SUBSPACE LEARNING DURING POPULATION-LOSS PLATEAUS

Anonymous authors
Paper under double-blind review

ABSTRACT

A nearly constant loss can conceal substantial feature learning. We prove this for two-layer ReLU and leaky-ReLU networks under simultaneous fixed-step population gradient descent from small IID Gaussian initialization. For positive mixtures of damped cubic links in Gaussian Sobolev space, loss stays near one while minimum alignment of the predictor's top-r average gradient outer product (AGOP) subspace with the teacher subspace improves by at least $1/2$. Identically budgeted refitting improves population MSE by more than 0.399, and the unchanged trajectory later reduces its own loss. A complementary theorem covers small additive Sobolev perturbations of unequal-weight cubic teachers using projected features. For SwiGLU, at fixed width and dimension as initialization vanishes, we prove leading-AGOP-direction alignment during a loss plateau for square-integrable teachers with a nonzero Hermite component of degree one, two, or three; unrestricted rank-one cubic refit gains hold at a specified width. These guarantees require structured teachers and restrictive parameters. An approximation result further distinguishes subspace recovery from neuron specialization: some interaction targets retain unavoidable prediction error when ridge neurons specialize along shared orthogonal axes within the teacher subspace. Population-moment experiments across 21 teacher links and 50 initializations compare alignment, refit improvement, and trained-loss reduction.

1 INTRODUCTION

A nearly constant prediction loss can conceal substantial changes in a network's representation. Early ReLU alignment (Maennel et al., 2018), *silent alignment* of tangent-kernel eigenstructure (Atanasov et al., 2022), and hidden feature amplification in parity learning (Barak et al., 2022) provide precedents. We ask: can an ordinary nonlinear training trajectory recover *every* direction of a target subspace while its population loss remains high, and can the learned features already support a substantially better predictor?

We study this question for two-layer ReLU and leaky-ReLU networks trained by simultaneous population gradient descent on all parameters. Inputs are Gaussian and the target depends on an unknown r-dimensional subspace. This setting connects to representation learning for Gaussian multi-index models (Damian et al., 2022; Bietti et al., 2025), while its small initialization connects to early alignment and alternating descriptions of feature acquisition (Boursier & Flammarion, 2025; Kunin et al., 2025). We measure the weakest recovered direction through the leading r-dimensional eigenspace of the current predictor's average gradient outer product (AGOP), including its heads, gates, and cross-neuron terms. Hidden-weight concentration along one teacher direction is insufficient. Geometry also does not determine predictive quality, so we separately refit the actual feature bank under identical normalization and coefficient budgets before and after learning.

All heads, weights and biases train together. The refit is only a measurement of the learned features and never enters training. Small initialization separates output amplitude from relative feature geometry; the proof follows both through the coupled updates. Our first result combines these requirements on one trajectory: all-direction recovery and improved refitting during a high-loss window, followed by a quantified decrease in the trained loss. We establish this for positive mixtures of Gaussian-damped cubic links, including h_3 and nonpolynomial links, as defined in (6).

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.

Informal Theorem 4.1 (ReLU and leaky ReLU). *For our positive-mixture teachers, fixed-step population GD from small Gaussian initialization learns the teacher subspace while loss stays near one. With high probability, minimum AGOP alignment improves by at least $1/2$ and same-budget refitting reduces MSE by more than 0.399 during the plateau. The same trajectory later lowers its own loss.*

To examine this separation beyond kink activations and the preceding teacher family, we next study the smooth gated architecture SwiGLU. Here a broader teacher class supports a complementary geometric guarantee: a leading AGOP direction approaches the teacher subspace.

Informal Theorem 5.1 (SwiGLU). *For square-integrable teachers with a nonzero Hermite component of degree one, two, or three, SwiGLU exhibits a related separation. At fixed width, dimension and GD step, with the intercept fitted exactly, its leading AGOP direction can approach the teacher subspace arbitrarily closely as Gaussian initialization shrinks, while loss remains near its initial value.*

The formal statements specify the teacher and parameter restrictions. Extensions of the ReLU/leaky-ReLU family appear in Theorems D.1 and E.1. For a narrower rank-one teacher family and specified width, Corollary 5.2 adds a simultaneous unrestricted-refit gain during the SwiGLU plateau; its pure-h_3 example gives alignment and refit gains of at least 0.5 and 0.7 with high asymptotic probability.

Figure 1 illustrates the separation at accessible scales across fifty initializations per teacher. For the normalized $\mathrm{SiLU}(X_1)X_2$ teacher, mean minimum-AGOP alignment rises from 0.028 to 0.986 and mean refit MSE falls from 0.829 to 0.455 by update 200, while loss remains near one. The neuron-angle and axis-snapping diagnostics show why subspace recovery need not imply specialization to teacher axes: the observed trajectories retain neurons that mix teacher coordinates. These interaction targets are Gaussian-Sobolev examples outside the stated plateau theorems. Figure 2 separately reports SwiGLU students over twenty seeds per teacher.

2 Related work

Feature changes before loss reduction. Small-initialization ReLU networks can align before appreciable loss reduction (Maennel et al., 2018). Silent alignment describes a related separation through the evolving tangent kernel, analytically in linear networks and experimentally in nonlinear networks (Atanasov et al., 2022). Nonlinear early-alignment guarantees cover classification under label-correlation conditions (Min et al., 2024), empirical gradient flow (Boursier & Flammarion, 2025), and orthogonal-input regression under balanced initialization (Boursier et al., 2022). Hidden progress in parity learning instead amplifies Fourier features (Barak et al., 2022). Our result connects this early progress to every-direction AGOP recovery, a bounded-readout gain, and later loss decrease on one simultaneous-GD trajectory.

Plateaus, scale, and escape. Delayed transitions occur in exact deep-linear dynamics (Saxe et al., 2014). Alternating directional acquisition and feature growth have a proved small-initialization limit in diagonal linear networks (Kunin et al., 2025); the first escape in deep ReLU networks exhibits a low-rank bias in deeper layers (Bantzis et al., 2026). Univariate ReLU plateaus also reflect activation-pattern dynamics (Ainsworth & Shin, 2021). Scaling determines whether training is lazy (Chizat et al., 2019), so our initialization and GD step are substantive hypotheses.

Gaussian multi-index learning. Staged representation learning for polynomial targets (Damian et al., 2022) assumes an expected Hessian of rank r, which excludes our central odd links. Gaussian multi-index population flow with infinitely faster nonparametric link fitting exhibits Hermite-dependent phases (Bietti et al., 2025). For suitable orthogonal additive targets and initialization, correlation-loss flow gives all-index recovery with order-$r \log r$ neurons (Şimşek et al., 2025). Unlike those decoupled neurons, ours remain coupled through square loss, with trained biases and heads; recovery concerns the current predictor's AGOP. The emphasis here is the joint dynamical and spectral guarantee under simultaneous training; our width prescriptions are sufficient existence bounds and do not improve that coverage rate. Large first-layer steps followed by readout fitting provide another regime (Dandi et al., 2024).

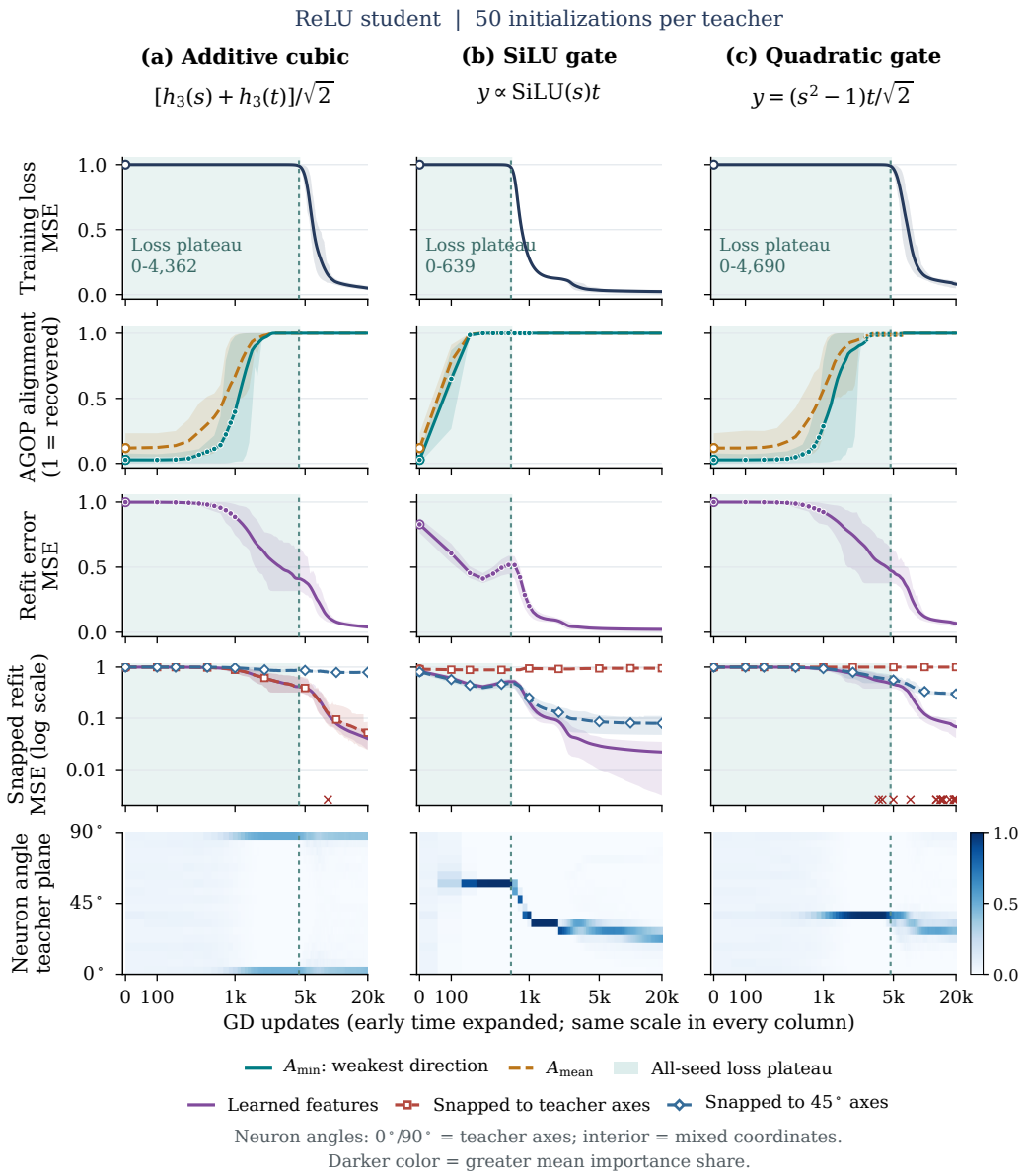


Figure 1: **Features improve before loss falls; gated teachers retain mixed neurons.** Fifty initializations of one ReLU-student configuration per normalized teacher ($s = X_1$, $t = X_2$). *Read down each column:* training loss stays near one, AGOP subspace alignment rises, and a same-budget readout refit improves before loss release. Shading marks the loss-only plateau shared by all fifty seeds; the labeled boundaries are the earliest individual plateau endpoints. Horizontal position is $\log(1 + n/100)$: early time is expanded, ticks show actual updates, and initialization and all 20,000 updates remain visible. For SiLU, mean $A_{\min}$ rises $0.028 \to 0.986$ and refit MSE falls $0.829 \to 0.455$ by update 200 while loss remains near one; refit improvement is not monotone. Row four projects each weight's teacher-plane component onto its nearest axis in the indicated frame, retaining outside-plane components and biases, then refits. Bottom-row color shows mean importance share: $0°/90°$ are teacher axes; interior angles mix coordinates. Curves use all fifty seeds; light bands are pointwise 10th–90th percentiles, not confidence intervals. Dotted alignment bounds retain unresolved spectra; red crosses mark snapped-refit solver tolerance misses. Diagnostics are recorded every 100 updates (early dots); connecting lines add no observations. Appendix A gives numerical brackets, the plateau rule and eighteen further teachers.

Hermite structure. Information exponents govern high-dimensional online search times (Ben Arous et al., 2021); leap complexity describes successive multi-index feature acquisition (Abbe et al., 2023). Our mixture and Sobolev-neighborhood hypotheses likewise control both directional signal and approximation by learned profiles; regularity alone is insufficient.

Gradient-based subspaces and AGOP. Gradient outer products support active-subspace approximation and dimension reduction (Constantine et al., 2014; Yuan et al., 2025). The neural feature ansatz relates weight geometry and AGOP (Radhakrishnan et al., 2024), with exact identities under balanced deep-linear flow and nonlinear counterexamples (Tansley et al., 2026). The features-at-convergence identity concerns critical points with nonzero weight decay (Boix-Adserà et al., 2026), rather than our unregularized finite-time trajectory. AGOP progress during flat loss is observed in recursive feature machines for modular arithmetic (Mallinar et al., 2025); that alignment uses the final learned matrix, whereas ours uses the known target subspace. We control the full current-head AGOP and cutoff eigengap without assuming proportionality to a weight Gram matrix; bounded refitting separately measures predictive information.

3 Population training and representation diagnostics

Let $X \sim \gamma_d = N(0, I_d)$, and let $U = (u_1, \dots, u_r) \in \mathbb{R}^{d\times r}$ have orthonormal columns, with $1 \le r < d$. The teacher depends only on $U^T X$; $P_U = UU^T$ and $P_{U^\perp} = I_d - P_U$ project onto its subspace and complement. For functions, $\|g\|_2^2 = \mathbb{E}[g(X)^2]$ and $\|g\|_{H^1}^2 = \|g\|_2^2 + \mathbb{E}[\|\nabla g(X)\|^2]$, with weak derivatives when needed. Unless stated otherwise, $\|y\|_2 = 1$. Vector norms $\|\cdot\|_p$ are ℓ^p norms; unadorned norms are Euclidean for vectors and operator norms for matrices. For symmetric M, $\lambda_k(M)$ are decreasing eigenvalues, $\lambda_{\min}(M)$ is the least, and $M \succeq H$ means $M - H$ is positive semidefinite. Write $\varphi(t) = e^{-t^2/2}/\sqrt{2\pi}$, $\Phi(t) = \int_{-\infty}^t \varphi(z)\,dz$, and $h_k(t) = (-1)^k \varphi^{(k)}(t)/(\sqrt{k!}\varphi(t))$ for the orthonormal probabilists' Hermite polynomials; thus $h_3(t) = (t^3 - 3t)/\sqrt{6}$.

Network and optimization. For a fixed $0 \le \alpha < 1$, put $J_\alpha = 1 - \alpha$ and $\sigma_\alpha(t) = \alpha t + J_\alpha \max\{t, 0\}$; $\alpha = 0$ gives ReLU. With width m, scalar heads $A_{j,n}$, weights $W_{j,n} \in \mathbb{R}^d$, and biases $B_{j,n}$ at update n, the network and loss are

$$f_n(x) = \frac{1}{m}\sum_{j=1}^m A_{j,n}\sigma_\alpha(W_{j,n}^T x + B_{j,n}), \qquad L_n = \mathbb{E}[(y(X) - f_n(X))^2]. \tag{1}$$

Every scalar entry of $A_{j,0}, W_{j,0}, B_{j,0}$ is independently $N(0, s^2/d)$ for the same public $s > 0$. Thus the output heads are random and signed. All parameters undergo simultaneous Euclidean gradient descent with fixed force step $h > 0$ and raw step $\eta = mh/2$:

$$\Theta_{n+1} = \Theta_n - \eta \nabla_\Theta L_n, \qquad \Theta_n = ((A_{j,n}, W_{j,n}, B_{j,n}))_{j=1}^m. \tag{2}$$

For example, $A_{j,n+1} = A_{j,n} + h\mathbb{E}[(y - f_n)\sigma_\alpha(W_{j,n}^T X + B_{j,n})]$. Expectations are over X unless initialization is specified; training uses exact population gradients. The appendix gives the remaining coordinate updates and kink convention.

Full predictor AGOP. Define

$$\mathsf{G}_n = \mathbb{E}[\nabla_x f_n(X) \nabla_x f_n(X)^T], \qquad \nabla_x f_n(X) = \frac{1}{m}\sum_j A_{j,n}\sigma'_\alpha(W_{j,n}^T X + B_{j,n}) W_{j,n}. \tag{3}$$

This includes all pairs of neurons, with their current heads. Whenever $\lambda_r(\mathsf{G}_n) > \lambda_{r+1}(\mathsf{G}_n)$, let P_n be its unique leading rank-r orthogonal projector and set

$$A_{\min,n} = \lambda_{\min}(U^T P_n U), \qquad A_{\mathrm{mean},n} = \frac{1}{r}\operatorname{tr}(U^T P_n U). \tag{4}$$

These are the minimum and mean squared cosines of the principal angles. The minimum score requires recovery of every teacher direction. The theorems establish the required spectral gaps at both compared checkpoints. A null eigenspace is never completed with an arbitrary basis to define a favorable alignment. Multiplying f_n by a nonzero constant rescales its AGOP but preserves its eigenspaces. These scores therefore measure directional geometry separately from output amplitude.

Normalized-feature refit. The main diagnostic uses the full unprojected feature bank:

$$\phi_{j,n}(x) = \frac{\sigma_\alpha(W_{j,n}^T x + B_{j,n})}{\sqrt{\|W_{j,n}\|^2 + B_{j,n}^2}}, \qquad \mathcal{R}(n) = \inf_{\substack{\|v\|_2 \le 32/J_\alpha \\ \|v\|_1 \le 64\sqrt{r}/J_\alpha}} \left\| y - \sum_j v_j \phi_{j,n} \right\|_2^2. \tag{5}$$

A zero augmented row gives the zero feature. Both coefficient caps are identical at every checkpoint and independent of s; there is no additional intercept. The readout is fitted only to measure information available in the features. It does not replace the trained heads in (1). By positive homogeneity, scaling a weight and its bias by the same positive factor leaves the normalized feature unchanged. The diagnostic thus compares feature shapes at a fixed readout budget.

