(x_i-x_j)(y_i-y_j). \tag{101}$$

In Eq. (101), take $x_i = \rho_i$, $y_i = \arcsin\rho_i - \rho_i$. Every term is ≥ 0, because $x \mapsto \arcsin x - x$ is nondecreasing. Finally $1 - s \ge 0$ by (iii), and $|P_c\rho|^2 = |\xi|^2/|w|^2$.

Part (iii): Bounds on the radial coefficient. Proposition 24(a) gives $s \in [|\rho|^2/2, |\rho|^2]$, and $|\rho| = |\hat w_U| \le 1$. □

Lemma 26 (Nondegenerate whitening at $\alpha = 0$)**.** *For $\alpha = 0$, every step $\bar\eta > 0$, every $\sigma_0 > 0$ and every $t \ge 0$, ν_t is invariant under the group $\mathcal{G} \cong S_r \times O(D)$ of Proposition 21, absolutely continuous, and $\Sigma_{\nu_t} \succ 0$.*

Proof. Induction on t; ν_0 is $O(d)$-invariant and absolutely continuous. Let $\nu := \nu_t$ be $\mathcal{G}$-invariant and absolutely continuous. At $\alpha = 0$, $F = g(\hat w)$, and Proposition 21(b) gives $\Sigma_\nu = a_1 P_1 + a_c P_c + a_\perp P_\perp$ with $a_1 = \mathbb{E}[(\bar e^\top g)^2]$, $a_c = \mathbb{E}[|P_c g|^2]/(r-1)$ and $a_\perp = \mathbb{E}[|P_\perp g|^2]/D$.

Step 1 (The mean block is positive). We have $\bar e^\top g = c_0 f(\rho)/\sqrt{r} \ge c_0 f_{\min}/\sqrt{r} > 0$ everywhere (Lemma 25(i)), so $a_1 > 0$.

Step 2 (The perpendicular block is positive). The force $P_\perp g = -c_0 s P_\perp \hat w$ with $s \ge |\rho|^2/2$ is nonzero unless $w_U = 0$ or $w_\perp = 0$, a null set; so $a_\perp > 0$.

Step 3 (The contrast block is positive). By Lemma 25(ii), $P_c\rho \cdot P_c g_U \ge c_0(1-s)|P_c\rho|^2$, and $s < 1$ unless $\hat w = \pm u_i$: since $x \mapsto 1 - \sqrt{1-x}$ is strictly convex and vanishes at 0,

$$s \le 1 - \sqrt{1-|\rho|^2} \le 1,$$

with equality in both only if $|\rho| = 1$ and at most one $\rho_i \ne 0$. So $P_c g \ne 0$ off the null set $\{P_c w_U = 0\} \cup \{\hat w = \pm u_i\}$, and $a_c > 0$.

The three positive blocks give $\Sigma_\nu \succ 0$; ν_{t+1} is $\mathcal{G}$-invariant (Proposition 21(c)) and absolutely continuous (Lemma 20). □

G.2.5 THE MF AGOP AND THE CRITERION FOR (F)

For a $\mathcal{G}$-invariant ν with $\nu(\{0\}) = 0$ and $\mathbb{E}[|w|^2] < \infty$ put

$$\mathsf{v}(X) := \mathbb{E}_{w\sim\nu}\big[\mathbf{1}\{w^\top X > 0\}\, w\big], \qquad \mathcal{M}(\nu) = \mathbb{E}_X\big[\mathsf{v}(X)\mathsf{v}(X)^\top\big] = \mathbb{E}_{w,w'}\big[H(w,w')\, ww'^\top\big],$$

where the second equality is Fubini with independent $w, w' \sim \nu$ and $\mathbb{P}_X(w^\top X > 0, w'^\top X > 0) = (\pi - \angle(w,w'))/(2\pi)$. The two statistics of the criterion are

$$\Gamma_U := \frac{\mathbb{E}[|\xi|^2/|w|]}{r-1}, \qquad \Gamma_\perp := \frac{\mathbb{E}[|\zeta|]}{\sqrt{D-1/2}}.$$

Proposition 27 (MF AGOP: block form and the (F) criterion)**.** *Let ν be $\mathcal{G}$-invariant with $\nu(\{0\}) = 0$ and $\mathbb{E}[|w|^2] < \infty$, with a reduced representation (ω, ζ, n).*

(0) (Block form.) $\mathcal{M}(\nu) = \lambda_1 P_1 + \lambda_c P_c + \lambda_\perp P_\perp$, with

$$\lambda_1 = \mathbb{E}_X\big[(\bar{e}^\top \mathsf{v}(X))^2\big],$$
$$\lambda_c = \frac{\mathbb{E}_X\big[|\mathbb{E}_w[(\mathbf{1}\{w^\top X > 0\} - \frac{1}{2})\,\xi]|^2\big]}{r-1},$$
$$\lambda_\perp = \frac{\mathbb{E}_X[|P_\perp \mathsf{v}(X)|^2]}{D}.$$

Consequently $A_r(\mathcal{M}(\nu)) = 1$ with a strict gap $\lambda_r > \lambda_{r+1}$ if and only if $\min(\lambda_1, \lambda_c) > \lambda_\perp$. In that case the top-$r$ space is U and $\lambda_r/\lambda_{r+1} = \min(\lambda_1, \lambda_c)/\lambda_\perp$.

(1) $\lambda_c \geq \Gamma_U^2/(2\pi)$.

(2) $\lambda_\perp \leq \Gamma_\perp^2/(2\pi)$.

(3) Consequently $A_r = 1$ with a strict gap as soon as $\Gamma_U > \Gamma_\perp$ and $\lambda_1 > \lambda_\perp$.

(4) If ν is $O(d)$-invariant (for instance $\nu = \nu_0$), then $\mathcal{M}(\nu)$ is a multiple of I_d.

Proof. Part (0): Block form and eigenspace criterion. For $Q \in \mathcal{G}$, substitute $w = Qw'$ with $w' \sim \nu$. This gives

$$\mathsf{v}(QX) = \mathbb{E}_{w'}[\mathbf{1}\{w'^\top X > 0\} Qw'] = Q\mathsf{v}(X).$$

Since $X \overset{d}{=} QX$,

$$Q\mathcal{M}(\nu)Q^\top = \mathbb{E}_X[\mathsf{v}(QX)\mathsf{v}(QX)^\top] = \mathcal{M}(\nu). \tag{102}$$

As in the proof of Proposition 21(b), a symmetric matrix commuting with $\mathcal{G}$ has the form $\lambda_1 P_1 + \lambda_c P_c + \lambda_\perp P_\perp$, with

$$\lambda_1 = \bar{e}^\top \mathcal{M} \bar{e},$$
$$\lambda_c = \mathrm{tr}(P_c\mathcal{M})/(r-1) = \mathbb{E}_X[|P_c\mathsf{v}|^2]/(r-1),$$
$$\lambda_\perp = \mathrm{tr}(P_\perp\mathcal{M})/D.$$

Next, $P_c\mathsf{v}(X) = \mathbb{E}_w[\mathbf{1}\{w^\top X > 0\}\xi]$, and $\mathbb{E}_w[\xi] = 0$: the vector $\mathbb{E}[\omega]$ is permutation-invariant, hence a multiple of $\mathbf{1}$, and $\mathbb{E}[\xi] = P_c\mathbb{E}[\omega] = 0$. So $\frac{1}{2}\mathbb{E}_w[\xi]$ may be subtracted.

For the criterion: if $\min(\lambda_1, \lambda_c) > \lambda_\perp$, the r largest eigenvalues (with multiplicity) are λ_1 (on $\bar{e}$) and λ_c (on the contrast space, multiplicity $r-1$). Their span is U, so $P_r = P_U$ is unique, $A_r = 1$, and $\lambda_r = \min(\lambda_1, \lambda_c)$, $\lambda_{r+1} = \lambda_\perp$.

Conversely, suppose $\lambda_r > \lambda_{r+1}$ and $A_r = 1$ for the unique selector. Then $\mathrm{tr}(P_U P_r) = r$ with both projectors of rank r, so $P_r = P_U$; the eigenvalues on U are therefore the r largest, and $\lambda_{r+1} = \lambda_\perp < \min(\lambda_1, \lambda_c)$.

Part (1): Lower-bound the contrast eigenvalue. Put $A(X) := P_c\mathsf{v}(X)$ and use the test function $P_c X_U$, where $X_U := P_U X$. By the exact formula of Section G.1.1, $\mathbb{E}_X[X_U \mathbf{1}\{w^\top X > 0\}] =$

$\hat{w}_U/\sqrt{2\pi} = \omega/(\sqrt{2\pi}|w|)$, hence by Fubini

$$\begin{aligned}\mathbb{E}_X[A(X)\cdot P_c X_U] &= \mathbb{E}_w\Big[\xi\cdot\frac{P_c\omega}{\sqrt{2\pi}|w|}\Big]\\ &= \frac{\mathbb{E}[|\xi|^2/|w|]}{\sqrt{2\pi}} = \frac{(r-1)\Gamma_U}{\sqrt{2\pi}}.\end{aligned} \tag{103}$$

Since $\mathbb{E}_X[|P_cX_U|^2] = \operatorname{tr} P_c = r-1$, Cauchy–Schwarz applied to Eq. (103) gives

$$(r-1)\lambda_c = \mathbb{E}_X[|A|^2] \geq (r-1)^2\Gamma_U^2/(2\pi(r-1)),$$

that is, $\lambda_c \geq \Gamma_U^2/(2\pi)$.

Part (2): Upper-bound the perpendicular eigenvalue. Write $X_\perp := P_\perp X$, $\hat{X}_\perp := X_\perp/|X_\perp|$ ($X_\perp \neq 0$ a.s.) and $q := n^\top \hat{X}_\perp$. Since $w^\top X = \omega^\top X_U + \sqrt{D}\zeta|X_\perp|q$ and n is independent of (ω,ζ),

$$P_\perp \mathsf{v}(X) = \mathbb{E}_{(\omega,\zeta)}\Big[\sqrt{D}\,\zeta\;\mathbb{E}_n\big[\mathbf{1}\{\omega^\top X_U + \sqrt{D}\zeta|X_\perp|q>0\}\,n\big]\Big].$$

Fix X and (ω,ζ) and decompose $n = q\hat{X}_\perp + n'$. Given q, the vector n' is symmetric in $\hat{X}_\perp^\perp$, while the gate depends on n only through q. Hence $\mathbb{E}_n[\mathbf{1}\{\cdot\}\,n] = \hat{X}_\perp\,\psi$ with $\psi := \mathbb{E}_q[\mathbf{1}\{\cdot\}\,q]$ and $|\psi| \leq \mathbb{E}[q_+] = \frac{1}{2}\mathbb{E}[|q|]$, because $-\mathbb{E}[q_-] \leq \mathbb{E}[\mathbf{1}_E q] \leq \mathbb{E}[q_+]$ for any event E and q is symmetric. Therefore, for every X,

$$\begin{aligned}|P_\perp \mathsf{v}(X)| &\leq \sqrt{D}\,\mathbb{E}[|\zeta|]\,\mathbb{E}[|q|]/2,\\ \lambda_\perp &\leq (\mathbb{E}[|\zeta|])^2(\mathbb{E}[|q|])^2/4.\end{aligned} \tag{104}$$

To bound Eq. (104), use that q is the first coordinate of a uniform point of S^{D-1}, so $\mathbb{E}[|q|] = \Gamma(\frac{D}{2})/(\sqrt{\pi}\,\Gamma(\frac{D+1}{2}))$. Kershaw's form of Gautschi's inequality, $\Gamma(x+1)/\Gamma(x+\frac{1}{2}) > (x+\frac{1}{4})^{1/2}$ for $x>0$, applied with $x = \frac{D-1}{2}$ gives $\mathbb{E}[|q|] \leq (2/(\pi(D-\frac{1}{2})))^{1/2}$ for $D \geq 2$. Substituting in Eq. (104) gives

$$\lambda_\perp \leq (\mathbb{E}[|\zeta|])^2/(2\pi(D-\frac{1}{2})) = \Gamma_\perp^2/(2\pi).$$

Part (3): Combine the scalar bounds. By (1) and (2),

$$\lambda_c \geq \Gamma_U^2/(2\pi) > \Gamma_\perp^2/(2\pi) \geq \lambda_\perp.$$

Together with $\lambda_1 > \lambda_\perp$, (0) applies.

Part (4): Isotropy at initialization. As in Eq. (102), $\mathcal{M}(\nu)$ commutes with every $Q \in O(d)$; commuting with every reflection, it is a multiple of I_d. □

G.2.6 Continuity in the head scale and in the width

Here $\bar{\eta} > 0$ is fixed and the head scale $\alpha \geq 0$ varies; F_α, Σ_α, $T_{\mu,\alpha}$ and Φ_α are the objects of Section G.2.1 at head scale α.

Lemma 28 (Continuity of Φ in the law and the head scale)**.** *Let $\alpha_k \to \alpha \geq 0$, let $\mu_0 \in \mathcal{P}_2$ with $\mu_0(\{0\}) = 0$ and $\Sigma_\alpha(\mu_0) \succ 0$, and let $W_2(\mu_k,\mu_0) \to 0$. Then $\Sigma_{\alpha_k}(\mu_k) \to \Sigma_\alpha(\mu_0)$, $\sup_w |T_{\mu_k,\alpha_k}(w) - T_{\mu_0,\alpha}(w)| \to 0$ and $W_2(\Phi_{\alpha_k}(\mu_k), \Phi_\alpha(\mu_0)) \to 0$.*

Proof. Write $F_k := F_{\alpha_k}(\cdot;\mu_k)$, $F_0 := F_\alpha(\cdot;\mu_0)$, $T_k := T_{\mu_k,\alpha_k}$ and $T_0 := T_{\mu_0,\alpha}$.

Step 1 (The force converges uniformly). First separate the change in head scale from the change in law:

$$F_k - F_0 = -(\alpha_k - \alpha)\,S(\hat{w};\mu_k) - \alpha\big(S(\hat{w};\mu_k) - S(\hat{w};\mu_0)\big).$$

Lemma 19(b) gives

$$\sup_w |F_k(w) - F_0(w)| \leq |\alpha_k - \alpha| \int |w'|\,d\mu_k(w') + 2\alpha\sqrt{d}\,W_2(\mu_k,\mu_0) \;\to\; 0, \tag{105}$$

because $\int|w'|\,d\mu_k \leq (\int|w'|^2 d\mu_k)^{1/2}$ stays bounded (W_2-convergence implies convergence of second moments). The F_k are uniformly bounded: $|F_k| \leq \sup|g| + \alpha_k \int|w'|\,d\mu_k$.

Step 2 (The whitening converges). Decompose the covariance difference as

$$\Sigma_{\alpha_k}(\mu_k) - \Sigma_\alpha(\mu_0) = \int (F_k F_k^\top - F_0 F_0^\top)\, d\mu_k + \Big(\int F_0 F_0^\top d\mu_k - \int F_0 F_0^\top d\mu_0 \Big). \tag{106}$$

The first term tends to 0 by Eq. (105), the second by the portmanteau theorem, as $F_0 F_0^\top$ is bounded and continuous off the μ_0-null set $\{0\}$ (Lemma 19(a),(c)).

Step 3 (The maps converge). By Eq. (106), $\Sigma_{\alpha_k}(\mu_k) \succ 0$ eventually and $\Sigma_{\alpha_k}(\mu_k)^{-1/2} \to \Sigma_\alpha(\mu_0)^{-1/2}$. Together with Eq. (105), this gives $T_k \to T_0$ uniformly ($T_{\mu,\alpha}(0) = 0$ for every μ and α).

Step 4 (The laws converge). Coupling $T_k(w)$ with $T_0(w)$, $w \sim \mu_k$, gives

$$W_2(\Phi_{\alpha_k}(\mu_k), \Phi_\alpha(\mu_0)) \le \sup|T_k - T_0| + W_2((T_0)_\# \mu_k, (T_0)_\# \mu_0). \tag{107}$$

It remains to control the second term in Eq. (107). Now $T_0 = \mathrm{id} + b_0$ with b_0 bounded and continuous μ_0-a.e., so $(T_0)_\# \mu_k \to (T_0)_\# \mu_0$ weakly (continuous mapping theorem). In

$$|T_0(w)|^2 = |w|^2 + 2w^\top b_0 + |b_0|^2,$$

the integral of the first term converges, the second is continuous μ_0-a.e. with linear growth, hence uniformly integrable under (μ_k), and the third is bounded and continuous μ_0-a.e. Weak convergence together with convergence of second moments is W_2-convergence. □

Proposition 29 (Small head and large width)**.** *Fix $\bar\eta, \sigma_0 > 0$ and $T \in \mathbb{N}$, and let $\nu_\alpha(t)$ be the MF trajectory from $\nu_0 = N(0, \sigma_0^2 I_d/d)$ at head scale $\alpha \ge 0$.*

- *(a) (Small head.) As $\alpha \to 0$, for every $t \le T$: $W_2(\nu_\alpha(t), \nu_0(t)) \to 0$, $\Sigma_\alpha(\nu_\alpha(t)) \to \Sigma_0(\nu_0(t)) \succ 0$ and $\mathcal{M}(\nu_\alpha(t)) \to \mathcal{M}(\nu_0(t))$. In particular there is $\alpha_T > 0$ such that $\Sigma_\alpha(\nu_\alpha(t)) \succ 0$ for all $t \le T$ and $0 \le \alpha \le \alpha_T$.*
- *(b) (Large width.) Fix $\alpha \ge 0$ and assume $\Sigma_\alpha(\nu_\alpha(t)) \succ 0$ for every $t < T$. Let $(w_j(0))_{j \ge 1}$ be IID $N(0, \sigma_0^2 I_d/d)$, $W^{(m)}(0) := [w_1(0), \dots, w_m(0)]$, and run Eq. (87) at width m with $\eta = \bar\eta\sqrt{m}$ and head α. Then almost surely, for every $t \le T$: $W_2(\mu_{W^{(m)}(t)}, \nu_\alpha(t)) \to 0$, $L(W^{(m)}(t)) \to L(\nu_\alpha(t))$ and $M(W^{(m)}(t)) \to \alpha^2 \mathcal{M}(\nu_\alpha(t))$. If moreover $\alpha > 0$ and $\lambda_r > \lambda_{r+1}$ for $\mathcal{M}(\nu_\alpha(t))$, then the top-$r$ eigenspace of $M(W^{(m)}(t))$ converges to that of $\mathcal{M}(\nu_\alpha(t))$.*

Proof. The common ingredient is continuity of $\mathcal{L}$ and $\mathcal{M}$. Their integrands extend continuously by 0 to $w = 0$ or $w' = 0$ and grow at most quadratically, and $W_2(\mu_k \otimes \mu_k, \nu \otimes \nu) \to 0$ when $W_2(\mu_k, \nu) \to 0$. Hence $\mathcal{L}$ and $\mathcal{M}$ are continuous on $(\mathcal{P}_2, W_2)$.

Part (a): Small head. Let $\alpha_k \to 0$. By Lemmas 20 and 26, every $\nu_0(t)$ lies in $\mathcal{P}_2$, has no atom at 0 and satisfies $\Sigma_0(\nu_0(t)) \succ 0$, and every $\nu_{\alpha_k}(t)$ lies in $\mathcal{P}_2$. We show $W_2(\nu_{\alpha_k}(t), \nu_0(t)) \to 0$ by induction on $t \le T$. At $t = 0$ the laws coincide. If it holds at some $t < T$, Lemma 28 with $\mu_k = \nu_{\alpha_k}(t)$, $\mu_0 = \nu_0(t)$ and $\alpha = 0$ gives

$$W_2(\nu_{\alpha_k}(t+1), \nu_0(t+1)) \to 0.$$

The same lemma at every $t \le T$, followed by continuity of $\mathcal{M}$, gives

$$\Sigma_{\alpha_k}(\nu_{\alpha_k}(t)) \to \Sigma_0(\nu_0(t)),$$
$$\mathcal{M}(\nu_{\alpha_k}(t)) \to \mathcal{M}(\nu_0(t)).$$

The sequence $\alpha_k \to 0$ was arbitrary, so the limits hold as $\alpha \to 0$. Finally the positive definite matrices form an open set and t ranges over a finite set, which gives α_T.

Part (b): Large width. By Lemma 17, $\mu_{W^{(m)}(t)} = \Phi_\alpha^t(\mu_{W^{(m)}(0)})$. Varadarajan's theorem and the strong law for $m^{-1}\sum_j |w_j(0)|^2$ give

$$W_2(\mu_{W^{(m)}(0)}, \nu_0) \to 0 \tag{108}$$

almost surely. By Lemma 20, $\nu_\alpha(s) \in \mathcal{P}_2$ and $\nu_\alpha(s)(\{0\}) = 0$; with $\Sigma_\alpha(\nu_\alpha(s)) \succ 0$ for $s < T$, Lemma 28 (with $\alpha_k = \alpha$) applies at each $\nu_\alpha(s)$, $s < T$, and induction on t gives W_2-continuity of Φ_α^t at ν_0 for $t \le T$. Apply this continuity to Eq. (108). With Lemma 17(d) and the continuity of $\mathcal{L}$ and $\mathcal{M}$, this gives the convergence of the loss and of the AGOP.

The last claim is the Davis–Kahan theorem: with a strict gap, the spectral projector onto the top-r eigenspace is continuous at $\mathcal{M}(\nu_\alpha(t))$ (for every choice of top-r eigenvectors of the approximants), and the factor $\alpha^2 > 0$ changes no eigenspace. □

G.3 The plateau and the loss drop

What this section proves. (P) (Theorem 42) and (R) (Corollary 45), for every head scale $0 < \alpha \le \alpha^*(t_P)$. Only three scalar statistics of the particle law enter: the output scales y, y_U and the mean mode x (Definition 30). The loss is sandwiched as $1 - 2\sqrt{\mathsf{k}_1}\, y_U \le L/V \le 1 - c_x x + \mathsf{k}_1 y^2$ (Lemma 36); one whitened step raises y and y_U by at most the clock increment (Lemma 31); and the mean mode increases on the small-output window used below (Lemma 39).

Where it is used. Steps 1 and 5 of the proof of Theorem 16 (Section G.7).

Standing hypothesis. Apart from Lemmas 31 and 39, nothing in Sections G.3.1–G.3.3 involves the dynamics: each statement holds for every particle law ν with finite second moment and $\nu(\{0\}) = 0$, so that $\hat{w}$ is defined ν-almost surely. Along the MF trajectory this holds at every time (Lemma 20).

G.3.1 Output scales and one MF step

We retain the teacher amplitude γ, variance $V = \mathrm{Var}(Y)$, mean direction $\bar{e}$ and constant $c_0 = \gamma/(2\pi)$ from Section G.1.1. Here $\|A\|$ for a matrix is the operator norm; all expectations carrying ν average over $w \sim \nu$, while moments of functions of X average over the Gaussian input. For an input function h, $\|h\|_2 := (\mathbb{E}_X[|h(X)|^2])^{1/2}$. $E_U := [u_1, \dots, u_r] \in \mathbb{R}^{d\times r}$ and, for $w \in \mathbb{R}^d$, $w_U := E_U^\top w \in \mathbb{R}^r$ (teacher coordinates) and $w_\perp := (I - P_U)w$. For $w \neq 0$ put $\rho = \rho(w) := E_U^\top \hat{w} \in \mathbb{R}^r$, so $|\rho| = |w_U|/|w| \le 1$. For $v \in \mathbb{R}^r$ with $|v| \le 1$ let

$$s(v) := \sum_{i=1}^r \big(1 - \sqrt{1 - v_i^2}\big), \qquad \tfrac{1}{2}|v|^2 \le s(v) \le |v|^2 \le 1,$$

since $t/2 \le 1 - \sqrt{1-t} \le t$ on $[0,1]$. The constant $\mathsf{k}_1 = (\pi - 1)/(2\pi) = 0.3408450\ldots$ appears throughout. Because $V = \gamma^2 r \mathsf{k}_1$,

$$\frac{\gamma\sqrt{r}}{\sqrt{V}} = \frac{1}{\sqrt{\mathsf{k}_1}}, \qquad \frac{\sqrt{V}\sqrt{r}}{\gamma} = r\sqrt{\mathsf{k}_1}. \tag{109}$$

Definition 30 (Output scales and clock)**.** *For a law ν on $\mathbb{R}^d$ with finite second moment,*

$$y(\nu) := \frac{\alpha\|\mathbb{E}_\nu[ww^\top]\|^{1/2}}{\sqrt{V}},$$

$$y_U(\nu) := \frac{\alpha\|\mathbb{E}_\nu[w_U w_U^\top]\|^{1/2}}{\sqrt{V}},$$

$$x(\nu) := \frac{\alpha\,\mathbb{E}_\nu[\bar{e}^\top w]}{\sqrt{V}} \quad \textit{(the mean mode)}.$$

Along the MF trajectory we write $y(t), y_U(t), x(t)$, $y_0 := y(0)$ and $y_{U0} := y_U(0)$. The clock *is $T(t) := \alpha\bar{\eta}t/\sqrt{V}$, and $\Delta T := T(1) = \alpha\bar{\eta}/\sqrt{V}$.*

Always $|x| \le y_U \le y$, since $|\mathbb{E}_\nu[\bar{e}^\top w]| \le (\mathbb{E}_\nu[(\bar{e}^\top w)^2])^{1/2} \le \|\mathbb{E}_\nu[w_U w_U^\top]\|^{1/2}$. At width m these are the familiar quantities $\alpha\|W\|/\sqrt{mV}$, $\alpha\|E_U^\top W\|/\sqrt{mV}$ and $\alpha\,\bar{e}^\top W\mathbf{1}_m/(m\sqrt{V})$ evaluated at μ_W.

Lemma 31 (One MF step)**.** *Let ν have finite second moment, put $\delta w := \bar{\eta}\,\Sigma_\nu^{\dagger 1/2} F(w;\nu)$ and $\nu^+ := \Phi(\nu)$, and let P be the orthogonal projector onto the range of Σ_ν.*

(a) $\mathbb{E}_\nu[\delta w\,\delta w^\top] = \bar{\eta}^2 P$*; in particular* $\mathbb{E}_\nu[|\delta w|^2] = \bar{\eta}^2 \operatorname{rank}\Sigma_\nu \le \bar{\eta}^2 d$.

(b) $\|\mathbb{E}_{\nu^+}[ww^\top]\|^{1/2} \le \|\mathbb{E}_\nu[ww^\top]\|^{1/2} + \bar{\eta}$*, and the same holds for* w_U*. Hence along the MF trajectory* $y(t) \le y_0 + T(t)$ *and* $y_U(t) \le y_{U0} + T(t)$.

(c) $x(\nu^+) - x(\nu) = \Delta T \cdot p_{11}(\nu)$ *with* $p_{11}(\nu) := \mathbb{E}_\nu\big[\bar{e}^\top \Sigma_\nu^{\dagger 1/2} F(w;\nu)\big]$*, and* $|p_{11}| \le 1$*. Along the trajectory we write* $p_{11}(t) := p_{11}(\nu_t)$.

Proof. Part (a): The increment covariance. The whitening identity is

$$\mathbb{E}_\nu[\delta w \delta w^\top] = \bar{\eta}^2 \Sigma_\nu^{\dagger 1/2} \Sigma_\nu \Sigma_\nu^{\dagger 1/2} = \bar{\eta}^2 P. \tag{110}$$

Taking the trace proves the stated RMS bound.