Teacher family for the main theorem. For any probability measure μ on $[1/3, 1]$, define

$$\begin{aligned} g_\mu(t) &= \int_{1/3}^{1} v^{-7/2}(t^3 - 3vt) e^{-(v^{-1}-1)t^2/2}\, d\mu(v), \\ N_\mu &= \|g_\mu\|_{L^2(\gamma_1)}, \qquad q_\mu = g_\mu / N_\mu, \qquad y(x) = \frac{1}{\sqrt{r}} \sum_{i=1}^{r} q_\mu(u_i^T x). \end{aligned} \tag{6}$$

These links belong to $H^1(\gamma_1)$: their mean and first Hermite coefficient vanish, $\sqrt{6} \le N_\mu < 9$, and $\|y\|_2 = 1$. Writing δ_v for a unit point mass at v, $\mu = \delta_1$ gives $q_\mu = h_3$. Whenever $\mu([1/3, 1)) > 0$, the link has an infinite Hermite expansion, starting at degree three. For atomic measures we abbreviate g_{δ_v} as g_v. The entire link is used in the objective and gradients. The positive optimized signal and the later-loss scale are

$$K_* = \frac{J_\alpha}{2\sqrt{r}\, N_\mu} \max_{B \ge 0} \left\{ \sqrt{\frac{B}{1+B}} \int v^{-3/2} \varphi(\sqrt{B/v})\, d\mu(v) \right\}, \qquad G_c = \frac{K_*^2}{32768} > 0. \tag{7}$$

SwiGLU convention. Section 5.1 instead uses $\mathrm{SiLU}(t) = t/(1 + e^{-t})$ and

$$f_\theta(x) = b_0 + \frac{\kappa}{m} \sum_{j=1}^{m} a_j\, \mathrm{SiLU}(w_j^T x + b_j)(z_j^T x + c_j), \qquad \kappa > 0. \tag{8}$$

Here θ collects scalar heads a_j, gate/value weights $w_j, z_j \in \mathbb{R}^d$ and biases b_j, c_j; κ is a fixed output scale. Only the intercept is fitted exactly: $b_0 = \mathbb{E}[y - \widetilde{f}_\theta]$, where $\widetilde{f}_\theta = f_\theta - b_0$. The profiled loss is $L(\theta) = \mathrm{Var}(y - \widetilde{f}_\theta)$. This removes the error in fitting the target mean; the plateau concerns learning its input-dependent variation. All other parameters train jointly by $\dot\theta = -m\nabla L$ or $\theta^{n+1} = \theta^n - mh\nabla L(\theta^n)$: the SwiGLU raw GD step is mh, fixed independently of ε. For $\theta_j = (a_j, (w_j, b_j), (z_j, c_j))$, initialize $\theta_j(0) = \varepsilon\vartheta_j$ with independent $\vartheta_j \sim N(0, \mathrm{diag}(\beta_0^2, s_0^2 I_{2d+2}))$ and fixed $\beta_0, s_0 > 0$. The illustrations use $\beta_0 = 1$, $s_0 = 1/\sqrt{d+1}$, and $\kappa = 1$. The SwiGLU theorem normalizes $\mathrm{Var}(y) = 1$; its illustrations retain $\mathbb{E}[y^2] = 1$ and plot $\ell = L/\mathrm{Var}(y)$. For its full AGOP G, $A_{\mathrm{top}} = \min\{\|P_U e\|^2 : \|e\| = 1,\ \mathsf{G}e = \lambda_1(\mathsf{G})e\}$.

4 Subspace acquisition during a loss plateau

We state the equal-coefficient, common-link case; Appendix D.1 gives the broader statement. The theorem assumes Gaussian inputs, exact population gradients, the specified teacher, and the appendix's explicit size, initialization and step conditions.

The acquisition checkpoint N lies inside the loss-controlled window $0 \le n \le N_c$. At N, the features already support a better bounded refit; at a later N_2, the original predictor beats every loss value in that window. Thus improved representation and improved trained prediction are quantified separately along the same trajectory.

Parameter scope. The appendix prescription is part of the theorem's hypotheses. It supplies explicit occupancy, dimension, mesh, force, and initialization inequalities in an acyclic order before the

random draw. In particular $m \geq r$ and $d \geq 16r/\delta$, but these two inequalities alone are insufficient. Compatible widths may exceed d; arbitrary triples (r, d, m) are not covered. Width must populate rare Gaussian categories, and the raw scale and mesh can be exponentially small in an already large acquisition horizon. The result gives no practical complexity or sample-size bound.

Theorem 4.1 (Population subspace learning on a loss plateau)**.** *Fix $r \geq 1$, μ, and $0 \leq \alpha < 1$ as above, and tolerances $0 < \delta < 1/2$, $0 < \epsilon_L \leq 1$, $0 < \epsilon_G \leq 1/4$. Choose m, d, s, h and the public checkpoints $N < N_c$ according to the complete parameter prescription in Appendix B.1, specialized to $\mu_i = \mu$, $\lambda_i = r^{-1/2}$ and residual $e = 0$. Train (1) by (2) from the stated IID Gaussian initialization. With probability at least $1 - 7\delta/8$, the following hold simultaneously:*

1. ***Loss plateau.*** *Throughout $0 \leq n \leq N_c$, the original loss obeys*
$$1 - \epsilon_L/16 \leq L_n \leq 1 + \epsilon_L/16, \quad L_n \geq 1 - G_c, \quad \frac{\max_{n \leq N_c} L_n}{\min_{n \leq N_c} L_n} \leq 1 + \epsilon_L/4. \tag{9}$$

2. ***Subspace recovery.*** *The full AGOP has a positive rank-r cutoff gap at 0 and N, and its principal-angle scores satisfy*
$$\begin{aligned} &A_{\min,0} \leq A_{\mathrm{mean},0} \leq \tfrac{1}{4}, \qquad A_{\min,N} \geq 1 - \epsilon_G, \qquad A_{\mathrm{mean},N} \geq 1 - \epsilon_G/r, \\ &A_{\min,N} - A_{\min,0} \geq \tfrac{3}{4} - \epsilon_G \geq \tfrac{1}{2}. \end{aligned} \tag{10}$$

3. ***Refit improvement.*** *For the identical bounded diagnostic (5),*
$$\mathcal{R}(0) - \mathcal{R}(N) > 0.399. \tag{11}$$

4. ***Later loss decrease.*** *At a finite subsequent checkpoint $N_2 > N_c$, the same unchanged GD trajectory satisfies*
$$L_{N_2} \leq 1 - 2G_c, \qquad L_n - L_{N_2} \geq G_c \quad (0 \leq n \leq N_c). \tag{12}$$

The teacher subspace $\mathrm{span}(U)$ is minimal. The times N, N_c are fixed before initialization; N_2 may depend on the realized trajectory.

Proof outline. The complete discrete-time proof appears in Appendix D. We summarize the four interfaces that connect acquisition to the stated observable.

1. Directional acquisition. Gaussian occupancy supplies initial neurons in suitable axis and bias regions. A coupled induction on the actual GD updates shows that these neurons acquire every teacher direction by a public time. The induction retains the forces generated by all neurons. A subsequent interval makes their signal dominate the total signed defect, including adverse heads and components outside U. The small raw scale keeps the full loss within (9) while these relative changes occur.

2. From gradients to the AGOP. The proof controls a matrix of second-Hermite coefficients of the *complete* input gradient. Write $Z = U^T X$, $e_r = r^{-1/2}(1, \ldots, 1)^T$, and $\xi = (e_r^T Z)Z - e_r$, so $\mathbb{E}[\xi\xi^T] = I + e_r e_r^T \preceq 2I$. For the symmetric frame $D = \mathbb{E}[(U^T \nabla f_N)\xi^T]$, the signed-defect estimate gives a positive lower bound on $\lambda_{\min}(D)$, as well as a small complete outside-gradient energy. Bessel's inequality yields
$$U^T \mathsf{G}_N U \succeq \tfrac{1}{2} D D^T. \tag{13}$$
Minmax and the positive-semidefinite comparison $P_N \preceq \mathsf{G}_N / \lambda_r(\mathsf{G}_N)$ turn these two estimates into the spectral gap and both angle bounds. This step retains the AGOP cross terms. Initially, rotational invariance and almost-sure simplicity of the positive spectrum imply $\mathbb{E}[A_{\mathrm{mean},0}] = r/d$, averaging over initialization.

3. Prediction from the learned features. The teacher's absence of Hermite degrees below three bounds its correlation with an initial normalized feature by the cube of that feature direction's overlap with U. A simultaneous Gaussian bound and the fixed ℓ^1 cap imply $\mathcal{R}(0) \geq 1 - 10^{-4}/8$. The acquired biased profiles admit a readout satisfying both caps with $\mathcal{R}(N) < (\sqrt{3/5} + 10^{-4}/64)^2$. Their difference exceeds 0.399.

4. Loss release on the same trajectory. A full-energy continuation estimate turns the acquired signal into later loss decrease under the original step size. The acquisition, initial-refit, and initial-projector events have failure probabilities at most $3\delta/8$, $\delta/4$, and $\delta/4$, respectively. A union bound proves the simultaneous claim without assuming these events are independent.

Broader mixture teachers. The full theorem in Appendix D.1 permits different measures μ_i, positive coefficients λ_i with $\sum_i \lambda_i^2 = 1$, and a normalized target $y = \sum_i \lambda_i q_{\mu_i}(u_i^T x) + e(U^T x)$. The residual obeys every explicit H^1 bound in the parameter prescription and may contain interactions within U. Each directional signal K_i^* is defined by (7) with $(r^{-1/2}, \mu)$ replaced by (λ_i, μ_i); the assumption is $\min_i K_i^* \geq (5/6) \max_i K_i^*$. For a common link this allows a coefficient ratio at most $6/5$. Neither arbitrary signed mixtures nor arbitrary H^1 teachers satisfy this hypothesis.

Unequal coefficients near an additive cubic. A complementary theorem, proved in Appendix E, starts from

$$y_c(x) = \sum_{i=1}^{r} a_i h_3(u_i^T x), \qquad a_i > 0, \quad \sum_i a_i^2 = 1, \qquad \|y - y_c\|_{H^1} \leq \nu/2, \tag{14}$$

where $\nu > 0$ is the prescribed perturbation tolerance and the unit-norm teacher y must remain *additive* on U. No ratio bound is imposed on the positive a_i; the smallest coefficient enters the public costs. Under the complete prescription in Appendix B.2, with probability at least $1 - \delta$, its public acquisition checkpoint N_A has minimum-AGOP gain at least $1/2$ during a high-loss plateau. At that checkpoint, the projected diagnostic

$$\mathcal{R}_B^P(n) = \inf_{\|v\|_2 \leq B} \left\| y - \sum_j v_j \sigma_\alpha((P_n W_{j,n})^T X + B_{j,n}) \right\|_2^2 \tag{15}$$

improves by more than $1/2$, both for $B = C_\alpha/(7s)$ and for $B = \infty$. The explicit profile constant C_α and the much stricter admissible angle tolerance are given in that appendix. Projection is before activation, biases are retained, and rows are not renormalized. The same unchanged GD later reduces loss below every plateau value by at least $G_A = 1/(8192 \overline{T}_A^2) > 0$. Here $N_A, \overline{T}_A$ denote the quantities $N, \overline{T}$ in that parameter prescription. Thus this extension uses a different feature class and budget from (5).

5 Population illustrations across teacher links

Twenty-one normalized teachers share a ReLU-student configuration: $r = 2$, $d = 16$, $m = 32$, $s = 10^{-4}$. All parameters train from IID $N(0, s^2/d)$ entries, with fifty predetermined seeds per teacher and a 20,000-update population-GD budget. These are illustrative parameters, outside the theorem's sufficient prescriptions. Labeled followups use $d = 64$, $s = 0.01$ (Appendix A.2).

Full teachers and population evaluation. For each of the thirteen scalar raw links q, we retain its actual mean and set

$$y(X) = \frac{\sum_{i=1}^{r} q(X_i)}{\sqrt{r}\sqrt{\mathbb{E}[q(Z)^2] + (r-1)(\mathbb{E}[q(Z)])^2}}, \qquad Z \sim N(0,1). \tag{16}$$

Six product targets and two additive SiLU controls complete the family (Appendix A); the interactions are exploratory $H^1(\gamma_2)$ examples outside the plateau theorems. Gradients use analytic student moments and validated deterministic teacher quadrature, without sampling.

What the averaged curves measure. Figure 1 shows three teachers; the appendix gives the others. Means and 10th–90th percentiles retain all fifty seeds, including bounds for unresolved AGOP scores. The shared display expands early time without seed-specific alignment or smoothing.

Each seed's loss-only plateau is the maximal initial prefix with MSE in $[0.95, 1.05]$ and running maximum/minimum at most 1.05, checked at every update. Shading marks the common prefix; training runs to the fixed horizon.

Prediction and geometry are separate diagnostics. Refitting uses the unprojected, normalized bank in (5), with fixed caps $\|v\|_2 \leq 32$, $\|v\|_1 \leq 64\sqrt{2}$ and no intercept; it never enters training. Joint success requires resolved minimum-AGOP gain ≥ 0.5 and conservative refit-MSE gain ≥ 0.399 at

one in-plateau checkpoint. Separate ensemble means do not establish this event; the appendix reports continuous gains and later original loss.

Neuron angles and axis-snapped refits probe basis-dependent specialization: *monosemantic* means concentration near teacher axes; *polysemantic* means mixing their coordinates. Snapping projects each teacher-plane component onto its nearest frame axis, preserving bias, outside-plane component and normalization (Appendix A.1). A small snapping penalty alone does not prove specialization. The quadratic-gate target remains at squared L^2 distance at least $1/4$ from every orthonormal two-axis additive class (Proposition A.1).

Optimization. Runs start at $h = 0.05$ (raw step $\eta = mh/2 = 0.8$). Armijo backtracking halves rejected steps and carries the accepted step forward (Appendix A); the theorems use fixed steps.

5.1 SwiGLU: feature learning through a smooth gate

SwiGLU's smooth gate in (8) permits directional learning while its nonconstant output remains small.

Theorem 5.1 (SwiGLU plateau and leading AGOP alignment)**.** *Fix finite $m \geq 1$, $1 \leq r < d$, and $\kappa, \beta_0, s_0, h > 0$. Let $y = F(U^T X)$ with $F \in L^2(\gamma_r)$, $U^T U = I_r$, and $\mathrm{Var}(y) = 1$. Assume that F has a nonzero Hermite component of total degree one, two, or three. Put $k = 3$ if degree one or two is nonzero, and $k = 4$ otherwise; higher degrees are unrestricted. Use the SwiGLU model, profiled intercept, Gaussian initialization and updates of Section 3, with the same Gaussian draw ϑ coupled across ε.*

There is an initialization event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geq 1 - 2^{-m}$ such that, almost surely on $\mathcal{E}$, for every $\delta \in (0,1)$ there are $\tau_1, C > 0$ and $\varepsilon_0 = \varepsilon_0(\vartheta, \delta, h) > 0$ for which every $0 < \varepsilon < \varepsilon_0$ satisfies, at $n_1 = \lfloor \tau_1/(h\varepsilon^{k-2}) \rfloor$,

$$\max_{0 \leq n \leq n_1} |L_n - L_0| \leq C\varepsilon^k, \quad L_0 = 1 + O_\vartheta(\varepsilon^k), \quad A_{\mathrm{top},n_1} \geq 1 - \delta. \tag{17}$$

For gradient flow the same conclusions hold at $t_1 = \tau_1 \varepsilon^{-(k-2)}$, with the maximum replaced by the supremum over $0 \leq t \leq t_1$. The constants τ_1, C depend on (ϑ, δ); the initialization-dependent cutoff gives no explicit finite-scale confidence bound.

Theorem 5.1 controls geometry. To also quantify prediction, we restrict to rank-one teachers with cubic signal and no linear or quadratic Hermite component. At a specified width, the next result gives alignment and unrestricted-refit gains at the same checkpoint during the plateau.

Corollary 5.2 (Joint SwiGLU alignment and refit gains)**.** *In Theorem 5.1, take $r = 1$, $d \geq 2$ and $m = D = (d+1)(d+2)/2$. Let $y = F(u^T X)$ have unit variance, vanishing first and second Hermite coefficients, and $a_3 = \mathbb{E}[F(Z)h_3(Z)] \neq 0$, $Z \sim N(0,1)$. Write $\mathcal{R}_{\mathrm{free}}$ for minimum population MSE after unrestricted refitting of the frozen SwiGLU features and an intercept as defined in (79). Fix $q_{\mathrm{A}}, q_{\mathrm{R}} \in (0,1)$ and $\delta_{\mathrm{A}}, \delta_{\mathrm{R}} > 0$ with $q_{\mathrm{A}} + \delta_{\mathrm{A}} < 1$. For flow or any fixed-step GD, at a common endpoint $*$ the joint event*

$$A_{\mathrm{top},*} - A_{\mathrm{top},0} \geq 1 - \delta_{\mathrm{A}} - q_{\mathrm{A}}, \qquad \mathcal{R}_{\mathrm{free}}(0) - \mathcal{R}_{\mathrm{free}}(*) \geq a_3^2(1 - q_{\mathrm{R}}) - \delta_{\mathrm{R}}$$

along with the entire-prefix loss bound $\max |L - L_0| = O(\varepsilon^4)$ has limiting inferior probability, as $\varepsilon \downarrow 0$, at least

$$\max\left\{0, \Pr(B_d \leq q_{\mathrm{A}}) - 2^{-D} - \frac{3}{q_{\mathrm{R}} d(d+2)}\right\}, \qquad B_d \sim \mathrm{Beta}(1/2, (d-1)/2).$$

The endpoint time scales as ε^{-2}; its rescaled time, constants and initialization cutoff depend on the Gaussian draw. The rescaled time is fixed before ε shrinks.