Part (b): Growth of the output scales. For a unit $a \in \mathbb{R}^d$, $(\mathbb{E}_{\nu^+}[(a^\top w)^2])^{1/2}$ is the $L^2(\nu)$ norm of $a^\top w + a^\top \delta w$. By Minkowski's inequality and Eq. (110), it is at most

$$\|a^\top w\|_{L^2(\nu)} + \bar{\eta}|Pa| \le \|a^\top w\|_{L^2(\nu)} + \bar{\eta}.$$

Take the supremum over unit $a \in \mathbb{R}^d$ (respectively $a \in U$), and multiply by $\alpha/\sqrt{V}$; induct on t.

Part (c): Growth of the mean mode. By definition,

$$x(\nu^+) - x(\nu) = \alpha\,\mathbb{E}_\nu[\bar{e}^\top \delta w]/\sqrt{V} = \Delta T\, p_{11}(\nu).$$

Cauchy–Schwarz and Eq. (110) give

$$|p_{11}| \le (\mathbb{E}_\nu[(\bar{e}^\top \Sigma_\nu^{\dagger 1/2} F)^2])^{1/2} = |P\bar{e}| \le 1. \qquad \square$$

G.3.2 The loss sandwich

Lemma 32 (Kernel expansion)**.** $\mathsf{k}(\rho) = \frac{\rho}{4} + \frac{k_2(\rho)}{2\pi}$ *with* $k_2(\rho) := \sqrt{1-\rho^2} + \rho \arcsin\rho - 1 = \sum_{n\ge 1} b_n \rho^{2n}$*, where every* $b_n > 0$*,* $b_1 = \frac{1}{2}$ *and* $\sum_n b_n = k_2(1) = \frac{\pi}{2} - 1$*. Hence*

$$\tfrac{1}{2}\rho^2 \le k_2(\rho) \le (\tfrac{\pi}{2} - 1)\rho^2 \qquad (|\rho| \le 1),$$

and $\mathsf{k}(\rho) = \sum_{k\ge 1} q_k \rho^k$ *with* $q_1 = \frac{1}{4}$*,* $q_{2n} = b_n/(2\pi)$ *and* $q_k = 0$ *otherwise. So all* $q_k \ge 0$ *and* $\sum_k q_k = \mathsf{k}_1$.

Proof. Insert $\pi - \arccos\rho = \frac{\pi}{2} + \arcsin\rho$ in the formula for k of Section G.1.1. Next,

$$k_2'(\rho) = \arcsin\rho, \qquad k_2''(\rho) = (1-\rho^2)^{-1/2} = \sum_{n\ge 0} \binom{2n}{n} 4^{-n} \rho^{2n}.$$

Integrating twice from $k_2(0) = k_2'(0) = 0$ gives

$$b_{n+1} = \binom{2n}{n} 4^{-n}/((2n+1)(2n+2)) > 0,$$

so $b_1 = \frac{1}{2}$.

The series converges absolutely on $[-1,1]$, with sum $k_2(1) = \frac{\pi}{2} - 1$ at $\rho = 1$. Since $k_2(\rho)/\rho^2 = \sum_n b_n \rho^{2n-2}$ is nondecreasing in $|\rho|$, it lies between $b_1 = \frac{1}{2}$ and $k_2(1) = \frac{\pi}{2} - 1$. The coefficients q_k are read off, and $\sum_k q_k = \mathsf{k}(1) = \mathsf{k}_1$. $\square$

Lemma 33 (Loss decomposition)**.** *Put* $g_\nu(x) := \alpha \mathbb{E}_{w\sim\nu}[\mathrm{ReLU}(w^\top x)]$ *as in Section G.2.1. Then* $L(\nu) = \mathrm{Var}(Y - g_\nu(X))$ *and*

$$L(\nu) = V - 2\alpha\Phi(\nu) + \alpha^2 Q(\nu), \qquad \Phi(\nu) := \mathbb{E}_\nu[\mathrm{Cov}(Y, \mathrm{ReLU}(w^\top X))] = \Phi_{\mathrm{lin}}(\nu) + \Phi_{\mathrm{nl}}(\nu),$$

$$\Phi_{\mathrm{lin}}(\nu) := \tfrac{\gamma}{4}\sqrt{r}\,\mathbb{E}_\nu[\bar{e}^\top w], \qquad \Phi_{\mathrm{nl}}(\nu) := \tfrac{\gamma}{2\pi}\mathbb{E}_\nu\Big[|w| \sum_i k_2(\rho_i(w))\Big] \ge 0,$$

$$Q(\nu) := \mathrm{Var}(\mathbb{E}_\nu[\mathrm{ReLU}(w^\top X)]) = \mathbb{E}_{w,w'\sim\nu}[|w||w'|\mathsf{k}(\hat{w}^\top \hat{w}')].$$

In particular, exactly,

$$\frac{2\alpha\Phi_{\mathrm{lin}}(\nu)}{V} = c_x\, x(\nu), \qquad c_x := \frac{\gamma\sqrt{r}}{2\sqrt{V}} = \frac{1}{2\sqrt{\mathsf{k}_1}} = 0.8564292\ldots \tag{111}$$

Proof. $\mathcal{L}(\nu) = \mathrm{Var}(Y - g_\nu(X))$ is the definition of $\mathcal{L}$ in Section G.2.1 written out with the covariances $\mathrm{Cov}(Y, \mathrm{ReLU}(w^\top X)) = \gamma|w| \sum_i \mathsf{k}(\rho_i)$ and $\mathrm{Cov}(\mathrm{ReLU}(w^\top X), \mathrm{ReLU}(w'^\top X)) = |w||w'|\mathsf{k}(\hat{w}^\top \hat{w}')$ (homogeneity and the kernel k; Fubini's theorem applies by finite second moments). Expanding the variance gives the first display. Insert $\mathsf{k}(\rho) = \rho/4 + k_2(\rho)/(2\pi)$ and use $|w| \sum_i \rho_i = \sum_i u_i^\top w = \sqrt{r}\, \bar{e}^\top w$. Nonnegativity of Φ_{nl} is $k_2 \geq 0$. For Eq. (111), $2\alpha \frac{\gamma}{4}\sqrt{r}\, \mathbb{E}_\nu[\bar{e}^\top w]/V = \frac{\gamma\sqrt{r}}{2\sqrt{V}} x(\nu)$, and $\gamma\sqrt{r}/\sqrt{V} = 1/\sqrt{\mathsf{k}_1}$ by Eq. (109). □

Lemma 34 (Quadratic term)**.** *For every law ν with finite second moment, $Q(\nu) \leq \mathsf{k}_1 \|\mathbb{E}_\nu[ww^\top]\|$.*

Proof. By Lemma 32, $Q(\nu) = \sum_k q_k\, \mathbb{E}_{w,w'}[(w^\top w')(\hat{w}^\top \hat{w}')^{k-1}]$. Let $\phi(\hat{w}) := \hat{w}^{\otimes(k-1)}$ with coordinates ϕ_ι, so that $(\hat{w}^\top \hat{w}')^{k-1} = \sum_\iota \phi_\iota(\hat{w})\phi_\iota(\hat{w}')$ and $\sum_\iota \phi_\iota(\hat{w})^2 = 1$. Then

$$\begin{aligned} \mathbb{E}_{w,w'}[(w^\top w')(\hat{w}^\top \hat{w}')^{k-1}] &= \sum_\iota |\mathbb{E}_\nu[\phi_\iota(\hat{w})w]|^2 \\ &\leq \sum_\iota \|\mathbb{E}_\nu[ww^\top]\|\, \mathbb{E}_\nu[\phi_\iota(\hat{w})^2] = \|\mathbb{E}_\nu[ww^\top]\|, \end{aligned} \tag{112}$$

because $|\mathbb{E}[fw]|^2 = \sup_{|a|=1}(\mathbb{E}[f\, a^\top w])^2 \leq \mathbb{E}[f^2]\|\mathbb{E}[ww^\top]\|$ by Cauchy–Schwarz. Sum Eq. (112) over k, using $q_k \geq 0$ and $\sum_k q_k = \mathsf{k}_1$. □

Lemma 35 (Covariance bounds)**.** $\Phi(\nu) \leq \mathsf{k}_1\gamma\sqrt{r}\, \|\mathbb{E}_\nu[w_U w_U^\top]\|^{1/2}$ *and*

$$0 \leq \Phi_{\mathrm{nl}}(\nu) \leq (\mathsf{k}_1 - \tfrac{1}{4})\gamma\, \mathbb{E}_\nu[|w_U|^2/|w|] \leq (\mathsf{k}_1 - \tfrac{1}{4})\gamma\sqrt{r}\, \|\mathbb{E}_\nu[w_U w_U^\top]\|^{1/2}.$$

Proof. Note $\mathsf{k}_1 - \frac{1}{4} = \frac{\pi/2-1}{2\pi}$. By Lemma 32, $\sum_i k_2(\rho_i) \leq (\frac{\pi}{2} - 1)|\rho|^2$ and $|w||\rho|^2 = |w_U|^2/|w| \leq |w_U|$. Moreover $\mathbb{E}_\nu[|w_U|] \leq (\mathrm{tr}\, \mathbb{E}_\nu[w_U w_U^\top])^{1/2} \leq \sqrt{r}\, \|\mathbb{E}_\nu[w_U w_U^\top]\|^{1/2}$. This gives the bound on Φ_{nl}. With $|\mathbb{E}_\nu[\bar{e}^\top w]| \leq \|\mathbb{E}_\nu[w_U w_U^\top]\|^{1/2}$, $\Phi \leq \frac{\gamma}{4}\sqrt{r}\, \|\mathbb{E}_\nu[w_U w_U^\top]\|^{1/2} + (\mathsf{k}_1 - \frac{1}{4})\gamma\sqrt{r}\, \|\mathbb{E}_\nu[w_U w_U^\top]\|^{1/2}$. □

Lemma 36 (Loss sandwich)**.** *For every law ν with finite second moment and $\nu(\{0\}) = 0$, with y, y_U, x from Definition 30:*

(a) $1 - 2\sqrt{\mathsf{k}_1}\, y_U \leq L/V \leq 1 - c_x\, x + \mathsf{k}_1\, y^2$*;*

(b) $L/V \geq 1 - c_x\, x - e_{\mathrm{nl}}$*, where* $e_{\mathrm{nl}} := 2\alpha\Phi_{\mathrm{nl}}(\nu)/V$ *satisfies* $0 \leq e_{\mathrm{nl}} \leq c_{\mathrm{nl}}\, y_U$ *with* $c_{\mathrm{nl}} := 2(\mathsf{k}_1 - \frac{1}{4})/\sqrt{\mathsf{k}_1}$*;*

(c) for the Gaussian law $\nu_0 = N(0, \sigma_0^2 I_d/d)$*:* $e_{\mathrm{nl}}(\nu_0) \leq c_{\mathrm{nl}}\sqrt{r/d}\, y_0'$ *with* $y_0' := \alpha\sigma_0/(\sqrt{d}\sqrt{V})$*.*

Numerically $2\sqrt{\mathsf{k}_1} = 1.1676387\ldots$*,* $c_x = 0.8564292\ldots$ *and* $c_{\mathrm{nl}} = 0.3112094\ldots$*.*

Proof. Throughout, use Lemma 33.

Part (a): The two-sided loss bound. For the lower bound, drop $\alpha^2 Q \geq 0$ and use Lemma 35. By Eq. (109),

$$2\alpha\mathsf{k}_1\gamma\sqrt{r}\, \|\mathbb{E}[w_U w_U^\top]\|^{1/2}/V = 2\mathsf{k}_1 \frac{\gamma\sqrt{r}}{\sqrt{V}} y_U = 2\sqrt{\mathsf{k}_1} y_U.$$

For the upper bound, $\Phi \geq \Phi_{\mathrm{lin}}$ with Eq. (111), and $\alpha^2 Q/V \leq \mathsf{k}_1 y^2$ by Lemma 34.

Part (b): Keep the nonlinear covariance explicit. Drop $\alpha^2 Q \geq 0$ and write $\Phi = \Phi_{\mathrm{lin}} + \Phi_{\mathrm{nl}}$ with Eq. (111). By Lemma 35,

$$e_{\mathrm{nl}} \leq 2\alpha(\mathsf{k}_1 - \frac{1}{4})\gamma\sqrt{r}\,\|\mathbb{E}[w_U w_U^\top]\|^{1/2}/V$$
$$= 2(\mathsf{k}_1 - \frac{1}{4})\frac{\gamma\sqrt{r}}{\sqrt{V}} y_U = c_{\mathrm{nl}} y_U.$$

Part (c): Use the Gaussian initial law. Under ν_0, $w = |w|\hat{w}$ with $\hat{w}$ uniform on the sphere and independent of $|w|$. Thus

$$\mathbb{E}[|w_U|^2/|w|] = \mathbb{E}[|w|] \cdot \mathbb{E}[|E_U^\top \hat{w}|^2] = (r/d)\mathbb{E}[|w|] \leq \sigma_0 r/d. \tag{113}$$

Substituting Eq. (113) into the covariance bound gives

$$e_{\mathrm{nl}} \leq 2\alpha(\mathsf{k}_1 - \frac{1}{4})\gamma\sigma_0 r/(dV) = 2(\mathsf{k}_1 - \frac{1}{4})\frac{\gamma\sqrt{r}}{\sqrt{V}}\sqrt{\frac{r}{d}}\,\frac{\alpha\sigma_0}{\sqrt{d}\sqrt{V}}.$$ □

G.3.3 THE SELF-FORCE AND THE MEAN MODE

Lemma 37 (Self-force)**.** *For every law ν with finite second moment and every $w \neq 0$,*

$$F(w;\nu) - g(\hat{w}) = -\mathbb{E}\big[(g_\nu(X) - \mathbb{E}[g_\nu])\, X\, \mathbf{1}\{w^\top X > 0\}\big],$$
$$|F(w;\nu) - g(\hat{w})| \leq \sqrt{\mathsf{k}_1/2}\; y(\nu)\sqrt{V}.$$

Here $\sqrt{\mathsf{k}_1/2} = 0.4128226\ldots$ In units of $c_0/\sqrt{r}$, the natural unit of the mean force (Lemma 38), the bound reads

$$\frac{\sqrt{r}}{c_0}\sqrt{\mathsf{k}_1/2}\, y\sqrt{V} = e(y) := \sqrt{2}\,\pi\mathsf{k}_1\, r\, y \approx 1.5143347\, r\, y. \tag{114}$$

The displayed decimal coefficient is rounded upward, so using it in a small-output threshold is conservative.

Proof. The identity is $F(w;\nu) = \mathbb{E}[R_\nu X \mathbf{1}\{w^\top X > 0\}]$ with $R_\nu = (Y - \mathbb{E}[Y]) - (g_\nu - \mathbb{E}[g_\nu])$ (Lemma 17(a) and its proof), whose teacher part is $g(\hat{w})$. For a unit $a \in \mathbb{R}^d$, Cauchy–Schwarz gives

$$|a^\top(F - g)| \leq \|g_\nu - \mathbb{E}[g_\nu]\|_2\, (\mathbb{E}[(a^\top X)^2 \mathbf{1}\{w^\top X > 0\}])^{1/2}. \tag{115}$$

The second factor is $\sqrt{1/2}$ by the symmetry $X \mapsto -X$, and Lemma 34 gives

$$\|g_\nu - \mathbb{E}[g_\nu]\|_2^2 = \alpha^2 Q(\nu) \leq \alpha^2 \mathsf{k}_1 \|\mathbb{E}_\nu[ww^\top]\| = \mathsf{k}_1 y^2 V.$$

Taking the supremum over unit a in Eq. (115) proves the norm bound. For Eq. (114), use $\sqrt{V}\sqrt{r}/\gamma = r\sqrt{\mathsf{k}_1}$ from Eq. (109) and $c_0 = \gamma/(2\pi)$. □

For $\rho \in \mathbb{R}^r$ with $|\rho| \leq 1$ let

$$f(\rho) := \sum_{i=1}^r \Big(\frac{\pi}{2} + \arcsin\rho_i\Big) - s(\rho)\sum_{i=1}^r \rho_i,$$
$$\Psi = \Psi(r) := \frac{\pi}{2} - 1 + \frac{4}{3\sqrt{6}}\sqrt{r},$$

and $\underline{f} := \frac{r\pi}{2} - \Psi$, $\overline{f} := \frac{r\pi}{2} + \Psi$. ($f$ is also used in Section G.2.4.)

Lemma 38 (The mean force)**.** *(a) $\bar{e}^\top g(\hat{w}) = \frac{c_0}{\sqrt{r}} f(\rho(w))$. (b) $|f(\rho) - \frac{r\pi}{2}| \leq \Psi(r)$ for every $|\rho| \leq 1$; so $0 < \underline{f} \leq f \leq \overline{f}$ for $r \geq 2$.*

Proof. Part (a): Project the teacher force. By Eq. (89), $g(\hat{w}) = c_0\big[E_U(\frac{\pi}{2}\mathbf{1}_r + \arcsin\rho) - s(\rho)\hat{w}\big]$, and $\bar{e}^\top E_U = \mathbf{1}_r^\top/\sqrt{r}$, $\bar{e}^\top\hat{w} = \sum_i \rho_i/\sqrt{r}$.

Part (b): Bound the deviation from the central value. Let $h(t) := \arcsin t - t = \sum_{n\geq 1} a_n t^{2n+1}$ with $a_n \geq 0$. Since $h(t)/t^3$ is nondecreasing in $|t|$, $|h(t)| \leq h(1)|t|^3 = (\frac{\pi}{2} - 1)|t|^3$. Write $a := |\rho| \leq 1$. Then

$$f(\rho) - \frac{r\pi}{2} = (1 - s(\rho))\sum_i \rho_i + \sum_i h(\rho_i),$$

with $0 \leq 1 - s \leq 1 - a^2/2$, $|\sum_i \rho_i| \leq \sqrt{r}\, a$ and $\sum_i|\rho_i|^3 \leq a^3$. Hence

$$|f(\rho) - \frac{r\pi}{2}| \leq \sqrt{r}\, a(1 - a^2/2) + (\frac{\pi}{2} - 1)a^3. \tag{116}$$

The first term on the right of Eq. (116) is at most $(\frac{2}{3})^{3/2}\sqrt{r}$ (maximum at $a = \sqrt{2/3}$) and the second at most $\frac{\pi}{2} - 1$. Finally $\underline{f} > 0$ for $r \geq 2$: $\frac{r\pi}{2} - \Psi(r)$ is increasing in r and equals $1.80\ldots$ at $r = 2$. □

Lemma 39 (MF mean-mode increment)**.** *Let $\mathcal{G} \cong S_r \times O(D)$ be the group of Proposition 21, let ν be $\mathcal{G}$-invariant with finite second moment and $\nu(\{0\}) = 0$, and put $y := y(\nu)$, $e := e(y)$ as in Eq. (114) and $u := \Psi/\underline{f}$. Every law ν_t of the MF trajectory satisfies these hypotheses, for every $\alpha \geq 0$ (Proposition 21(c) and Lemma 20). The mean-mode increment of a particle is $\delta\omega_1(w) := \bar{e}^\top(T_\nu(w) - w) = \bar{\eta}\,\bar{e}^\top\Sigma_\nu^{\dagger 1/2}F(w;\nu)$.*

(a) $\bar{e}$ is an eigenvector of Σ_ν with eigenvalue $\lambda_1 := \mathbb{E}_\nu[(\bar{e}^\top F)^2]$. If $e < \underline{f}$, then $\lambda_1 > 0$ and every particle has

$$\frac{\delta\omega_1(w)}{\bar{\eta}} = \frac{\bar{e}^\top F(w;\nu)}{\sqrt{\lambda_1}} \;\geq\; \frac{\underline{f} - e}{\overline{f} + e} \;> 0.$$

(b) If $e \leq \underline{f}$, then

$$p_{11}(\nu) = \frac{\mathbb{E}_\nu[\delta\omega_1]}{\bar{\eta}} \;\geq\; c(y) := \frac{1 - e(y)/\underline{f}}{\sqrt{1 + u^2} + e(y)/\underline{f}} \;\geq 0.$$

Set $p_{\mathrm{MF}}(y) := c(y)$ if $e(y) \leq \underline{f}$ and $p_{\mathrm{MF}}(y) := -1$ otherwise. Then p_{MF} is nonincreasing and $p_{11}(\nu) \geq p_{\mathrm{MF}}(y(\nu))$ for every such ν.

(c) $p_{\mathrm{MF}}(y; r)$ is nondecreasing in r. A sufficient condition for $e(y) < \underline{f}$ is $y < \underline{f}/(1.5143347\, r)$; the right side is $.5946\ldots$ at $r = 2$ and increases with r.

Proof. Part (a): Positivity of every mean increment. By $\mathcal{G}$-invariance, $Q\Sigma_\nu Q^\top = \Sigma_\nu$ for $Q \in \mathcal{G}$ (Proposition 21(a)), so $Q\Sigma_\nu\bar{e} = \Sigma_\nu Q\bar{e} = \Sigma_\nu\bar{e}$. The $\mathcal{G}$-fixed vectors are exactly $\mathrm{span}\{\bar{e}\}$: permutations fix only multiples of $\mathbf{1}_r$ in U, and $-I_{U^\perp} \in \mathcal{G}$ fixes no nonzero vector of $U^\perp$. Therefore

$$\Sigma_\nu\bar{e} = (\bar{e}^\top\Sigma_\nu\bar{e})\bar{e} = \lambda_1\bar{e},$$

and $\Sigma_\nu^{\dagger 1/2}\bar{e} = \lambda_1^{-1/2}\bar{e}$ if $\lambda_1 > 0$ (and $= 0$ if $\lambda_1 = 0$).

Write the mean force in its natural units:

$$a(w) := \frac{\sqrt{r}}{c_0}\bar{e}^\top F(w;\nu) = f(\rho(w)) + \delta a(w). \tag{117}$$

By Lemmas 38(a) and 37, $|\delta a| \leq e$, and $f \in [\underline{f}, \overline{f}]$ by Lemma 38(b). So $a \in [\underline{f} - e, \overline{f} + e]$. If $e < \underline{f}$ then $a > 0$ everywhere, and

$$\lambda_1 = \frac{c_0^2}{r}\mathbb{E}_\nu[a^2] > 0, \qquad \sqrt{\lambda_1} \leq \frac{c_0}{\sqrt{r}}(\overline{f} + e).$$

Part (b): Lower-bound the average increment. If $e = \underline{f}$, then $a \geq 0$, so $\delta\omega_1 \geq 0$ and the bound $c = 0$ holds. Let $e < \underline{f}$. With a from Eq. (117), $\mathbb{E}_\nu[\delta\omega_1]/\bar{\eta} = \mathbb{E}_\nu[a]/\|a\|_{L^2(\nu)}$. Let $M := \mathbb{E}_\nu[f] \geq \underline{f}$ and $\mathrm{sd}_\nu(f) := (\mathbb{E}_\nu[(f - \mathbb{E}_\nu[f])^2])^{1/2}$. Since f takes values in an interval of length 2Ψ, $\mathrm{sd}_\nu(f) \leq \Psi$; we omit the subscript ν below. Also $\mathbb{E}_\nu[a] \geq M - e$ and, by Minkowski's inequality,

$$\|a\|_{L^2(\nu)} \leq \|f\|_{L^2(\nu)} + e = \sqrt{M^2 + \mathrm{sd}(f)^2} + e.$$

Hence

$$\frac{\mathbb{E}_\nu[a]}{\|a\|_{L^2(\nu)}} \geq \frac{1 - e/M}{\sqrt{1 + \mathrm{sd}(f)^2/M^2} + e/M} \geq \frac{1 - e/\underline{f}}{\sqrt{1 + u^2} + e/\underline{f}}, \tag{118}$$

because $(1 - \epsilon)/(\sqrt{1 + v^2} + \epsilon)$ is decreasing in $\epsilon \in [0, 1]$ and in $v \geq 0$. The right side of Eq. (118) is $c(y)$. This function is decreasing in y and $c \geq 0$ on $\{e \leq \underline{f}\}$, so p_{MF} is nonincreasing; $p_{11} \geq -1$ always (Lemma 31(c)).

Part (c): Monotonicity in the teacher rank. The two parameters are

$$e/\underline{f} = \sqrt{2}\,\pi \mathsf{k}_1\, y/(\frac{\pi}{2} - \Psi/r), \qquad u = (\Psi/r)/(\frac{\pi}{2} - \Psi/r).$$

Both decrease in r, because $\Psi/r = (\frac{\pi}{2} - 1)/r + (\frac{2}{3})^{3/2}/\sqrt{r}$ does, and c decreases in $e/\underline{f}$ and in u. The set $\{e \leq \underline{f}\}$ grows with r for the same reason. At $r = 2$, $\underline{f}/(2 \cdot 1.5143347) \approx 1.800997/3.028669 \approx .59465$. □

G.3.4 The plateau

By the loss sandwich (Lemma 36), $L/V - 1$ is controlled by the output scales y, y_U, x. By Lemma 31, these grow at most by the clock $T(t)$, which is linear in αt. Two further facts give the constant: in MF the initial scales are known exactly, and the mean mode x does not decrease on the plateau window, so the linear term of the loss can only lower it.

Lemma 40 (Initial scales)**.** *For $\nu_0 = N(0, \sigma_0^2 I_d/d)$: $\mathbb{E}_{\nu_0}[ww^\top] = (\sigma_0^2/d) I_d$ and $\mathbb{E}_{\nu_0}[w] = 0$. So, exactly,*

$$y_0 = y_{U0} = y_0' := \frac{\alpha\sigma_0}{\sqrt{d}\sqrt{V}} = \frac{\alpha\bar{\eta}\kappa}{\sqrt{V}}, \qquad x(0) = 0.$$

Proof. Immediate from Definition 30 and $\sigma_0 = \kappa\bar{\eta}\sqrt{d}$ Eq. (91). □

Lemma 41 (Mean-mode monotonicity)**.** *If $y(t) < \underline{f}/(1.5143347\, r)$ (which is approximately .5946 at $r = 2$ and increases with r), then every particle's $\bar{e}^\top w$ increases at step t, and $x(t + 1) > x(t)$.*

Proof. Lemma 39(a) gives $\delta\omega_1 > 0$ for every particle, so $p_{11}(t) > 0$ and $x(t + 1) - x(t) = \Delta T\, p_{11}(t) > 0$ by Lemma 31(c). □

Theorem 42 (Plateau)**.** *Let $r \geq 2$, $d > r$, $\kappa > 0$ and $t_P \in \mathbb{N}$. Let $\bar{y}$ be the positive root of $1.17933\,\bar{y} + .34085\,\bar{y}^2 = .01$, so $\bar{y} = .0084587115\ldots$, and*

$$\alpha^*(t_P) := \frac{\bar{y}\sqrt{V}}{\bar{\eta}(\kappa + t_P)}.$$

For every $\alpha \in (0, \alpha^*(t_P)]$, the MF trajectory Eq. (92) from $\nu_0 = N(0, \sigma_0^2 I_d/d)$ satisfies

$$\max_{s,t \leq t_P} \frac{L(t)}{L(s)} \leq 1.01 \qquad \text{and} \qquad L(t) \geq .99V \quad (0 \leq t \leq t_P),$$

and its mean mode is nondecreasing on $[0, t_P + 1]$, with $x \geq 0$ there.

Proof. Step 1 (Control the output scales throughout the window). By Lemma 40, $y_0 = y_{U0} = \alpha\bar{\eta}\kappa/\sqrt{V}$ and $x(0) = 0$, and $T(t_P) = \alpha\bar{\eta}t_P/\sqrt{V}$. Put

$$Y := y_0 + T(t_P) = y_{U0} + T(t_P) = \alpha\bar{\eta}(\kappa + t_P)/\sqrt{V} \le \bar{y}. \tag{119}$$

By Lemma 31(b), $y(t) \le Y \le \bar{y} < .5946$ and $y_U(t) \le Y$ for $t \le t_P$. So by Lemma 41 the mean mode increases at every step $t \le t_P$, and $x(t) \ge x(0) = 0$ for $t \le t_P + 1$.

Step 2 (Apply the loss sandwich). By Lemma 36(a), for $s, t \le t_P$,

$$\begin{aligned} \frac{L(t)}{V} &\le 1 - c_x x(t) + \mathsf{k}_1 y(t)^2 \le 1 + \mathsf{k}_1 Y^2, \\ \frac{L(s)}{V} &\ge 1 - 2\sqrt{\mathsf{k}_1}\, y_U(s) \ge 1 - 2\sqrt{\mathsf{k}_1}\, Y. \end{aligned} \tag{120}$$

Now $2\sqrt{\mathsf{k}_1}Y \le 1.1676388\,\bar{y} < .00988$, which gives $L \ge .99V$.

Step 3 (Verify the one-percent ratio). Using Eq. (120), the required ratio is bounded by

$$\frac{L(t)}{L(s)} \le \frac{1 + \mathsf{k}_1 Y^2}{1 - 2\sqrt{\mathsf{k}_1}Y} \le 1.01 \iff \mathsf{k}_1 Y^2 + 1.01 \cdot 2\sqrt{\mathsf{k}_1}\, Y \le .01. \tag{121}$$

Since $\mathsf{k}_1 = .3408451\ldots \le .34085$ and $1.01 \cdot 2\sqrt{\mathsf{k}_1} = 1.1793151\ldots \le 1.17933$, the last inequality in Eq. (121) follows from $1.17933\,Y + .34085\,Y^2 \le .01$. This holds because the left side is increasing in $Y \ge 0$ and $Y \le \bar{y}$ by Eq. (119). □

Remark 43 (The constant $\bar{y}$)**.** *The window enters only through $\kappa + t_P$: $\alpha^*(t_P) = \bar{y}\sqrt{V}/(\bar{\eta}\kappa(1+\tau_P))$ with $\tau_P := t_P/\kappa$, and $\bar{y}$ does not depend on t_P, r or d. The decimal .008459, the root rounded to nearest, is* not *admissible:* $1.17933\,(.008459) + .34085\,(.008459)^2 = .0100003\ldots > .01$. *The plateau estimate uses the exact positive root of the displayed quadratic; its decimal expansion is descriptive and is not substituted for that root.*

G.3.5 The loss drop

The mechanism is the mean mode. By Lemma 39, x grows by at least $p_{\mathrm{MF}}(y)\,\Delta T$ per step. This lowers the upper loss bound linearly, while the only penalty, $\mathsf{k}_1 y^2$, is quadratic in the output scale. The linear part of the initial covariance cancels exactly.

Theorem 44 (Loss-drop bound)**.** *Consider the MF trajectory Eq. (92) from $\nu_0 = N(0, \sigma_0^2 I_d/d)$ with $\alpha > 0$, and let $t \in \mathbb{N}$ and $T := T(t)$. Then*

$$\frac{L(0) - L(t)}{V} \;\ge\; \mathcal{D}(T) := c_x \int_0^T p_{\mathrm{MF}}(y_0 + u)\, du - \mathsf{k}_1 (y_0 + T)^2 - e_{\mathrm{nl}},$$

with p_{MF} from Lemma 39(b) and $0 \le e_{\mathrm{nl}} \le c_{\mathrm{nl}}\sqrt{r/d}\, y_0$ (Lemma 36(c)).

Proof. Step 1 (Cancel the initial linear covariance). Lemma 36(b) at time 0 and (a) at time t give $L(0)/V \ge 1 - c_x x(0) - e_{\mathrm{nl}}$ and $L(t)/V \le 1 - c_x x(t) + \mathsf{k}_1 y(t)^2$. Subtracting,

$$\frac{L(0) - L(t)}{V} \ge c_x\big(x(t) - x(0)\big) - \mathsf{k}_1 y(t)^2 - e_{\mathrm{nl}}. \tag{122}$$

The linear part of the initial covariance has cancelled exactly by Eq. (111).

Step 2 (Integrate the mean-mode lower bound). By Lemma 31(c) and Lemma 39(b), which applies at every step because every ν_s is $\mathcal{G}$-invariant with $\nu_s(\{0\}) = 0$,

$$\begin{aligned} x(t) - x(0) = \Delta T \sum_{s<t} p_{11}(s) \;&\ge\; \Delta T \sum_{s<t} p_{\mathrm{MF}}(y(s)) \\ &\ge\; \Delta T \sum_{s<t} p_{\mathrm{MF}}(y_0 + s\Delta T) \;\ge\; \int_0^{t\Delta T} p_{\mathrm{MF}}(y_0 + u)\, du. \end{aligned} \tag{123}$$

The second inequality uses $y(s) \le y_0 + s\Delta T$ (Lemma 31(b)) and that p_{MF} is nonincreasing; the third is the left Riemann sum of a nonincreasing function. Substitute Eq. (123) into Eq. (122), using $y(t) \le y_0 + T$ and $c_x > 0$. □

Corollary 45 (Loss drop)**.** *Let $r \geq 2$, $d \geq 64r$, $\kappa \geq 1$, $t_P \in \mathbb{N}$ and $\alpha \in (0, \alpha^*(t_P)]$, with α^* and $\bar{y}$ as in Theorem 42. Put $T_2^* := .05086$. Then*

$$L(0) - L(t_2) \geq .03\, V \qquad \text{at} \qquad t_2 := \left\lceil \frac{T_2^* \sqrt{V}}{\alpha \bar{\eta}} \right\rceil,$$

that is, once the clock $T(t)$ has reached T_2^*. At $\alpha = \alpha^*(t_P)$ this gives $t_2 \leq C_R(\kappa + t_P) + 1$ with $C_R := T_2^*/\bar{y} = 6.0127\ldots \leq 6.013$; in particular $t_2 \leq 6.62\,\kappa + 1$ whenever $t_P \leq .1007\,\kappa$. For $0 < \alpha < \alpha^*(t_P)$, $t_2 = \lceil (\alpha^*(t_P)/\alpha)\, C_R(\kappa + t_P) \rceil$.

Proof. Step 1 (Reduce to the largest initial scale and smallest rank). Write $y_0' = y_0 = \alpha\bar{\eta}\kappa/\sqrt{V}$ (Lemma 40); $y_0' \leq \bar{y}$ because $\alpha \leq \alpha^*(t_P) \leq \bar{y}\sqrt{V}/(\bar{\eta}\kappa)$. Since $\sqrt{r/d} \leq \frac{1}{8}$, Theorem 44 gives $(L(0) - L(t))/V \geq \mathcal{D}(T(t); y_0', r)$, where

$$\mathcal{D}(T; y_0', r) := c_x \int_0^T p_{\mathrm{MF}}(y_0' + u; r)\, du - \mathsf{k}_1 (y_0' + T)^2 - \frac{c_{\mathrm{nl}}}{8}\, y_0'.$$

This function is nondecreasing in r, because $p_{\mathrm{MF}}(\cdot; r)$ is (Lemma 39(c)), and pointwise nonincreasing in y_0', because p_{MF} is nonincreasing and the last two terms decrease in y_0'. Hence

$$\mathcal{D}(T; y_0', r) \geq \mathcal{D}^*(T) := \mathcal{D}(T; \bar{y}, 2). \tag{124}$$

The function $\mathcal{D}^*$ is concave (the integral of a nonincreasing function is concave, and $-\mathsf{k}_1(\bar{y} + T)^2$ is concave), so its superlevel set $\{\mathcal{D}^* \geq .03\}$ is an interval. It therefore suffices to certify two endpoints for the lower bound in Eq. (124).

Step 2 (Bound the mean increment by a rational function). We certify the endpoints using rational bounds. Put

$$Y := .008458712, \qquad a_+ := 1.681668, \qquad b_+ := 1.246628,$$
$$p_-(y) := \frac{1 - a_+ y}{b_+ + a_+ y}. \tag{125}$$

The defining quadratic for $\bar{y}$ gives $\bar{y} < Y$. At $r = 2$, write $a = (e(y)/y)/\underline{f}$ and $b = \sqrt{1 + (\Psi/\underline{f})^2}$, so that $p_{\mathrm{MF}}(y; 2) = (1 - ay)/(b + ay)$ on the interval used here. The bounds

$$\sqrt{2} < 1.414214, \qquad \pi > 3.14159265358979323846264338327950288,$$
$$\Psi(2) \leq .570797 + .544332\sqrt{2}, \qquad e(y)/y \leq 3.028672$$

imply $a \leq a_+$ and $b \leq b_+$. These last two inequalities follow by rational cross-multiplication (and squaring for the second). Since $a_+(Y + .1) < 1$, the numerator stays positive and $p_{\mathrm{MF}}(y; 2) \geq p_-(y)$ for $0 \leq y \leq Y + .1$.

Step 3 (Integrate and bound the logarithm). Integrating the rational function in Eq. (125) and using $c_x \geq .856429$, $\mathsf{k}_1 \leq .340846$ and $c_{\mathrm{nl}}/8 \leq .0389013$, we obtain, for $0 \leq T \leq .1$,

$$\mathcal{D}^*(T) \geq .856429 \left[\frac{1 + b_+}{a_+} \log \frac{b_+ + a_+(Y + T)}{b_+ + a_+ Y} - T \right] - .340846(Y + T)^2 - .0389013Y. \tag{126}$$

For $z(T) := a_+ T/[2(b_+ + a_+ Y) + a_+ T]$, let

$$\ell_{12}(T) := 2 \sum_{j=0}^{11} \frac{z(T)^{2j+1}}{2j + 1},$$
$$Q(T) := .856429 \left[\frac{1 + b_+}{a_+} \ell_{12}(T) - T \right] - .340846(Y + T)^2 - .0389013Y. \tag{127}$$

The positive-term expansion

$$\log((1 + z)/(1 - z)) = 2 \sum_{j \geq 0} z^{2j+1}/(2j + 1),$$

valid for $0 \leq z < 1$, gives $\mathcal{D}^*(T) \geq Q(T)$ when substituted in Eq. (126).

Step 4 (Certify the two endpoints). All quantities defining Q at rational T are rational, and direct cross-multiplication in Eq. (127) gives

$$Q(.05086) > .030007, \qquad Q(.1) > .053264. \tag{128}$$

By Eq. (128) and concavity, $\mathcal{D}^*(T) \geq .03$ throughout $[.05086, .1]$.

Step 5 (Convert the clock bound to an integer time). By the definition of t_2,

$$T(t_2) \in [T_2^*, T_2^* + \Delta T], \qquad \Delta T = \alpha\bar{\eta}/\sqrt{V} \leq \bar{y}/(\kappa + t_P) < .0085.$$

Thus $T(t_2)$ lies in the certified interval and $L(0) - L(t_2) \geq .03V$. At $\alpha = \alpha^*(t_P)$, $\sqrt{V}/(\alpha\bar{\eta}) = (\kappa + t_P)/\bar{y}$, which gives the bound on t_2; for $t_P \leq .1007\kappa$, $6.0128 \cdot 1.1007 < 6.62$. The case $\alpha < \alpha^*(t_P)$ is the same identity. □

G.4 The $\alpha = 0$ trajectory from the Gaussian initialization

What this section proves. Facts about the $\alpha = 0$ trajectory from $\nu_0 = N(0, \kappa^2 I_d)$ that hold uniformly in $r \geq 2$, D and κ: the initial statistics (Proposition 47); per-particle monotonicity, in particular $|\xi|$ never decreases (Lemma 48); the $\perp$-statistic $\Gamma_\perp$ grows only through rare overshoots (the kick budget, Lemma 49); the mean mode grows linearly and lifts λ_1 (Lemma 50); and an exact three-part decomposition of the one-step change of $\Phi = (r-1)\Gamma_U$ (Lemma 51).

Where it is used. These are the inputs (F1)–(F7) of the certificate (Section G.5.2); Lemma 50 and Lemma 48(e) are also used in Step 2 of Section G.7. The conventions of Section G.2.4 are in force.

G.4.1 Setup and initial law

Recall $D = d - r$, $P_1 = \bar{e}\bar{e}^\top$ and $P_c = P_U - P_1$. In the reduced coordinates of Definition 23, $w = (\omega, \sqrt{D}\,\zeta n)$ with $n \in S^{D-1}$, $\omega_1 = \bar{e}^\top\omega$ and $\xi = P_c\omega$; teacher-space vectors and projectors are identified with their r-dimensional coordinate representations. Unsubscripted expectations and L^2 norms below use the particle law. In units $\bar{\eta} = 1$ the initial law is $\nu_0 = N(0, \kappa^2 I_d)$. Here χ_k denotes a chi random variable with k degrees of freedom, so $\mathbb{E}[\chi_k] = \sqrt{2}\,\Gamma(\frac{k+1}{2})/\Gamma(\frac{k}{2})$, and we put $v := \chi_D/\sqrt{D}$. Tilde units are $\tilde{\omega} := \omega/\kappa$, $\tilde{\zeta} := \zeta/\kappa$ and $\tau := t/\kappa$.

Lemma 46 (Initial reduced law)**.** *ν_0 is $O(d)$-invariant, hence $\mathcal{G}$-invariant, so the reduction of Section G.2.4 applies from $t = 0$. A reduced representation of $w(0)$ is*

$$\omega(0) = \kappa z, \qquad \zeta(0) = \kappa v, \qquad n,$$

with $z \sim N(0, I_r)$, $v = \chi_D/\sqrt{D}$ and n uniform on S^{D-1}, all independent. Equivalently $w(0) = \kappa\mathsf{g}$ with $\mathsf{g} \sim N(0, I_d)$. At $\alpha = 0$, every ν_t is $\mathcal{G}$-invariant (Proposition 21(c)) and $\Sigma_{\nu_t} \succ 0$ for every $t \geq 0$ (Lemma 26). So Propositions 24 and 27 hold at every $t \geq 0$.

Proof. $w_U(0) = \kappa z$ and $w_\perp(0) = \kappa\mathsf{g}_\perp$ with $\mathsf{g}_\perp \sim N(0, I_D)$ independent of z. The radius $|\mathsf{g}_\perp| = \chi_D$ and the direction $n := \mathsf{g}_\perp/|\mathsf{g}_\perp|$ are independent, and n is uniform. So $\zeta(0) = |w_\perp(0)|/\sqrt{D} = \kappa v$. □

Notation. Per particle we use the quantities of Definition 23: $|w|$, ρ, s, $b = s/|w|$, ω_1, $\xi = P_c\omega$, and the *contrast cosine* $k_p := |P_c\rho| = |\xi|/|w|$. By Proposition 24(b) the updates are

$$\begin{aligned} \delta\omega_1 &= \frac{f}{\|f\|}, & \delta\xi &= \frac{h}{\sqrt{\bar{a}}}, \quad \bar{a} = \frac{\mathbb{E}[|h|^2]}{r-1}, \\ \zeta^+ &= \zeta\Big(1 - \frac{b}{N}\Big), & N &= \|b\zeta\|. \end{aligned} \tag{129}$$

Normalizations at time t carry a subscript: N_t, $\bar{a}_t$, and b_t for the per-particle b. The statistics are

$$\Phi := \mathbb{E}\Big[\frac{|\xi|^2}{|w|}\Big] = (r-1)\Gamma_U, \qquad \Gamma_\perp := \frac{\mathbb{E}[|\zeta|]}{\sqrt{D - 1/2}}, \qquad \sigma_c^2 := \frac{\mathbb{E}[|\xi|^2]}{r-1},$$

and $\lambda_1, \lambda_c, \lambda_\perp$ are the eigenvalues of $\mathcal{M}(\nu_t)$ (Proposition 27).

Proposition 47 (Initial values)**.**

$$\Gamma_U(0) = \kappa\,\frac{\mathbb{E}[\chi_d]}{d}, \qquad \Phi(0) = \kappa(r-1)\frac{\mathbb{E}[\chi_d]}{d}, \qquad \Gamma_\perp(0) = \frac{\kappa\,\mathbb{E}[v]}{\sqrt{D-1/2}}, \qquad \sigma_c(0) = \kappa,$$

$$R_0 := \frac{\Gamma_U(0)}{\Gamma_\perp(0)} = \frac{\mathbb{E}[\chi_d]}{d}\cdot\frac{\sqrt{D}\sqrt{D-1/2}}{\mathbb{E}[\chi_D]}.$$

Proof. By Lemma 46, $w(0) = \kappa \mathrm{g}$. Write $\mathrm{g} = |\mathrm{g}|\,e$ with e uniform on S^{d-1} and independent of $|\mathrm{g}| = \chi_d$. The contrast statistic separates into a radius and a direction:

$$|\xi|^2/|w| = \kappa|\mathrm{g}|\,|P_c e_U|^2, \qquad \mathbb{E}[|P_c e_U|^2] = \mathrm{tr}(P_c)/d = (r-1)/d.$$

Thus $\Phi(0) = \kappa\,\mathbb{E}[\chi_d]\,(r-1)/d$. The remaining initial values follow from

$$\mathbb{E}[|\zeta(0)|] = \kappa\mathbb{E}[v], \qquad \sigma_c(0)^2 = \kappa^2\mathbb{E}[|P_c z|^2]/(r-1) = \kappa^2. \qquad \square$$

Lemma 48 (Per-particle facts)**.** *For every $t \geq 0$:*

(a) $\delta\omega_1(t) \in [c_1, 1/c_1]$ for every particle, with $c_1 = f_{\min}/f_{\max} \geq (\sqrt{r}-1)/(\sqrt{r}+1)$;

(b) $|\xi(t)|$ is nondecreasing in t for every particle;

(c) $\mathbb{E}[|\delta\xi(t)|^2] = r-1$, hence $\sigma_c(t) \leq \kappa + t$;

(d) $\mathbb{E}[(\delta\zeta(t))^2] = 1$, hence $\|\zeta(t) - \zeta(0)\| \leq t$;

(e) $|\xi(t)| > 0$ and $P_c\rho \neq 0$ for almost every particle, hence $\Phi(t) > 0$ and $\Gamma_U(t) > 0$.

Proof. Part (a): Mean increment. This is Lemma 25(i).

Part (b): Contrast monotonicity. The contrast update in Eq. (129) and Lemma 25(ii) give

$$|\xi^+|^2 - |\xi|^2 = 2\,\xi\cdot h/\sqrt{a} + |\delta\xi|^2 \geq 0. \tag{130}$$

Parts (c) and (d): Second-moment bounds. The identities are Proposition 24(c). By Minkowski,

$$\|\xi(t)\| \leq \|\xi(0)\| + t\sqrt{r-1} = \sqrt{r-1}(\kappa + t),$$
$$\|\zeta(t) - \zeta(0)\| \leq \sum_{s<t}\|\delta\zeta(s)\| = t.$$

Part (e): Strict positivity. Initially $\xi(0) = \kappa P_c z \neq 0$ almost surely. By Eq. (130), $|\xi(t)| \geq |\xi(0)|$; also $P_c\rho = \xi/|w|$. $\square$

G.4.2 The kick budget and the mean mode

Define the *overshoot mass* and the *kick budget*

$$\nu_{\mathrm{os}}(t) := \nu_t\big(b_t > 2N_t\big), \qquad K_\nu(t) := \frac{1}{\kappa}\sum_{s<t}\sqrt{\nu_{\mathrm{os}}(s)}.$$

Lemma 49 (Kick budget: $\Gamma_\perp$ grows only through overshoot)**.** *Every particle with $b_t \leq 2N_t$ has $|\zeta(t+1)| \leq |\zeta(t)|$. Moreover $\mathbb{E}[|\zeta(t+1)|] \leq \mathbb{E}[|\zeta(t)|] + \sqrt{\nu_{\mathrm{os}}(t)}$, hence $\Gamma_\perp(t) \leq \Gamma_\perp(0)\big(1 + K_\nu(t)/\mathbb{E}[v]\big)$.*

Proof. Step 1 (Isolate the overshoots). By Eq. (129), $\zeta^+ = \zeta(1 - b/N)$ with $b \geq 0$. If $b \leq 2N$ then $|1 - b/N| \leq 1$ and $|\zeta^+| \leq |\zeta|$. If $b > 2N$ then

$$|\zeta^+| - |\zeta| = |\zeta|(b/N - 2) \leq b|\zeta|/N.$$

Thus the pointwise bound is

$$|\zeta^+| - |\zeta| \leq \mathbf{1}\{b > 2N\}\,b|\zeta|/N. \tag{131}$$

Step 2 (Average and sum). Cauchy–Schwarz bounds the numerator in Eq. (131):

$$\mathbb{E}[\mathbf{1}\{b > 2N\}\, b|\zeta|] \le \sqrt{\nu_{\rm os}}\, \|b\zeta\| = \sqrt{\nu_{\rm os}}\, N.$$

Summing over $s < t$ gives

$$\mathbb{E}[|\zeta(t)|] \le \mathbb{E}[|\zeta(0)|] + \kappa K_\nu(t) = \kappa(\mathbb{E}[v] + K_\nu(t)).$$

Divide by $\sqrt{D - 1/2}$ and use $\Gamma_\perp(0) = \kappa\mathbb{E}[v]/\sqrt{D - 1/2}$ (Proposition 47). □

Lemma 50 (Mean mode)**.** *For every $t \ge 0$, $\mathbb{E}[\delta\omega_1(t)] \ge 2\sqrt{f_{\min} f_{\max}}/(f_{\min} + f_{\max}) \ge \sqrt{1 - 1/r}$. Hence $\mathbb{E}[\omega_1(t)] \ge \sqrt{1 - 1/r}\; t$ and $\lambda_1(t) \ge (\mathbb{E}[\omega_1(t)])^2/4$.*

Proof. Step 1 (The Pólya–Szegő bound). Let $Z \in [a, b]$ a.s. with $0 < a \le b$. Then $(Z-a)(b-Z) \ge 0$ gives $\mathbb{E}[Z^2] \le (a + b)\mathbb{E}[Z] - ab$. Together with $((a + b)\mathbb{E}[Z] - 2ab)^2 \ge 0$, this yields

$$(a + b)^2(\mathbb{E}[Z])^2 \ge 4ab[(a + b)\mathbb{E}[Z] - ab] \ge 4ab\,\mathbb{E}[Z^2].$$

Equivalently,

$$\mathbb{E}[Z]/\sqrt{\mathbb{E}[Z^2]} \ge 2\sqrt{ab}/(a + b). \tag{132}$$

Step 2 (Accumulate the mean increment). Apply Eq. (132) to $Z = f(\rho)$ under ν_t, with $a = f_{\min}$ and $b = f_{\max}$ (Lemma 25(i)). Then

$$\mathbb{E}[\delta\omega_1] = \mathbb{E}[f]/\|f\| \ge 2\sqrt{c_1}/(1 + c_1).$$

The map $c \mapsto 2\sqrt{c}/(1 + c)$ is nondecreasing on $(0, 1]$, and $c_1 \ge c_r := (\sqrt{r} - 1)/(\sqrt{r} + 1)$, for which $2\sqrt{c_r}/(1 + c_r) = \sqrt{(r - 1)/r}$. Since $\mathbb{E}[\omega_1(0)] = \kappa\,\mathbb{E}[\bar{e}^\top z] = 0$, we obtain

$$\mathbb{E}[\omega_1(t)] = \sum_{s<t} \mathbb{E}[\delta\omega_1(s)] \ge \sqrt{1 - 1/r}\, t. \tag{133}$$

Step 3 (Pass to the AGOP eigenvalue). To turn Eq. (133) into an eigenvalue bound, test with the constant function:

$$\lambda_1 = \mathbb{E}_X[(\bar{e}^\top \mathsf{v}(X))^2] \ge (\mathbb{E}_X[\bar{e}^\top \mathsf{v}(X)])^2,$$

$$\mathbb{E}_X[\bar{e}^\top \mathsf{v}(X)] = \mathbb{E}_w[\mathbb{P}_X(w^\top X > 0)\, \omega_1] = \frac{1}{2}\mathbb{E}[\omega_1]. \quad \square$$

G.4.3 The contrast step

Write $\varphi(x, y) := x^2/\sqrt{x^2 + y^2}$ for $x, y \ge 0$ (and $\varphi(0, 0) := 0$). For each particle put

$$y := (\omega_1^2 + D\zeta^2)^{1/2}, \qquad y_m := ((\omega_1 + \delta\omega_1)^2 + D\zeta^2)^{1/2}, \qquad y^+ := ((\omega_1 + \delta\omega_1)^2 + D(\zeta^+)^2)^{1/2}.$$

Since $|w|^2 = \omega_1^2 + |\xi|^2 + D\zeta^2$, we have $|\xi|^2/|w| = \varphi(|\xi|, y)$ and $|\xi^+|^2/|w^+| = \varphi(|\xi^+|, y^+)$. Define the contrast, mean-mode and $\perp$ steps

$$dU := \varphi(|\xi^+|, y) - \varphi(|\xi|, y),$$

$$dM := \varphi(|\xi^+|, y_m) - \varphi(|\xi^+|, y),$$

$$dP := \varphi(|\xi^+|, y^+) - \varphi(|\xi^+|, y_m).$$

For particles with $P_c\rho \ne 0$ put $F_\rho := P_c\rho \cdot h/k_p^2$ and $\hat{h}(k) := \min\big(k(2 - k^2), 1\big)$, and let $\ell := (1 - \rho_\infty^2)^{-1/2} - 1$.

Lemma 51 (One-step decomposition of Φ)**.** *For one step $t \to t + 1$, $\Phi(t + 1) - \Phi(t) = \mathbb{E}[dU + dM + dP]$ exactly, and:*

(i) $dU \ge \hat{h}(k_p)\, \Delta u$, where $\Delta u := \hat{\xi} \cdot \delta\xi = F_\rho k_p/\sqrt{\bar{a}} \ge 0$ and $\hat{\xi} := \xi/|\xi|$.

(ii) For every ρ with $P_c\rho \ne 0$: $F_\rho \ge 1 - s \ge 1 - |\rho|^2$, $(1 - |\rho|^2 - \ell)k_p \le |h| \le (1 + \ell)k_p$, and always $|h| \le \frac{\pi}{2} + 1$.

(iii) $dP \ge 0$ for every particle with $|\zeta^+| \le |\zeta|$, in particular for every particle with $b \le 2N$. Let $k' := |\xi^+|/(|\xi^+|^2 + y^2)^{1/2}$ and $y > 0$. Then

$$dM \;\ge\; -k'^2\,(y_m - y)_+, \qquad (y_m - y)_+ \le \min\Big(\delta\omega_1,\; \frac{\delta\omega_1(|\omega_1| + \delta\omega_1/2)}{y}\Big).$$

Proof. Step 1 (Telescope the three changes). The three steps telescope:

$$dU + dM + dP = |\xi^+|^2/|w^+| - |\xi|^2/|w|.$$

Since $\nu_{t+1} = (T_{\nu_t})_{\#}\nu_t$, taking ν_t-expectations gives $\Phi(t+1) - \Phi(t)$.

Step 2 (Record the derivatives). With $k := x/\sqrt{x^2+y^2} \in [0,1]$,

$$\partial_x \varphi = k(2-k^2) \geq 0, \qquad \partial_y \varphi = -k^2 y/\sqrt{x^2+y^2} \in [-k^2, 0]. \tag{134}$$

For fixed y, k is nondecreasing in x; for fixed x, k is nonincreasing in y. The function $k \mapsto k(2-k^2)$ increases on $[0, \sqrt{2/3}]$ and decreases on $[\sqrt{2/3}, 1]$ to the value 1, so on any interval $[k_0, 1]$ its minimum is $\hat{h}(k_0)$.