A zero bound is vacuous; high probability requires suitable dimensions and thresholds. For pure h_3, $d = 16$ and $m = 153$, the gains can be at least 0.5 and 0.7, with asymptotic probability lower bound approximately 0.9519 (Corollary T.14). No practical initialization cutoff is supplied; refit coefficients may diverge as $\varepsilon \to 0$. Later decrease of the trained loss remains open.

Proofs. Appendix T.2 proves Theorem 5.1 and further gate/value alignment bounds; Corollary T.14 proves Corollary 5.2.

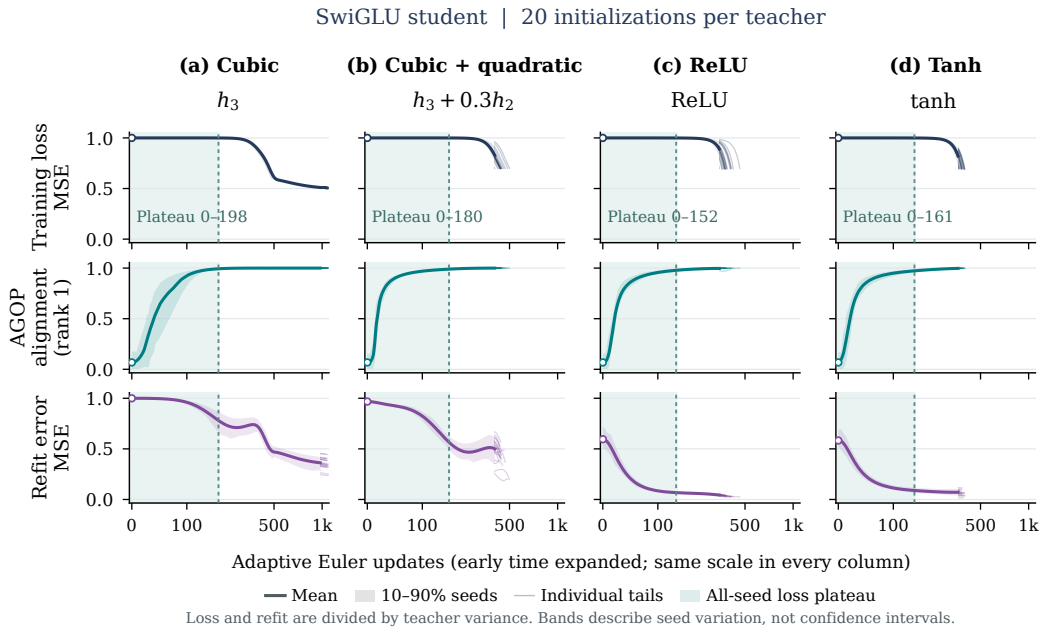


Figure 2: **SwiGLU students: twenty initializations per teacher.** All seeds 0–19; $d = 16$, $m = 64$, $\varepsilon = .05$. Loss and numerical unrestricted-refit MSE are divided by teacher variance; rank-one AGOP alignment is $A_{\text{top}} = A_{\min}$. Solid means and light 10th–90th percentile bands use all twenty seeds on their shared support; later individual tails remain visible. Shading marks the shared loss-only prefix $|\ell - \ell(0)| \leq 10^{-3}$. The axis counts adaptive Euler updates, with early updates expanded. Diagnostic interpolation is for display only; bands show seed variation, not confidence intervals or integration-error bounds. Crosses, if present, mark the last finite point of censored or failed trajectories.

6 Discussion

Theorem 4.1 connects every-direction AGOP recovery, bounded-refit gain, and later loss reduction: feature geometry can improve during a high-loss window. Refitting measures predictive information in the learned features; it does not intervene in the training trajectory.

Gaussian Sobolev regularity includes our positive-mixture and additive-cubic families, and interaction targets $g(S)T/\|g\|_2$, where $S = u_1^T X$, $T = u_2^T X$, and g is SiLU, ReLU, GELU, $\tanh$, or h_2. Their angle and axis-snapping diagnostics illustrate mixed neurons (Figures 1, 3, and 4). For $g = h_2$, Proposition A.1 gives MSE at least $1/4$ when every neuron's teacher-plane component lies on a shared orthonormal pair of axes. This distinguishes subspace recovery from axis specialization as approximation objectives, without proving their training dynamics.

The mechanisms suggest further extensions. For SiLU and GELU, trainable biases can expose cubic teacher correlations, and collective neuron responses can reveal teacher directions while output remains small. For broader Sobolev links, one can seek to control the effect of additional components on the acquisition gradient. High-degree components may exert weak forces yet carry substantial target energy, so prediction needs separate control. SwiGLU uses gating to separate directional learning from output growth, as illustrated by its leading-direction theorem here. These observations motivate extensions, not additional guarantees in this paper.

Gaussian Sobolev regularity alone is insufficient: our guarantees require structured teachers and restrictive initialization and size conditions. Priorities include sharper parameter costs, persistence after acquisition, and finite-sample counterparts to the population analysis.

Conclusion. A loss plateau is an incomplete measure of learning progress. Jointly tracking loss, AGOP geometry and refit error under the same budget can reveal feature acquisition during a nearly flat loss window.

AI USE DISCLOSURE

Generative AI tools assisted with problem formulation and hypothesis development, mathematical modeling and theorem formulation, proof development and writing, experimental design, implementation and debugging, numerical diagnostics, interpretation of results, literature discovery, and manuscript and figure preparation. The authors are responsible for the correctness, originality, and final content of this work. AI-assisted reviews and numerical checks do not constitute formal verification.

REPRODUCIBILITY STATEMENT

The appendix gives all technical assumptions and supporting proofs, including the analytic scalar inequalities used in the prediction bounds. No theorem depends on executing software or consulting auxiliary research notes. The population-moment implementation, fixed numerical protocols, recorded per-seed diagnostics, and scripts that reproduce the figures without retraining are available in the anonymous code and data repository: https://anonymous.4open.science/r/subspace-learning-plateaus-0D78/. The supplied training commands regenerate raw parameter archives omitted from the compact release. The separate typesetting archive contains the manuscript sources and rendered figures. The experiments use deterministic Gaussian moment evaluations in floating-point arithmetic; they are not finite-sample simulations or interval-certified computations.

REFERENCES

Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2552–2623. PMLR, 2023. URL https://proceedings.mlr.press/v195/abbe23a.html.

Mark Ainsworth and Yeonjong Shin. Plateau phenomenon in gradient descent training of ReLU networks: Explanation, quantification, and avoidance. *SIAM Journal on Scientific Computing*, 43 (5):A3438–A3468, 2021. doi: 10.1137/20M1353010. URL https://epubs.siam.org/doi/10.1137/20M1353010.

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.

Ioannis Bantzis, James B. Simon, and Arthur Jacot. Saddle-to-saddle dynamics in deep ReLU networks: Low-rank bias in the first saddle escape. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=B4zcoLvjw0.

Boaz Barak, Benjamin L. Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21750–21764, 2022. doi: 10.52202/068431-1581. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/884baf65392170763b27c914087bde01-Abstract-Conference.html.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021. URL https://jmlr.org/papers/v22/20-1288.html.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning Gaussian multi-index models with gradient flow part I: General properties and two-timescale learning. *Communications on Pure and Applied Mathematics*, 78(12):2354–2435, 2025. doi: 10.1002/cpa.70006. URL https://doi.org/10.1002/cpa.70006.

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.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20105–20118, 2022. doi: 10.52202/068431-1462. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7eeb9af3eb1f48e29c05e8dd3342b286-Abstract-Conference.html.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/ae614c557843b1df326cb29c57225459-Abstract.html.

Paul G. Constantine, Eric Dow, and Qiqi Wang. Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014. doi: 10.1137/130916138. URL https://doi.org/10.1137/130916138.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5413–5452. PMLR, 2022. URL https://proceedings.mlr.press/v178/damian22a.html.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349):1–65, 2024. URL https://www.jmlr.org/papers/v25/23-1543.html.

Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James B. Simon, Michael R. DeWeese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0156. URL https://papers.nips.cc/paper_files/paper/2025/hash/06cbd2e81dfbd3bb4cb0abce95b32584-Abstract-Conference.html.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes ReLU network features, 2018. URL https://arxiv.org/abs/1803.08367.

Neil Rohit Mallinar, Daniel Beaglehole, Libin Zhu, Adityanarayanan Radhakrishnan, Parthe Pandit, and Mikhail Belkin. Emergence in non-neural models: grokking modular arithmetic via average gradient outer product. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42834–42856. PMLR, 2025. URL https://proceedings.mlr.press/v267/mallinar25a.html.

Hancheng Min, Enrique Mallada, and Rene Vidal. Early neuron alignment in two-layer ReLU networks with small initialization. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0f07eb4d39094358667640e4bd9f2c9d-Abstract-Conference.html.

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://doi.org/10.1126/science.adi5639.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6120.

Berfin Şimşek, Amire Bendjeddou, and Daniel Hsu. Learning Gaussian multi-index models with gradient flow: Time complexity and directional convergence. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4204–4212. PMLR, 2025. URL https://proceedings.mlr.press/v258/simsek25a.html.

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.

Gan Yuan, Mingyue Xu, Samory Kpotufe, and Daniel Hsu. Efficient estimation of the central mean subspace via smoothed gradient outer products. *SIAM Journal on Mathematics of Data Science*, 7(3):1241–1264, 2025. doi: 10.1137/23M1626700. URL https://doi.org/10.1137/23M1626700.

TABLE OF CONTENTS

1 Introduction 1
2 Related work 2
3 Population training and representation diagnostics 4
4 Subspace acquisition during a loss plateau 5
5 Population illustrations across teacher links 7
6 Discussion 9
A Numerical protocol and complete teacher comparison 14
B Teacher families and complete parameter prescriptions 27
C Exact update and minimal teacher rank 29
D Full mixture theorem and proof 30
E The additive cubic-neighborhood extension 32
F Gaussian kernels and compatible simultaneous updates 34
G Original Gaussian initialization and prediction baselines 38
H Mixture profile acquisition 41
I Control of every signed row for mixture teachers 46
J The complete symmetric derivative frame 50
K A bounded readout for the full mixture links 52
L Acquisition of full-bank signal before loss release 55
M Same-step release of the original population risk 57
N Cubic row growth and selected-cone invariance 61
O Cubic normalized correlation and complete-row geometry 63
P Simultaneous maturation of the cubic categories 74
Q Complete head-weighted outside control for cubic teachers 76
R Original-coordinate anchors and the cubic signed frame 79
S The identical projected-bank diagnostic 86
T SwiGLU students: leading-direction learning during a plateau 90

A NUMERICAL PROTOCOL AND COMPLETE TEACHER COMPARISON

Run matrix and reproducible initialization. The original baseline comparison has twenty-one teacher functions, one ReLU student configuration, and the fixed initialization seeds 1–50 for each teacher: 1,050 trajectories. The sizes are $r = 2, d = 16, m = 32$, with all raw heads, spatial weights and biases drawn independently from $N(0, 10^{-8}/d)$. The same seed produces the same initial parameter arrays across teachers; seed outcomes across different teachers are therefore paired, not independent additional trials. The configuration is inherited from the earlier illustrations. Seeds 1–3 were previously inspected; the remaining forty-seven extend that comparison. No seed is selected or removed because of its trajectory.

The endpoint-gain table below retains this original baseline. Four replacement teacher panels use the separate followup protocol in Section A.2.

All parameters train throughout. Every trajectory has a fixed budget of 20,000 updates; neither the loss plateau nor feature learning stops training. The earlier 114-run ReLU/leaky-ReLU/Adam study and the separate fifteen-run gated extension remain archived under their original protocols. They are not pooled as extra seeds in these averages.

Computational environment. Seeds 1–20 use Python 3.11.4 on macOS and fresh seeds 21–50 use Python 3.11.10 on Linux, with identical frozen numerical source and NumPy 1.24.3, SciPy 1.10.1, and Clarabel 0.11.1. Disjoint scheduling uses six local and twenty-four remote CPU workers, with one numerical thread per worker. Three fixed 1,000-update portability checks (additive cubic, SiLU product, and quadratic product at seed 1) agree with the corresponding local trajectory prefixes to within 2×10^{-15} in the compared loss, step, path, and parameter arrays. This check does not assert bitwise equality of all subsequent trajectories across platforms. Completed transfers retain source hashes; no partial trajectory is resumed in a different runtime.

Population gradient descent. For full MSE L and parameter vector θ, write $F = -(m/2)\nabla L(\theta)$. The nominal force step is $h = 0.05$, or raw learning rate $\eta = mh/2 = 0.8$. A trial is accepted if

$$L(\theta + hF) \le L(\theta) - 10^{-4}\frac{2h}{m}\|F\|^2 + 10^{-13}\max\{1, |L(\theta)|\}. \tag{18}$$

Otherwise h is halved; the accepted step persists and cannot increase. At most forty halvings are allowed per update. This safeguard is used uniformly from initialization, rather than switched on at an observed plateau endpoint. A divergent or nonfinite run is a recorded failure, not an excluded seed. The update axis does not represent a common physical-time horizon when accepted steps differ. Cumulative raw GD time, accepted steps and backtracking counts are saved separately. These variable-step experiments are distinct from the fixed-step learning theorem.

Complete teacher definitions. For $Z \sim N(0,1)$ let $h_k = \mathrm{He}_k/\sqrt{k!}$ and use the positive-mixture building block g_v defined in the main theorem. The thirteen scalar raw links are

$$\begin{gathered} h_3, \quad g_{2/3}, \quad g_{1/3}, \quad (g_1 + g_{1/3})/2, \quad g_{1/2}, \quad 0.2g_{0.4} + 0.3g_{0.7} + 0.5g_1, \\ h_3 + 0.1h_5, \quad h_3 + 0.1\sin, \quad h_2, \quad h_4, \quad z_+, \quad |z|, \quad \sin z. \end{gathered}$$

Each gives the full additive target in (16), including its mean and every Hermite component. The positive-mixture links are in the central teacher family; the displayed perturbation magnitude 0.1 is not certified by the small-neighborhood theorem. For independent $S, T \sim N(0,1)$, the six product targets are

$$Y_g = \frac{g(S)T}{\sqrt{\mathbb{E}[g(S)^2]}}, \qquad g \in \{\mathrm{SiLU},\ z_+,\ z\Phi(z),\ \tanh,\ h_2,\ z\}. \tag{19}$$

Here $z\Phi(z)$ is exact GELU and $\mathrm{SiLU}(z) = z/(1 + e^{-z})$. The two further controls are

$$\begin{aligned} Y_{\mathrm{raw}} &= \frac{\mathrm{SiLU}(S) + \mathrm{SiLU}(T)}{\sqrt{2\mathbb{E}[\mathrm{SiLU}(S)^2] + 2(\mathbb{E}[\mathrm{SiLU}(S)])^2}}, \\ Y_{\mathrm{centered}} &= \frac{\mathrm{SiLU}(S) + \mathrm{SiLU}(T) - 2\mathbb{E}[\mathrm{SiLU}(S)]}{\sqrt{2\,\mathrm{Var}(\mathrm{SiLU}(S))}}. \end{aligned} \tag{20}$$

All targets have unit second moment. Products have zero mean but can retain a linear component in T. For nonzero $g \in H^1(\gamma)$, $\|Y_g\|^2_{H^1(\gamma_2)} = 2 + \|g'\|_2^2/\|g\|_2^2$. This ambient regularity is not a plateau guarantee for these standalone interaction targets. Expectations use analytic student moments and analytic or deterministic one-dimensional teacher quadrature, without finite training or test samples.

Per-seed plateaus and common observation times. For each seed separately, the plateau ends immediately before the first update at which MSE leaves $[0.95, 1.05]$ or its running maximum/minimum ratio exceeds 1.05. Loss is checked at every update. If no exit occurs by update 20,000, the plateau is right-censored at that horizon. AGOP and refit do not select the interval. The numerical tables use the exact endpoint state, while ensemble curves use the common grid $n = 0, 100, \ldots, 20000$. Initialization, the endpoint and its successor are retained separately. A first qualifying saved checkpoint is not an exact hitting time between saved states.

Averaging and uncertainty. Every seed has weight $1/50$. Identified values are summarized by arithmetic means and pointwise empirical 10th–90th percentiles across seeds. Where a score is only known within a numerical interval, the figure instead propagates lower and upper bounds on its mean and percentile band. These are variability bands, not confidence intervals for a population mean. Loss uses its full every-update record; the other curves join observed common-grid diagnostics without smoothing or imputing an unmeasured diagnostic. Neuron-angle heatmaps average the separately normalized importance distributions of all seeds. They do not pool unnormalized neuron mass in a way that gives high-norm runs more weight.

Let E_s be the endpoint update of initialization seed s. Pale background marks $[0, \min_s E_s]$, the intersection of the individual loss-selected plateaus; endpoint quantiles additionally summarize their variation. The panel abbreviation "cens." counts prefixes that have not ended by the fixed horizon. A plateau of the average loss would not establish the same property on every run and is not used. Likewise, separately high mean alignment and mean refit improvement do not imply that they occur together within each run. The archived event counts are computed from individual trajectories at common checkpoints within each run's own plateau, including its exact endpoint.

Full AGOP and unresolved spectra. We use current trained heads, activation gates and all signed cross-neuron terms in the population AGOP. Positive scalar rescaling for eigendecomposition leaves its eigenspaces unchanged. A rank-r score is numerically resolved only if both $\lambda_r/\lambda_1 > 10^{-9}$ and $(\lambda_r - \lambda_{r+1})/\lambda_1 > 10^{-9}$. This threshold is a reporting convention, not a uniform error theorem. For averaging, an unresolved score contributes its possible interval $[0, 1]$: its mean and percentile uncertainty are propagated, and the resolved fraction is reported. It is never set to zero, treated as a proved alignment, or silently removed from a fifty-seed denominator. The hidden-weight energy fraction $A_{\mathrm{sub}} = \sum_j \|UU^T W_j\|^2 / \sum_j \|W_j\|^2$ (undefined for a zero denominator) is saved but does not replace the full-predictor AGOP statistic.