Step 3 (Bound the contrast change in part (i)). By Proposition 24(b), $\delta\xi = h/\sqrt{\bar{a}}$. Since $\hat{\xi} = P_c\rho/k_p$, $\hat{\xi} \cdot h = F_\rho k_p$, so $\Delta u = F_\rho k_p/\sqrt{\bar{a}}$, which is ≥ 0 by (ii). Next, $|\xi^+| \geq \hat{\xi} \cdot (\xi + \delta\xi) = |\xi| + \Delta u$, and φ is nondecreasing in x by Eq. (134), so

$$dU \geq \int_{|\xi|}^{|\xi|+\Delta u} \partial_x \varphi(x, y)\, dx.$$

Along the segment the cosine k increases from $k(|\xi|, y) = |\xi|/|w| = k_p$ and stays ≤ 1, so $\partial_x \varphi \geq \hat{h}(k_p)$ there.

Step 4 (Bound the contrast force in part (ii)). Dividing the identity of Lemma 25(ii) by $c_0|w|$ gives

$$P_c\rho \cdot h = \sum_i (\rho_i - \bar{\rho})(\arcsin \rho_i - \rho_i) + (1-s)k_p^2 \geq (1-s)k_p^2,$$

so $F_\rho \geq 1 - s \geq 1 - |\rho|^2$.

For the norm bound, put $\tilde{\phi}(x) := \arcsin x - x$ and $e := P_c\tilde{\phi}(\rho)$ (coordinatewise). Then

$$h = (1-s)P_c\rho + e. \tag{135}$$

On $[-\rho_\infty, \rho_\infty]$, $0 \leq \tilde{\phi}'(x) = (1-x^2)^{-1/2} - 1 \leq \ell$, so $\tilde{\phi}$ is ℓ-Lipschitz there. Using $|P_c x|^2 = \frac{1}{2r}\sum_{i,j}(x_i - x_j)^2$ gives $|e| \leq \ell k_p$. Therefore Eq. (135) implies

$$(1 - s - \ell)k_p \leq |h| \leq (1 - s + \ell)k_p,$$

and $0 \leq s \leq |\rho|^2$ gives the two bounds. Finally P_c is a contraction, $|\arcsin x| \leq \frac{\pi}{2}|x|$ and $s \leq |\rho|^2 \leq 1$, so $|h| \leq |\arcsin \rho| + s|\rho| \leq \frac{\pi}{2} + 1$.

Step 5 (Control the remaining changes in part (iii)). If $|\zeta^+| \leq |\zeta|$ then $y^+ \leq y_m$. By Eq. (134), φ is nonincreasing in y, so $dP \geq 0$; by Lemma 49 this holds whenever $b \leq 2N$.

For dM, fix $x := |\xi^+|$. If $y_m \leq y$ then $dM \geq 0$. If $y_m > y$, then

$$|\partial_y \varphi(x, y')| \leq k(x, y')^2 \leq k(x, y)^2 = k'^2 \qquad (y' \geq y),$$

so $dM \geq -k'^2(y_m - y)$. Finally, $y_m \leq y + \delta\omega_1$ by the triangle inequality ($\delta\omega_1 > 0$); if $y_m > y$,

$$y_m - y = \delta\omega_1(2\omega_1 + \delta\omega_1)/(y_m + y) \leq \delta\omega_1(2|\omega_1| + \delta\omega_1)/(2y). \qquad \square$$

(Only the one-sided bound on dM is used, and only it is claimed: when $y_m < y$, dM can be positive and larger than $k'^2\delta\omega_1(|\omega_1| + \delta\omega_1/2)/y$.)

G.5 The good-set certificate at $\alpha = 0$

What this section proves. Theorem 52: for the $\alpha = 0$ dynamics, at $t_F = \lfloor \tau_F \kappa \rfloor$ the top-r space of the MF AGOP is exactly U, with an explicit eigengap, for all $r \geq 2$, $d \geq 64r$, $D \geq 400$ and $\kappa \geq \kappa_{\min}(r, D)$. This section states the theorem and proves its analytic core, Proposition 58 ("CERT"): a static good set G of initial conditions is fixed; each of its particles carries explicit

envelopes; three super-solution inequalities close an induction over the window; and on G the contrast statistic Φ grows at a rate that the $\perp$-statistic cannot match. Section G.6 evaluates the certificate: in closed form for large D (region A) and by 80 interval-arithmetic certificates otherwise (region N).

Where it is used. Step 2 of the proof of Theorem 16 (Section G.7).

The certificate budget Ψ in slot (C7) is a local quantity, distinct from $\Psi(r)$, the mean-force deviation bound of Lemma 38.

G.5.1 Statement

Theorem 52 (Alignment at t_F, $\alpha = 0$)**.** *Consider the $\alpha = 0$ MF dynamics of Definition 22 from $\nu_0 = N(0, \kappa^2 I_d)$ (units $\bar{\eta} = 1$). Let $r \geq 2$, $d \geq 64r$ and $D = d - r \geq 400$. There are explicit numbers $\tau_F(r, D) \in (0, 0.101]$, $\kappa_{\min}(r, D) < \infty$ and $\delta(r, D) > 0$ such that, for every $\kappa \geq \kappa_{\min}(r, D)$, with $t_F := \lfloor \tau_F \kappa \rfloor$ (so $t_F \geq 2$),*

$$\lambda_c(t_F) \geq (1 + \delta)\, \lambda_\perp(t_F) \qquad \text{and} \qquad \lambda_1(t_F) \geq (1 + \delta)\, \lambda_\perp(t_F).$$

Hence $A_r(t_F) = 1$ for every leading-r selector, with the strict gap $\lambda_r/\lambda_{r+1} \geq 1 + \delta$. The constants come from two regions, which together cover every (r, D) with $D \geq \max(400, 63r)$.

(a) Region N*: $2 \leq r \leq 80$ and $\max(400, 63r) \leq D \leq D_A(r)$, with $D_A(r)$ the right end of the last D-interval of r in Table 12. Then $(\tau_F, \kappa_{\min}, 1+\delta)$ are the $(\tau_F, \kappa_0, \text{margin})$ of the row of Table 12 whose D-interval contains D. In particular $\tau_F \in [0.013740, 0.088861]$, $\kappa_{\min} \in [91, 777]$ as tabulated (the certified real floors lie in $[90.88, 776.91]$), and $1 + \delta \geq 1.006223$; moreover $1 + \delta \geq 1.012711$ for $r \geq 3$ and $1 + \delta \geq 1.079513$ for $r \geq 10$.*

(b) Region A*: $D \geq D_A(r)$; for $r \geq 81$ this is every $D \geq 63r$. With $M = M(r, D)$ and $\bar{A} = \bar{A}(r, D)$ the explicit numbers of Section G.6.2,*

$$\tau_F = \tau_1 := \frac{151}{750\, \bar{A}} \in [0.0802648,\ 0.100667] \cdot \frac{\sqrt{r^2 - 1}}{M}, \qquad \kappa_{\min} = \frac{4}{\tau_1},$$

and, at t_F,

$$\log \frac{\lambda_c}{\lambda_\perp} \geq 0.077216\, \frac{\sqrt{r^2 - 1}}{M} - \frac{1.1(r + 1)}{D} \geq 0.0176\, \frac{\sqrt{r^2 - 1}}{M},$$

$$\frac{\lambda_1}{\lambda_\perp} \geq \left(\frac{0.0601986\, \sqrt{r^2 - 1}\, \sqrt{\frac{\pi}{2}(1 - \frac{1}{r})(D - \frac{1}{2})}}{1.00106\, M} \right)^2 > 1.$$

Here $M = (r+4) + 2\sqrt{(r+4)L} + 2L$ with $L = \ln(D/(r-1)) + 5$, so $M \approx r + 2\sqrt{r \ln D} + 2 \ln D$. Thus $\tau_F = \Theta(r/\log D)$ for fixed r and $D \to \infty$, and $\tau_F \approx 0.08$–0.1 for large r.

The proof is completed in Section G.6.4. In region A, define $1 + \delta := \min\{\exp(0.0176\sqrt{r^2 - 1}/M), B_1\}$, where $B_1 > 1$ is the displayed lower bound for $\lambda_1/\lambda_\perp$ in (b).

G.5.2 Setting, inputs and conventions

Conventions. We work in MF units $\bar{\eta} = 1$. The MF AGOP has the block form $\lambda_1 P_1 \oplus \lambda_c P_c \oplus \lambda_\perp I_D$ (Proposition 27); here $\lambda_1, \lambda_c, \lambda_\perp$ are the eigenvalues of the surrogate $\mathcal{M}$ at $\alpha = 0$. At positive head scale the actual AGOP includes the common positive factor $\alpha^2 \bar{\eta}^2$, which cancels in every ratio. *In Sections G.5 and G.6, M denotes the squared-radius cutoff of the good set below, not the AGOP.* The letter ν denotes a mass bound; the MF law of the particles at time t is written ν_t. Unless a subscript specifies otherwise, $\mathbb{E}[\,\cdot\,]$ and $\mathbb{P}$ use this law (or the initial law for particle labels), and $\|Z\| = (\mathbb{E}[|Z|^2])^{1/2}$. For an event B, $\mathbb{E}[Z; B] := \mathbb{E}[Z\mathbf{1}\{B\}]$ is a restricted, unnormalized expectation; $\mathbb{E}[Z \mid q]$ denotes conditional expectation given q. In particular, $\mathbb{E}[Z; B \mid q] = \mathbb{E}[Z\mathbf{1}\{B\} \mid q]$.

Particles. As in Section G.4.1, a particle is labelled by its initial condition $p = (z, v, n)$: $\omega(0) = \kappa z$ with $z \sim N(0, I_r)$; $\zeta(0) = \kappa v$ with $v = \chi_D/\sqrt{D}$ independent of z; and n uniform on S^{D-1}, independent of (z, v). We write

$$q := |z|^2, \qquad z_c := P_c z, \qquad z_1 := \bar{e} \cdot z,$$

and use the tilde variables $\tilde{\omega} = \omega/\kappa$, $\tilde{\zeta} = \zeta/\kappa$, $\tilde{\xi} = \xi/\kappa$, $\tilde{w} = w/\kappa$ and $\tau = t/\kappa$. For each particle, $|w|^2 = |\omega|^2 + D\zeta^2$, $\rho = \omega/|w| \in \mathbb{R}^r$ (the U-part of $\hat{w}$), $\rho_\infty := \max_i |\rho_i|$, $\omega_1 := \bar{e} \cdot \omega$, $\xi := P_c\omega$, $u := |\xi|$, $k_p := |P_c\rho| = u/|w|$ (the contrast cosine), $s := \sum_i (1 - \sqrt{1-\rho_i^2})$ and $b := s/|w|$.

Dynamics ($\alpha = 0$; Proposition 24). The exact reduced updates are

$$\delta\omega_1 = \frac{f}{\|f\|}, \qquad f(\rho) := \sum_i \Big(\frac{\pi}{2} + \arcsin\rho_i\Big) - s\sum_i \rho_i;$$

$$\delta\xi = \frac{h}{\sqrt{a}}, \qquad h(\rho) := P_c(\arcsin\rho - s\rho), \qquad a := \frac{\mathbb{E}[|h|^2]}{r-1};$$

$$\zeta^+ = \zeta\Big(1 - \frac{b}{N}\Big), \qquad N := \|b\zeta\| = \Big(\frac{\mathbb{E}[s^2(1-|\rho|^2)]}{D}\Big)^{1/2}.$$

The vector n of each particle never changes. A superscript $+$ denotes the next step, and $\delta Z := Z^+ - Z$. Section G.4 writes $\bar{a}$ for a.

Statistics. $\Phi := \mathbb{E}[u^2/|w|] = (r-1)\Gamma_U$, $\Gamma_\perp := \mathbb{E}[|\zeta|]/\sqrt{D - \frac{1}{2}}$, $\mathbb{E}[v] = \mathbb{E}[\chi_D]/\sqrt{D}$ and $R_0 := \Gamma_U(0)/\Gamma_\perp(0) = (\mathbb{E}[\chi_d]/d)\sqrt{D}\sqrt{D - \frac{1}{2}}/\mathbb{E}[\chi_D]$. We also use $\sigma_c := (\mathbb{E}[|\xi|^2]/(r-1))^{1/2}$ and $\Gamma_1 := \mathbb{E}[\omega_1^2/|w|]$. Here Φ is the contrast statistic; the law-update map of Section G.2.1 and the teacher–student covariance of Section G.3.2 are separate, locally defined uses of this letter.

Facts used (all proved in Sections G.2 and G.4).

(F1) (Lemma 25, Lemma 48.) $|\xi_p|$ is nondecreasing for every particle; $\delta\omega_1 > 0$ for every particle; and $\mathbb{E}[|\delta\xi|^2] = r - 1$, so $\sigma_c(t) \le \kappa + t$.

(F2) (Proposition 27.) $\lambda_c \ge \Gamma_U^2/(2\pi)$ and $\lambda_\perp \le \Gamma_\perp^2/(2\pi)$. Hence $A_r = 1$ with a strict gap, for every leading selector, iff $\min(\lambda_1, \lambda_c) > \lambda_\perp$; then $\lambda_r/\lambda_{r+1} = \min(\lambda_1, \lambda_c)/\lambda_\perp$.

(F3) (Lemma 49.) With $\nu_{\mathrm{os}}(t) := \mathbb{P}(b_t > 2N_t)$ and $K_\nu(t) := \kappa^{-1}\sum_{s<t}\sqrt{\nu_{\mathrm{os}}(s)}$: $\Gamma_\perp(t) \le \Gamma_\perp(0)(1 + K_\nu(t)/\mathbb{E}[v])$. Every particle with $b \le 2N$ has $|\zeta^+| \le |\zeta|$.

(F4) (Lemma 50.) $\mathbb{E}[\omega_1(t)] \ge \sqrt{1 - 1/r}\; t$.

(F5) (Lemma 51.) Write $F(u, W) := u^2/\sqrt{u^2 + W^2}$ with $W^2 := \omega_1^2 + D\zeta^2$, so that $u^2/|w| = F(u, W)$. The per-particle increment of $u^2/|w|$ is exactly $dU + dM + dP$, with $dU := F(u^+, W) - F(u, W)$, $dM := F(u^+, W_m) - F(u^+, W)$, $W_m^2 := (\omega_1 + \delta\omega_1)^2 + D\zeta^2$, and $dP := F(u^+, W^+) - F(u^+, W_m)$. Moreover:

- $dU \ge \hat{h}(k_p)F_\rho k_p/\sqrt{a}$, where $\hat{h}(k) := \min(k(2-k^2), 1)$ and $F_\rho := P_c\rho \cdot h/k_p^2$;
- $F_\rho \ge 1 - s \ge 1 - |\rho|^2$;
- $(1 - |\rho|^2 - \ell)k_p \le |h| \le (1+\ell)k_p$ with $\ell := (1-\rho_\infty^2)^{-1/2} - 1$, and $|h| \le \pi/2 + 1$ always;
- $dP \ge 0$ whenever $|\zeta^+| \le |\zeta|$;
- $dM \ge -k'^2(W_m - W)_+$, where $k' := u^+/\sqrt{u^{+2} + W^2}$ and $(W_m - W)_+ \le \min\big(\delta\omega_1,\ \delta\omega_1(|\omega_1| + \delta\omega_1/2)/W\big)$.

(F6) $|f/(r\pi/2) - 1| \le e_f(|\rho|) := |\rho|(1 + 2|\rho|^2/\pi)/\sqrt{r}$. Indeed $|\sum_i \arcsin\rho_i| \le \frac{\pi}{2}\sum_i|\rho_i| \le \frac{\pi}{2}\sqrt{r}|\rho|$ and $|s\sum_i \rho_i| \le |\rho|^2\sqrt{r}|\rho|$, because $|\arcsin x| \le \frac{\pi}{2}|x|$ and $s \le |\rho|^2$.

(F7) (Proposition 47.) $\Phi(0) = \kappa(r-1)\mathbb{E}[\chi_d]/d$ and $\Gamma_\perp(0) = \kappa\,\mathbb{E}[v]/\sqrt{D - \frac{1}{2}}$.

G.5.3 AUXILIARY LEMMAS

Lemma 53 (Exact split of λ_1 and the pairing bound)**.** *Let the particle law be $O(D) \times S_r$-invariant with n uniform (for instance ν_t). Put $\varepsilon_p(X) := \mathbf{1}\{w_p^\top X > 0\} - \frac{1}{2}$ and $A_1(X) := \mathbb{E}_p[\varepsilon_p(X)\,\omega_{1p}]$, where $\omega_{1p} = \bar{e} \cdot \omega_p$. Then*

$$\lambda_1 = \tfrac{1}{4}(\mathbb{E}[\omega_1])^2 + \mathbb{E}_X\big[A_1(X)^2\big] \;\geq\; \tfrac{1}{4}(\mathbb{E}[\omega_1])^2 + \frac{\Gamma_1^2}{2\pi}, \qquad \Gamma_1 = \mathbb{E}\big[\omega_1^2/|w|\big].$$

Consequently, along the MF trajectory, at every $t \geq 0$ with $\tau = t/\kappa$,

$$\frac{\lambda_1}{\lambda_\perp} \geq \frac{\pi}{2}\Big(1 - \frac{1}{r}\Big)\frac{\tau^2(D - \frac{1}{2})}{(\mathbb{E}[v] + K_\nu)^2} + \Big(\frac{\Gamma_1\sqrt{D - \frac{1}{2}}}{\kappa\,(\mathbb{E}[v] + K_\nu)}\Big)^2. \tag{136}$$

The pairing bound for the cross term is part (iii) of Proposition 58.

Proof. Let $v(X) := \mathbb{E}_p[\mathbf{1}\{w_p^\top X > 0\} w_p]$, so that the MF AGOP is $\mathbb{E}_X[v(X)v(X)^\top]$ and, by the block form, $\lambda_1 = \mathbb{E}_X[(\bar{e}\cdot v(X))^2]$. Now $\bar{e}\cdot v(X) = \frac{1}{2}\mathbb{E}[\omega_1] + A_1(X)$. Since $\varepsilon_p(-X) = -\varepsilon_p(X)$ for almost every X and X is symmetric, $\mathbb{E}_X[A_1(X)] = 0$; expanding the square gives the identity. For the bound, apply Cauchy–Schwarz with the test function $\varphi(X) := \bar{e}^\top X$, for which $\mathbb{E}_X[\varphi(X)^2] = 1$: $\mathbb{E}_X[A_1^2] \geq (\mathbb{E}_X[A_1\varphi])^2$. Since $\mathbb{E}_X[X\mathbf{1}\{w^\top X > 0\}] = \hat{w}/\sqrt{2\pi}$ and $\mathbb{E}_X[\varphi] = 0$,

$$\mathbb{E}_X[A_1\varphi] = \mathbb{E}_p\big[\omega_{1p}\,\bar{e}^\top \mathbb{E}_X[\varepsilon_p(X)X]\big] = \mathbb{E}_p\Big[\omega_{1p}\frac{\bar{e}^\top \hat{w}_p}{\sqrt{2\pi}}\Big] = \frac{\mathbb{E}[\omega_1^2/|w|]}{\sqrt{2\pi}},$$

because $\bar{e}^\top \hat{w}_p = \omega_{1p}/|w_p|$. For Eq. (136), divide by $\lambda_\perp \leq \Gamma_\perp^2/(2\pi)$ (F2), use $\Gamma_\perp(t) \leq \Gamma_\perp(0)(1 + K_\nu/\mathbb{E}[v]) = \kappa(\mathbb{E}[v] + K_\nu)/\sqrt{D - \frac{1}{2}}$ ((F3), (F7)), and $(\mathbb{E}[\omega_1])^2 \geq (1 - 1/r)t^2$ (F4). □

Lemma 54 (Lemma H: golden-ratio refinement)**.** *If $|\rho_p|^2 \leq 0.38 < ((\sqrt{5} - 1)/2)^2 = 0.3819\ldots$, then*

$$\hat{h}(k_p)\,F_\rho\,k_p \;\geq\; q_p\,k_p^2, \qquad q_p := (2 - |\rho_p|^2)(1 - |\rho_p|^2),$$

and $2 - q_p = 3|\rho_p|^2 - |\rho_p|^4$.

Proof. The function $k \mapsto k(2 - k^2)$ increases on $[0, \sqrt{2/3}]$ and equals 1 at $k = (\sqrt{5} - 1)/2$ (the root in $(0, 1)$ of $k^3 - 2k + 1 = (k - 1)(k^2 + k - 1)$). Since $k_p \leq |\rho_p| < (\sqrt{5} - 1)/2$, we get $\hat{h}(k_p) = k_p(2 - k_p^2) \geq k_p(2 - |\rho_p|^2)$. Multiply by $F_\rho k_p$ and use $F_\rho \geq 1 - |\rho_p|^2 \geq 0$ (F5). □

For $y \in (0, 1)$ put $\iota(y) := y - 1 - \ln y > 0$, and for $n \geq 1$

$$\mathcal{U}(n, y) := \frac{(y\,e^{1-y})^{n/2}}{\sqrt{\pi n}\,(1 - y)} = \frac{e^{-n\iota(y)/2}}{\sqrt{\pi n}\,(1 - y)}.$$

Lemma 55 (Lemma V: v-averages and the χ^2 lower tail)**.** *Let $v = \chi_D/\sqrt{D}$ (so $\mathbb{E}[v^2] = 1$) and $v_{\min} < 1$.*

- *(a) Let φ be convex and nonincreasing on $(0, \infty)$ with $0 \leq \varphi(x) \leq \bar{\varphi}(x)$ for $x < v_{\min}^2$. Then $\mathbb{E}[\varphi(v^2); v \geq v_{\min}] \geq \varphi(1) - \mathbb{E}[\bar{\varphi}(v^2); v < v_{\min}]$. It is used with $\varphi(x) = D^k/(c + Dx)^k$, $\bar{\varphi}(x) = x^{-k}$ ($k = 1, 2$, $c \geq 0$), and with $\varphi(x) = (c + Dx)^{-1/2}$, $\bar{\varphi}(x) = (Dx)^{-1/2}$.*
- *(b) For $0 < y < 1$ and $n \geq 1$: $\mathbb{P}(\chi_n^2 < ny) \leq \mathcal{U}(n, y)$. $\mathcal{U}$ is decreasing in n and increasing in y.*
- *(c) Exactly, for $k = 1, 2$: $\mathbb{E}[v^{-2k}; v < v_{\min}] = D^k\mathbb{P}(\chi_{D-2k}^2 < Dv_{\min}^2)/\prod_{j \leq k}(D - 2j)$, and $\mathbb{E}[v^{-1}; v < v_{\min}] = (\sqrt{D}/\mathbb{E}[\chi_{D-1}])\,\mathbb{P}(\chi_{D-1}^2 < Dv_{\min}^2)$.*

(d) Put $y := v_{\min}^2$ and assume $Dy/(D-4) < 1$. For $v = \chi_{D'}/\sqrt{D'}$ at every $D' \geq D$:

$$\nu_v := \mathbb{P}(v < v_{\min}) \leq \mathcal{U}(D, y),$$
$$m_1 := \mathbb{E}[v^{-2}; v < v_{\min}] \leq \tfrac{D}{D-2}\,\mathcal{U}\big(D-2, \tfrac{Dy}{D-2}\big),$$
$$m_2 := \mathbb{E}[v^{-4}; v < v_{\min}] \leq \tfrac{D^2}{(D-2)(D-4)}\,\mathcal{U}\big(D-4, \tfrac{Dy}{D-4}\big),$$
$$m_h := \mathbb{E}[v^{-1}; v < v_{\min}] \leq \sqrt{\tfrac{D}{D-3/2}}\,\mathcal{U}\big(D-1, \tfrac{Dy}{D-1}\big),$$

and

$$D'\nu_v(D') \leq \sup_{D'' \geq D} \frac{\sqrt{D''}\, e^{-D''\iota(y)/2}}{\sqrt{\pi}(1-y)} = \begin{cases} \sqrt{D}\, e^{-D\iota(y)/2}/(\sqrt{\pi}(1-y)), & D\iota(y) \geq 1, \\ \iota(y)^{-1/2} e^{-1/2}/(\sqrt{\pi}(1-y)), & \textit{otherwise.} \end{cases}$$

Proof. Part (a): Removing the lower tail. Jensen's inequality and the bound on φ give

$$\mathbb{E}[\varphi(v^2)] \geq \varphi(\mathbb{E}[v^2]) = \varphi(1), \qquad \mathbb{E}[\varphi(v^2); v < v_{\min}] \leq \mathbb{E}[\bar{\varphi}(v^2); v < v_{\min}].$$

In the three uses, φ is convex and nonincreasing, and $\varphi \leq \bar{\varphi}$ because $c \geq 0$.

Part (b): A lower-tail probability bound. With $a = n/2$ and $x = ay < a$, $\mathbb{P}(\chi_n^2 < ny) = \gamma(a, x)/\Gamma(a)$. The positive series satisfies

$$\gamma(a, x) = x^a e^{-x} \sum_{j \geq 0} \frac{x^j}{a(a+1)\cdots(a+j)} \leq \frac{x^a e^{-x}}{a(1 - x/a)}, \tag{137}$$

since each ratio of consecutive terms is at most x/a. Thus Eq. (137) gives $\mathbb{P}(\chi_n^2 < ny) \leq x^a e^{-x}/(\Gamma(a+1)(1-y))$, and Stirling's lower bound $\Gamma(a+1) \geq \sqrt{2\pi a}(a/e)^a$ gives $(ay)^a e^{-ay}/(\sqrt{2\pi a}(a/e)^a(1-y)) = \mathcal{U}(n, y)$. Since $\ln y < y - 1$, $ye^{1-y} < 1$, so $\mathcal{U}$ decreases in n; and ye^{1-y} and $1/(1-y)$ increase on $(0, 1)$.

Part (c): Inverse moments on the tail. If f_n is the χ_n^2 density, then

$$x^{-k} f_D(x) = f_{D-2k}(x)\, 2^{-k} \frac{\Gamma(\frac{D}{2} - k)}{\Gamma(\frac{D}{2})} = \frac{f_{D-2k}(x)}{\prod_{j \leq k}(D - 2j)},$$
$$x^{-1/2} f_D(x) = f_{D-1}(x)\, \frac{\Gamma(\frac{D-1}{2})}{\sqrt{2}\,\Gamma(\frac{D}{2})} = \frac{f_{D-1}(x)}{\mathbb{E}[\chi_{D-1}]}.$$

Use $v^{-2k} = D^k(\chi_D^2)^{-k}$ and integrate these identities over the lower tail.

Part (d): Uniformity over $D' \geq D$. Combine (b) and (c), and use $\mathbb{E}[\chi_{D-1}] \geq \sqrt{D - 3/2}$ (Lemma 57). Every prefactor decreases in D. In $\mathcal{U}(D-j, Dy/(D-j))$ the first argument increases and the second decreases with D, so each bound decreases in D and the value at D bounds the value at every $D' \geq D$. For the last claim, $D'\nu_v(D') \leq \sqrt{D'}\, e^{-D'\iota/2}/(\sqrt{\pi}(1-y))$, and $x \mapsto \sqrt{x}\, e^{-x\iota/2}$ increases up to $x = 1/\iota$ and decreases afterwards. □

Lemma 56 (Lemma D: discrete growth)**.** *For $x \geq 0$, $\kappa > 0$ and every integer $t \geq 0$: $\prod_{s<t}\big(1 + \frac{x}{\kappa+s}\big) \geq \exp\big(x\ln(1 + t/\kappa) - x^2 t/(2\kappa^2)\big)$.*

Proof. Use $\ln(1+y) \geq y - y^2/2$ for $y \geq 0$, together with

$$\sum_{s<t}(\kappa+s)^{-1} \geq \int_0^t (\kappa+s)^{-1} ds = \ln(1 + t/\kappa), \qquad \sum_{s<t}(\kappa+s)^{-2} \leq t/\kappa^2.$$

The first inequality is the left Riemann sum of a decreasing function. □

Lemma 57 (Moments of χ distributions)**.** *Let $c_k := \mathbb{E}[\chi_k]/\sqrt{k}$.*

(a) c_k is increasing in k, and $\sqrt{1-1/(2k)} \le c_k \le 1$; equivalently $\mathbb{E}[\chi_k] \ge \sqrt{k-\frac{1}{2}}$ (Kershaw).

(b) $\mathbb{E}[\chi_k]\,\mathbb{E}[\chi_{k+1}] = k$, and $\mathbb{E}[1/\chi_k] = 1/\mathbb{E}[\chi_{k-1}] \le (k-\frac{3}{2})^{-1/2}$ for $k \ge 2$.

(c) For integers $D \in [D_{\rm lo}, D_{\rm hi}]$ and $d = D + r$: $\mathbb{E}[v] = c_D \in [c_{D_{\rm lo}}, c_{D_{\rm hi}}] \subset [\sqrt{1-1/(2D_{\rm lo})}, 1]$ and

$$R_0 = \frac{c_d}{c_D}\sqrt{\frac{D-\frac{1}{2}}{d}} \;\ge\; \frac{c_{D_{\rm lo}+r}}{c_{D_{\rm hi}}}\sqrt{\frac{D_{\rm lo}-\frac{1}{2}}{D_{\rm lo}+r}} \;\ge\; \sqrt{1-\frac{1}{2(D_{\rm lo}+r)}}\sqrt{\frac{D_{\rm lo}-\frac{1}{2}}{D_{\rm lo}+r}}.$$

Proof. Part (a): Bounds and monotonicity. $c_k \le 1$ is Jensen. Kershaw's inequality $\Gamma(x+1)/\Gamma(x+s) > (x+s/2)^{1-s}$ $(x>0,\,0<s<1)$, with $s=\frac{1}{2}$, $x=\frac{k-1}{2}$, gives $\mathbb{E}[\chi_k] = \sqrt{2}\,\Gamma(\frac{k+1}{2})/\Gamma(\frac{k}{2}) > \sqrt{k-\frac{1}{2}}$ for $k \ge 2$; for $k=1$, $\sqrt{2/\pi} > \sqrt{1/2}$. For monotonicity, with ψ the digamma function, $\frac{d}{dk}\ln c_k = \frac{1}{2}[\psi(\frac{k}{2}+\frac{1}{2}) - \psi(\frac{k}{2})] - \frac{1}{2k}$, and for $x>0$ the series of ψ gives

$$\begin{aligned}\psi(x+\tfrac{1}{2}) - \psi(x) &= \sum_{n\ge 0}\frac{1/2}{(x+n)(x+n+\frac{1}{2})} > \frac{1}{2}\sum_{n\ge 0}\frac{1}{(x+n)(x+n+1)}\\ &= \frac{1}{2}\sum_{n\ge 0}\Big(\frac{1}{x+n} - \frac{1}{x+n+1}\Big) = \frac{1}{2x}.\end{aligned}$$

With $x = k/2$ this gives $\frac{d}{dk}\ln c_k > 0$.

Part (b): Moment identities. The gamma-function formulas give

$$\mathbb{E}[\chi_k]\mathbb{E}[\chi_{k+1}] = \frac{2\Gamma(\frac{k}{2}+1)}{\Gamma(\frac{k}{2})} = k, \qquad \mathbb{E}[1/\chi_k] = \frac{\Gamma(\frac{k-1}{2})}{\sqrt{2}\,\Gamma(\frac{k}{2})} = \frac{1}{\mathbb{E}[\chi_{k-1}]}.$$

Apply part (a) to the last denominator.

Part (c): Bounds on a D-interval. This follows from (a), since $\sqrt{(D-\frac{1}{2})/(D+r)}$ increases in D. □

G.5.4 The certificate (Proposition CERT)

Certificate data. The certificate consists of the following numbers and derived quantities.

(C1) Parameters: $r \ge 2$, an integer $D \ge 5$, $M > 0$, $\theta \in (0,1)$, $\tau_1 > 0$, and a triple $\mathbf{x} = (\Omega, A, C') \in (0,\infty)^3$.

(C2) $v_{\min} := (\theta^3 + 3A\tau_1)^{1/3}$, assumed < 1, and the *static good set* $G := \{p:\; q \le M,\; v \ge v_{\min}\}$, a set of initial conditions fixed once and for all. Put $\nu_q := \mathbb{P}(\chi^2_r > M)$, let ν_v be any upper bound on $\mathbb{P}(v < v_{\min})$, and $\nu := \nu_q + \nu_v \ge \mathbb{P}(G^c)$, with $0 \le \nu < 1$.

(C3) $\rho_G^2 := \Omega^2/(\Omega^2 + D\theta^2)$, assumed ≤ 0.38; $\ell_G := (1-\rho_G^2)^{-1/2} - 1$; $e_G := e_f(\rho_G)$, assumed $< \frac{1}{2}$; and $c_G := (1-e_G)\sqrt{1-\nu}/(1+e_G)$.

(C4) $I(\tau) := \int_0^\tau (v_{\min}^3 - 3A\sigma)^{-1/3} d\sigma = \big(v_{\min}^2 - (v_{\min}^3 - 3A\tau)^{2/3}\big)/(2A)$ for $0 \le \tau \le \tau_1$.

(C5) Envelopes, for $0 \le \bar q \le M$ and $0 \le \tau \le \tau_1$:

$$\hat\Omega^2(\tau;\bar q) := \Big(\sqrt{\bar q} + \frac{\tau}{c_G}\Big)^2 e^{2C' I(\tau)}, \qquad \hat\theta(\tau;\bar q) := \Big(v_{\min}^3 - 3\tau A\,\frac{\hat\Omega^2(\tau;\bar q)}{\Omega^2}\Big)^{1/3},$$

$$\hat\rho^2(\tau;\bar q) := \frac{\hat\Omega^2}{\hat\Omega^2 + D\hat\theta^2},$$

$$\hat\ell := (1-\hat\rho^2)^{-1/2} - 1,$$

$$\hat\delta_M(\tau;\bar q) := \frac{1}{c_G}\min\Big(1, \frac{\hat\Omega + 1/(2c_G\kappa)}{\sqrt{D}\,\hat\theta}\Big).$$

Under (S-Ω) below, $\hat{\Omega} \le \Omega$, so the base of $\hat{\theta}$ is at least $v_{\min}^3 - 3\tau_1 A = \theta^3 > 0$ and $\hat{\theta} \ge \theta$. The envelope $\hat{\delta}_M$ is the only one that depends on κ.

(C6) Slots. In the expectations below, each particle's envelopes are taken at its own $\bar{q} = q$ and at $\tau = \tau_1$. Let $\mathcal{N}, \mathcal{K}_a, \underline{K} > 0$ and $Q_2, H_2, W_2 \ge 0$ satisfy

$$\mathcal{N} \le \tfrac{1}{4}\mathbb{E}\Big[|z_c|^4(1-\hat{\rho}^2)\frac{D^2}{(\hat{\Omega}^2 + Dv^2)^2}; G\Big],$$

$$\mathcal{K}_a \le \mathbb{E}\Big[(1-\hat{\rho}^2-\hat{\ell})_+^2 \frac{|z_c|^2 D}{\hat{\Omega}^2 + Dv^2}; G\Big],$$

$$\underline{K} \le \mathbb{E}\Big[\frac{|z_c|^2}{\hat{\Omega}^2 + Dv^2}; G\Big],$$

$$Q_2^2 \ge \mathbb{E}\big[(3\hat{\rho}^2 - \hat{\rho}^4)^2\hat{\rho}^2; G\big],$$

$$H_2 \ge \mathbb{E}\Big[\frac{\hat{\rho}^4}{1-\hat{\rho}^2}; G\Big],$$

$$W_2^2 \ge \mathbb{E}\big[\hat{\rho}^2\hat{\delta}_M^2; G\big].$$

(C7) Derived constants:

$$\beta_G := \frac{(1 - Q_2/(2\sqrt{\underline{K}}))_+}{\sqrt{1+\Psi}}, \qquad \Psi := \frac{H_2 + (\pi/2+1)^2\nu}{\underline{K}},$$

$$\lambda_L := \frac{(1 + C'/(\kappa\theta))^2 W_2}{2\beta_G\sqrt{r-1}}, \qquad \beta' := \beta_G(1-\lambda_L);$$

$$e_0 := \Big[\mathbb{P}(\chi^2_{r+2} > M)\,\mathbb{E}[1/\chi_D] + \frac{\mathbb{E}[\chi_r]}{r}\,\nu_v\Big]\frac{d}{\mathbb{E}[\chi_d]},$$

$$R(\tau) := R_0(1-e_0)\frac{\exp\big(2\beta'\ln(1+\tau) - 2\beta'^2\tau/\kappa\big)}{1 + \tau\sqrt{\nu}/\mathbb{E}[v]}.$$

(C8) Mean-eigenvalue slots, at $\tau \le \tau_1$ (envelopes at τ): with m_h as in Lemma 55(d) and $\bar{\vartheta}_r := \mathbb{E}[|\vartheta_1|]$ for ϑ uniform on S^{r-1},

$$F_p(\tau) := \Big(1 - \frac{(\hat{\Omega}^2(\tau;q) - q)_+}{2(q + Dv_{\min}^2)}\Big)_+,$$

$$\mathrm{main}(\tau) := \mathbb{E}\Big[\frac{q}{r}\,F_p(\tau)\Big(\frac{\sqrt{D - \frac{1}{2}}}{\sqrt{q+D}} - m_h\Big)_+;\ q \le M\Big],$$

$$X_p(\tau) := \frac{1+e_f(\hat{\rho})}{\hat{\theta}} - (1-e_f(\hat{\rho}))\,F_p(\tau)\Big(\frac{\sqrt{D-\frac{1}{2}}}{\sqrt{q+D}} - m_h\Big)_+,$$

$$\mathrm{cross}(\tau) := \frac{\tau\,\bar{\vartheta}_r\,\mathbb{E}[\sqrt{q}\,X_p(\tau); q \le M]}{\sqrt{1-\nu}\,(1-e_G)},$$

and $\mathcal{G}_1(\tau) := \mathrm{main}(\tau) - \mathrm{cross}(\tau)$.

Self-consistency (super-solution) inequalities:

$$\text{(S-}\Omega\text{)}\ \ \Omega \ge \big(\sqrt{M} + \tau_1/c_G\big)\,e^{C' I(\tau_1)},$$

$$\text{(S-}A\text{)}\ \ A \ge \frac{\Omega^2}{(1+\sqrt{1-\rho_G^2})\sqrt{\mathcal{N}}}, \qquad \text{(S-}C\text{)}\ \ C' \ge \frac{\sqrt{r-1}\,(1+\ell_G)}{\sqrt{\mathcal{K}_a}}.$$

Proposition 58 (Proposition CERT). *Assume the certificate data (C1)–(C8) satisfy (S-Ω), (S-A), (S-C), and let $\kappa \ge A/\theta^3$. Then for every integer t with $0 \le t \le \tau_1\kappa$, every $p \in G$ and $\tau = t/\kappa$:*

(I1) $\tilde{\zeta}_p$ *is nonincreasing on* $[0,t]$ *and* $\tilde{\zeta}_p(t) \ge \hat{\theta}(\tau;q_p) \ge \theta$*; at every step* $s<t$ *the particle neither flips sign nor overshoots (*$0 \le b_p \le N$*);*

(I2) $|\tilde{\omega}_p(t)| \le \hat{\Omega}(\tau;q_p) \le \Omega$*;*

(I3) $|\tilde{\xi}_p(t)| \ge |z_c|$*; and* $\delta\omega_{1,p}(t) \in [(1-e_f(\hat{\rho}))s_t, (1+e_f(\hat{\rho}))s_t]$ *with* $\hat{\rho} = \hat{\rho}(\tau;q_p)$ *and* $s_t := (r\pi/2)/\|f(t)\| \le 1/((1-e_G)\sqrt{1-\nu})$*.*

Moreover, put $K_G(t) := \mathbb{E}[k_p(t)^2;G]$ *and* $\Phi_G(t) := \mathbb{E}[u^2/|w|;G]$*, and assume* $\beta' > 0$ *and* $0 \le e_0 < 1$*. Then:*

(i) (Φ-growth.) $\Phi_G(t+1) \ge \Phi_G(t)\big(1 + 2\beta'/(\kappa+t)\big)$ *whenever* $t+1 \le \tau_1\kappa$*.*

(ii) (Ratio.) $\Gamma_U(t)/\Gamma_\perp(t) \ge R(\tau)$*, hence* $\lambda_c(t)/\lambda_\perp(t) \ge R(\tau)^2$*.*

(iii) (Mean eigenvalue.) $\dfrac{\lambda_1(t)}{\lambda_\perp(t)} \ge \dfrac{\pi}{2}\Big(1-\dfrac{1}{r}\Big)\dfrac{\tau^2(D-\frac{1}{2})}{(\mathbb{E}[v]+\tau\sqrt{\nu})^2} + \Big(\dfrac{(\mathcal{G}_1(\tau))_+}{\mathbb{E}[v]+\tau\sqrt{\nu}}\Big)^2$.

Proof. We first prove (I1), (I2) and (I3) of Proposition 58 by induction. Let $\mathcal{H}(t)$ be the statement that, for every $p \in G$, $\tilde{\zeta}_p$ is nonincreasing on $[0,t]$ and

$$\tilde{\zeta}_p(t) \ge \hat{\theta}(t/\kappa;q_p), \qquad |\tilde{\omega}_p(t)| \le \hat{\Omega}(t/\kappa;q_p), \qquad |\tilde{\xi}_p(t)| \ge |z_c|.$$

The initial values establish $\mathcal{H}(0)$:

$$\tilde{\zeta}_p(0) = v \ge v_{\min} = \hat{\theta}(0;q), \qquad |\tilde{\omega}_p(0)| = \sqrt{q} = \hat{\Omega}(0;q), \qquad \tilde{\xi}_p(0) = z_c.$$

Fix t with $t+1 \le \tau_1\kappa$ and assume $\mathcal{H}(s)$ for all $s \le t$. We first prove four facts at every time $s \le t$, and then $\mathcal{H}(t+1)$. Note that the envelopes are nondecreasing in τ ($\hat{\Omega}$, $\hat{\rho}$) or nonincreasing ($\hat{\theta}$), so every bound below that uses envelopes at $\tau = \tau_1$ holds at $s/\kappa \le \tau_1$.

Step 1 (Lower bound for N*).* Since $1-\sqrt{1-x} \ge x/2$, we have $s \ge |\rho|^2/2$ and therefore

$$N^2 = \frac{\mathbb{E}[s^2(1-|\rho|^2)]}{D} \ge \frac{\mathbb{E}[|\rho|^4(1-|\rho|^2);G]}{4D}.$$

On G, the induction hypothesis gives the two bounds

$$\begin{aligned} |\rho_p|^2 &\ge \frac{|\tilde{\xi}_p|^2}{|\tilde{w}_p|^2} \ge \frac{|z_c|^2}{\hat{\Omega}^2 + Dv^2}, \\ |\rho_p|^2 &= \frac{|\tilde{\omega}_p|^2}{|\tilde{\omega}_p|^2 + D\tilde{\zeta}_p^2} \le \hat{\rho}^2. \end{aligned} \tag{138}$$

For the first line, $|\tilde{w}_p|^2 = |\tilde{\omega}_p|^2 + D\tilde{\zeta}_p^2 \le \hat{\Omega}^2 + Dv^2$, because $\tilde{\zeta}_p$ is nonnegative and nonincreasing from v. The second line uses $\tilde{\zeta}_p \ge \hat{\theta}$. Substituting Eq. (138) into the N-bound and using slot (C6) yields

$$D^3N^2 \ge \mathcal{N}, \qquad D^{3/2}N \ge \sqrt{\mathcal{N}}. \tag{139}$$

Step 2 (Erosion). On G, $|\rho_p|^2 \le \hat{\rho}^2 \le \rho_G^2$ (as $\hat{\Omega} \le \Omega$ and $\hat{\theta} \ge \theta$), so $s = \sum_i \rho_i^2/(1+\sqrt{1-\rho_i^2}) \le |\rho|^2/(1+\sqrt{1-\rho_G^2})$; and $|w| \ge \sqrt{D}|\zeta| = \kappa\sqrt{D}\tilde{\zeta}$. The step decreases $\tilde{\zeta}_p$ by $\delta\tilde{\zeta}_p := \tilde{\zeta}_p b_p/N$, and

$$\delta\tilde{\zeta}_p = \frac{\tilde{\zeta}_p\, s}{|w|\,N} \le \frac{|\omega|^2}{(1+\sqrt{1-\rho_G^2})|w|^3N}\,\tilde{\zeta}_p \le \frac{A_p(s)}{\kappa\tilde{\zeta}_p^2},$$

$$A_p(s) := \frac{|\tilde{\omega}_p(s)|^2}{(1+\sqrt{1-\rho_G^2})D^{3/2}N} \le \frac{A|\tilde{\omega}_p(s)|^2}{\Omega^2},$$

by Eq. (139) and (S-A). Since $|\tilde{\omega}_p| \le \Omega$, $\tilde{\zeta}_p \ge \theta$ and $\kappa \ge A/\theta^3$:

$$\delta\tilde{\zeta}_p \le \frac{\tilde{\zeta}_p A}{\kappa\theta^3} \le \tilde{\zeta}_p.$$

Thus $b_p/N \in [0,1]$: no flip and no overshoot at step s. Moreover $(\tilde{\zeta}-\delta)^3 \geq \tilde{\zeta}^3 - 3\tilde{\zeta}^2\delta$ for $0 \leq \delta \leq \tilde{\zeta}$, so

$$\tilde{\zeta}_p(s+1)^3 \geq \tilde{\zeta}_p(s)^3 - \frac{3A\,\hat{\Omega}^2(s/\kappa; q_p)}{\kappa\,\Omega^2}. \tag{140}$$

Step 3 (Contrast). By (F5), $|\delta\xi| = |h|/\sqrt{a} \leq (1+\ell)k_p/\sqrt{a} \leq (1+\ell_G)|\xi|/(|w|\sqrt{a})$ on G, since $\ell \leq (1-|\rho|^2)^{-1/2} - 1 \leq \ell_G$. For the normalization, again by (F5) and Eq. (138), $|h| \geq (1-|\rho|^2-\ell)k_p \geq (1-\hat{\rho}^2-\hat{\ell})k_p$ and $k_p^2 \geq |z_c|^2/(\hat{\Omega}^2 + Dv^2)$ on G, so

$$(r-1)a = \mathbb{E}[|h|^2] \geq \mathbb{E}\big[(1-\hat{\rho}^2-\hat{\ell})_+^2 k_p^2; G\big] \geq \mathcal{K}_a/D.$$

With $|w| \geq \kappa\sqrt{D}\tilde{\zeta}$ and (S-C), this gives $|\delta\tilde{\xi}_p| \leq C'|\tilde{\xi}_p|/(\kappa\tilde{\zeta}_p)$, hence

$$|\tilde{\xi}_p(s+1)| \leq |\tilde{\xi}_p(s)|\Big(1 + \frac{C'}{\kappa\tilde{\zeta}_p(s)}\Big). \tag{141}$$

Step 4 (Mean mode). On G, (F6) gives

$$\frac{f_p}{r\pi/2} \in [1-e_f(\hat{\rho}), 1+e_f(\hat{\rho})] \subset [1-e_G, 1+e_G],$$

because e_f is increasing and $|\rho_p| \leq \hat{\rho} \leq \rho_G$. Consequently

$$\|f\|^2 \geq \mathbb{E}[f^2; G] \geq (r\pi/2)^2(1-e_G)^2(1-\nu), \qquad s_s = \frac{r\pi/2}{\|f(s)\|} \leq \frac{1}{(1-e_G)\sqrt{1-\nu}}.$$

The resulting increment bound is

$$\delta\omega_{1,p} = s_s\,\frac{f_p}{r\pi/2} \in [(1-e_f(\hat{\rho}))s_s, (1+e_f(\hat{\rho}))s_s] \subset (0, 1/c_G]. \tag{142}$$

This is the second part of (I3) at time s.

Step 5 (Closing the induction). By Step 2, $\tilde{\zeta}_p$ is nonincreasing up to $t+1$. Summing Eq. (140) over $s = 0, \ldots, t$ and using that $\hat{\Omega}$ is nondecreasing in τ and $v \geq v_{\min}$,

$$\tilde{\zeta}_p(t+1)^3 \geq v^3 - \frac{3(t+1)A\,\hat{\Omega}^2((t+1)/\kappa; q_p)}{\kappa\,\Omega^2} \geq \hat{\theta}\big(\tfrac{t+1}{\kappa}; q_p\big)^3,$$

and $\hat{\theta} \geq \theta$. With $\hat{\Omega} \leq \Omega$ the same sum gives $\tilde{\zeta}_p(s) \geq \vartheta(s/\kappa) := (v_{\min}^3 - 3As/\kappa)^{1/3}$ for $s \leq t$. Since $1/\vartheta$ is increasing, its left Riemann sum is below its integral, and iterating Eq. (141) gives

$$\begin{aligned} |\tilde{\xi}_p(t+1)| &\leq |z_c| \prod_{s\leq t}\Big(1 + \frac{C'}{\kappa\vartheta(s/\kappa)}\Big) \\ &\leq |z_c| \exp\Big(C' \sum_{s\leq t} \frac{1}{\kappa\,\vartheta(s/\kappa)}\Big) \leq |z_c|\, e^{C' I((t+1)/\kappa)}. \end{aligned} \tag{143}$$

By Eq. (142) and $\tilde{\omega}_1(0) = z_1$, $|\tilde{\omega}_{1,p}(t+1)| \leq |z_1| + (t+1)/(\kappa c_G)$. With $\tau' = (t+1)/\kappa$, $|z_1|^2 + |z_c|^2 = q$ and $e^{2C'I} \geq 1$, combine this with Eq. (143) to obtain

$$\begin{aligned} |\tilde{\omega}_p(t+1)|^2 = \tilde{\omega}_{1,p}^2 + |\tilde{\xi}_p|^2 &\leq \big(|z_1| + \tau'/c_G\big)^2 e^{2C'I(\tau')} + |z_c|^2 e^{2C'I(\tau')} \\ &\leq \big(\sqrt{q} + \tau'/c_G\big)^2 e^{2C'I(\tau')} = \hat{\Omega}^2(\tau'; q), \end{aligned}$$

and $\hat{\Omega}(\tau'; q) \leq \hat{\Omega}(\tau_1; M) \leq \Omega$ by (S-Ω). Finally $|\tilde{\xi}_p|$ is nondecreasing (F1), so $|\tilde{\xi}_p(t+1)| \geq |z_c|$. This proves $\mathcal{H}(t+1)$, hence (I1)–(I3).

Part (i): Φ-growth. Fix t with $t+1 \le \tau_1\kappa$; everything below is at time t. Since G is a fixed set of particles, $\Phi_G(t+1) - \Phi_G(t) = \mathbb{E}[dU + dM + dP; G]$ by (F5). On G: $dP \ge 0$ (no overshoot); $dU \ge q_p k_p^2/\sqrt{a}$ by (F5) and Lemma 54 ($|\rho_p|^2 \le \rho_G^2 \le 0.38$); and $dM \ge -k'^2(W_m - W)_+$. For the last term, $\delta\omega_1 \le 1/c_G$, $|\omega_1| \le \kappa\hat{\Omega}$ and $W \ge \sqrt{D}|\zeta| \ge \kappa\sqrt{D}\hat{\theta}$ give $(W_m - W)_+ \le \hat{\delta}_M$; and $u^+ \le u(1 + C'/(\kappa\theta))$ (Step 3) with $u^+ \ge u$ (F1) gives $k' \le (1 + C'/(\kappa\theta))k_p$. Hence

$$\Phi_G(t+1) - \Phi_G(t) \ge \frac{\mathbb{E}[q_p k_p^2; G]}{\sqrt{a}} - \Big(1 + \frac{C'}{\kappa\theta}\Big)^2 \mathbb{E}[k_p^2 \hat{\delta}_M; G]. \tag{144}$$

We bound the gain and loss in Eq. (144) separately.

For the numerator, Cauchy–Schwarz gives

$$\mathbb{E}[q_p k_p^2; G] = 2K_G - \mathbb{E}[(2 - q_p)k_p \cdot k_p; G] \ge 2K_G - Q_2\sqrt{K_G}. \tag{145}$$

Indeed, $(2 - q_p)^2 k_p^2 = (3|\rho|^2 - |\rho|^4)^2 k_p^2 \le (3\hat{\rho}^2 - \hat{\rho}^4)^2\hat{\rho}^2$ ($x \mapsto 3x - x^2$ increases on $[0,1]$, and $k_p \le |\rho| \le \hat{\rho}$). Also $\mathbb{E}[q_p k_p^2; G] \ge 0$. For the normalization budget,

$$(r-1)a = \mathbb{E}[|h|^2] \le \mathbb{E}[(1 + \ell_p)^2 k_p^2; G] + (\pi/2 + 1)^2\nu.$$

Here $(1 + \ell_p)^2 - 1 = \rho_\infty^2/(1 - \rho_\infty^2) \le \hat{\rho}^2/(1 - \hat{\rho}^2)$ and $k_p^2 \le \hat{\rho}^2$, so

$$(r-1)a \le K_G + H_2 + (\pi/2 + 1)^2\nu. \tag{146}$$

Combining Eq. (145) and Eq. (146), and using the nonnegativity of the numerator, the gain is at least

$$\begin{aligned} &2\sqrt{r-1}\sqrt{K_G}\,\frac{(1 - Q_2/(2\sqrt{K_G}))_+}{\sqrt{1 + (H_2 + (\pi/2+1)^2\nu)/K_G}} \\ &\quad \ge 2\beta_G\sqrt{r-1}\sqrt{K_G}. \end{aligned} \tag{147}$$

Each factor is nondecreasing in K_G, and $K_G \ge \underline{K}$ because $k_p^2 \ge |z_c|^2/(\hat{\Omega}^2 + Dv^2)$ on G. For the mean-mode loss, Cauchy–Schwarz and $k_p \le \hat{\rho}$ give

$$\mathbb{E}[k_p \cdot k_p \hat{\delta}_M; G] \le \sqrt{K_G}\,(\mathbb{E}[\hat{\rho}^2 \hat{\delta}_M^2; G])^{1/2} \le \sqrt{K_G}\,W_2. \tag{148}$$

Substitute Eq. (147) and Eq. (148) into Eq. (144). By the definition of β',

$$\Phi_G(t+1) - \Phi_G(t) \ge 2\beta'\sqrt{r-1}\sqrt{K_G}.$$

Finally, (F1) bounds the current statistic by

$$\Phi_G = \mathbb{E}[u\,k_p; G] \le (\mathbb{E}[u^2])^{1/2} K_G^{1/2} = \sqrt{r-1}\,\sigma_c\sqrt{K_G} \le \sqrt{r-1}(\kappa + t)\sqrt{K_G}.$$

Since $\beta' > 0$, these two inequalities yield

$$\Phi_G(t+1) - \Phi_G(t) \ge \frac{2\beta'\Phi_G(t)}{\kappa + t}. \tag{149}$$

Part (ii): Ratio. Iterating Eq. (149) and applying Lemma 56 with $x = 2\beta'$ gives

$$\begin{aligned} \Phi(t) \ge \Phi_G(t) &\ge \Phi_G(0) \prod_{s<t} \Big(1 + \frac{2\beta'}{\kappa + s}\Big) \\ &\ge \Phi_G(0) \exp\big(2\beta' \ln(1+\tau) - 2\beta'^2\tau/\kappa\big). \end{aligned} \tag{150}$$

We next compare the initial good-set statistic with the full statistic:

$$\Phi_G(0) \ge \Phi(0)(1 - e_0). \tag{151}$$

To prove Eq. (151), at $t = 0$ we have $u^2/|w| = \kappa|z_c|^2/|g|$ with $g = (z, \sqrt{D}v\,n) \sim N(0, I_d)$, and $G^c \subset \{q > M\} \cup \{v < v_{\min}\}$. On $\{q > M\}$ use $|g| \ge \sqrt{D}v = \chi_D$ and the independence of z and v: $\mathbb{E}[|z_c|^2; q > M] = (r-1)\mathbb{P}(\chi^2_{r+2} > M)$, since $|z_c|^2/q$ has mean $(r-1)/r$ independently of q and $\mathbb{E}[q; q > M] = r\mathbb{P}(\chi^2_{r+2} > M)$. On $\{v < v_{\min}\}$ use $|g| \ge |z|$ and $\mathbb{E}[|z_c|^2/|z|] = (r-1)\mathbb{E}[\chi_r]/r$. Dividing by $\Phi(0) = \kappa(r-1)\mathbb{E}[\chi_d]/d$ (F7) proves Eq. (151).