Constrained refitting and its numerical error. At every checkpoint, use all unprojected features divided by their original augmented row norms $\sqrt{\|W_j\|^2 + B_j^2}$, with caps $\|v\|_2 \le 32$ and $\|v\|_1 \le 64\sqrt{2}$ and no extra intercept. For normalized features $\phi = (\phi_1, \ldots, \phi_m)^T$, let $H = \mathbb{E}[\phi(X)\phi(X)^T]$ and $c = \mathbb{E}[y(X)\phi(X)]$. Minimize $R(v) = 1 - 2c^T v + v^T H v$ on that fixed feasible set. Warm starts do not change this set. No ridge is added and no Gram eigenmode is discarded in the objective. Each feasible candidate gives an upper risk $\overline{R}$; a first-order support-function bound, with negative-eigenvalue and arithmetic allowances, supplies a numerical lower estimate $\underline{R}$. Original and snapped refit brackets remain separate from the between-seed bands. Solver-tolerance misses remain flagged; these are numerical optimization brackets for the computed moments, not rigorous interval bounds on integration error.

Per-run outcomes. Define the conservative numerical refit improvement $\widehat{\Delta}_n = \underline{R}_0 - \overline{R}_n - 10^{-5}$. The joint reference event requires resolved initial/current spectra, $A_{\min,n} - A_{\min,0} \ge 0.5$ and $\widehat{\Delta}_n \ge 0.399$ at one checkpoint in the same plateau. We distinguish the exact endpoint event from an event at any saved in-prefix checkpoint, and record later original-model loss separately. Failure of the absolute 0.399 test is not absence of useful learning when the initial refit risk is already small. Mean and linear target components remain in the optimization objective. Their contributions and nonlinear

residual errors are retained in the machine-readable results; large initial means are not removed to create a high-loss claim.

Table 1: Arithmetic means of per-seed endpoint gains and final original MSE. All means give equal weight to all 50 seeds. The endpoint is each seed's own exact loss-prefix endpoint. Brackets are mean lower and upper bounds, not confidence intervals; unresolved AGOPs contribute $[0, 1]$ before taking gain differences. Refits retain numerical lower and feasible upper risks. Final MSE is shown as mean and seed standard deviation.

Teacher	n	Mean $\Delta A_{\min}$	Mean refit decrease	Final MSE
$\lvert z\rvert$	50	$[-0.013, 0.727]$	-0.021	0.147 ± 0.106
st	50	0.972	0.769	0.003 ± 0.001
$g_{1/2}$	50	$[0.952, 0.972]$	0.697	0.099 ± 0.053
$g_{1/3}$	50	$[0.912, 0.972]$	0.473	0.173 ± 0.058
$g_{2/3}$	50	0.972	0.783	0.032 ± 0.003
$\mathrm{GELU}(s)t$	50	0.972	0.240	0.017 ± 0.005
h_2	50	0.972	0.468	0.005 ± 0.001
h_3	50	0.972	0.601	0.050 ± 0.014
$h_3 + 0.1h_5$	50	0.972	0.545	0.066 ± 0.018
$h_3 + 0.1 \sin$	50	0.972	0.259	0.190 ± 0.080
h_4	50	0.972	0.133	0.198 ± 0.069
$g_1 + g_{1/3}$	50	$[0.952, 0.972]$	0.656	0.089 ± 0.058
$h_2(s)t$	50	$[0.952, 0.972]$	0.553	0.080 ± 0.037
ReLU	50	$[-0.028, 0.972]$	0.110	0.000 ± 0.000
$\mathrm{ReLU}(s)t$	50	0.972	0.224	0.004 ± 0.000
Raw additive SiLU	50	$[-0.028, 0.972]$	0.141	0.001 ± 0.008
Centered additive SiLU	50	$[-0.028, 0.972]$	0.193	0.001 ± 0.001
$\mathrm{SiLU}(s)t$	50	0.972	0.307	0.023 ± 0.019
$\sin z$	50	$[-0.028, 0.972]$	0.209	0.061 ± 0.006
$\tanh(s)t$	50	0.972	0.819	0.002 ± 0.001
Three-scale mixture	50	0.972	0.808	0.012 ± 0.002

A.1 Neuron axes and frame interventions

Here u_1, u_2 are the orthonormal teacher directions, and A_j, W_j, B_j denote a neuron's trained head, spatial vector and bias at the displayed checkpoint. The two extra rows in the additive/gated panels distinguish subspace recovery from axis concentration. For a nonzero in-plane row set

$$\vartheta_j = \operatorname{atan2}(|W_j^T u_2|, |W_j^T u_1|) \in [0, 90^\circ], \qquad \omega_j = \frac{|A_j| \|W_j\|}{\sum_k |A_k| \|W_k\|}. \tag{21}$$

Five-degree bins contain the sum of these weights; zero in-plane rows are omitted and the omitted mass is recorded. The heatmap averages these histograms across seeds. Concentration near zero or ninety degrees describes teacher-axis alignment; interior angles mix teacher coordinates. This unsigned projection is not itself a prediction metric. The labels *monosemantic* and *polysemantic* refer to this chosen basis and observed time, not a permanent or semantic classification of every neuron.

For an orthonormal frame rotated within the teacher plane, project each row's in-plane component onto its closest unoriented frame axis. Preserve its bias, outside-plane component and *original* augmented denominator. Refit with exactly the same coefficient caps as for the original bank. The fourth row compares original features, teacher-axis snapping, 45° snapping, and the best of eighteen frames at $0, 5, \ldots, 85^\circ$. The two fixed frames are evaluated on the common hundred-update grid; the eighteen-frame comparison is evaluated at $0, 1000, \ldots, 20000$. For the best-grid numerical interval we minimize lower and feasible risks separately at each seed and checkpoint before averaging. The selected frame can therefore differ between seeds and times; it is not a single shared rotation. In all cases, markers denote evaluated states and connecting lines only guide the eye. For original and snapped brackets the gap interval is

$$[\underline{R}_F - \overline{R},\ \overline{R}_F - \underline{R}]. \tag{22}$$

A small gap means that the intervention preserves refit quality; it does not prove that the original neurons already specialized. Negative gaps can occur because original and snapped spans are not nested. A tested rotation grid is not an optimization over all rotations.

An exact obstruction for a fixed orthonormal frame. The following elementary population statement separates axis geometry from subspace recovery. It is an approximation bound, not a learning dynamics theorem.

Proposition A.1 (Orthogonal-frame floor)**.** *Let $X \sim N(0, I_d)$, $P = UU^T$, and let (a_1, a_2) be an orthonormal basis of* $\mathrm{span}(U)$*. Write $Z_k = a_k^T X$ and let Y be a square-integrable function of $U^T X$. Suppose every row of a finite ridge network satisfies $PW_j \in \mathrm{span}(a_1)$ or $PW_j \in \mathrm{span}(a_2)$, with arbitrary biases and outside-plane components, and square-integrable features. Then its population MSE is at least*

$$\mathcal{E}_F(Y) = \big\| Y - \mathbb{E}[Y \mid Z_1] - \mathbb{E}[Y \mid Z_2] + \mathbb{E}[Y] \big\|_2^2. \tag{23}$$

For independent $S = u_1^T X, T = u_2^T X \sim N(0,1)$ and $Y = g(S)T/\|g\|_2$, the teacher-frame floor is $\mathrm{Var}(g(S))/\mathbb{E}[g(S)^2]$*. The minimum floor over all orthonormal frames is $1/4$ when $g = h_2$, and zero when $g(z) = z$.*

Proof. Gaussian independence of PX and $(I - P)X$ makes the conditional expectation of each ridge feature given $U^T X$ a univariate function of its selected Z_k. Thus $\mathbb{E}[f(X) \mid U^T X]$ is additive in Z_1, Z_2. Conditional Jensen bounds the network's MSE below by the squared distance of Y from this additive space. Since Z_1, Z_2 are independent, their centered univariate spaces are orthogonal and the residual is exactly (23). In the teacher frame, the additive projection of $g(S)T/\|g\|_2$ is $(\mathbb{E}[g(S)])T/\|g\|_2$, giving the stated variance ratio. For a frame rotated by θ, put $c = \cos\theta$, $s = \sin\theta$. Hermite orthogonality gives captured additive energy $3c^2s^2$ for $h_2(S)T$ and $4c^2s^2$ for ST. Since $c^2s^2 \le 1/4$, their minimum residual energies are $1/4$ and zero, respectively, both attained at $45°$. □

This obstruction concerns every orthonormal frame, whereas a finite rotation-grid diagnostic is only a comparison over the tested frames. Neither statement excludes all nonorthogonal frames. For smooth gates, a computed $45°$ floor does not establish that this rotation is the best possible one. The proposition is an approximation bound; it adds no learning-dynamics theorem for interaction targets.

A.2 Denser followups for four short-prefix teachers

Why a separate followup is needed. The raw and centered additive SiLU, additive ReLU and additive sine teachers have short loss-selected prefixes under the original $d = 16$, $s = 10^{-4}$ protocol. Their early behavior is difficult to see on a 20,000-update axis. A display change alone also leaves a numerical problem: the original endpoint rank-two spectra are unresolved for these four cohorts. We therefore retain the original outcomes in the baseline tables and report a separate experiment with better resolved spectra and denser observations. These familiar links remain empirical cases outside the positive-family theorem.

Pilot and disclosed configuration choice. The pilot varied $s \in \{10^{-4}, 0.01, 0.1\}$ at $d = 64$, $m = 32$, $r = 2$, with force step $h = 0.05$, seeds 51–53 and 3,000 updates: 36 trajectories over the four teachers. No configuration attained the predeclared joint alignment/refit event on at least two of its three seeds. Consequently, the pilot did not identify a reliable regime of full-subspace learning during the plateau. For a common, interpretable comparison we subsequently chose $s = 0.01$ for all four teachers; all twelve pilot endpoints at this scale had resolved rank-two spectra. This choice was made after inspecting the pilot, not by maximizing the frozen joint-event score. All pilot outcomes are retained.

Fifty-seed followup and unchanged definitions. Each selected teacher uses seeds 1–50, 20,000 updates, $d = 64$, $m = 32$, $r = 2$, and IID $N(0, 0.01^2/d)$ entries for every raw head, weight and bias. These seeds are disjoint from the pilot but reuse the baseline seed identifiers; the followup is not an independent replication of the original protocol. All parameters train. The targets, persistent Armijo rule, plateau test, full-predictor AGOP, resolution convention and constrained refit are unchanged. No teacher mean is removed except in the explicitly centered SiLU control. Computation uses Linux CPUs, one numerical thread per worker, with the same numerical libraries and frozen population kernels as the baseline. No finite sample or stochastic gradient is used.

The loss is saved at every update. AGOP and refit diagnostics are saved every ten updates through 1,000, every fifty through 3,000, and every hundred thereafter, together with each exact plateau endpoint and its successor. For each teacher the figure pairs a linear 0–500-update view with a linear 0–20,000-update view. Both use all fifty seeds with equal weights and empirical 10th–90th percentile envelopes. Shading ends at $\min_s E_s$, the earliest individual loss-selected endpoint; neither alignment nor refit selects the shaded interval. An expanded early view is not a claim of a longer physical-time plateau.

Interpretation. Both $A_{\min}$ and A_{mean} are shown. An increase in the mean with little increase in the minimum describes partial geometric learning, not recovery of every teacher direction. Refit gains are measured separately and may include improved prediction of low-degree components. A successful refit therefore does not by itself establish full-subspace recovery. The table below retains the same simultaneous reference event used in the original comparison and reports later original-model loss separately.

All 200 initial states, endpoint states and saved in-prefix spectra are resolved under the stated numerical convention. The simultaneous reference event occurs in two of fifty raw-SiLU runs and in none of the other three cohorts. In contrast, median mean-alignment and refit gains are substantial across all four teachers, while weakest-direction gains are smaller. Thus these followups display partial feature and prediction improvement on the plateau without claiming uniform recovery of both teacher directions. The 281 unresolved checkpoints later in training remain visible as numerical uncertainty in the plots; no refit solver-tolerance misses occurred. Better visibility does not imply better final optimization: raw SiLU has median final MSE 0.0542 here, versus 0.000279 in the original baseline. The reduction in mean-plus-linear squared prediction error accounts for a median 90–98% of the original-loss decrease during these plateaus. Accordingly, the displayed refit gains should not be interpreted as isolated learning of higher Hermite components.

Table 2: Separate four-teacher followup, fifty seeds each. Endpoint columns report medians of individual-run endpoint changes, not changes at an averaged endpoint. E is the plateau endpoint update (median and range); $\Delta A = A(E) - A(0)$ for each alignment score; $\Delta R = \underline{R}_0 - \overline{R}_E - 10^{-5}$ is the conservative numerical refit gain. The last column counts the simultaneous reference event anywhere on a saved in-prefix checkpoint. Later loss is the original model at update 20,000.

Teacher	E [range]	$\Delta A_{\min}$	ΔA_{mean}	ΔR	L_{20000}	Joint
Raw additive SiLU	128 [122,137]	0.222	0.576	0.43	0.054	2/50
Centered additive SiLU	203 [190,227]	0.000411	0.476	0.589	0.00056	0/50
Additive ReLU	113 [108,120]	0.115	0.534	0.313	1.2e-06	0/50
Additive sine	203.5 [191,219]	0.0231	0.487	0.607	0.062	0/50

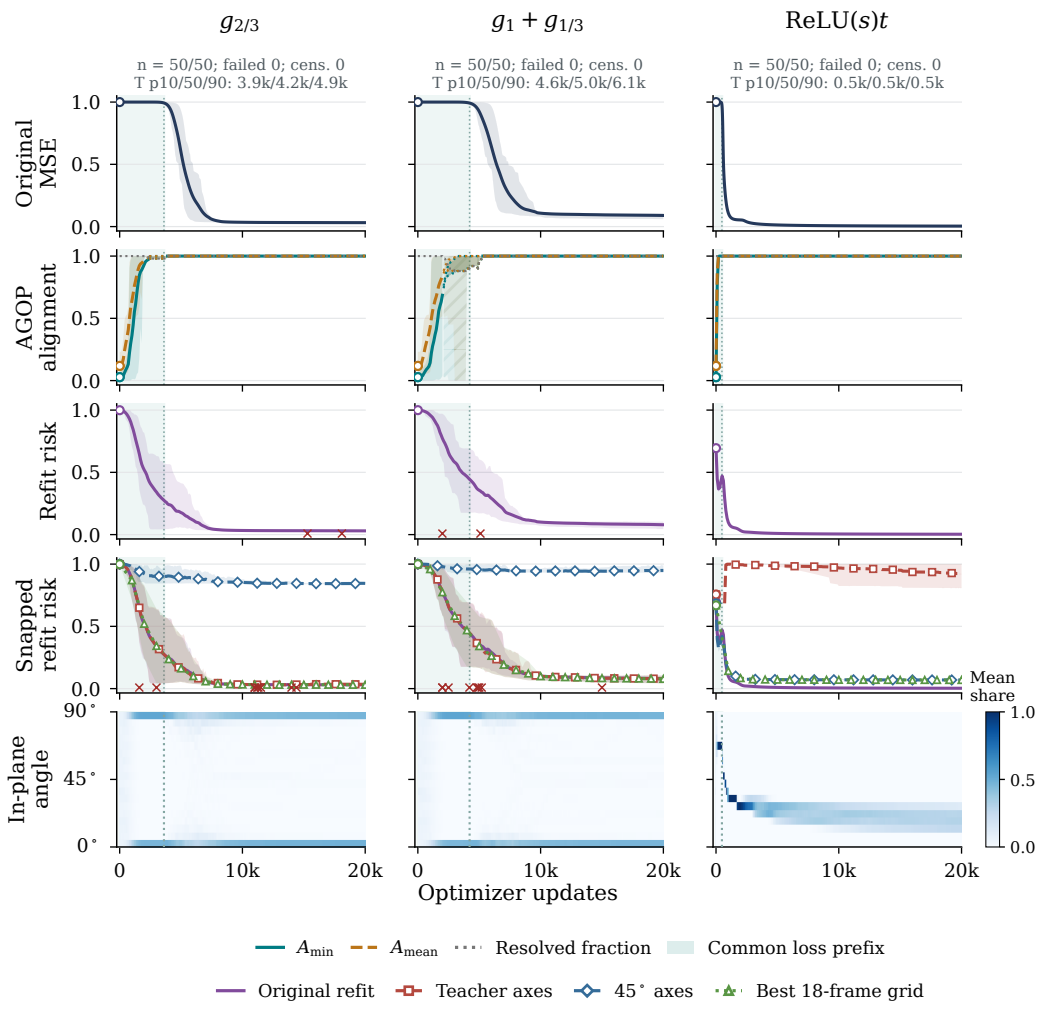


Figure 3: **Additive damped and mixture links, and the ReLU gate.** The two additive-family targets and the ReLU product use five diagnostic rows. Each teacher uses fifty fixed initialization seeds with identical optimization parameters. Means, numerical bounds, percentile bands and loss-prefix shading follow Figure 1.

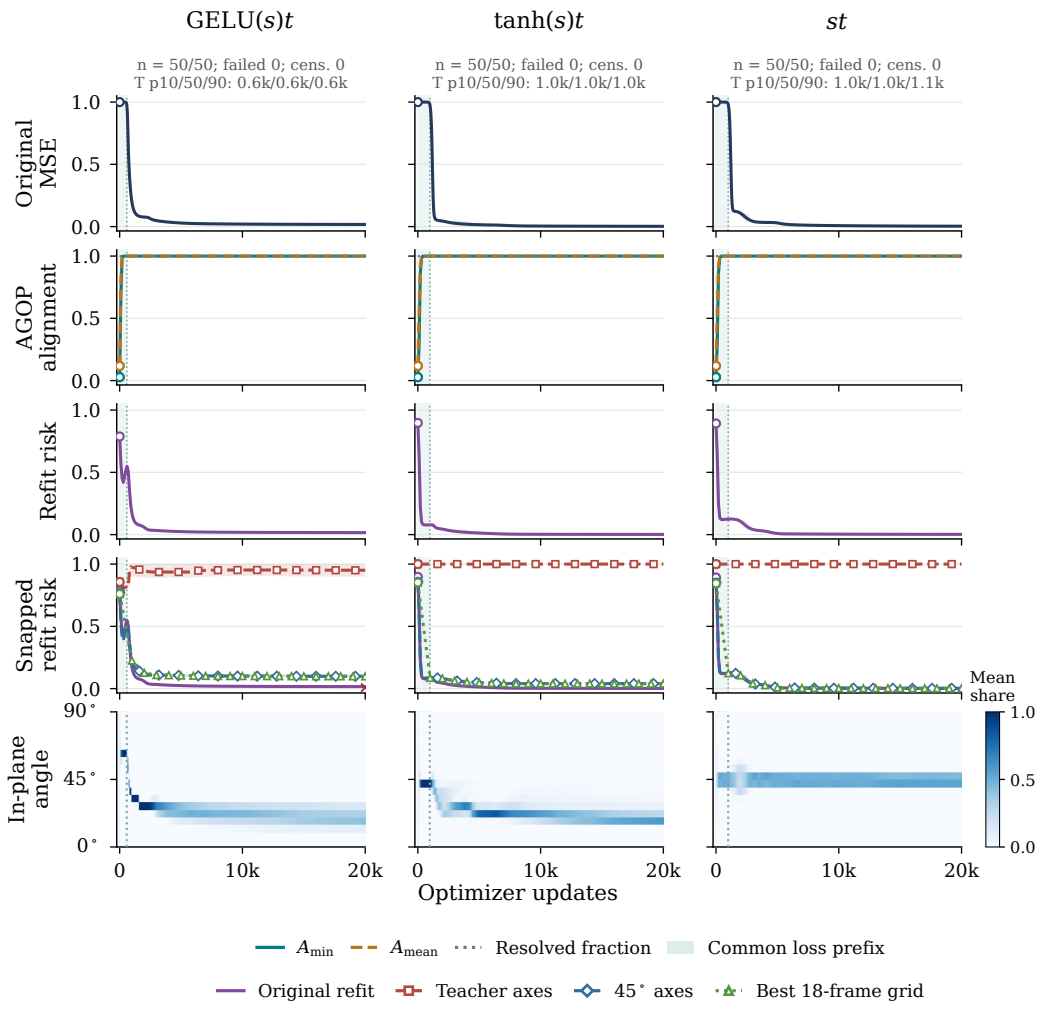


Figure 4: **Other gated teachers.** Exact GELU and tanh products, together with the bilinear rotation control, use the same five rows. Each teacher uses fifty fixed initialization seeds with identical optimization parameters. Means, numerical bounds, percentile bands and loss-prefix shading follow Figure 1.

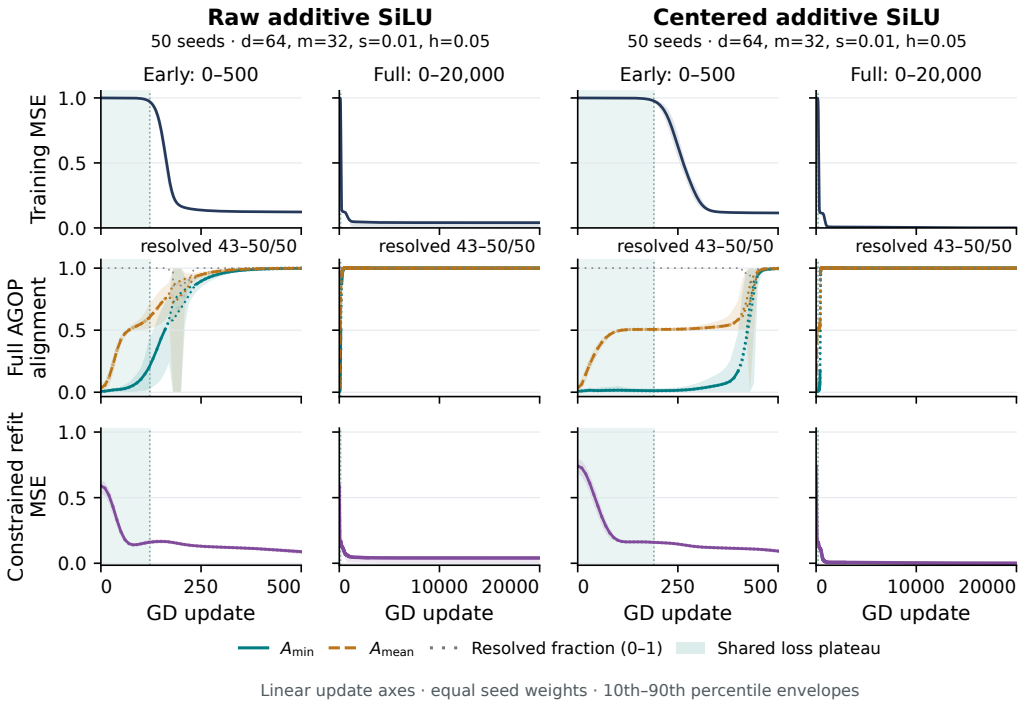


Figure 5: **Raw and centered additive SiLU: denser followups.** Each teacher has paired linear early (0–500 updates) and full (0–20,000 updates) views, with fifty seeds at $d = 64$, $m = 32$, $s = 0.01$. Rows show original loss, minimum and mean full-AGOP alignment, and constrained refit MSE. Bands are empirical 10th–90th percentiles; the gray trace reports the resolved fraction. Refit lower/upper numerical bounds remain separate. Shading marks the loss-prefix intersection across all fifty seeds. The common configuration was chosen after no pilot configuration passed the joint-event promotion criterion, as disclosed in Section A.2. The enlarged early view exposes partial alignment and refit changes; the averages do not establish full-subspace recovery for every seed.

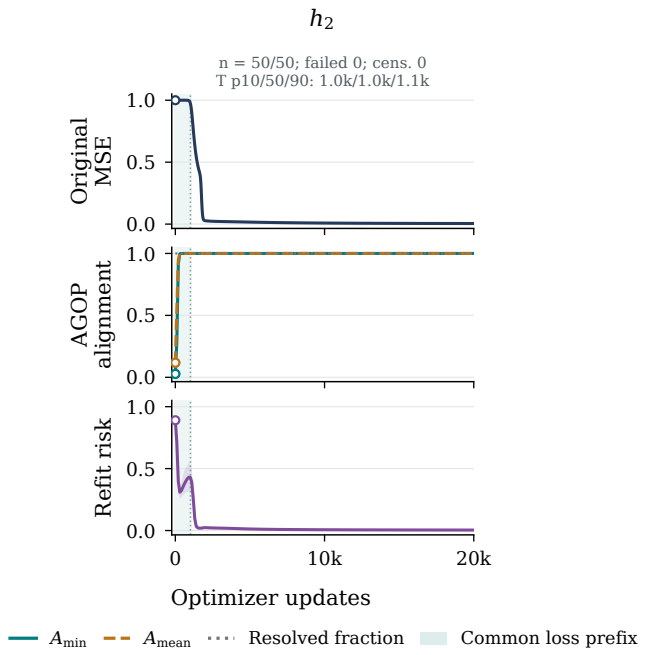


Figure 6: **Quadratic additive teacher.** The three rows display original loss, full-current-head AGOP alignment and same-budget refit MSE. Each teacher uses fifty fixed initialization seeds with identical optimization parameters. Means, numerical bounds, percentile bands and loss-prefix shading follow Figure 1.

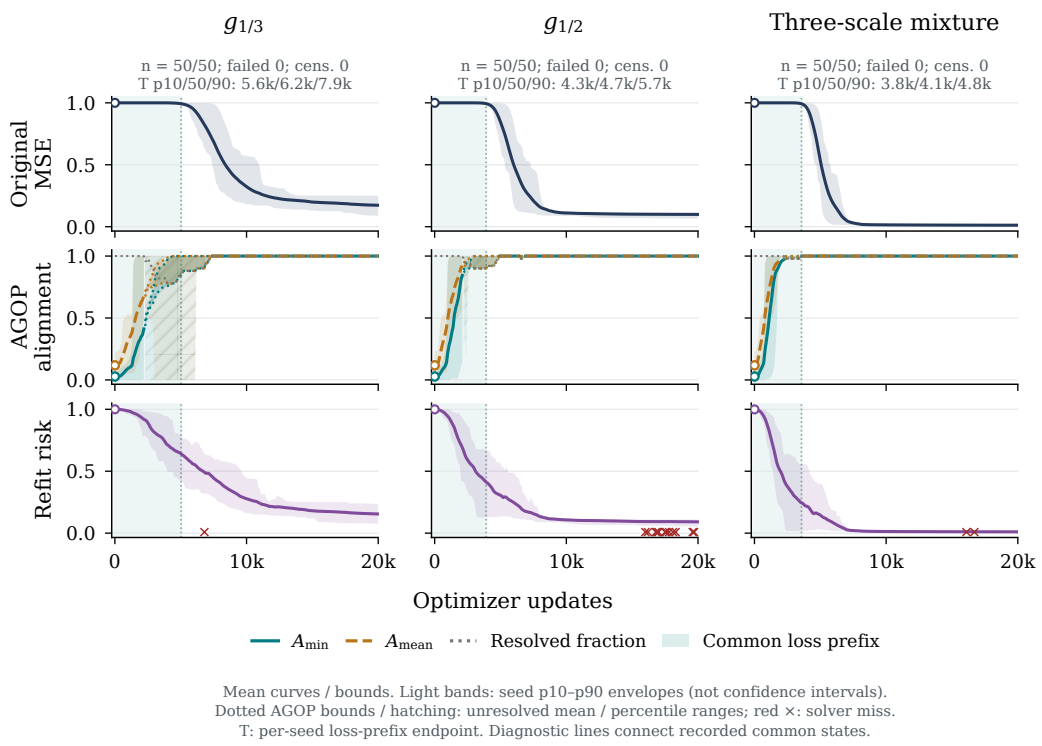
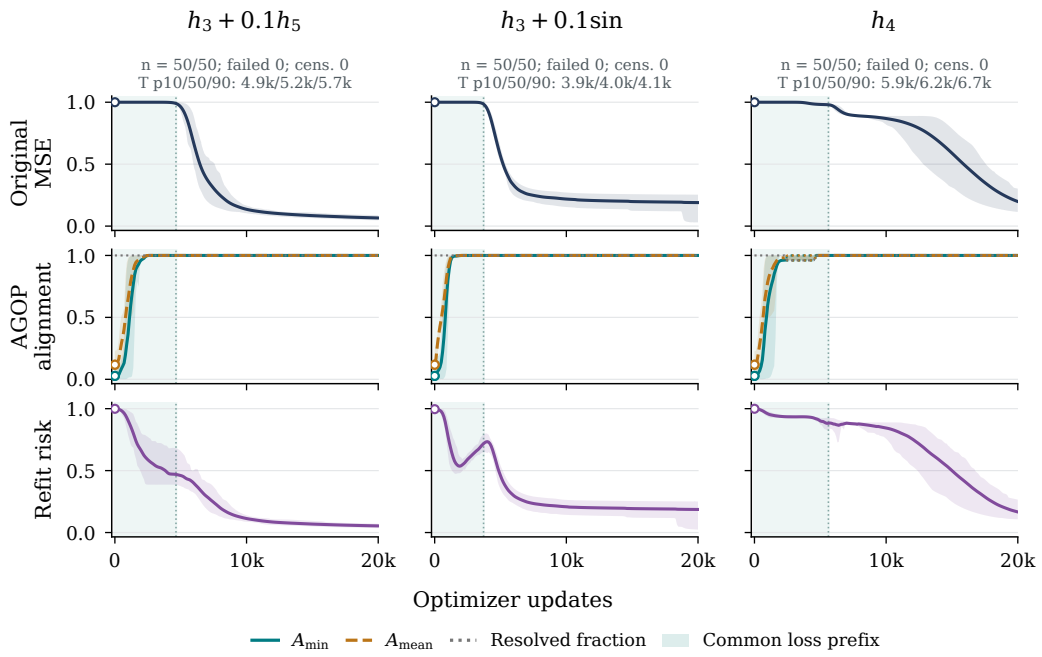


Figure 7: **Additional positive-mixture links.** Damped links and the three-scale mixture use the same three-row format. Each teacher uses fifty fixed initialization seeds with identical optimization parameters. Means, numerical bounds, percentile bands and loss-prefix shading follow Figure 1.



Mean curves / bounds. Light bands: seed p10–p90 envelopes (not confidence intervals).
Dotted AGOP bounds / hatching: unresolved mean / percentile ranges; red ×: solver miss.
T: per-seed loss-prefix endpoint. Diagnostic lines connect recorded common states.

Figure 8: **Perturbed cubic and fourth-Hermite teachers.** These standalone perturbation magnitudes and the fourth Hermite are empirical cases, not claims of the small-neighborhood theorem. Each teacher uses fifty fixed initialization seeds with identical optimization parameters. Means, numerical bounds, percentile bands and loss-prefix shading follow Figure 1.

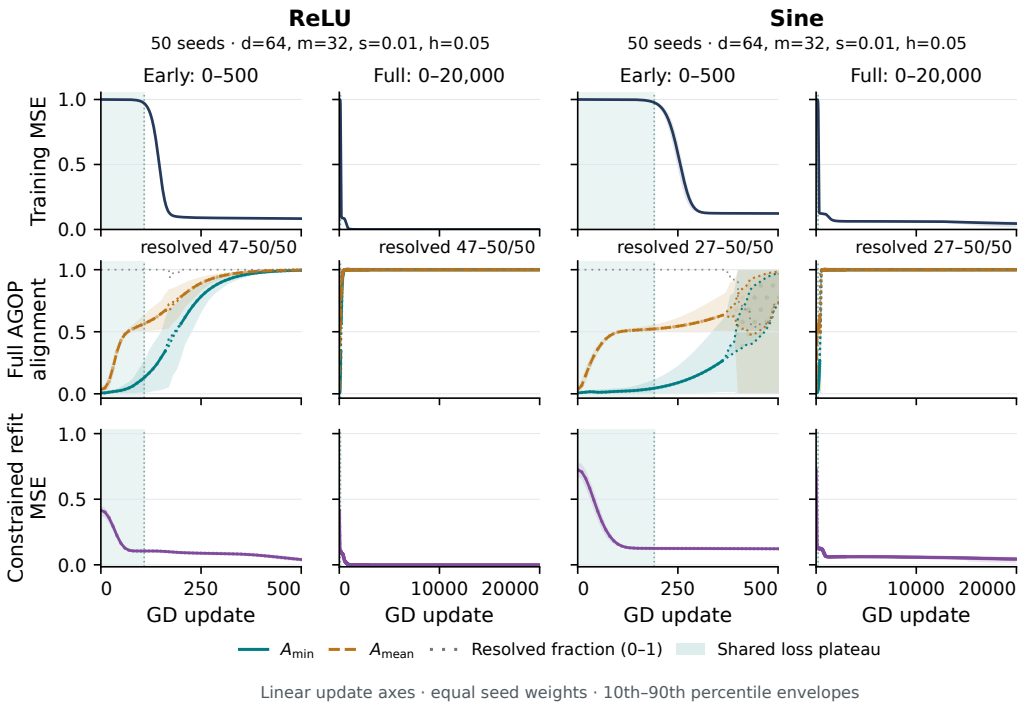


Figure 9: **Additive ReLU and sine: denser followups.** The paired early and full views use the same fifty-seed configuration, three rows, linear update axes, numerical uncertainty conventions and loss-prefix shading as Figure 5. Both teacher means and all Hermite components are retained. Partial geometric learning and refit gains are distinguished from the simultaneous weakest-direction criterion in Table 2.

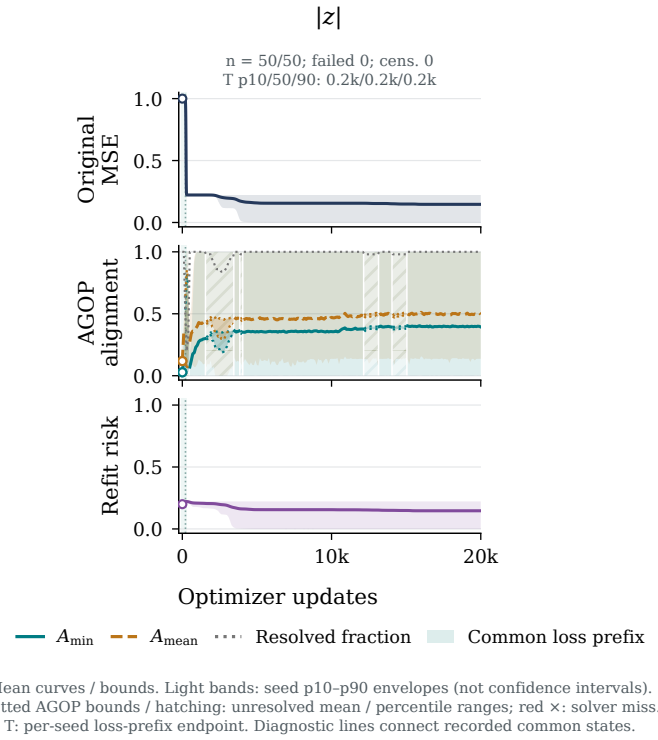


Figure 10: **Absolute-value teacher: original baseline.** This panel retains the original fifty seeds at $r = 2$, $d = 16$, $m = 32$, $s = 10^{-4}$ and initial force step $h = 0.05$; it was not rerun. The three rows show original loss, full AGOP alignment and constrained refit MSE over 20,000 updates. Means, percentiles, unresolved spectral intervals and shared loss-prefix shading follow the original baseline protocol. It is separated here so the four new followups do not displace the absolute-value control.