For the denominator, particles of G never overshoot by (I1). Thus

$$\nu_{\mathrm{os}}(s) \le \mathbb{P}(G^c) \le \nu, \qquad K_\nu(t) \le \tau\sqrt{\nu},$$

and (F3) yields

$$\Gamma_\perp(t) \le \Gamma_\perp(0)\Big(1 + \frac{\tau\sqrt{\nu}}{\mathbb{E}[v]}\Big). \tag{152}$$

Combine Eq. (150), Eq. (151) and Eq. (152). Since $\Gamma_U = \Phi/(r-1)$, $\Gamma_U(t)/\Gamma_\perp(t) \ge R(\tau) \ge 0$ (since $e_0 < 1$), and (F2) gives $\lambda_c/\lambda_\perp \ge (\Gamma_U/\Gamma_\perp)^2 \ge R(\tau)^2$.

Part (iii): Mean eigenvalue. We bound the terms of Eq. (136) with $K_\nu \le \tau\sqrt{\nu}$ (as in (ii)); it remains to show $\sqrt{D - \frac{1}{2}}\,\Gamma_1/\kappa \ge \mathcal{G}_1(\tau)$. First, $\Gamma_1/\kappa = \mathbb{E}[\tilde\omega_1^2/|\tilde w|] \ge \mathbb{E}[\tilde\omega_1^2/|\tilde w|; G]$. Write $\tilde\omega_{1,p}(t) = z_1 + m_p(t)$ with $m_p(t) := \kappa^{-1}\sum_{s<t}\delta\omega_{1,p}(s)$. By (I3) and the monotonicity of $\hat\rho$ in τ, $m_p \in [(1-e_f(\hat\rho))S_t, (1+e_f(\hat\rho))S_t]$, where $\hat\rho = \hat\rho(\tau; q_p)$ and $S_t := \kappa^{-1}\sum_{s<t} s_s \le \bar S := \tau/((1-e_G)\sqrt{1-\nu})$. Then $\tilde\omega_1^2 \ge z_1^2 + 2z_1 m_p$. We treat the main term and the cross term in this lower bound separately.

For the main term, put $|g|^2 := q + Dv^2$ and $\Delta_p := \hat\Omega^2(\tau; q) - q$. On G, $|\tilde w|^2 \le \hat\Omega^2 + Dv^2 = |g|^2 + \Delta_p$, so

$$\begin{aligned}\frac{1}{|\tilde w|} &\ge |g|^{-1}(1 + \Delta_p/|g|^2)^{-1/2} \\ &\ge |g|^{-1}(1 - \Delta_p/(2|g|^2))_+ \ge \frac{F_p(\tau)}{|g|},\end{aligned} \tag{153}$$

since $|g|^2 \ge q + Dv_{\min}^2$ on G. Given q, z_1^2 has mean q/r and v is independent; Lemma 55(a) with $\varphi(x) = (q+Dx)^{-1/2}$, $\bar\varphi(x) = (Dx)^{-1/2}$ gives $\mathbb{E}[(q+Dv^2)^{-1/2}; v \ge v_{\min}] \ge (q+D)^{-1/2} - m_h/\sqrt{D}$. Multiplying by $\sqrt{D - \frac{1}{2}}$ and using $\sqrt{(D-\frac{1}{2})/D} \le 1$, we obtain

$$\sqrt{D - \frac{1}{2}}\,\mathbb{E}[z_1^2/|\tilde w|; G] \ge \mathrm{main}(\tau). \tag{154}$$

For the cross term, use pairing: G and all envelopes depend on (q, v) only, and the initial law is invariant under $z_1 \mapsto -z_1$. Pair each particle p with $z_1 > 0$ with its mirror p' (same z_c, v, n, and $-z_1$): both lie in G and have the same envelopes. Using Eq. (153) for p and the lower bound $|\tilde w_{p'}| \ge \sqrt{D}\tilde\zeta_{p'} \ge \sqrt{D}\hat\theta$ for its mirror gives

$$\frac{m_p}{|\tilde w_p|} \ge (1-e_f)S_t\,\frac{F_p}{|g|}, \qquad \frac{m_{p'}}{|\tilde w_{p'}|} \le \frac{(1+e_f)S_t}{\sqrt{D}\hat\theta}.$$

Thus the pair contributes

$$\begin{aligned}2z_1\Big(\frac{m_p}{|\tilde w_p|} - \frac{m_{p'}}{|\tilde w_{p'}|}\Big) &\ge -2z_1 S_t\, B_p, \\ B_p &:= \frac{1+e_f}{\sqrt{D}\hat\theta} - (1-e_f)\frac{F_p}{|g|}.\end{aligned} \tag{155}$$

The bracket is pointwise nonnegative, because $F_p/|g| \le 1/|g| \le 1/(\sqrt{D}v_{\min}) \le 1/(\sqrt{D}\hat\theta)$ ($\hat\theta \le v_{\min}$). So S_t in Eq. (155) may be replaced by its upper bound $\bar S$. Averaging over v with Lemma 55(a) as in the main term (which only lowers the subtracted part), and multiplying by $\sqrt{D - \frac{1}{2}} \le \sqrt{D}$, gives $\sqrt{D - \frac{1}{2}}\,\mathbb{E}[B_p; v \ge v_{\min} \mid q] \le X_p(\tau)$. Finally $\mathbb{E}[2z_1\mathbf{1}\{z_1 > 0\} \mid q] = \mathbb{E}[|z_1| \mid q] = \sqrt{q}\,\bar\vartheta_r$. Hence

$$\sqrt{D - \frac{1}{2}}\,\mathbb{E}[2z_1 m_p/|\tilde w|; G] \ge -\mathrm{cross}(\tau). \tag{156}$$

Adding Eq. (154) and Eq. (156) gives $\sqrt{D - \frac{1}{2}}\,\Gamma_1/\kappa \ge \mathcal{G}_1(\tau)$, and Eq. (136) gives (iii). □

Remark 59 (Global and per-cell versions)**.** *Proposition 58 leaves the slots (C6), (C8) free. The* global *(single-cell) version replaces every envelope by its worst case over G and every expectation by exact χ^2 moments; it is used in Lemma 64. The* per-cell *version evaluates the expectations as rigorous Riemann sums over q-cells (Lemma 60); it is used for region N. Both are instances of Proposition 58.*

G.6 Evaluating the certificate

What this section proves. The hypotheses of Proposition 58 hold, with a margin > 1, for every (r, D) with $D \geq \max(400, 63r)$; this completes the proof of Theorem 52 (Section G.6.4). Section G.6.1 turns the certificate into finitely many checks: the expectations become Riemann sums over q-cells (Lemma 60), one set of numbers certifies a whole D-interval (Lemma 61), and one floor κ_0 certifies all $\kappa \geq \kappa_0$ (Lemma 62). Region A (large D) is handled in closed form (Lemma 64); region N by 80 certificates checked in interval arithmetic (Proposition 65).

What is checked numerically. Proposition 58 and Lemmas 53–57, 60, 61 and 62 are proved by hand. The finite Region A endpoint checks have explicit rational bounds in Section G.8.1. The 80 region-N certificates are verified with outward interval enclosures, using the exact inputs and formulas printed below.

G.6.1 Cells, D-intervals and the floor

Notation for the evaluation. We retain the certificate definitions of Section G.5.4: $D = d - r$, $q = |z|^2$ for $z \sim N(0, I_r)$, $z_c = P_c z$, and the independent radial variable $v = \chi_D/\sqrt{D}$. The good set is $G = \{q \leq M,\ v \geq v_{\min}\}$; M is its squared-radius cutoff. The hatted quantities are the envelopes in (C5), and $\mathcal{N}, \mathcal{K}_a, \underline{K}, Q_2, H_2, W_2$ are the moment bounds in (C6). The notation $\mathbb{E}[Z; B] = \mathbb{E}[Z\mathbf{1}\{B\}]$ retains its unnormalized meaning. The constants m_1, m_2, m_h bound the lower-tail inverse moments in Lemma 55(d).

Lemma 60 (Lemma QC: q-cells)**.** *Split $[0, M]$ into cells $[q_a^{(i)}, q_b^{(i)}]$, $i = 1, \dots, n$, with weights $w_i := \mathbb{P}(\chi_r^2 \in [q_a^{(i)}, q_b^{(i)}])$ (any split is valid; region N uses the deterministic lattice partition of Section G.6.3). Write $\hat{\Omega}_i, \hat{\theta}_i, \hat{\rho}_i, \hat{\ell}_i, \hat{\delta}_{M,i}$ for the envelopes at $(\tau_1; q_b^{(i)})$, and put $b_1 := (r-1)/r$ and $b_2 := (r^2-1)/(r(r+2))$. With m_1, m_2, m_h from Lemma 55(d), the following are admissible slots:*

$$\begin{aligned}
\mathcal{N} &:= \tfrac{b_2}{4} \textstyle\sum_i w_i\, (q_a^{(i)})^2 (1 - \hat{\rho}_i^2) \big(D^2/(\hat{\Omega}_i^2 + D)^2 - m_2\big)_+, \\
\mathcal{K}_a &:= b_1 \textstyle\sum_i w_i (1 - \hat{\rho}_i^2 - \hat{\ell}_i)_+^2\, q_a^{(i)} \big(D/(\hat{\Omega}_i^2 + D) - m_1\big)_+, \\
D\underline{K} &:= b_1 \textstyle\sum_i w_i\, q_a^{(i)} \big(D/(\hat{\Omega}_i^2 + D) - m_1\big)_+, \\
Q_2^2 &:= \textstyle\sum_i w_i (3\hat{\rho}_i^2 - \hat{\rho}_i^4)^2 \hat{\rho}_i^2, \\
H_2 &:= \textstyle\sum_i w_i\, \hat{\rho}_i^4/(1 - \hat{\rho}_i^2), \\
W_2^2 &:= \textstyle\sum_i w_i\, \hat{\rho}_i^2 \hat{\delta}_{M,i}^2.
\end{aligned}$$

Likewise, at time τ (envelopes at $(\tau; q_b^{(i)})$), with $F_i := \big(1 - (\hat{\Omega}^2(\tau; q_b^{(i)}) - q_a^{(i)})_+/(2(q_a^{(i)} + D v_{\min}^2))\big)_+$ and $L_i := F_i\big(\sqrt{D - \tfrac{1}{2}}/\sqrt{q_b^{(i)} + D} - m_h\big)_+$,

$$\mathrm{main}(\tau) \geq \sum_i w_i \frac{q_a^{(i)}}{r} L_i, \qquad \mathrm{cross}(\tau) \leq \frac{\tau \bar{\vartheta}_r}{\sqrt{1 - \nu}\,(1 - e_G)} \sum_i w_i \sqrt{q_b^{(i)}}\, X_i,$$

$$X_i := \frac{1 + e_f(\hat{\rho})}{\hat{\theta}} - (1 - e_f(\hat{\rho})) L_i,$$

with $e_f(\hat{\rho})$ and $\hat{\theta}$ at $(\tau; q_b^{(i)})$. These are rigorous lower (resp. upper) Riemann sums for every n.

Proof. Every envelope is monotone in $\bar{q}$: $\hat{\Omega}$, $\hat{\rho}$ and $\hat{\delta}_M$ increase and $\hat{\theta}$ decreases. Given q, $|z_c|^2/q$ has the Beta$(\frac{r-1}{2}, \frac{1}{2})$ law, independent of q, with first two moments b_1 and b_2; and v is independent of z. Thus

$$\mathbb{E}[|z_c|^4 \mathbf{1}\{q \in \mathrm{cell}_i\}] = b_2 \mathbb{E}[q^2 \mathbf{1}\{\mathrm{cell}_i\}] \geq b_2 (q_a^{(i)})^2 w_i,$$

with analogous bounds for $|z_c|^2$ and z_1^2 ($\mathbb{E}[z_1^2 \mid q] = q/r$).

In each cell, bound the integrand below at the worst corner: use q_a in polynomial factors, and the envelopes at q_b in denominators and in $1-\hat{\rho}^2$. Lemma 55(a),(d) removes the v-average:

$$\begin{aligned}\mathbb{E}[D/(\hat{\Omega}^2+Dv^2);v\ge v_{\min}]&\ge D/(\hat{\Omega}^2+D)-m_1,\\ \mathbb{E}[D^2/(\hat{\Omega}^2+Dv^2)^2;v\ge v_{\min}]&\ge D^2/(\hat{\Omega}^2+D)^2-m_2.\end{aligned}\tag{157}$$

These cellwise moment bounds and Eq. (157) give the three lower slots.

For the upper slots, the integrands of Q_2^2, H_2 and W_2^2 are increasing in $\hat{\rho}^2$ and $\hat{\delta}_M$ and do not depend on v; bound them above by their values at q_b and drop the restriction $v\ge v_{\min}$. In cross, $X_p\ge 0$, so $\sqrt{q}\,X_p\le\sqrt{q_b}\,X_i$. □

In region N, $n=32000$ except in row $2b$, where $n=64000$; any partition gives a valid bound.

Lemma 61 (Lemma MON: one certificate per D-interval)**.** *Fix r, integers $D_{\rm hi}\ge D_{\rm lo}\ge 5$ and the numbers $(M,\theta,\tau_1,\mathbf{x})$ with $D_{\rm lo}v_{\min}^2/(D_{\rm lo}-4)<1$, and use the q-cells of Lemma 60 (independent of D). Evaluate every slot of (C3)–(C8) at $D=D_{\rm lo}$, with ν_v,m_1,m_2,m_h the bounds of Lemma 55(d) at $D_{\rm lo}$, except in two places, which use $D_{\rm hi}$:*

(a) in Ψ, the budget term $(\pi/2+1)^2\nu/\underline{K}=(\pi/2+1)^2D\nu/(D\underline{K})$ uses $D\nu\le\nu_qD_{\rm hi}+\sup_{D'\ge D_{\rm lo}}D'\nu_v(D')$;

(b) in the λ_1 bound (iii), $\mathbb{E}[v]$ is replaced by its upper bound $c_{D_{\rm hi}}\le 1$;

and $\mathbb{E}[v]$ in $R(\tau)$ is replaced by its lower bound $c_{D_{\rm lo}}$, and R_0 by the lower bound of Lemma 57(c). If (S-Ω), (S-A), (S-C) hold for these values, then for every integer $D\in[D_{\rm lo},D_{\rm hi}]$ they hold at D. Subject to the certificate-data domains, conclusions (I1), (I2) and (I3) of Proposition 58 hold at D; conclusions (i)–(iii) additionally require $\beta'>0$ and $0\le e_0<1$ for the conservative slot values computed in this way.

Proof. The parameters $(M,\theta,\tau_1,\mathbf{x})$, $v_{\min}$, G (as a set of (q,v)), $I(\tau)$ and the cell weights do not depend on D. We check the five groups of slots in turn.

Step 1 (Super-solution map). $\rho_G^2=\Omega^2/(\Omega^2+D\theta^2)$ decreases in D, hence so do ℓ_G and $e_G=e_f(\rho_G)$. The bound $\nu(D)=\nu_q+\mathcal{U}(D,v_{\min}^2)$ decreases, so $c_G(D)=(1-e_G)\sqrt{1-\nu}/(1+e_G)$ increases. Therefore $\hat{\Omega}^2=(\sqrt{\bar{q}}+\tau/c_G)^2e^{2C'I}$ decreases in D, and $\hat{\theta}^3=v_{\min}^3-3\tau A\hat{\Omega}^2/\Omega^2$ increases (this is the only, implicit, way $\hat{\theta}$ depends on D: through $c_G(D)$). So $\hat{\rho}^2$ and $\hat{\ell}$ decrease. The factors $D/(\hat{\Omega}^2+D)-m_1$ and $D^2/(\hat{\Omega}^2+D)^2-m_2$ increase, where m_1,m_2 are the $D_{\rm lo}$ bounds, valid at every $D\ge D_{\rm lo}$ by Lemma 55(d). Hence $\mathcal{N}$ and $\mathcal{K}_a$ of Lemma 60 increase, and the right-hand sides of (S-Ω), (S-A), (S-C) decrease in D: a super-solution at $D_{\rm lo}$ is one at every $D\ge D_{\rm lo}$. The side conditions $\rho_G^2\le 0.38$, $e_G<\frac{1}{2}$, $v_{\min}<1$ are preserved.

Step 2 (Budget ratios). Write them with D-scaled quantities. $D\underline{K}$ increases. For the Q_2 slot, along $x(D)=a/(a+bD)$ (the explicit D-dependence of $\hat{\rho}^2$ at fixed $a=\hat{\Omega}^2$, $b=\hat{\theta}^2$) one has $dx/dD=-x(1-x)/D$ and

$$\frac{d}{dD}\ln\big(Dx^3(3-x)^2\big)=\frac{1}{D}\Big[1-\Big(3-\frac{2x}{3-x}\Big)(1-x)\Big]<0\quad(0\le x\le\tfrac{1}{2}),\tag{158}$$

because the bracketed product is at least 1.3 there; and the implicit dependence ($\hat{\Omega}\downarrow$, $\hat{\theta}\uparrow$) only lowers x, while $Dx^3(3-x)^2$ increases in x on $[0,\frac{1}{2}]$. Together with Eq. (158), this shows that DQ_2^2 decreases. Likewise $\frac{d}{dD}\ln(Dx^2/(1-x))=(x-1)/D<0$, so DH_2 decreases. Hence $Q_2/\sqrt{\underline{K}}$ and $H_2/\underline{K}$ are largest at $D_{\rm lo}$. The only increasing term is $D\nu_q$ in Ψ, which is taken at $D_{\rm hi}$; $D\nu_v(D)$ is bounded by its supremum over $D\ge D_{\rm lo}$.

Step 3 (Loss). $\hat{\delta}_M$ decreases in D ($c_G\uparrow$, $\hat{\Omega}\downarrow$, $\hat{\theta}\uparrow$, $\sqrt{D}\uparrow$), and so do $\hat{\rho}$ and W_2. Hence β_G increases and λ_L decreases, so β' increases.

Step 4 (Ratio). R_0 is bounded below by Lemma 57(c). In e_0, Lemma 57(b) gives

$$\mathbb{P}(\chi^2_{r+2} > M)\,\mathbb{E}[1/\chi_D]\frac{d}{\mathbb{E}[\chi_d]} \le \mathbb{P}(\chi^2_{r+2} > M)\sqrt{\frac{d}{D-\frac{3}{2}}}\sqrt{\frac{d}{d-\frac{1}{2}}},$$

$$\nu_v\frac{d}{\mathbb{E}[\chi_d]} \le \frac{e^{-D\iota(y)/2}\sqrt{1+r/D}\sqrt{d/(d-\frac{1}{2})}}{\sqrt{\pi}(1-y)},$$

and both right-hand sides decrease in D. Also ν decreases, $\mathbb{E}[v] \ge c_{D_{\rm lo}}$ in $R(\tau)$, and β' increases.

Step 5 (Mean eigenvalue). $D-\frac{1}{2}$ increases; $\mathbb{E}[v] \le c_{D_{\rm hi}} \le 1$ in the denominators (this is the second slot that uses $D_{\rm hi}$); $\sqrt{D-\frac{1}{2}}/\sqrt{q+D}$ and F_p increase in D, while X_p, ν and e_G decrease. So $\mathcal{G}_1$ increases in D. □

Lemma 62 (The floor κ_0 and the certified margin)**.** *Let a certificate on $[D_{\rm lo}, D_{\rm hi}]$ satisfy the hypotheses of Lemma 61, set $\tau_F := \tau_1$, let $\kappa_0 \ge \max(A/\theta^3, 2/\tau_1, 2)$ (the certificates use this maximum, rounded up), and put $[\tau_a, \tau_b] := [\tau_F - 1/\kappa_0, \tau_F]$. Evaluate β' and $\hat\delta_M$ at $\kappa = \kappa_0$, and suppose $\beta' > 0$ and $0 \le e_0 < 1$. Define*

$$R_{\rm low} := R_0(1-e_0)\frac{\exp\big(2\beta'\ln(1+\tau_a) - 2\beta'^2\tau_b/\kappa_0\big)}{1+\tau_b\sqrt{\nu}/(\mathbb{E}[v])_{\rm lo}}, \tag{159}$$

$$\Lambda_{\rm low} := \frac{\pi}{2}\Big(1-\frac{1}{r}\Big)\frac{\tau_a^2(D_{\rm lo}-\frac{1}{2})}{((\mathbb{E}[v])_{\rm hi}+\tau_a\sqrt{\nu})^2} + \Big(\frac{\mathcal{G}_1(\tau_b)_+}{(\mathbb{E}[v])_{\rm hi}+\tau_b\sqrt{\nu}}\Big)^2, \tag{160}$$

with the slot values of Lemma 61 ($(\mathbb{E}[v])_{\rm lo} = c_{D_{\rm lo}}$, $(\mathbb{E}[v])_{\rm hi} = c_{D_{\rm hi}}$ or 1), and $\text{margin} := \min(R^2_{\rm low}, \Lambda_{\rm low})$. *Then for every integer $D \in [D_{\rm lo}, D_{\rm hi}]$ and every $\kappa \ge \kappa_0$, at $t_F := \lfloor \tau_F\kappa \rfloor \ge 2$: $\lambda_c(t_F)/\lambda_\perp(t_F) \ge \text{margin}$ and $\lambda_1(t_F)/\lambda_\perp(t_F) \ge \text{margin}$.*

Proof. $\kappa \ge \kappa_0 \ge A/\theta^3$, so Proposition 58 applies at every such D (Lemma 61), and $\tau_F\kappa \ge 2$.

Every κ-dependent factor improves as κ grows: $e^{-2\beta'^2\tau/\kappa}$, $(1+C'/(\kappa\theta))^2$ in λ_L, and $1/(2c_G\kappa)$ in $\hat\delta_M$. Thus $\beta'(\kappa) \ge \beta'(\kappa_0)$. At $t = t_F$, the floor is accounted for by

$$\tau = \frac{t_F}{\kappa} \in (\tau_F - 1/\kappa, \tau_F] \subset [\tau_a, \tau_b], \qquad t_F \le \tau_1\kappa.$$

Applying Lemma 56 with $x = 2\beta'(\kappa_0)$ in part (ii) of the proof of Proposition 58 gives $\Gamma_U/\Gamma_\perp \ge R_{\rm low}$, with $R_{\rm low}$ as in Eq. (159).

In part (iii), the mean part increases in τ, while $\mathcal{G}_1(\tau)$ decreases in τ: $\hat\Omega$ and $\hat\rho$ increase and $\hat\theta$ decreases, so F_p decreases, X_p increases and cross increases. Evaluate these two terms at τ_a and τ_b, respectively, to obtain $\Lambda_{\rm low}$ in Eq. (160). Conclude with (F2). □

G.6.2 Region A: the closed-form certificate

Parameter rule. For $r \ge 2$ and $D \ge 63r$ put $\theta := \frac{1}{2}$, $\bar v := \frac{9}{10}$,

$$L := \ln\frac{D}{r-1} + 5, \qquad M := (r+4) + 2\sqrt{(r+4)L} + 2L, \qquad \bar\Omega^2 := 2M, \qquad \varepsilon := \frac{2M}{2M + D/4},$$

$\ell := (1-\varepsilon)^{-1/2} - 1$ and $\nu_0 := e^{-L} = e^{-5}(r-1)/D$. Let ν_v, m_1, m_2, m_h be the bounds of Lemma 55(d) at this D and $y = \bar v^2 = 0.81$, and

$$\bar C' := \frac{1+\ell}{(1-\varepsilon-\ell)\sqrt{(1-\nu_0)\big(\frac{D}{\bar\Omega^2+D} - m_1\big)}}, \qquad \bar A := \frac{\bar\Omega^2}{(1+\sqrt{1-\varepsilon})\sqrt{\bar{\mathcal{N}}}},$$

$$\bar{\mathcal{N}} := \frac{(1-\varepsilon)(r^2-1)(1-\nu_0)}{4}\Big(\frac{D^2}{(\bar\Omega^2+D)^2} - m_2\Big).$$

Finally $\tau_1 := (\bar v^3 - \theta^3)/(3\bar A) = 151/(750\bar A) = 0.201333\ldots/\bar A$ (so that $v_{\min} = \bar v$), $\tau_F := \tau_1$ and $\kappa_A := 4/\tau_1$.

Conditions.

(R1) $D \geq 1000$;

(R2) $D \geq 32M$;

(R3) $0.06015\sqrt{r^2-1}\sqrt{\frac{\pi}{2}(1-\frac{1}{r})(D-\frac{1}{2})}/M \geq 1.0011$.

Lemma 63 (Validity range of the rule)**.** *(a) (R1)–(R3) are upward closed in D (at fixed r).*

(b) For every $r \geq 81$, (R1)–(R3) hold at $D = 63r$; put $D_A(r) := 63r$.

(c) For $2 \leq r \leq 80$, (R1)–(R3) hold at $D = D_A(r)$ (Appendix G.9).

Hence (R1)–(R3) hold for all $D \geq D_A(r)$, for every $r \geq 2$.

Proof. Part (a): Upward closure. $L'(D) = 1/D$, so $D\,dM/dD = \sqrt{(r+4)/L} + 2$. Since $L \geq 5$, $\sqrt{(r+4)/L} < r+4$ and $2 < 2L$, hence $M > D\,dM/dD$ and D/M increases. Also $2(\sqrt{(r+4)/L}+2) < M$ (as $2\sqrt{(r+4)/5} < r+4$ and $4 < 2L$), so D/M^2 increases. (R1) is trivially upward closed, (R2) is $D/M \geq 32$, and the left side of (R3) is a constant times $\sqrt{(D-\frac{1}{2})/D}\sqrt{D/M^2}$.