B TEACHER FAMILIES AND COMPLETE PARAMETER PRESCRIPTIONS

B.1 POSITIVE MIXTURES OF GAUSSIAN THIRD DERIVATIVES

Here $X \sim \gamma_d = N(0, I_d)$, $U = (u_1, \dots, u_r)$, $P_U = UU^T$ and $P_{U^\perp} = I_d - P_U$. Function norms are Gaussian: $\|g\|_2^2 = \mathbb{E}[g^2]$ and $\|g\|_{H^1}^2 = \mathbb{E}[g^2 + \|\nabla g\|^2]$. Write $h_k = \mathrm{He}_k / \sqrt{k!}$ for the normalized probabilists' Hermite polynomial, and Φ for the standard normal CDF. Expectations without an indicated law are over the displayed Gaussian input; expectations of initialization statistics are over the original parameter draw. Let $J = 1 - \alpha > 0$, $U^T U = I_r$, and let μ_i be probability measures on $[1/3, 1]$. Throughout this appendix φ_v denotes the density of $N(0, v)$ and $\varphi = \varphi_1$. Define

$$g_v(t) = v^{-7/2}(t^3 - 3vt)e^{-(v^{-1}-1)t^2/2}, \qquad g_i = \int g_v \, d\mu_i(v), \qquad N_i = \|g_i\|_2,$$

$$y_c(X) = \sum_{i=1}^{r} \lambda_i g_i(u_i^T X)/N_i, \qquad \lambda_i > 0, \qquad \sum_i \lambda_i^2 = 1. \tag{24}$$

The target is $y = y_c + e = y(U^T X)$ with $\|y\|_2 = 1$ and $\|e\|_{H^1} \le \zeta$. The residual may contain interactions on U. The central common-link theorem takes $\mu_i = \mu$, $\lambda_i = r^{-1/2}$ and $e = 0$. For the more general theorem require

$$K_i^* = \max_{B \ge 0} \frac{J\lambda_i}{2N_i} \sqrt{\frac{B}{1+B}} \int v^{-3/2} \varphi(\sqrt{B/v}) \, d\mu_i(v), \qquad K_* := \max_i K_i^*, \qquad \min_i K_i^* \ge \frac{5}{6} K_*. \tag{25}$$

Here and below, all quantities are public choices made before the independent Gaussian draw. The following list prints the entire prescription, including the conditions needed for later loss release. There is no further implicit width or dimension condition.

Structural constants and sizes. Fix $0 < \delta < 1/2$, $0 < \epsilon_L \le 1$ and $0 < \epsilon_G \le 1/4$. Set

$$\lambda_* := \min_i \lambda_i, \quad \chi_0 = 10^{-4}, \quad e_* = \frac{J\chi_0}{4096\sqrt{r}}, \quad \varepsilon_{\mathrm{BM}} = e_*/8, \quad \kappa_{\mathrm{BM}} = J\lambda_*/1024,$$

$$\vartheta = \min\left\{\lambda_*/1000, e_*/16, \sqrt{\kappa_{\mathrm{BM}}\varepsilon_{\mathrm{BM}}/6400}\right\}, \quad R = \max\left\{8, 16/e_*, \sqrt{25600/(\kappa_{\mathrm{BM}}\varepsilon_{\mathrm{BM}})}\right\}. \tag{26}$$

With $p_I = \Phi(1) - \Phi(1/2)$ and $p_G = \Phi(2) - \Phi(1)$, put

$$p_{\mathrm{rect}} = p_G p_I^2 \begin{cases} 1, & r = 1, \\ [2\Phi(\vartheta/(4\sqrt{r-1})) - 1]^{r-1}, & r > 1, \end{cases} \qquad m \ge \max\{r, p_{\mathrm{rect}}^{-1} \log(16r/\delta)\}. \tag{27}$$

Choose integer $d > r$ satisfying all of

$$\ell = \log(64m/\delta), \quad B_1 = 64\sqrt{r}/J, \quad \ell_I = \log(8m/\delta), \quad Q_I = r + 2\sqrt{r\ell_I} + 2\ell_I,$$
$$d \ge \max\{r, 64, 128\ell, 16r/\delta, 16\ell_I, 2Q_I(32B_1/\chi_0)^{2/3}\}. \tag{28}$$

These conditions allow compatible $m \ge d$. Once d is fixed set $x_0 = d^{-1/2}$, $k = J\lambda_*/(4096d^2)$, and $Q_* = k/(10\sqrt{d})$.

Acquisition and spectral clocks. Define

$$T_1 = \kappa_{\mathrm{BM}}^{-1} \log(4/\varepsilon_{\mathrm{BM}}), \quad T_0 = 4R/k + T_1 + 1, \quad S_0 = 5me^{4T_0}, \quad v_* = 4R^2,$$
$$c_G = \frac{J\varphi(1)\sqrt{\epsilon_G}}{128\sqrt{r}}, \quad \Gamma = \frac{3S_0/4 + 400m}{v_*}, \quad T_G = \frac{100}{K_*} \log \frac{2\Gamma}{c_G},$$
$$T = T_0 + T_G + 1, \quad S = S_0 e^{4(T_G+1)}, \quad M = 2\sqrt{S}. \tag{29}$$

Choose one positive mesh h and one positive force allowance ν obeying *every* cutoff in

$$h \le \min\left\{2^{-12}, \frac{k}{512M}, \frac{1}{4M^2T}, \frac{\kappa_{\mathrm{BM}}\varepsilon_{\mathrm{BM}}}{3200}, \frac{K_*}{10000}, \frac{1}{1000r}, 10^{-4}, \frac{c_G v_*}{20000TS}\right\}, \tag{30}$$

$$\nu \le \min\left\{\begin{array}{c} \frac{Q_*}{4}, \frac{1}{400T}, \frac{x_0}{4MT}, \frac{\vartheta\lambda_* k x_0^2}{2048M^3}, \frac{k}{512M}, \frac{2k}{3M^2}, \\ \frac{1}{80M^2T}, \frac{\kappa_{\mathrm{BM}}\varepsilon_{\mathrm{BM}}}{320}, \frac{K_*}{10000}, \sqrt{\frac{K_*}{1000T}}, \frac{c_G v_*}{80TS} \end{array}\right\}. \tag{31}$$

Set

$$N_* = \lceil 4R/(hk) \rceil + \lceil T_1/h \rceil, \qquad N = N_* + \lceil T_G/h \rceil,$$
$$t_G = (N - N_*)h, \qquad v_G = v_* e^{3.32 K_* t_G}. \tag{32}$$

In particular $N_* h \le T_0$, $T_G \le t_G \le T_G + h$ and $Nh \le T$.

Continuation, residual and original Gaussian scale. After choosing the same h, define

$$\gamma = K_*/2, \quad k_c = \left\lceil \frac{2}{h\gamma} \log \frac{32S}{\gamma v_*} \right\rceil, \quad T_c = h + \frac{2}{\gamma} \log \frac{32S}{\gamma v_*},$$
$$S_c = S e^{4T_c}, \quad N_c = N + k_c, \quad R_c^2 = \gamma m/64, \quad G_c = \gamma^2/8192 = K_*^2/32768,$$
$$0 < \nu_c \le \min\{\gamma/16, \sqrt{\gamma/(8T_c)}\}. \tag{33}$$

The allowable residual and raw scale satisfy

$$0 \le \zeta \le \min\left\{\nu/2, \chi_0/128, \lambda_*/100, \chi_0/(32B_1), \nu_c/2\right\}, \tag{34}$$

$$0 < s^2 \le \min\left\{\frac{m\nu}{S}, \frac{m\epsilon_L}{20M^2}, \frac{m\nu_c}{S_c}, \frac{R_c^2}{4S_c}, \frac{mG_c}{S_c}, \frac{m\epsilon_L}{20S_c}\right\}. \tag{35}$$

All raw coordinates are independently $N(0, s^2/d)$ and the single Euclidean step on full MSE is $\eta = mh/2$. Smaller positive h, ν, s are allowed, with the downstream quantities recomputed in the printed order. No parameter is chosen from an observed favorable trajectory. The very small scale and mesh are part of the theorem.

B.2 Additive cubic neighborhood: complete parameter package

Let $X \sim N(0, I_d)$, let $U = (u_1, \dots, u_r)$ have orthonormal columns, and let

$$y_c = \sum_{i=1}^{r} a_i h_3(u_i^T X), \qquad a_i > 0, \quad \sum_i a_i^2 = 1, \quad a_0 = \min_i a_i, \quad a_{\max} = \max_i a_i. \tag{AJ1}$$

In this subsection the local $N, \overline{T}, G_{\mathrm{rel}}$ are the main-text $N_A, \overline{T}_A, G_A$, respectively. Here $h_3(t) = (t^3 - 3t)/\sqrt{6}$, $\gamma_d = N(0, I_d)$, $\|g\|_2^2 = \mathbb{E}[g(X)^2]$, and $\|g\|_{H^1}^2 = \mathbb{E}[g(X)^2 + \|\nabla g(X)\|^2]$. Nonzero coefficient signs can be absorbed into the u_i. For $0 \le \alpha < 1$, train

$$f_n(X) = \frac{1}{m} \sum_{j=1}^{m} A_{j,n} \sigma_\alpha(W_{j,n}^T X + B_{j,n}), \qquad \sigma_\alpha(t) = \alpha t + (1-\alpha) t_+, \qquad L_n = \mathbb{E}[(y - f_n)^2] \tag{AJ2}$$

by simultaneous Euclidean GD in every raw coordinate at one fixed step $\eta = mh/2$. All initial raw coordinates, including heads and biases, are independent $N(0, s^2/d)$. There is no screening, resetting, freezing, sign selection by the algorithm, or refitting during training.

Fix $0 < \delta < 1/2$ and $0 < \delta_L \le 1$, and put $\delta_0 = \delta/2$ and $\zeta = \delta/2$. The following scalar constants depend only on the indicated public parameters, not on a realized initialization:

$$b_*^2 = (\sqrt{5} - 1)/2, \quad p = 2\Phi(b_*) - 1, \quad v_g = p - 2b_*\varphi(b_*) + b_*^2(1-p) - p^2,$$
$$t_g = -2b_*\varphi(b_*)/\sqrt{6}, \quad K_* = t_g^2/v_g > 2/3, \quad D_\alpha = \sqrt{(1-\alpha)^{-2} + p^2/(1+\alpha)^2},$$
$$C_\alpha = |t_g| D_\alpha / v_g, \qquad D_g = |t_g| \max\{p, 1-p\}/v_g,$$
$$\xi = \min\{(16\sqrt{r})^{-1}, (128 C_\alpha \sqrt{r})^{-1}\}, \quad \vartheta = \min\{a_0/8, \xi/32\}, \quad R_D = 64/\xi. \tag{AJ3}$$

Here φ, Φ are the standard normal density and distribution function. Set

$$p_C = [\Phi(2) - \Phi(1)][\Phi(2) - \Phi(1/2)]^2 [2\Phi(\vartheta/(4\sqrt{r-1})) - 1]^{r-1}, \tag{AJ4}$$

omitting the final factor when $r = 1$. Choose integer sizes satisfying

$$m \ge \max\{4r, \log(16r/\delta_0)/p_C\}, \quad \ell = \log(96m/\delta_0), \quad H = r + 2\sqrt{r\ell} + 2\ell,$$
$$d \ge \max\{256, 128\ell, 16H, (r+4)(128/\delta_0)^{1/3}, 16r/\delta\}. \tag{AJ5}$$

In particular $r < d$ and $m \geq r$. There is no $m < d$ requirement. The restriction on width is stronger than $m \geq r$: it pays for all four axis/bias categories under the original Gaussian draw.

Define the original-coordinate constants

$$g_0 = \frac{\zeta a_0}{4mr^2\sqrt{d}}, \quad b_0 = a_{\max}\sqrt{H/d}, \quad D_a = \sum_i a_i^{-2}, \quad L_a = (a_{\max}^2 D_a)^{1/3},$$
$$Z_* = 40 + 640\sqrt{D_a}\{b_0^2/g_0 + (1+L_a)b_0\}, \quad \Gamma = 3Z_* + 182. \tag{AJ6}$$

Choose any $0 < \varepsilon \leq \min\{1/4, (768D_g^2)^{-1}\}$ and then a public radius R with

$$R \geq \max\left\{R_D, 16\sqrt{m}, \sqrt{m\Gamma/18}, \sqrt{\frac{m\sqrt{1522/\varepsilon}}{18(1-\alpha)\varphi(1)}}\right\}. \tag{AJ7}$$

Only after this choice set

$$c = (1-\alpha)/\sqrt{6}, \quad k = c\varphi(1)a_0/(64d^2), \quad T = 1 + 8R/k, \quad \overline{T} = T + 1,$$
$$M^2 = 20me^{4\overline{T}}, \quad \overline{S} = 4\log(8M), \quad 0 < h \leq \min\{1/512, \xi/1024, (16\overline{T})^{-1}\}, \quad N = \lceil T/h \rceil. \tag{AJ8}$$

Choose $\nu > 0$ at most the minimum of

$$\frac{1}{64}, \quad \frac{k}{10^4M^2}, \quad \frac{1}{1024\overline{T}^2}, \quad \frac{k}{64M^2\overline{S}\sqrt{d}}, \quad \frac{k\vartheta a_0}{1024M^3d}, \quad \frac{k\xi}{12800M}, \quad \frac{m}{1024\overline{T}M^2}, \quad \frac{g_0e^{-4\overline{T}}}{10^5\overline{T}M}. \tag{AJ9}$$

The full target is any additive function on U^TX with

$$\|y\|_2 = 1, \qquad y \in H^1(\gamma_d), \qquad \|y - y_c\|_{H^1} \leq \nu/2. \tag{AJ10}$$

Its actual mean, linear component and higher tails remain in all gradients, losses and comparisons. Since $\nu/2 < a_0$, its minimal index subspace is U, by the cubic tensor argument in Appendix C. Thus r here is a true teacher rank, not only a representation. Finally put

$$a_{\rm rel} = 1/(64m\overline{T}), \quad R_{\rm rel}^2 = a_{\rm rel}m^2, \quad G_{\rm rel} = 1/(8192\overline{T}^2),$$
$$0 < s^2 \leq \min\left\{\frac{m\nu}{M^2}, \frac{m\delta_L}{50M^2}, \frac{R_{\rm rel}^2}{4M^2}, \frac{mG_{\rm rel}}{M^2}\right\}. \tag{AJ11}$$

This order of choices is acyclic: Z_* is determined before R, and neither Z_* nor R depends on the ensuing clock or raw scale. All quantities are finite and positive at every allowed finite rank and coefficient vector. No practical scale is asserted.

C Exact update and minimal teacher rank

Here $X \sim N(0, I_d)$, $f = m^{-1}\sum_j A_j\sigma_\alpha(W_j^TX + B_j)$, $\sigma_\alpha(t) = \alpha t + (1-\alpha)t_+$, and a superscript $+$ denotes the next simultaneous iterate. The teacher axes are the columns of U, and $h_3(t) = (t^3 - 3t)/\sqrt{6}$. The coefficient families a_i and λ_i/N_i are those in AJ1 and (24), respectively. For $z_j = W_j^TX + B_j$, the simultaneous full-MSE step $\eta = mh/2$ is

$$A_j^+ = A_j + h\mathbb{E}[(y-f)\sigma_\alpha(z_j)],$$
$$W_j^+ = W_j + hA_j\mathbb{E}[(y-f)\sigma'_\alpha(z_j)X],$$
$$B_j^+ = B_j + hA_j\mathbb{E}[(y-f)\sigma'_\alpha(z_j)]. \tag{36}$$

Every right-hand side uses the same current state. At a nonzero spatial row the kink has Gaussian probability zero. At a pure-bias row with nonzero bias the derivative is the corresponding affine derivative. The identically zero joint row is absorbing and may use any fixed subgradient convention. Initial nonzero balanced hidden states remain nonzero by the compatible balance identity proved below.

For a unit vector u, the normalized cubic $h_3(u^T X)$ corresponds isometrically to the symmetric third-chaos tensor $u^{\otimes 3}$. The central cubic tensor is $\sum_i c_i u_i^{\otimes 3}$, with $c_i = a_i$ in AJ and $c_i = \sqrt{6}\lambda_i/N_i$ in the mixture theorem. Its mode-one matricization has singular values c_i, because both u_i and $u_i \otimes u_i$ are orthonormal families. Orthogonal chaos projection and matricization give an operator perturbation at most $\|y - y_c\|_2$. In AJ this is less than a_0; in SW it is at most $\lambda_*/100 < \sqrt{6}\lambda_*/9$. Thus the cubic matricization of the actual target has rank at least r. If the target were measurable on a smaller linear subspace, all modes of every chaos tensor would lie in that subspace, a contradiction. Since the target depends on $U^T X$ by hypothesis, its minimal index subspace is exactly U.

For $Z \sim N(0,1)$, the mixture identity $g_v \varphi = -\varphi_v'''$ implies

$$\mathbb{E}[g_v(Z) e^{tZ - t^2/2}] = t^3 e^{-(1-v)t^2/2}.$$

The resulting coefficients are printed in MP2. Their squared sums and Hermite-degree-weighted squared sums converge uniformly for $v \in [1/3, 1]$. Indeed their generating function is $K(t) = 3(2+3t)(1-t)^{-7/2}$ at $t = (1-v)^2 \le 4/9$; the weighted sum is $2tK'(t) + 3K(t)$. Minkowski's inequality therefore justifies integration against arbitrary probability measures, including non-atomic ones, in both L^2 and H^1. It gives $\sqrt{6} \le N_i < 9$ and $\|g_i/N_i\|_{H^1} < 13$. The coefficient at every odd degree $2j+3$ is nonzero whenever $\mu_i([1/3,1)) > 0$. No truncation enters the updates or the proofs.

D Full mixture theorem and proof

Use the teacher and parameter choices of Appendix B.1, with $J = 1 - \alpha$ and $U = (u_1, \dots, u_r)$. The full raw student is $f_n = m^{-1} \sum_j A_{j,n} \sigma_\alpha(W_{j,n}^T X + B_{j,n})$, with $\sigma_\alpha(t) = \alpha t + J t_+$ and $X \sim N(0, I_d)$. For clarity, $L_n = \mathbb{E}[(y - f_n)^2]$, $\mathsf{G}_n = \mathbb{E}[\nabla_x f_n \nabla_x f_n^T]$, and P_n denotes its leading rank-r spectral projector at the theorem's checkpoints. The alignment scores are $A_{\min,n} = \lambda_{\min}(U^T P_n U)$ and $A_{\mathrm{mean,n}} = r^{-1} \operatorname{tr}(U^T P_n U)$. The diagnostic below uses $\phi_{j,n} = \sigma_\alpha(W_{j,n}^T X + B_{j,n}) / \sqrt{\|W_{j,n}\|^2 + B_{j,n}^2}$, with a zero augmented row assigned zero feature: $\mathcal{R}(n)$ is the infimum of $\|y - \sum_j v_j \phi_{j,n}\|_2^2$ under the two stated coefficient budgets.

Theorem D.1 (Comparable mixture signals)**.** *Under all conditions in Appendix B.1, let $\mathcal{R}(n)$ be the normalized, unprojected, full-bank diagnostic with ℓ^2 budget $32/J$ and ℓ^1 budget $64\sqrt{r}/J$. With probability at least $1 - 7\delta/8$ over the original Gaussian draw, the following conclusions hold simultaneously. Both leading rank-r AGOP projectors P_0, P_N have positive cutoff gaps, and*

$$A_{\min,0} \le A_{\mathrm{mean},0} \le 1/4, \qquad A_{\min,N} \ge 1 - \epsilon_G, \qquad A_{\mathrm{mean,N}} \ge 1 - \epsilon_G/r, \tag{37}$$

$$\lambda_r(\mathsf{G}_N) \ge a_G := \tfrac{1}{2}[J\varphi(1) s^2 v_G/(64m\sqrt{r})]^2 > 0, \qquad \lambda_{r+1}(\mathsf{G}_N) \le \epsilon_G a_G, \tag{38}$$

$$\mathcal{R}(0) \ge 1 - \chi_0/8, \qquad \mathcal{R}(N) < (\sqrt{3/5} + \chi_0/64)^2, \qquad \mathcal{R}(0) - \mathcal{R}(N) > 0.399. \tag{39}$$

For the entire prefix $0 \le n \le N_c$,

$$1 - \epsilon_L/16 \le L_n \le 1 + \epsilon_L/16, \qquad L_n \ge 1 - G_c, \qquad \frac{\max_{n \le N_c} L_n}{\min_{n \le N_c} L_n} \le 1 + \epsilon_L/4. \tag{40}$$

There is a finite later $N_2 > N_c$ with $L_{N_2} \le 1 - 2G_c$, and hence $L_n - L_{N_2} \ge G_c$ for every $n \le N_c$, under the same original GD update. The physical clocks obey

$$\frac{m}{2}(4R/k + T_1 + T_G) \le \eta N \le mT/2, \qquad \eta(N - N_*) \ge mT_G/2, \qquad \eta N_c \le (m/2)(T + T_c). \tag{41}$$

One may take N_2 to be the first subsequent hit of joint raw radius R_c. If $\mathfrak{R}_{N_c}^2 = \sum_j (A_j^2 + \|W_j\|^2 + B_j^2)$ at N_c and $a_{\mathrm{sig}} = \gamma/(64m)$, then

$$\frac{m}{4} \log \frac{R_c}{\mathfrak{R}_{N_c}} \le \eta(N_2 - N_c) \le \eta + a_{\mathrm{sig}}^{-1} \log \frac{R_c}{\mathfrak{R}_{N_c}}. \tag{42}$$

No angle or diagnostic persistence after N is required or asserted.

Proof. The complete proof is organized into the lemmas of Appendices F–M. We give the join here, including the numerical constants that determine the spectral conclusion.

Original event and actual profile acquisition. Appendix H proves that the Gaussian category occupancy and simultaneous row-norm event fails with probability at most $3\delta/8$. On that event its compatible-force induction controls all original rows through N, with $\sum_j V_j \le S$ and $\|E_j\| \le \nu$. Four distinct original rows per teacher axis retain calibrated profiles throughout $[N_*, N]$. The notation is $(A, W, B) = s(q, w, b)$, $\rho = \|w\|$, $V = q^2 + \rho^2 + b^2$, $C = \mathbb{E}[y_c \sigma_\alpha(w^T X + b)]$, and $Q = qC/V$. The selected rows satisfy

$$\rho > 2R, \quad |q|/\rho > 19/20, \quad 1/16 \le b^2/\rho^2 \le 3/4, \quad \|(w/\rho, b/\rho) - (\epsilon u_i, \tau\sqrt{B_i^*})\| \le e_*. \tag{43}$$

Their sign is $\mathrm{sign}(q) = -\epsilon\tau$. Appendix I proves, using the optimized signal condition, that each selected $V_j \ge v_G$ at N. It simultaneously controls every unselected or adversely signed row by

$$\Xi_N := \sum_j |q_j| \left(\|P_{U^\perp} w_j\|^2 + \sum_i \min\{-\mathrm{sign}(q_j b_j) u_i^T w_j, 0\}^2 \right)^{1/2} \le c_G v_G. \tag{44}$$

At $q_j b_j = 0$ either orientation is allowed. This is an all-row estimate for the actual coupled simultaneous candidates, including all entry and exit steps.

Full current-head AGOP. Put $Z = U^T X$, $e_r = r^{-1/2}(1, \dots, 1)^T$ and $\xi_i = (e_r^T Z) Z_i - (e_r)_i$. Then $\mathbb{E}[\xi\xi^T] = I + e_r e_r^T \preceq 2I$. Appendix J gives the exact symmetric frame

$$D = \mathbb{E}[(U^T \nabla f_N)\xi^T] = \sum_j \omega_j (e_r^T x_j) x_j x_j^T, \qquad \omega_j = \frac{J s^2}{m} |q_j \rho_j z_j| \varphi(z_j), \tag{45}$$

where $x_j = -\mathrm{sign}(q_j b_j) U^T w_j/\rho_j$ and $z_j = b_j/\rho_j$. Rows with $\rho_j = 0$ contribute zero; when $q_j b_j = 0$, set $\omega_j = 0$ and choose either orientation. Its complete negative part has norm at most $(J\varphi(1)s^2/m)\Xi_N$. Choose one category per axis. By (43), $\|x_j - e_i\| \le e_*$, $|q_j|\rho_j \ge V_j/4$, and $|z_j|\varphi(z_j) \ge \varphi(1)/4$. Consequently $\omega_j \ge J\varphi(1)s^2 v_G/(16m)$ and

$$\lambda_{\min}(D) \ge \frac{J\varphi(1)s^2}{m} \left(\frac{v_G}{32\sqrt{r}} - \Xi_N \right) \ge \frac{J\varphi(1)s^2 v_G}{64 m\sqrt{r}}. \tag{46}$$

Here $(r^{-1/2} - e_*)(1 - \sqrt{r}e_*)^2 \ge 1/(2\sqrt{r})$ and $c_G \le 1/(64\sqrt{r})$. Every other positive frame summand is positive semidefinite and is retained. Bessel's inequality gives $U^T \mathsf{G}_N U \succeq DD^T/2$. The full outside derivative obeys

$$\mathrm{tr}(P_{U^\perp} \mathsf{G}_N P_{U^\perp}) \le \beta_G := (s^4/m^2)\Xi_N^2, \qquad \beta_G/a_G \le \epsilon_G/2. \tag{47}$$

Minmax proves (38). Since $P_N \preceq \mathsf{G}_N/a_G$, equal-rank principal-angle identities prove the endpoint bounds in (37), retaining all cross blocks.

Initial scores and the same diagnostic. Appendix G proves almost-sure simplicity and rotational invariance, so $\mathbb{E}[A_{\mathrm{mean},0}] = r/d$ and $\Pr(A_{\mathrm{mean},0} > 1/4) \le 4r/d \le \delta/4$. The central mixture has no chaos below degree three. For each initial normalized feature its target correlation is at most $\|U^T w/\|w\|\|^3 + \zeta$. The simultaneous Gaussian estimate and (28)–(34) make this at most $\chi_0/(16B_1)$ except on probability $\delta/4$. The unchanged ℓ^1 cap then gives $\mathcal{R}(0) \ge 1 - \chi_0/8$. Appendix K constructs the terminal witness from the four profiles, with both coefficient budgets paid and with the full infinite Hermite tail included in its risk. Its error debit is $C_1 e_* + \zeta$, with $C_1 < 12\sqrt{r}/J$, so $C_1 e_* < \chi_0/128$ and $\zeta \le \chi_0/128$. This proves (39).

Loss band and continuation. At N a selected actual row has $Q > K_*/2$ and $V > v_*$. Appendix L, applied with $(\gamma, U, v) = (K_*/2, S, v_*)$, pays the remaining adverse energy through N_c. The envelope is S_c and $\|f_n\|_2 \le s^2 S_c/(2m)$ throughout this prefix. The scale bounds give the band and baseline (40). The same-step release lemma in Appendix M then gives (42) and the subsequent loss decrease. The integer ceiling definitions give (41). Finally the union of the three original-law failure events is at most $3\delta/8 + \delta/4 + \delta/4 = 7\delta/8$. No independence between these events is used. □

E THE ADDITIVE CUBIC-NEIGHBORHOOD EXTENSION

All parameters, the full teacher y and the raw student f_n are those of AJ1–11 in Appendix B.2. Here $(A_j, W_j, B_j) = s(q_j, w_j, b_j)$, $\theta_j = (w_j, b_j)$, $\rho_j = \|w_j\|$, and $V_n = \sum_j (q_{j,n}^2 + \|\theta_{j,n}\|^2)$ is complete normalized joint energy. Write $P_U = UU^T$, $P_{U^\perp} = I_d - P_U$ and $\|g\|_2^2 = \mathbb{E}[g(X)^2]$ for $X \sim N(0, I_d)$. Define the complete spatial-energy statistic and full student AGOP by

$$A_{\mathrm{sub},n} = \frac{\sum_j \|UU^T W_{j,n}\|^2}{\sum_j \|W_{j,n}\|^2}, \qquad \mathsf{G}_n = \mathbb{E}[\nabla_x f_n \nabla_x f_n^T]. \tag{AJ12}$$

At $n = 0, N$ let P_n be its genuine leading rank-r projector, and set

$$A_{\min,n} = \lambda_{\min}(U^T P_n U), \qquad A_{\mathrm{mean,n}} = r^{-1} \operatorname{tr}(U^T P_n U). \tag{AJ13}$$

The theorem proves that these projectors are well defined; it does not complete a deficient eigenspace arbitrarily.

Theorem E.1 (Simultaneous actual AGOP acquisition during a high-loss window)**.** *Under AJ1–11, one event of the original initialization, of probability at least $1-\delta$, has all of the following properties.*

Throughout $0 \le n \le N$, the full original loss satisfies

$$1 - \delta_L/20 \le L_n \le 1 + \delta_L/20, \qquad L_n \ge 1 - G_{\mathrm{rel}}, \qquad \frac{\max_{n \le N} L_n}{\min_{n \le N} L_n} \le 1 + \delta_L/4. \tag{AJ14}$$

At initialization $A_{\mathrm{sub},0} \le 1/4$ and $A_{\min,0} \le A_{\mathrm{mean},0} \le 1/4$. At the common public endpoint,

$$A_{\mathrm{sub},N} > .9, \qquad A_{\min,N} \ge 1 - \varepsilon, \qquad A_{\mathrm{mean,N}} \ge 1 - \varepsilon/r. \tag{AJ15}$$

In particular the complete spatial-energy improvement exceeds .65, and the minimum principal-angle improvement is at least $3/4 - \varepsilon \ge 1/2$. With

$$a_* = 324(1-\alpha)^2 \varphi(1)^2 s^4 R^4 / m^2, \tag{AJ16}$$

the full endpoint spectrum obeys

$$\lambda_r(\mathsf{G}_N) \ge a_*, \quad \lambda_{r+1}(\mathsf{G}_N) < 1522 s^4 \le \varepsilon a_*, \quad \lambda_r - \lambda_{r+1} \ge (1-\varepsilon) a_* > 0. \tag{AJ17}$$

Fix any public feature multiplier $\kappa > 0$. At both times compare the identical population diagnostic class

$$\psi_{j,n}^P(X) = \kappa \sigma_\alpha((P_n W_{j,n})^T X + B_{j,n}), \qquad B_\kappa = C_\alpha/(7\kappa s),$$
$$\mathcal{R}_{B_\kappa}^P(n) = \inf_{\|v\|_2 \le B_\kappa} \left\| y - \sum_j v_j \psi_{j,n}^P \right\|_2^2. \tag{AJ18}$$

Projection precedes the activation, and no projected weight is renormalized. The same definition with no coefficient constraint is $\mathcal{R}_\infty^P$. Then

$$\mathcal{R}_{B_\kappa}^P(0),\ \mathcal{R}_\infty^P(0) \ge 1 - 9/256,$$
$$\mathcal{R}_{B_\kappa}^P(N),\ \mathcal{R}_\infty^P(N) \le (1/\sqrt{3} + 11/128)^2,$$
$$\mathcal{R}_{B_\kappa}^P(0) - \mathcal{R}_{B_\kappa}^P(N), \quad \mathcal{R}_\infty^P(0) - \mathcal{R}_\infty^P(N) \ \ge\ 1 - 9/256 - (1/\sqrt{3} + 11/128)^2 > 1/2. \tag{AJ19}$$

These are squared population prediction-risk comparisons, not classification-accuracy or finite-sample guarantees. Refit is a diagnostic and is not an update of the trained heads. Its budget scales as $1/s$ and is identical at the two checkpoints.

The actual optimization clock satisfies

$$mT/2 \le \eta N \le m\overline{T}/2. \tag{AJ20}$$

There is an acquired dictionary time $J \le N$ with $\eta J \ge (m/8)\log(1024/5)$; complete $A_{\mathrm{sub}} > .9$ persists on $J \le n \le N$. The minimum-AGOP claim is at the public endpoint N; no earlier hitting-time or whole-suffix minimum-angle assertion is made. The unchanged GD run later reaches $L \le 1 - 2G_{\mathrm{rel}}$, a decrease at least G_{rel} from every loss in the plateau. This later original-loss bound may be very small and is distinct from the constant diagnostic gain in AJ19.

Proof. Use the category and all-row Gaussian event of Appendix P with confidence parameter δ_0. Its coupon failure is at most $\delta_0/4$ and its simultaneous row-tail failure at most $7\delta_0/96$. The initial projected-risk bound PJ21 with $\delta_I = \delta_0/4$ costs at most another $\delta_0/4$. AJ5 gives $q_{r,d} \le ((r+4)/d)^3 \le \delta_0/128$. Thus their joint failure is at most $55\delta_0/96$. Its Gaussian coupon and all-row estimates, actual coupled force induction, profile maturation and complete spatial-energy bounds use no $m < d$ assumption. Its sole replaced step is the initial raw-span comparator. It supplies the initial projected-risk event instead. The actual force induction in AC17–27 uses the chosen R, T, M and the bounds in AJ8–11. Therefore AJ7–11 give the same actual conclusions. The failure of this event is at most $55\delta_0/96$. In particular all $4r$ original category representatives mature at N, and the complete outside bound CW19 holds with trace and operator norm below $1522s^4$.

Intersect with the original weighted-coordinate gap event FA6–8. Its additional failure is below $\zeta/2$, with no conditioning or independence substitution. FA4 is exactly the last cutoff in AJ9 and the last step cutoff in AJ8. Thus the full forced-anchor proof gives FA24–28 for every original row through N.

Write the full signed second-Hermite gradient frame at N as

$$\begin{aligned} \mathsf{D}_{ki} &= \mathbb{E}[(u_k^T \nabla f_N) h_2(u_i^T X)] = \sum_j t_j v_{jk} v_{ji}^2, \\ v_j &= U^T w_j/\rho_j, \quad t_j = -\frac{(1-\alpha)s^2}{\sqrt{2}m} q_j \rho_j z_j \varphi(z_j). \end{aligned} \tag{AJ21}$$

To verify this identity directly, put $Z = w_j^T X/\rho_j$, $z_j = b_j/\rho_j$, and $h_2(t) = (t^2-1)/\sqrt{2}$. Gaussian regression gives

$$\begin{aligned} \mathbb{E}[h_2(u_i^T X) \mid Z] &= v_{ji}^2 h_2(Z), \\ \mathbb{E}[\sigma_\alpha'(\rho_j Z + b_j) h_2(Z)] &= \frac{1-\alpha}{\sqrt{2}} \int_{-z_j}^{\infty} (t^2-1)\varphi(t)\, dt = -\frac{1-\alpha}{\sqrt{2}} z_j \varphi(z_j). \end{aligned}$$

The first equality follows by writing $u_i^T X = v_{ji} Z + G_\perp$ with independent centered Gaussian $G_\perp$ of variance $1 - v_{ji}^2$; the last follows from $(t\varphi(t))' = (1-t^2)\varphi(t)$. Multiplying by the actual gradient row $(s^2/m) q_j w_j$ and summing proves AJ21, equivalently the full frame (FA25). All current trained heads and all self/cross effects remain present. FA28 decomposes this matrix as a nonnegative diagonal matrix plus an error whose operator norm is at most $(1-\alpha)\varphi(1)s^2\Gamma$. A favorable row's diagonal axis is its original oriented weighted extremum. For each category row this is its assigned axis: its original selected coordinate has weighted magnitude at least $a_i/\sqrt{d} \ge a_0/\sqrt{d}$, whereas each competitor has magnitude at most $\vartheta/(4\sqrt{r-1}\sqrt{d}) < a_0/\sqrt{d}$. The rank-one case has no competitors. Coordinate order is retained by FA12, and the category's favorable phase and orientation persist by AC. Consequently none of the chosen diagonal mass is lost in the complete FA decomposition.