Part (b): Ranks $r \geq 81$. At $D = 63r$ with $r \geq 81$, $L = \ln(63r/(r-1)) + 5 \leq \ln(63 \cdot 81/80) + 5 \leq 9.1556$. (R2) reads $31r \geq 128 + 64L + 64\sqrt{(r+4)L}$; since $32(4+2L) \leq 714$ and $64\sqrt{L} \leq 193.7$, it suffices that $31r \geq 714 + 193.7\sqrt{r+4}$. This holds at $r = 81$, since $\sqrt{85} < 9.22$ gives $714 + 193.7\sqrt{85} < 2500 < 2511$, and the difference increases in r, since $193.7/(2\sqrt{85}) \leq 31$. (R1): $63r \geq 5103$. (R3): by (R2), $M \leq 63r/32$, so its left side is at least $0.06015 \cdot \frac{32}{63}\frac{\sqrt{r^2-1}}{r}\sqrt{\frac{\pi}{2}(1-\frac{1}{r})(63r-\frac{1}{2})}$, which is ≥ 2.7 at $r = 81$ and increasing in r.

Part (c): Ranks $2 \leq r \leq 80$. This follows from the finite rational endpoint certificate in Section G.8.1. The numerical inequalities in (b) also follow from the elementary bounds specified there. □

Lemma 64 (Lemma A: closed-form region A)**.** *Assume $d \geq 64r$ and (R1)–(R3). Then for every $\kappa \geq \kappa_A$, at $t_F = \lfloor \tau_1 \kappa \rfloor \geq 4$:*

$$\log\frac{\lambda_c}{\lambda_\perp} \geq 0.077216\,\frac{\sqrt{r^2-1}}{M} - \frac{1.1(r+1)}{D} \;\geq\; 0.0176\,\frac{\sqrt{r^2-1}}{M} > 0,$$

$$\frac{\lambda_1}{\lambda_\perp} \geq \left(\frac{0.0601986\,\sqrt{r^2-1}\,\sqrt{\frac{\pi}{2}(1-\frac{1}{r})(D-\frac{1}{2})}}{1.00106\,M}\right)^2 > 1,$$

and $\tau_1 \in [0.0802648,\ 0.100667] \cdot \sqrt{r^2-1}/M$.

Proof. We apply Proposition 58 with one cell and worst-case envelopes (the global version of Remark 59), with M, $\theta = \frac{1}{2}$, τ_1 and $\mathbf{x} = (\bar{\Omega}, \bar{A}, \bar{C}')$, at the given D (no D-interval is needed). Every terminating decimal used below as an upper or lower bound denotes the corresponding exact rational number. The following inequalities establish the required directions of these bounds.

Step 1 (A-1: Tails and ε). Laurent–Massart gives $\mathbb{P}(\chi^2_k - k \geq 2\sqrt{kx} + 2x) \leq e^{-x}$; with $k = r+4$, $x = L$ this is $\mathbb{P}(\chi^2_{r+4} > M) \leq e^{-L} = \nu_0$, and by stochastic order the same bound holds for χ^2_r and χ^2_{r+2}. Since $D \geq 63r \geq 63(r-1)$, $\nu_0 \leq e^{-5}/63 \leq 1.06952 \cdot 10^{-4}$. By (R2), $\varepsilon \leq 2M/(2M+8M) = 0.2$; hence $\ell \leq 0.118034$, $1 - \varepsilon - \ell \geq 0.681965$, and $e_f(\rho_G)\sqrt{r} = \sqrt{\varepsilon}(1 + 2\varepsilon/\pi) \leq 0.504155$, where $\rho_G^2 = \bar{\Omega}^2/(\bar{\Omega}^2 + D\theta^2) = \varepsilon$. Also $L \geq \ln 63 + 5 \geq 9.14313$, so $M \geq r + 22.2862$.

Step 2 (A-2: v-tails). By Lemma 55(d) at $D = 1000$, all bounds decrease in D and $1/\iota(0.81) \leq 48.2602 < 1000$. Thus, for every $D \geq 1000$,

$$\nu_v \leq 2.973 \cdot 10^{-6}, \qquad m_1 \leq 3.710 \cdot 10^{-6}, \qquad m_2 \leq 4.629 \cdot 10^{-6}, \qquad D\nu_v \leq 2.973 \cdot 10^{-3}.$$

In particular,

$$\nu \le \nu_0 + \nu_v \le 1.0993 \cdot 10^{-4}, \qquad \sqrt{\nu} \le 0.0104845. \tag{161}$$

Step 3 (A-3: Constants). $c_G \ge 0.474366$ (the worst case is $r = 2$, as $e_G \le 0.504155/\sqrt{r}$). (R2) gives $D/(\bar\Omega^2 + D) \ge 16/17$, hence

$$\bar C' \le 1.68998, \qquad \frac{2M}{\sqrt{r^2-1}} \le \bar A \le \frac{2.50837\,M}{\sqrt{r^2-1}}.$$

The lower bound on $\bar A$ uses that the square root in its denominator is at most 1 and $1 + \sqrt{1-\varepsilon} \le 2$. Therefore

$$\tau_1 \in [0.0802648, 0.100667]\frac{\sqrt{r^2-1}}{M}, \qquad \frac{\sqrt{r^2-1}}{M} < 1. \tag{162}$$

Step 4 (A-4: Super-solution). With one cell and $\hat\Omega \le \bar\Omega$, $\hat\rho^2 \le \varepsilon$, choose

$$\begin{aligned} \mathcal{N} &:= \frac{(1-\varepsilon)(r^2-1)(1-\nu_0)}{4}\Big(\frac{D^2}{(\bar\Omega^2 + D)^2} - m_2\Big), \\ \mathcal{K}_a &:= (1-\varepsilon-\ell)^2(r-1)(1-\nu_0)\Big(\frac{D}{\bar\Omega^2 + D} - m_1\Big). \end{aligned} \tag{163}$$

These slots are admissible because

$$\mathbb{E}[|z_c|^4; q \le M] = (r^2-1)\mathbb{P}(\chi^2_{r+4} \le M), \qquad \mathbb{E}[|z_c|^2; q \le M] = (r-1)\mathbb{P}(\chi^2_{r+2} \le M),$$

and both probabilities are at least $1 - \nu_0$. With Eq. (163), (S-A) and (S-C) hold with equality by the definitions of $\bar A$ and $\bar C'$.

It remains to check (S-Ω). We have $I(\tau_1) = (\bar v^2 - \theta^2)/(2\bar A) = 0.28/\bar A \le 0.14\sqrt{r^2-1}/M \le 0.14$, so

$$\frac{(\sqrt{M} + \tau_1/c_G)e^{\bar C' I(\tau_1)}}{\sqrt{M}} \le \Big(1 + \frac{0.100667}{0.474366\sqrt{24.2862}}\Big)e^{1.68998 \cdot 0.14} \le 1.32149 < \sqrt{2} = \frac{\bar\Omega}{\sqrt{M}}.$$

Also $\rho_G^2 = \varepsilon \le 0.2 < 0.38$, $1 - \varepsilon - \ell > 0.05$, $e_G < \frac{1}{2}$, $v_{\min} = 0.9 < 1$, and $\kappa_A = 4/\tau_1 = (3000/151)\bar A \ge 8\bar A = \bar A/\theta^3$.

Step 5 (A-5: Gain and loss). Take

$$\underline{K} := (r-1)(1-\nu_0)\Big(\frac{1}{\bar\Omega^2 + D} - \frac{m_1}{D}\Big) \ge \frac{0.9410(r-1)}{D}.$$

In the single-cell setting we bound the two budget slots of part (i) of the proof of Proposition 58 pointwise instead of through Q_2 and H_2 (the pointwise bounds are used inside that proof, before its Cauchy–Schwarz step): on G, $|\rho_p|^2 \le \varepsilon$ gives $2 - q_p \le 3\varepsilon - \varepsilon^2$ and $(1+\ell_p)^2 - 1 \le \varepsilon/(1-\varepsilon)$. Consequently

$$\mathbb{E}[q_p k_p^2; G] \ge (2 - 3\varepsilon + \varepsilon^2)K_G, \qquad (r-1)a \le \frac{K_G}{1-\varepsilon} + (\pi/2+1)^2\nu.$$

The gain is therefore at least $2\beta_G\sqrt{r-1}\sqrt{K_G}$ with

$$\beta_G := \frac{(1 - 1.5\varepsilon + 0.5\varepsilon^2)\sqrt{1-\varepsilon}}{\sqrt{1 + (1-\varepsilon)\Psi_0}} \;\ge\; \frac{(1 - 1.5\varepsilon + 0.5\varepsilon^2)\sqrt{1-\varepsilon}}{\sqrt{1 + \Psi_0}}, \tag{164}$$

where $\Psi_0 := (\pi/2+1)^2\nu/\underline{K}$. Now $(\pi/2+1)^2 \le 6.61$, and $D\nu \le D\nu_0 + D\nu_v \le e^{-5}(r-1) + 2.973 \cdot 10^{-3}$, so $\Psi_0 \le 6.61(e^{-5} + D\nu_v)/((1-\nu_0)(16/17 - m_1)) \le 0.068207$. Since $(1 - 1.5\varepsilon + 0.5\varepsilon^2)\sqrt{1-\varepsilon}$ decreases in ε, Eq. (164) gives $\beta_G \ge 0.643987/1.033541 \ge 0.623088$.

For the loss, $\kappa \ge \kappa_A$ gives $\kappa\theta \ge 2/\tau_1 \ge 19.8675$, so

$$\Big(1 + \frac{\bar C'}{\kappa\theta}\Big)^2 \le 1.17736, \qquad \hat\delta_M \le \frac{2\sqrt{\bar\Omega^2/D} + 1/(c_G\kappa\sqrt{D})}{c_G} \le \frac{0.5 + 0.00168}{c_G} \le 1.05758,$$

because $2\sqrt{\bar\Omega^2/D} = 2\sqrt{2M/D} \le \frac{1}{2}$ by (R2), $\kappa \ge 4/0.100667$ and $D \ge 1000$. With $W_2 := \sqrt{\varepsilon} \cdot 1.05758$, this yields

$$\lambda_L\sqrt{r-1} \le 0.446844, \qquad \beta' = \beta_G(1 - \lambda_L) \ge 0.344664. \tag{165}$$

The worst case is $r = 2$.

Step 6 (A-6: Ratio). First control the initial ratio. Lemma 57 gives

$$R_0^2 \geq \Big(1 - \frac{1}{2d}\Big)\frac{D - \frac{1}{2}}{d} = \frac{(2d-1)(2D-1)}{4d^2} \geq \frac{D-1}{D+r} = 1 - x,$$

where $x := (r+1)/(D+r)$; the middle inequality is equivalent to $2r + 1 \geq 0$. By (R1) and $D \geq 63r$,

$$x \leq \max(16/1015, 17/1024) \leq 0.016602$$

(the maximum is at $r = 16$). Therefore

$$-\log R_0^2 \leq \frac{x}{1-x} \leq 1.0169x \leq \frac{1.0169(r+1)}{D}. \tag{166}$$

Next control the initial good-set loss e_0. By Lemma 57(b), its first part is at most

$$\nu_0 \sqrt{\frac{d}{D - \frac{3}{2}}}\sqrt{\frac{d}{d - \frac{1}{2}}} \leq 1.0090\,\nu_0.$$

Indeed,

$$\frac{d}{D - \frac{3}{2}} = 1 + \frac{r + \frac{3}{2}}{D - \frac{3}{2}} \leq 1 + \frac{D/63 + \frac{3}{2}}{D - \frac{3}{2}}.$$

The last expression decreases in D, so it is at most $1 + (1000/63 + \frac{3}{2})/998.5 \leq 1.017400$. Also $d/(d - \frac{1}{2}) \leq 1000/999.5 \leq 1.000501$. Thus the product of the square roots is at most $(1.017400 \cdot 1.000501)^{1/2} \leq 1.00892$.

The second part satisfies

$$e_0^v := (\mathbb{E}[\chi_r]/r)\nu_v d/\mathbb{E}[\chi_d] \leq \frac{e^{-D\iota(0.81)/2}\sqrt{1 + 1/63}\sqrt{1000/999.5}}{\sqrt{2}\sqrt{\pi} \cdot 0.19}.$$

Here $De_0^v \leq 0.06702$ at $D = 1000$ and decreases in D. Combining the two parts gives

$$e_0 \leq 1.0090 \cdot 1.06952 \cdot 10^{-4} + 0.06702/1000 \leq 4 \cdot 10^{-4},$$

and hence

$$-2\log(1 - e_0) \leq 2.001\, e_0 \leq \frac{0.01361(r-1)}{D} + \frac{0.135}{D}. \tag{167}$$

Since $0.135 + 0.01361(r-1) \leq 0.0831(r+1)$ for $r \geq 2$ and $1.0169 + 0.0831 \leq 1.1$, the three D-terms in Eq. (166) and Eq. (167) sum to at most $1.1(r+1)/D$.

We now include the discrete growth and the floor. With $\kappa \geq \kappa_A = 4/\tau_1$,

$$\tau_a := \tau_1 - 1/\kappa \geq \frac{3}{4}\tau_1, \qquad \frac{4\beta'^2\tau_1}{\kappa} \leq \beta'^2\tau_1^2.$$

Apply Lemma 62 with $\kappa_0 = \kappa_A \geq \max(\bar{A}/\theta^3, 2/\tau_1, 2)$ and use

$$\ln(1 + \tau_a) \geq \tau_a - \tau_a^2/2 \geq \frac{3}{4}\tau_1 - \frac{9}{32}\tau_1^2.$$

Together with the preceding initial-loss bounds, this yields

$$2\log R \geq \tau_1\Big[4\beta'\Big(\frac{3}{4} - \frac{9\tau_1}{32}\Big) - \beta'^2\tau_1 - \frac{2\sqrt{\nu}}{\mathbb{E}[v]}\Big] - \frac{1.1(r+1)}{D}. \tag{168}$$

The bracket increases in β' and decreases in τ_1. At $\beta' = 0.344664$, $\tau_1 = 0.100667$, $\sqrt{\nu} \leq 0.0104845$ and $\mathbb{E}[v] \geq \sqrt{1 - 1/2000} \geq 0.9997499$, it is ≥ 0.962025. These values are supplied by Eq. (165), Eq. (162) and Eq. (161). Using the lower time bound in Eq. (162) and $0.962025 \cdot 0.0802648 \geq 0.077216$, Eq. (168) becomes

$$\log(\lambda_c/\lambda_\perp) \geq 2\log R \geq 0.077216\frac{\sqrt{r^2 - 1}}{M} - \frac{1.1(r+1)}{D}. \tag{169}$$

Finally, (R2) gives

$$\frac{1.1(r+1)}{D} \le \frac{1.1}{32}\sqrt{\frac{r+1}{r-1}}\,\frac{\sqrt{r^2-1}}{M}, \qquad \sqrt{\frac{r+1}{r-1}} \le \sqrt{3}.$$

Thus the lower bound in Eq. (169) is at least

$$\Big(0.077216 - \frac{1.1}{32}\sqrt{3}\Big)\frac{\sqrt{r^2-1}}{M} \ge 0.0176\frac{\sqrt{r^2-1}}{M}.$$

In particular, this bound is positive if $D > 14.25M\sqrt{(r+1)/(r-1)}$, which (R2) implies since $14.25\sqrt{3} < 32$.

Step 7 (A-7: Mean eigenvalue). The mean part of Proposition 58(iii) suffices on its own. The time and mass bounds Eq. (162) and Eq. (161) give

$$\tau_a \ge \frac{3}{4}\tau_1 \ge 0.0601986\,\frac{\sqrt{r^2-1}}{M}, \qquad \mathbb{E}[v] + \tau_a\sqrt{\nu} \le 1 + 0.100667 \cdot 0.0104845 \le 1.00106,$$

where $\frac{3}{4} \cdot 0.0802648 = 0.0601986$. These give the displayed eigenvalue bound. Condition (R3), with $0.0601986 \ge 0.06015$ and $1.0011 > 1.00106$, makes it > 1. Finally $t_F \ge \lfloor 4 \rfloor = 4$. □

G.6.3 Region N: an explicit finite certificate

This part is a computer-assisted proof of finitely many inequalities. All of its real inputs, its finite sums and its target inequalities are specified here; the existence or output of a particular program is not a hypothesis. Region N consists of $2 \le r \le 80$ and $\max(400, 63r) \le D \le D_A(r)$, where $D_A(r)$ is the final endpoint for rank r in Table 12. The two intervals for $r = 2$ are indexed by $2a, 2b$; the other rows are indexed by r. Tables 13–16 specify $(M, \theta, \tau_1, \Omega, A, C')$ as dyadic rationals. Set $\tau_F = \tau_1$ and take the integer κ_0 and the rational six-decimal margin μ from Table 12.

A deterministic partition. For a row with integer $K = 2^{20}M$ and refinement index j, set $n = 32000\,2^j$ and

$$k_i = \left\lfloor \frac{iK}{n} \right\rfloor, \qquad q_i = 2^{-20}k_i \quad (0 \le i \le n). \tag{170}$$

Thus $q_0 = 0$, $q_n = M$, and every boundary is specified by integer arithmetic. The exact weights are

$$w_i = F_r(q_i) - F_r(q_{i-1}), \qquad F_r(q) = \Pr(\chi^2_r \le q).$$

Use the cells $[q_{i-1}, q_i]$ from Eq. (170) in Lemma 60. There are 64000 cells only in row $2b$ and 32000 in every other row. These uniform lattice cells replace the approximate equal-probability cells used to discover the original candidates; no quantile calculation is needed in this certificate.

Exact scalar choices. For each row let $D_- = D_{\mathrm{lo}}$, $D_+ = D_{\mathrm{hi}}$, $d_- = D_- + r$, and $y = v^2_{\min}$ with $v_{\min}$ from (C2). Choose

$$\nu_q = 1 - F_r(M), \qquad \nu_v = \mathcal{U}(D_-, y), \qquad \nu = \nu_q + \nu_v.$$

Use the explicit bounds m_1, m_2, m_h of Lemma 55(d), and use its displayed supremum for $\sup_{D \ge D_-} D\mathcal{U}(D, y)$. Evaluate all cell slots in Lemma 60 at D_-, with the $D\nu$ replacement in Lemma 61. Use

$$(\mathbb{E}[v])_{\mathrm{lo}} = \sqrt{1 - \frac{1}{2D_-}}, \quad (\mathbb{E}[v])_{\mathrm{hi}} = 1, \quad R_{0,\mathrm{lo}} = \sqrt{1 - \frac{1}{2d_-}}\sqrt{\frac{D_- - 1/2}{d_-}}.$$

In every lower bound involving $1 - e_0$, replace e_0 by the following explicit upper bound:

$$\begin{aligned} E_0 = (1 - F_{r+2}(M))&\sqrt{\frac{d_-}{D_- - 3/2}}\sqrt{\frac{d_-}{d_- - 1/2}} \\ &+ \frac{e^{-D_-\iota(y)/2}}{\sqrt{\pi r}(1-y)}\sqrt{1 + \frac{r}{D_-}}\sqrt{\frac{d_-}{d_- - 1/2}}. \end{aligned} \tag{171}$$

The bound Eq. (171) is proved in Lemma 61, using $\mathbb{E}[\chi_r]/r \le r^{-1/2}$. The row's integer κ_0 is used throughout the loss slot and the floor calculation, so $\tau_a = \tau_F - 1/\kappa_0$ and $\tau_b = \tau_F$. All remaining quantities are now fixed by (C3)–(C8) and Lemmas 60 and 62.

Explicit probability enclosures. There is no numerical quadrature over an unspecified probability law. For $a = r/2$, $x = q/2$, set

$$t_0 = 1, \qquad t_{m+1} = \frac{x}{a+m+1} t_m, \qquad S_N = \sum_{m=0}^{N} t_m, \qquad z_N = \frac{x}{a+N+1}.$$

With any integer $N \geq \lceil 2x \rceil + 40$ (the computation uses the common row value $N = \lceil M \rceil + 40$) one has $z_N < 1/2$, and the positive series gives

$$\frac{e^{-x}x^a}{\Gamma(a+1)} S_N \;\leq F_r(q) \leq\; \frac{e^{-x}x^a}{\Gamma(a+1)} \left(S_N + \frac{t_N z_N}{1-z_N} \right). \tag{172}$$

At $q = 0$ use $F_r(0) = 0$. For $r = 2m$, $\Gamma(a+1) = m!$; for $r = 2m+1$, $\Gamma(a+1) = (2m+2)!\sqrt{\pi}/(4^{m+1}(m+1)!)$. Thus all cell masses are differences of the explicit positive-series enclosures Eq. (172). Use any integer $N \geq \lceil 2x \rceil + 60$ at the single tail endpoint $q = M$ when evaluating $1 - F_r(M)$ and $1 - F_{r+2}(M)$.

The finite sums of Lemma 60 are enclosed using these weights and outward interval arithmetic. The computation uses binary64 intervals for cell arrays: after each correctly rounded elementary arithmetic operation or square root, move each endpoint one representable number outward. Cubic-root endpoints are accepted only when outward cubing proves the requisite inequalities. For a sum of n nonnegative interval terms, inflate the rounded lower and upper sums by

$$1 - 2n\,2^{-53} \qquad \text{and} \qquad 1 + 2n\,2^{-53},$$

respectively, and round outward; here $n \leq 64000$.

Scalar exponentials, logarithms, square roots and π are enclosed at 120-bit precision: each interval contains the exact real value, with its endpoints rounded outward, rather than merely giving a 120-bit point approximation. The following rational formulas allow independent implementation of these enclosures.

For the exponential, take $z \geq 0$ and $N + 2 > z$:

$$E_N(z) := \sum_{k=0}^{N} \frac{z^k}{k!} \leq e^z \leq E_N(z) + \frac{z^{N+1}}{(N+1)!} \frac{1}{1 - z/(N+2)};$$

Use reciprocals for negative arguments.

For the logarithm, take $x > 0$, write $x = 2^m u$ with $1 \leq u < 2$ and set $s = (u-1)/(u+1)$. Then

$$0 \leq \log u - 2 \sum_{k=0}^{N-1} \frac{s^{2k+1}}{2k+1} \leq \frac{2s^{2N+1}}{(2N+1)(1-s^2)}, \qquad \log x = m \log 2 + \log u,$$

where $\log 2$ uses the same series at $s = 1/3$.

For π, use that, for $0 < z < 1$, consecutive partial sums of $\sum_{k\geq 0} (-1)^k z^{2k+1}/(2k+1)$ enclose $\arctan z$, and $\pi = 16 \arctan(1/5) - 4 \arctan(1/239)$.

Positive square and cube roots are enclosed by rational bisection with the corresponding power comparisons. The factorial formulas above enclose every gamma value needed here; in particular $\bar{\vartheta}_r = \Gamma(r/2)/(\sqrt{\pi}\,\Gamma((r+1)/2))$.

Increase truncation orders and bisection precision until the desired scalar interval width is reached. These are explicit alternatives for independently reproducing the scalar enclosures, not additional assumptions on a software library. When evaluating the growth exponential, fix a certified lower endpoint of β' and apply Lemma 56 to that value, rather than independently optimizing its positive and negative occurrences.

Proposition 65 (The finite region-N check)**.** *For every exact input row, the finite calculations just specified verify the certificate-data domains, (S-Ω), (S-A), (S-C), and*

$$\kappa_0 \geq \max(A/\theta^3, 2/\tau_1, 2), \qquad \beta' > 0, \qquad 0 \leq E_0 < 1, \\ R_{\mathrm{low}}^2 \geq \mu > 1, \qquad \Lambda_{\mathrm{low}} \geq \mu. \tag{173}$$

The intervals in Table 12 cover each required integer interval without gaps. Consequently the row proves both eigenvalue ratios are at least its displayed μ for every $D \in [D_{\mathrm{lo}}, D_{\mathrm{hi}}]$ and every $\kappa \geq \kappa_0$.

Computer-assisted verification and deduction. Step 1 (Enclose the cell slots). Substitute the exact integer tables into the finite expressions above. Enclose each cell mass with Eq. (172), then form the lower slots $\mathcal{N}, \mathcal{K}_a, \underline{K}$ and the upper slots Q_2, H_2, W_2 of Lemma 60.

Step 2 (Verify each certificate). Evaluate the three super-solution right sides and compare their upper endpoints with the exact (Ω, A, C'). Form the remaining scalar and mean-eigenvalue bounds with the conservative D choices above. The resulting outward enclosures certify all inequalities in Eq. (173); the aggregate bounds are recorded in the numerical summary below, and every row's lower margin is recorded in Table 12. This is the finite computer-assisted step, not an analytic replacement for it.

Step 3 (Cover the full parameter range). For coverage, the first left endpoint for each r equals $\max(400, 63r)$, consecutive endpoints satisfy $D_{\mathrm{lo},i+1} = D_{\mathrm{hi},i} + 1$, and the last right endpoint defines $D_A(r)$. Finally apply Proposition 58 and Lemmas 61 and 62. These analytic results turn the finite checks into the claimed continuum of κ and integer intervals of D. □

Enclosure summary. The following bounds hold simultaneously over all 80 rows. Each decimal bound is rounded outward. The three super-solution ratios mean the enclosed right side divided by its exact left side.

quantity	upper bound	quantity	lower bound
$\max \mathrm{rhs}_\Omega/\Omega$	0.999999903	$\min \beta'$	0.415294
$\max \mathrm{rhs}_A/A$	0.999997516	$\min(1-\rho_G^2-\ell_G)$	0.353108
$\max \mathrm{rhs}_{C'}/C'$	0.999990967	$\min_i X_i$	0.032598
$\max v_{\min}$	0.968938	$\min\{R_{\mathrm{low}}^2, \Lambda_{\mathrm{low}}\}$	1.006240
$\max D_- v_{\min}^2/(D_- - 4)$	0.939091		
$\max \rho_G^2$	0.378463		
$\max e_G$	0.472221		
$\max E_0$	0.002185		

In addition $\nu \le 0.000327 < 1/2$, all three lower slots are strictly positive, every published row margin is no larger than its individually checked bound, and every printed integer floor exceeds its required real floor. The stronger common lower bound in this display is not needed: the conservative row-specific margins in Table 12 are retained.

G.6.4 Proof of Theorem 52

Proof of Theorem 52. Let $r \ge 2$ and $D \ge \max(400, 63r)$ (i.e. $d \ge 64r$ and $D \ge 400$).

- If $D \ge D_A(r)$ (which includes every such D when $r \ge 81$), then (R1)–(R3) hold by Lemma 63, and Lemma 64 applies with $\tau_F = \tau_1$ and $\kappa_{\min} = \kappa_A = 4/\tau_1$. It gives $\log(\lambda_c/\lambda_\perp) \ge 0.0176\sqrt{r^2-1}/M > 0$ and $\lambda_1/\lambda_\perp > 1$ at t_F; set $1+\delta := \min\{\exp(0.0176\sqrt{r^2-1}/M), B_1\}$, where $B_1 > 1$ is the displayed lower bound for $\lambda_1/\lambda_\perp$ in Lemma 64.
- Otherwise $2 \le r \le 80$ and $\max(400, 63r) \le D \le D_A(r)$. By Proposition 65, D lies in the D-interval of one of the 80 rows. Proposition 58, Lemma 60, Lemma 61 and Lemma 62 then give, for every $\kappa \ge \kappa_0$ of that row, $\lambda_c/\lambda_\perp \ge \mathrm{margin}$ and $\lambda_1/\lambda_\perp \ge \mathrm{margin}$ at $t_F = \lfloor \tau_F \kappa \rfloor$, with $\mathrm{margin} > 1$ the IA value of the row. Set τ_F, $\kappa_{\min} := \kappa_0$ and $1+\delta := \mathrm{margin}$ from that row.

In both cases $\min(\lambda_1, \lambda_c) \ge (1+\delta)\lambda_\perp > \lambda_\perp$ at t_F, so by (F2) (Proposition 27) the top-r eigenspace of the MF AGOP is U, it is unique, $A_r(t_F) = 1$ for every leading selector, and $\lambda_r/\lambda_{r+1} = \min(\lambda_1, \lambda_c)/\lambda_\perp \ge 1+\delta$. The numerical ranges in (a) are read off Table 12; those in (b) are Lemma 64. □

G.7 Proof of Theorem 16

What this section proves. Theorem 16. Steps 1 and 5 use only $\alpha \le \alpha^*(t_P)$. Steps 2–3 prove (F): first for the $\alpha = 0$ dynamics, by the certificate, and then for $0 < \alpha\bar{\eta} \le a_0$, by continuity in the head scale. Step 4 is the isotropy at $t = 0$.

Notation for the assembly. We use $D = d - r$, $P_\perp = I_d - P_U$, $P_1 = \bar{e}\bar{e}^\top$ and $P_c = P_U - P_1$ from Sections G.1.1 and G.2.3. The law-update map Φ_α, force covariance Σ_α and AGOP surrogate $\mathcal{M}$ are those of Section G.2.1; the actual AGOP is $\alpha^2\mathcal{M}$ in the original coordinates. The times t_P, t_F, head threshold $\alpha^*(t_P)$ and gap parameter δ are those of Theorem 16.

Proof of Theorem 16. Step 0 (The trajectory). With $\sigma_0 = \kappa\bar{\eta}\sqrt{d}$, the law $\nu_0 = N(0, \sigma_0^2 I_d/d)$ is the initialization of every result below, and $\nu_{t+1} = \Phi_\alpha(\nu_t)$ with Φ_α as in Section G.2.1; this is Eq. (92), with the Moore–Penrose convention, so the trajectory is defined for all t and every $\alpha \geq 0$.

By Lemma 20, every ν_t is absolutely continuous and has finite second moment, so $\nu_t(\{0\}) = 0$; by Proposition 21(c), every ν_t is $\mathcal{G}$-invariant. These are the standing hypotheses of Section G.3 and of Proposition 27.

Step 1 (P). Theorem 42 assumes $r \geq 2$, $d > r$, $\kappa > 0$, $t_P \in \mathbb{N}$ and $0 < \alpha \leq \alpha^*(t_P)$. It gives $\max_{s,t \leq t_P} L(t)/L(s) \leq 1.01$ and $L(t) \geq .99V$ for $t \leq t_P$, which is (P).

Step 2 ((F) for the $\alpha = 0$ dynamics). By Lemma 18, the trajectory with head α and step $\bar{\eta}$ is the image under $w \mapsto \bar{\eta}w$ of the trajectory $\tilde{\nu}_a(t)$ with head $a := \alpha\bar{\eta}$ and step 1 from $N(0, \kappa^2 I_d)$; the two have the same whitening matrices, and their MF AGOPs differ by the factor $\bar{\eta}^2$. So it suffices to prove (F) for $\tilde{\nu}_a$.

For $a = 0$, $\tilde{\nu}_0(t)$ is the trajectory of Definition 22. Our hypotheses $r \geq 2$, $d \geq 64r$, $D \geq 400$ and $\kappa \geq \kappa_{\min}$ are those of Theorem 52, which gives, at $t_F = \lfloor \tau_F \kappa \rfloor \geq 2$,

$$\begin{aligned} \mathcal{M}(\tilde{\nu}_0(t_F)) &= \lambda_1^0 P_1 + \lambda_c^0 P_c + \lambda_\perp^0 P_\perp, \\ \min(\lambda_1^0, \lambda_c^0) &\geq (1+\delta)\,\lambda_\perp^0. \end{aligned} \tag{174}$$

Both teacher-space block eigenvalues are strictly positive. Indeed, Lemma 50 gives the first inequality below, and Proposition 27(1) with Lemma 48(e) gives the second:

$$\lambda_1^0 \geq \frac{(\mathbb{E}[\omega_1(t_F)])^2}{4} \geq \frac{(1-1/r)\,t_F^2}{4} > 0, \qquad \lambda_c^0 \geq \frac{\Gamma_U(t_F)^2}{2\pi} > 0.$$

Thus $\min(\lambda_1^0, \lambda_c^0) > 0$.

Step 3 (Transfer to $0 < a \leq a_0$). For a symmetric $d \times d$ matrix A put

$$\ell_1(A) := \bar{e}^\top A \bar{e}, \qquad \ell_c(A) := \frac{\operatorname{tr}(P_c A)}{r-1}, \qquad \ell_\perp(A) := \frac{\operatorname{tr}(P_\perp A)}{D},$$

$$\mathrm{g}(A) := \min\big(\ell_1(A), \ell_c(A)\big) - \big(1 + \tfrac{\delta}{2}\big)\,\ell_\perp(A). \tag{175}$$

These functions are continuous, and on a matrix of the block form $\lambda_1 P_1 + \lambda_c P_c + \lambda_\perp P_\perp$ they return λ_1, λ_c and $\lambda_\perp$. The gap Eq. (174) and positivity from Step 2 give

$$\mathrm{g}\big(\mathcal{M}(\tilde{\nu}_0(t_F))\big) \geq \min(\lambda_1^0, \lambda_c^0)\Big(1 - \frac{1 + \delta/2}{1+\delta}\Big) > 0. \tag{176}$$

Apply Proposition 29(a) in units $\bar{\eta} = 1$, with $\sigma_0 = \kappa\sqrt{d}$ and $T = t_F$. As $a \to 0$,

$$\mathcal{M}(\tilde{\nu}_a(t_F)) \to \mathcal{M}(\tilde{\nu}_0(t_F)),$$

and there is $a_T > 0$ with $\Sigma_a(\tilde{\nu}_a(t)) \succ 0$ for all $t \leq t_F$ and $a \leq a_T$. By continuity, the strict inequality Eq. (176) persists: there is $a_0 \in (0, a_T]$ such that

$$\mathrm{g}(\mathcal{M}(\tilde{\nu}_a(t_F))) > 0 \qquad \text{for every } a \in (0, a_0]. \tag{177}$$

The trajectories $\tilde{\nu}_a$ are determined by (r, d, κ, a), and $\delta = \delta(r, D)$, so a_0 can be chosen as a function of (r, d, κ) alone.

Fix $a = \alpha\bar{\eta} \in (0, a_0]$. By Step 0 and Proposition 27(0), $\mathcal{M}(\tilde{\nu}_a(t_F))$ has the block form, and its eigenvalues are $\lambda_j = \ell_j(\mathcal{M}(\tilde{\nu}_a(t_F)))$, $j \in \{1, c, \perp\}$. Use Eq. (177) in the definition Eq. (175), together with $\lambda_\perp \geq 0$ ($\mathcal{M}$ is positive semidefinite), to obtain

$$\min(\lambda_1, \lambda_c) \;>\; \big(1 + \tfrac{\delta}{2}\big)\lambda_\perp \;\geq\; \lambda_\perp.$$

By Proposition 27(0), the top-r eigenspace of $\mathcal{M}(\tilde{\nu}_a(t_F))$ is U and is unique, $A_r(t_F) = 1$ for every leading selector, and $\lambda_r = \min(\lambda_1, \lambda_c) \geq (1 + \delta/2)\lambda_{r+1}$.

The MF AGOP of the original trajectory is

$$\alpha^2 \mathcal{M}(\nu_{t_F}) = \alpha^2 \bar{\eta}^2 \mathcal{M}(\tilde{\nu}_a(t_F)),$$

which has the same eigenspaces and eigenvalue ratios. Also $\Sigma_{\nu_t} = \Sigma_a(\tilde{\nu}_a(t)) \succ 0$ for $t \leq t_F$. This is (F) at t_F.

Step 4 ((F) at $t = 0$). ν_0 is $O(d)$-invariant, so $\mathcal{M}(\nu_0)$ is a multiple of I_d by Proposition 27(4).

Step 5 (R). Corollary 45 assumes $r \geq 2$, $d \geq 64r$, $\kappa \geq 1$, $t_P \in \mathbb{N}$ and $0 < \alpha \leq \alpha^*(t_P)$. Our hypotheses include these: $\kappa \geq \kappa_{\min} \geq 2/\tau_F > 19$, because $\kappa_0 \geq 2/\tau_1$ in region N (Lemma 62) and $\kappa_A = 4/\tau_1$ in region A. The corollary gives

$$L(0) - L(t_2) \geq .03V, \qquad t_2 = \left\lceil \frac{T_2^* \sqrt{V}}{\alpha \bar{\eta}} \right\rceil, \qquad T_2^* = .05086.$$

At $\alpha = \alpha^*(t_P)$ it also gives $t_2 \leq C_R(\kappa + t_P) + 1$ with $C_R \leq 6.013$; for $t_P \leq .1007\kappa$ this is at most $6.62\kappa + 1$. □

Remark 66 (Finite width)**.** *Under (F), $\Sigma_{\nu_t} \succ 0$ for $t \leq t_F$, so Proposition 29(b) applies with $T = t_F + 1$: the width-m trajectory at fixed $\bar{\eta}$ converges to the MF one on $[0, t_F + 1]$, and its top-r AGOP space at t_F converges to U as $m \to \infty$. Given ν_0, the MF trajectory is deterministic, so (P), (F) and (R) hold surely.*

G.8 Finite verification details

This section specifies the finite inequalities used above.

G.8.1 The Region A endpoints

To prove Lemma 63(c), the following finite rational certificate applies to the right endpoints $D_A(r)$ of Table 12. Write H_r for the left side of (R3) at $D = D_A(r)$. Each row gives lower bounds valid for every integer rank in its first column:

r	lower bound on $D_A(r)$	$D_A(r)/M \geq$	$H_r^2 \geq$
2:10	7917	125.18416249	1.00220175
11:20	2596	34.75697688	1.00224498
21:30	2429	32.00130160	1.00385938
31:40	2905	32.00055080	1.86188294
41:50	3363	32.00109878	2.83732405
51:60	3808	32.00078823	3.89645378
61:70	4243	32.00079990	5.01895909
71:80	4671	32.00051696	6.19274707

In particular $D_A(r) \geq 1000$, $D_A(r)/M \geq 32$, and $H_r^2 \geq 1.00220175 > 1.0011^2 = 1.00220121$, proving (R1)–(R3). The entries follow from three rational enclosures, described below.

Logarithm enclosure. For $0 \leq z \leq 1/3$, put

$$S(z) := 2 \sum_{j=0}^{31} \frac{z^{2j+1}}{2j+1}, \qquad E(z) := \frac{2z^{65}}{65(1 - z^2)}.$$

The positive logarithm series and a geometric tail bound give $S(z) \leq \log((1 + z)/(1 - z)) \leq S(z) + E(z)$. For rational $x \geq 1$, write $x = 2^k y$ with $1 \leq y < 2$ and define

$$\mathsf{L}(x) := k\big[S(1/3) + E(1/3)\big] + S\left(\frac{y-1}{y+1}\right) + E\left(\frac{y-1}{y+1}\right). \tag{178}$$

Thus $\log x \leq \mathsf{L}(x)$, with a rational right side.

Square-root enclosure. For rational $v \geq 0$, choose the integer q characterized by

$$q^2 \leq 10^{30} v < (q+1)^2, \qquad \mathsf{S}(v) := \frac{q+1}{10^{15}}. \tag{179}$$

Then $\sqrt{v} < \mathsf{S}(v)$, certified by the two rational inequalities in Eq. (179).

Endpoint comparison. For each of the 79 exact endpoints in the complete certificate table (not the eight blockwise minima above), use Eq. (178) and Eq. (179) to set

$$\begin{aligned} L_r^+ &:= 5 + \mathsf{L}(D_A(r)/(r-1)), \\ M_r^+ &:= r + 4 + 2\mathsf{S}((r+4)L_r^+) + 2L_r^+. \end{aligned} \tag{180}$$

With $\pi_- := 3.14159265358979323846264338327950288 < \pi$, the third and fourth columns of the table are lower bounds on the following rational numbers, using Eq. (180), respectively:

$$\frac{D_A(r)}{M_r^+}, \qquad \frac{(.06015)^2(r^2-1)(\pi_-/2)(1-1/r)(D_A(r)-1/2)}{(M_r^+)^2}.$$

All table inequalities follow by integer cross-multiplication, with the results rounded downward to eight decimal places.

The bound on π can itself be checked from Machin's identity $\pi = 16\arctan(1/5) - 4\arctan(1/239)$ using the first 64 terms of the alternating power series for each arctangent and its next-term remainder bound. This proves (c) using only the printed endpoints and finite rational inequalities. The numerical inequalities in (b) follow from the same elementary bounds.

G.9 Complete certificate data

The exact dyadic input tables and Table 12 together specify every parameter of the region-N certificate. The former determine $M, \theta, \tau_F, \Omega, A, C'$ and the deterministic cell partition; the latter give the integer D-interval, integer floor κ_0 and rational lower margin for each row. The six-decimal τ_F in the results table is a display only: its exact value is $T2^{-59}$ in the input table. No externally stored parameter values are needed.

The result row $(r, [D_{\mathrm{lo}}, D_{\mathrm{hi}}])$ proves both eigenvalue ratios are at least its displayed margin at $t_F = \lfloor \tau_F \kappa \rfloor$, for every integer D in that interval and every $\kappa \geq \kappa_0$. The margins retained here are conservative six-decimal lower bounds; the deterministic-cell verification can yield slightly stronger bounds. For $2 \leq r \leq 80$, define $D_A(r)$ to be the last right endpoint for r in this table. For $r \geq 81$, define $D_A(r) = 63r$. The region-A argument uses only the stated inequalities (R1)–(R3) at these endpoints and their upward closure, not the historical method used to select them.

Table 12: The 80 region-N certificates at $\alpha = 0$: D-interval, τ_F, floor κ_0 and certified margin $\min(\lambda_1, \lambda_c)/\lambda_\perp$ at t_F.

r	$[D_{\mathrm{lo}}, D_{\mathrm{hi}}]$	τ_F	κ_0	margin	r	$[D_{\mathrm{lo}}, D_{\mathrm{hi}}]$	τ_F	κ_0	margin
2	$[400, 14930]$	0.013740	336	1.006223	41	$[2583, 3363]$	0.078835	126	1.228737
2	$[14931, 465350]$	0.014026	777	1.048716	42	$[2646, 3408]$	0.079339	125	1.230167
3	$[400, 119688]$	0.014191	286	1.012711	43	$[2709, 3453]$	0.079824	124	1.231501
4	$[400, 55355]$	0.017549	172	1.025338	44	$[2772, 3497]$	0.080300	124	1.232778
5	$[400, 32467]$	0.022016	136	1.029847	45	$[2835, 3542]$	0.080758	123	1.233975
6	$[400, 21679]$	0.025939	119	1.033781	46	$[2898, 3587]$	0.081197	122	1.235096
7	$[441, 15709]$	0.027378	110	1.039982	47	$[2961, 3631]$	0.081637	121	1.236213
8	$[504, 12040]$	0.030688	99	1.048280	48	$[3024, 3675]$	0.082058	121	1.237254
9	$[567, 9613]$	0.039394	125	1.057938	49	$[3087, 3720]$	0.080968	122	1.238515
10	$[630, 7917]$	0.039459	127	1.079513	50	$[3150, 3764]$	0.081371	122	1.239536
11	$[693, 6681]$	0.045293	120	1.089914	51	$[3213, 3808]$	0.081765	121	1.240522
12	$[756, 5748]$	0.046299	119	1.101188	52	$[3276, 3852]$	0.082149	121	1.241473
13	$[819, 5025]$	0.048771	113	1.109930	53	$[3339, 3896]$	0.082525	120	1.242394
14	$[882, 4452]$	0.051069	109	1.117599	54	$[3402, 3939]$	0.082900	119	1.243321
15	$[945, 3989]$	0.053212	104	1.124371	55	$[3465, 3983]$	0.083257	119	1.244175
16	$[1008, 3608]$	0.053451	104	1.130751	56	$[3528, 4027]$	0.083605	118	1.244999
17	$[1071, 3290]$	0.055296	101	1.136427	57	$[3591, 4070]$	0.083944	118	1.245786
18	$[1134, 3022]$	0.057027	98	1.141549	58	$[3654, 4113]$	0.084282	117	1.246585
19	$[1197, 2793]$	0.058666	96	1.146243	59	$[3717, 4157]$	0.084603	117	1.247310
20	$[1260, 2596]$	0.060199	93	1.150495	60	$[3780, 4200]$	0.084923	116	1.248040
21	$[1323, 2429]$	0.061591	91	1.154202	61	$[3843, 4243]$	0.085235	116	1.248739
22	$[1386, 2478]$	0.067373	140	1.164924	62	$[3906, 4286]$	0.085546	115	1.249446
23	$[1449, 2526]$	0.062740	141	1.162169	63	$[3969, 4329]$	0.085839	115	1.250079
24	$[1512, 2574]$	0.066769	140	1.176584	64	$[4032, 4372]$	0.086132	114	1.250719
25	$[1575, 2622]$	0.070623	139	1.179868	65	$[4095, 4415]$	0.086416	114	1.251325
26	$[1638, 2670]$	0.066155	140	1.182391	66	$[4158, 4458]$	0.086699	114	1.251938
27	$[1701, 2717]$	0.069717	139	1.193370	67	$[4221, 4500]$	0.086974	113	1.252516
28	$[1764, 2764]$	0.072271	138	1.188541	68	$[4284, 4543]$	0.087239	113	1.253060
29	$[1827, 2811]$	0.068865	139	1.197385	69	$[4347, 4586]$	0.087505	112	1.253610
30	$[1890, 2858]$	0.073818	135	1.200144	70	$[4410, 4628]$	0.086480	114	1.254297
31	$[1953, 2905]$	0.074541	134	1.204589	71	$[4473, 4671]$	0.086736	114	1.254845
32	$[2016, 2951]$	0.075237	133	1.208340	72	$[4536, 4713]$	0.086992	113	1.255396
33	$[2079, 2997]$	0.075905	131	1.211545	73	$[4599, 4755]$	0.087239	113	1.255911
34	$[2142, 3044]$	0.074843	133	1.214607	74	$[4662, 4797]$	0.087487	112	1.256433
35	$[2205, 3090]$	0.075475	132	1.217328	75	$[4725, 4840]$	0.087725	112	1.256919
36	$[2268, 3135]$	0.076079	131	1.219727	76	$[4788, 4882]$	0.087963	112	1.257411
37	$[2331, 3181]$	0.076674	130	1.221887	77	$[4851, 4924]$	0.088192	111	1.257865
38	$[2394, 3227]$	0.077242	129	1.223829	78	$[4914, 4966]$	0.088421	111	1.258326
39	$[2457, 3272]$	0.077791	128	1.225601	79	$[4977, 5008]$	0.088640	111	1.258751
40	$[2520, 3317]$	0.078322	127	1.227230	80	$[5040, 5050]$	0.088860	111	1.259180

Exact encoding. For each row let $M = K2^{-20}$, $\tau_1 = \tau_F = T2^{-59}$, $\Omega = O2^{-50}$, $A = B2^{-51}$, and $C' = C2^{-52}$. The index h defines θ by

$$\theta(3) = 5404319552844595\,2^{-54},$$
$$\theta(35) = 3152519739159347\,2^{-53},$$
$$\theta(4) = 3602879701896397\,2^{-53}.$$

The cell count is $n = 32000 \cdot 2^j$. Rows $2a, 2b$ refer to the two $r = 2$ intervals in the results table; all other rows are indexed by r. Every integer below is exact.

Table 13: Exact partition inputs for region N (rows 1–40).

row	K	h	j	T
2a	21600665	35	0	7920963464610295
2b	36059480	3	1	8085720989539134
3	29042409	35	0	8180717210882235
4	35333865	4	0	10116431691206438
5	34521219	4	0	12691884803174992
6	35458646	4	0	14952843575107892
7	38181797	4	0	15782428537537256
8	39344668	4	0	17690797072963136
9	35233202	35	0	22709285047772204
10	39985348	35	0	22746804483491268
11	41193308	35	0	26110149771855272
12	44290801	35	0	26689626722048196
13	45557481	35	0	28114570042194732
14	46829404	35	0	29439767329931008
15	48102375	35	0	30674718207391344
16	51380224	35	0	30812462728338840
17	52688846	35	0	31876420407381584
18	53996421	35	0	32873880731484156
19	55301898	35	0	33819093133848028
20	56606326	35	0	34702557992338884
21	57912852	35	0	35505276062688100
22	55029268	3	0	38838108865442336
23	58798899	3	0	36167659532130680
24	60316188	3	0	38489789387184296
25	61824040	3	0	40711645453042408
26	65633517	3	0	38136192341073864
27	67151855	3	0	40189166235655360
28	68661805	3	0	41661607666473440
29	72502738	3	0	39698352425382656
30	71661780	3	0	42553516633528128
31	73151807	3	0	42970444493867288
32	74635542	3	0	43371539650649280
33	76111937	3	0	43756802103874080
34	80035708	3	0	43144604232996312
35	81522589	3	0	43508756414811544
36	83004227	3	0	43857075893069584
37	84480622	3	0	44200117803475232
38	85952823	3	0	44527327010323696
39	87418732	3	0	44843981081467368
40	88880447	3	0	45150080016906256

Table 14: Exact partition inputs for region N (rows 41–80).

row	K	h	j	T
41	90337968	3	0	45445623816640352
42	91790245	3	0	45735890048522056
43	93238329	3	0	46015601144698960
44	94680121	3	0	46290034673023472
45	96119816	3	0	46553913065643208
46	97556365	3	0	46807236322558152
47	98987671	3	0	47060559579473088
48	100414783	3	0	47303327700683232
49	104529395	3	0	46675297126248288
50	105968041	3	0	46907510111753648
51	107402493	3	0	47134445529406608
52	108832751	3	0	47356103379207184
53	110260912	3	0	47572483661155360
54	111684878	3	0	47788863943103536
55	113105698	3	0	47994689089346928
56	114524422	3	0	48195236667737920
57	115937902	3	0	48390506678276520
58	117350334	3	0	48585776688815120
59	118759620	3	0	48770491563648928
60	120164712	3	0	48955206438482736
61	121567707	3	0	49134643745464144
62	122968604	3	0	49314081052445568
63	124366356	3	0	49482963223722192
64	125762011	3	0	49651845394998824
65	127154520	3	0	49815449998423056
66	128544931	3	0	49979054601847280
67	129932197	3	0	50137381637419120
68	131317366	3	0	50290431105138560
69	132700438	3	0	50443480572858008
70	137042591	3	0	49852392973389808
71	138436149	3	0	50000164873256856
72	139826561	3	0	50147936773123904
73	141213827	3	0	50290431105138560
74	142600044	3	0	50432925437153216
75	143984164	3	0	50570142201315480
76	145366188	3	0	50707358965477736
77	146745065	3	0	50839298161787600
78	148122894	3	0	50971237358097464
79	149498626	3	0	51097898986554928
80	150872260	3	0	51224560615012400

Table 15: Exact supersolution inputs for region N (rows 1–40).

row	O	B	C
2a	5292723980082718	32428288108055624	5608012396498452
2b	6759749330873526	47234666444723024	4567102900135213
3	6150004491384073	27557958287837056	5960833993768631
4	6813178771523468	24649586303920624	5919624845377477
5	6805068924342156	19562686959486248	5979297203968494
6	6961059657208525	17116075937116040	6087232431130972
7	7244570957128253	15814519020401872	6084414177483384
8	7397342834379401	14265978139366600	5982018411495578
9	7165248123289302	11994103524559812	6108689384936562
10	7624330648560800	12188339050011954	6107855295699327
11	7816992396520535	11584092604917802	6047903646975958
12	8114724398401507	11408304546243588	6026939544538555
13	8265067162849759	10878388211123606	5983459685048755
14	8412778197264684	10428066953059118	5946617038309694
15	8557675136672041	10039938774911952	5915045146480910
16	8842706773114639	10037296873231022	5908968387516019
17	8982767256514036	9724656117379244	5883696393520844
18	9120312371788476	9447363298365294	5861474521720000
19	9255597473357322	9199936591472062	5841814212876274
20	9388473355124492	8977254943204165	5824337885408719
21	9518708692360278	8775033938432237	5808753889297893
22	9475197000037700	8488557454026792	6097795241830543
23	9713465021536114	8525516794517812	6100380809618886
24	9901178661896920	8482673800498109	6089131885914375
25	10084705286757180	8443389019409702	6079829751118036
26	10310197713544460	8480045370510340	6082520858566669
27	10486821438934066	8443337268778820	6073497625162175
28	10646271907328426	8386871754573909	6065406584806843
29	10874580845956476	8445379180561290	6068856206736862
30	10904259663039498	8197305009682678	6047007800350284
31	11030272289731130	8110909709240749	6038778496457271
32	11154461939110860	8029491184468583	6031111719018059
33	11276785963160284	7952426465393715	6023943385245597
34	11535056041251706	8077705013149340	6029545993731846
35	11653987293091250	8004452736735269	6022722784745882
36	11771298843906270	7934791483279786	6016304083349640
37	11887327152694930	7868826098726293	6010267694267958
38	12001848803664564	7805896294084367	6004573184671171
39	12114905764571344	7745765307089613	5999189004889925
40	12226654541612578	7688369272956011	5994099997093083

Table 16: Exact supersolution inputs for region N (rows 41–80).

row	O	B	C
41	12337113433222418	7633481250065179	5989283289725050
42	12446397115720632	7581022616188417	5984719452569983
43	12554428182284166	7530683027299399	5980389152653382
44	12661257956522734	7482341945772299	5976273220784853
45	12766998498356294	7435959038899650	5972366854639561
46	12871597276411164	7391309963979724	5968652097336570
47	12975276608248382	7348516067938118	5965118537347517
48	13077781077239216	7307133485710754	5961749265244648
49	13310520257288940	7416909742886614	5965930815396461
50	13410953891133806	7376288651918253	5962574449157443
51	13510434200087326	7337061796677928	5959366216036499
52	13608974335688814	7299138910873558	5956297449397509
53	13706707044802002	7262568106318256	5953365894217178
54	13803700236105944	7227330813318137	5950562068819066
55	13899670972862910	7193028672532432	5947873578315527
56	13994864905308404	7159850747766006	5945300518889238
57	14089114945790088	7127546338099018	5942829274854179
58	14182846904304308	7096492258904211	5940467714917040
59	14275590210831866	7066127608059542	5938198062975710
60	14367659848247368	7036721999401950	5936020736832794
61	14458998169125878	7008154097169425	5933931755850891
62	14549797373180440	6980562010179316	5931930015212254
63	14639638500452052	6953473216162533	5930002652690999
64	14728956487954490	6927273706497871	5928154447869270
65	14817516618845280	6901684965020106	5926376715333673
66	14905570296063234	6876907518315706	5924670920206436
67	14992881188422112	6852671138826332	5923029152365483
68	15079510739209174	6828996676976786	5921450034687528
69	15165655963302758	6806033357063836	5919933626041060
70	15376809956722276	6898256573472885	5922142225437706
71	15461648514619670	6875227199849992	5920568180005206
72	15545977330830428	6852809223160655	5919051220385149
73	15629607382902882	6830796914691879	5917586011992899
74	15712847109331618	6809441371700104	5916176202553731
75	15795451327347682	6788490936913212	5914815157153569
76	15877623668380466	6768102119292371	5913503719228300
77	15959118060443858	6748030252218483	5912235937372244
78	16040243809047156	6728523450096594	5911015169338573
79	16120752682242078	6709340075482683	5909835739195862
80	16200851077071876	6690638106199919	5908698653676628