At N, each of the four chosen rows per axis satisfies $|q_j| > 6R$, $|q_j| \le \sqrt{2}\rho_j$, and $1/2 < |z_j| < 1$. These are the acquired AC conclusions used in the calibrated category profiles. Each contributes at least $9(1-\alpha)\varphi(1)s^2R^2/m$ to its assigned diagonal. Therefore

$$\begin{aligned} \lambda_{\min}(\mathsf{D}_+) &\ge 36(1-\alpha)\varphi(1)s^2R^2/m, \\ \sigma_{\min}(\mathsf{D}) &\ge (1-\alpha)\varphi(1)s^2(36R^2/m - \Gamma) \ge 18(1-\alpha)\varphi(1)s^2R^2/m. \end{aligned} \tag{AJ22}$$

The last inequality is the third radius condition in AJ7. Extra favorable rows only add nonnegative diagonal mass; their errors and every nonfavorable row have already been charged by FA28. Thus this is a bound for the complete student, not its chosen part.

Bessel's inequality for the orthonormal tests $h_2(u_i^T X)$ gives $U^T \mathsf{G}_N U \succeq \mathsf{D}\mathsf{D}^T$, so its least eigenvalue is at least a_*. The last radius condition in AJ7 gives $1522s^4 \le \varepsilon a_*$. Minmax bounds the first omitted eigenvalue by the outside block and the rth by the inside block, proving AJ17. Spectral decomposition also gives $P_N \preceq \mathsf{G}_N/a_*$. Taking its outside norm or trace and using equal-rank principal-angle identities gives

$$1 - A_{\min,N} = \|P_{U^\perp} P_N P_{U^\perp}\| \le \varepsilon, \qquad 1 - A_{\mathrm{mean,N}} = r^{-1} \operatorname{tr}(P_{U^\perp} P_N) \le \varepsilon/r. \tag{AJ23}$$

No cross block was discarded in these positive-semidefinite inequalities.

At initialization IA gives an almost surely unique positive top-r projector with Haar orientation, hence $\mathbb{E}[A_{\mathrm{mean},0}] = r/d$. Markov and AJ5 give

$$\mathbb{P}(A_{\mathrm{mean},0} > 1/4) \le 4r/d \le \delta/4. \tag{AJ24}$$

The total failure of the three original-law events is at most $55\delta_0/96 + \zeta/2 + \delta/4 < \delta$. Their dependence does not affect the union bound. AC supplies the initial and final complete A_{sub} claims, the full loss band and the clocks.

For prediction, AJ15 and the choice of ε imply PJ22 with $\mu = 1 - \varepsilon$. PJ9 constructs a legal readout in exactly AJ18 using coefficients proportional to $(s\rho_j)^{-1}$; these cancel the learned radii when spatial projection precedes activation. PJ14–17 bound the projection distortion by $D_g\sqrt{\varepsilon(1 + (1-\varepsilon)^{-1})} \le 1/16$. The normalized-profile and full-teacher costs are $1/64$ and $1/128$. PJ21 and AJ5 supply the initial projected oracle lower bound on the same event. This proves AJ19, with no changed comparator or actual head reset. Finally, the complete endpoint energy exceeds twice its initial value. NC20–23 give $\mathbb{E}[yf_N]/(s^2V_N) \ge 1/(16m\overline{T}) = 4a_{\mathrm{rel}}$. The target has H^1 norm below 3, all rows remain balanced, and AJ11 implies $s\sqrt{V_N} \le R_{\mathrm{rel}}/2$. The same-step lemma in Appendix M, with network factor $1/m$ and half-MSE rate mh, yields a later checkpoint with $L \le 1 - 2G_{\mathrm{rel}}$. Throughout the prefix, $\|f_n\|_2 \le s^2M^2/(2m) \le \delta_L/100$ and $\|f_n\|_2 \le G_{\mathrm{rel}}/2$; expanding $\|y - f_n\|_2^2$ proves the claimed full loss band, ratio and baseline. Thus the later decrease is at least G_{rel} from every prefix checkpoint.

□

F Gaussian kernels and compatible simultaneous updates

F.1 Conventions and the normalized additive teacher

Let $X \sim \gamma_d = N(0, I_d)$, $0 \le \alpha < 1$, $\sigma(t) = \alpha t + (1-\alpha)t_+$, and $\widetilde{X} = (X, 1)$. Here $\|g\|_2$ is its Gaussian L^2 norm; φ and Φ are the standard normal density and CDF. Derivatives of the teacher are with respect to X, while ∇C_y and D^2C_y differentiate the augmented parameter θ. Write $\theta = (w, b)$ and

$$C_y(\theta) = \mathbb{E}[y(X)\sigma(w^TX + b)], \quad A = \|y\|_2, \quad S = \mathbb{E}[\|\nabla y\|^2], \quad H_y = (A^2 + S)^{1/2}. \tag{HK1}$$

The weak Gaussian Sobolev space here is the closure of polynomials in the norm H_y (equivalently, the usual weak-derivative space). All norms of parameter derivatives below are Euclidean operator norms.

For orthonormal $u_1, \dots, u_r$ and real c_i, let $y = \sum_i c_i g_i(u_i^TX)$ with $g_i \in H^1(\gamma_1)$. Use $h_k = \mathrm{He}_k/\sqrt{k!}$ and $g_i = \sum_{k\ge0}\gamma_{ik}h_k$. Exactly,

$$\begin{aligned} \mu &= \mathbb{E}[y] = \sum_i c_i\gamma_{i0}, & \ell &= \mathbb{E}[yX] = \sum_i c_i\gamma_{i1}u_i, \\ A^2 &= \mu^2 + \sum_i c_i^2 \sum_{k\ge1}\gamma_{ik}^2, & S &= \sum_i c_i^2 \sum_{k\ge1} k\gamma_{ik}^2. \end{aligned} \tag{HK2}$$

Full population normalization means $A = 1$ in these formulas. In particular, a common link with $c_i = r^{-1/2}$ has unnormalized energy $r\gamma_0^2 + \sum_{k\ge1}\gamma_k^2$, not $\|g\|_2^2$.

An optional sharper directional size is

$$\Lambda_y^2 = \|\mathbb{E}[\nabla y \nabla y^T]\|_{\mathrm{op}} \le S. \tag{HK3}$$

For the additive teacher its matrix on $U = \mathrm{span}(u_i)$ is $\mathrm{diag}\{c_i^2\,\mathrm{Var}(g_i')\} + \ell\ell^T$; thus $\Lambda_y^2 \le \max_i c_i^2\,\mathrm{Var}(g_i') + \|\ell\|^2$. These expressions contain no extraneous factor of r or d.

F.2 A Gaussian Hilbert-space trace bound

Lemma F.1 (One-dimensional weighted trace)**.** *Let $f \in H^1(\gamma_1; \mathcal{H})$ for a real Hilbert space $\mathcal{H}$, with $a = \|f\|_{L^2(\gamma;\mathcal{H})}$ and $b = \|f'\|_{L^2(\gamma;\mathcal{H})}$. Its continuous local representative satisfies, at every real t,*

$$\varphi(t)\|f(t)\|_{\mathcal{H}}^2 \le a\sqrt{b^2 + a^2/2} \le (a^2 + b^2)/\sqrt{2}. \tag{HK4}$$

Consequently, for a unit $v \in \mathbb{R}^d$, the canonical L^2 trace of any $y \in H^1(\gamma_d)$ obeys

$$\|y|_{v^T X=t}\|_{L^2(\gamma_{v^\perp})} \le \varphi(t)^{-1/2} A^{1/2} (\|\partial_v y\|_2^2 + A^2/2)^{1/4} \le 2^{-1/4} \varphi(t)^{-1/2} H_y. \tag{HK5}$$

Proof. First take smooth f with sufficient decay and set $h(t) = \sqrt{\varphi(t)} f(t)$. Gaussian integration by parts gives

$$\|h\|_{L^2(dt)} = a, \qquad \|h'\|_{L^2(dt)}^2 = b^2 + a^2/2 - \|tf\|_{L^2(\gamma)}^2/4 \le b^2 + a^2/2.$$

There is no unproved multiplication hypothesis: integration by parts also gives $\|tf\|_2^2 = a^2 + 2\langle tf, f'\rangle$, hence $\|tf\|_2 \le b + \sqrt{b^2 + a^2}$. Density extends this multiplier estimate and the displayed identity to $H^1(\gamma; \mathcal{H})$. For an $H^1(\mathbb{R}; \mathcal{H})$ function, integrate the derivative of $\|h\|^2$ on the two sides of t and average the results. This gives $\|h(t)\|^2 \le \int \|h(s)\| \|h'(s)\| ds \le \|h\|_2 \|h'\|_2$. The final inequality in (HK4) follows from $a^2(b^2 + a^2/2) \le (a^2 + b^2)^2/2$. Rotate Gaussian space so its first coordinate is $v^T X$ and take $\mathcal{H} = L^2(\gamma_{v^\perp})$ to obtain (HK5). □

F.3 Uniform full-parameter curvature, including pure-bias rows

Lemma F.2 (Dimension-free H^1 kink regularity)**.** *For every $y \in H^1(\gamma_d)$, C_y is positively homogeneous of degree one and belongs to $C^2(\mathbb{R}^{d+1} \setminus \{0\})$. For all $\theta \ne 0$,*

$$|C_y(\theta)| \le A\|\theta\|, \qquad \|\nabla C_y(\theta)\| \le A, \qquad \theta^T \nabla C_y(\theta) = C_y(\theta), \qquad D^2 C_y(\theta)\theta = 0, \tag{HK6}$$

$$\|\theta\| \, \|D^2 C_y(\theta)\| \le B_y := 4(1-\alpha) H_y. \tag{HK7}$$

One may instead use the sharper constant

$$B_y^{\mathrm{sharp}} = K_0 (1-\alpha) A^{1/2} (\Lambda_y^2 + A^2/2)^{1/4}, \qquad K_0 = \sqrt{3}(2\pi)^{-1/4} 6^{3/2} e^{-5/4} < 4.61. \tag{HK8}$$

When $\rho = \|w\| > 0$, $v = w/\rho$, and $z = b/\rho$, the exact Hessian is

$$D^2 C_y(\theta) = (1-\alpha) \frac{\varphi(z)}{\rho} \mathbb{E}[y(X) \widetilde{X} \widetilde{X}^T \mid v^T X = -z], \tag{HK9}$$

where the conditional expectation is the canonical trace in (HK5). At $w = 0, b \ne 0$ its values are exactly

$$C_y(0,b) = \mu \sigma(b), \qquad \nabla C_y(0,b) = \sigma'(b)(\ell, \mu), \qquad D^2 C_y(0,b) = 0. \tag{HK10}$$

In particular, no uniform lower bound on spatial radii, no bound on $|b|/\|w\|$, and no exclusion of aligned teacher directions is needed.

Proof. For polynomial y, differentiate the Gaussian half-space integral to obtain (HK9). For any unit augmented parameter vector a, conditional on $v^T X = -z$, the scalar $a^T \widetilde{X}$ is Gaussian with mean m and variance s^2 satisfying $m^2 + s^2 \le 1 + z^2$. Consequently

$$\mathbb{E}[(a^T \widetilde{X})^4 \mid v^T X = -z] = m^4 + 6m^2 s^2 + 3s^4 \le 3(1+z^2)^2.$$

Testing the symmetric matrix (HK9) on a, using (HK5), and noting $\|\theta\|/\rho = \sqrt{1+z^2}$, gives

$$\|\theta\| \, \|D^2 C_y(\theta)\| \le (1-\alpha)\sqrt{3}(1+z^2)^{3/2} \sqrt{\varphi(z)} A^{1/2} (\|\partial_v y\|_2^2 + A^2/2)^{1/4}. \tag{HK11}$$

The maximum of $(1+z^2)^{3/2} e^{-z^2/4}$ occurs at $z^2 = 5$. This proves (HK8); (HK4) gives the simpler constant $2^{-1/4} K_0 < 3.88 < 4$ in (HK7).

For polynomial y, the same formula and its derivatives have a polynomial factor in z times $\varphi(z)/\rho$. Thus they extend at $w = 0, b \ne 0$ with zero Hessian and the affine values in (HK10). Now approximate arbitrary y by its multivariate Hermite polynomials $y_N \to y$ in H^1. Cauchy–Schwarz and $|\sigma(t)| \le |t|$, $|\sigma'| \le 1$ give, for any such difference e,

$$|C_e(\theta)| \le \|e\|_2 \|\theta\|, \qquad \|\nabla C_e(\theta)\| \le \|e\|_2. \tag{HK12}$$

Indeed $\mathbb{E}[(a^T \widetilde{X})^2] = 1$ for unit a. Together with (HK7), applied to polynomial differences, these estimates make C_{y_N}, their gradients, and their Hessians uniformly Cauchy on each compact subset of $\theta \ne 0$. Their limit is C_y and is C^2 there, with all the asserted bounds and values. This also proves continuity through changes of hyperplane direction and through the pure-bias rows; pointwise weak derivatives of the original link never need to be evaluated. At each fixed hyperplane, (HK5) makes the traces converge in L^2, proving (HK9) for y. Homogeneity gives the two Euler identities in (HK6). □

The linear dependence on a general unnormalized H_y is natural by scaling. With $A = 1$, (HK8) improves the large-derivative dependence to $O(\Lambda_y^{1/2})$. This power cannot be decreased uniformly: take a nonnegative smooth bump $g_\epsilon(t)$ supported on $[-\epsilon, \epsilon]$, of value $\asymp \epsilon^{-1/2}$ at zero, and normalize its Gaussian L^2 norm to one. Then $\|g_\epsilon'\|_2 \asymp \epsilon^{-1}$, whereas at $\theta = (u, 0)$ the bias-bias Hessian is $(1-\alpha)\varphi(0)g_\epsilon(0) \asymp \epsilon^{-1/2}$. In particular no bound depending only on $\|y\|_2 = 1$ is possible.

F.4 Full uncentered Hermite kernel and meaning of differentiation

For $Z \sim \gamma_1$ define $\eta_k(z) = \mathbb{E}[h_k(Z)\sigma(Z+z)]$. Gaussian integration by parts, in the distributional sense for the kink, gives

$$\begin{aligned}
\eta_0(z) &= \alpha z + (1-\alpha)\{\varphi(z) + z\Phi(z)\},\\
\eta_1(z) &= \alpha + (1-\alpha)\Phi(z),\\
\eta_k(z) &= (1-\alpha)\varphi(z)\frac{h_{k-2}(-z)}{\sqrt{k(k-1)}} \quad (k \geq 2),\\
\eta_0' &= \eta_1, \qquad \eta_k' = (1-\alpha)\varphi(z)h_{k-1}(-z)/\sqrt{k} \quad (k \geq 1),\\
\eta_k'' &= (1-\alpha)\varphi(z)h_k(-z) \quad (k \geq 0).
\end{aligned} \tag{HK13}$$

For $a_i = u_i^T v$ the exact full correlation is

$$C_y(w, b) = \rho F(v, z), \quad F(v, z) = \mu\eta_0(z) + \eta_1(z)\ell^T v + \sum_i \sum_{k\geq 2} c_i \gamma_{ik} a_i^k \eta_k(z). \tag{HK14}$$

The series is absolutely convergent at every (v, z), including $a_i = \pm 1$, by Cauchy–Schwarz and Parseval for $\sigma(Z + z)$. It is the original correlation, not a truncated or centered replacement. Finite Hermite truncations of (HK14), composed with the actual θ coordinates, converge together with all parameter derivatives through order two uniformly on compact subsets of $\theta \neq 0$, by Theorem F.2. This is a rigorous termwise-differentiation meaning even at aligned directions.

For first derivatives one may also use the ambient expression

$$J(v, z) = \mathbb{E}[\nabla y(X)\sigma'(v^T X + z)] = \sum_i c_i u_i \sum_{k\geq 1} k\gamma_{ik} a_i^{k-1} \eta_k(z), \qquad D(v, z) = v^T J(v, z). \tag{HK15}$$

Here $\|J\| \leq \sqrt{S}$ and $|D| \leq \|\partial_v y\|_2$. The ith scalar sum is the covariance derivative of $\mathbb{E}[g_i(T)\sigma(Z + z)]$, where $\mathrm{Corr}(T, Z) = a_i$. In the open interval this derivative is $\mathbb{E}[g_i'(T)\sigma'(Z + z)]$ by Gaussian integration by parts; its magnitude is at most $\|g_i'\|_2$. Approximation in H^1 extends this formula and uniform convergence to $a_i = \pm 1$.

Weak Gaussian integration by parts also gives the exact formulas

$$H(v, z) = (1-\alpha)\varphi(z)\mathbb{E}[y \mid v^T X = -z], \quad \nabla_w C_y = J + Hv, \quad \partial_b C_y = F_z = \mathbb{E}[y\sigma'(v^T X + z)]. \tag{HK16}$$

They can alternatively be proved first for polynomials and then by (HK5), (HK12), and H^1 convergence. In spherical coordinates,

$$\nabla_w C_y = v(F - zF_z) + (I - vv^T)J, \qquad F - zF_z = D + H. \tag{HK17}$$

No claim is made that the artificial off-sphere partial derivatives $\partial_{a_i}^2 F$ or $\partial_{a_i}\partial_z F$ exist at $|a_i| = 1$ under H^1 alone. Their apparent singularities can cancel in the actual θ derivatives. For example, for a compactly supported link agreeing near zero with $|t|^\beta$, $1/2 < \beta < 1$, the link is H^1 but g' has no finite trace at zero; the aligned mixed covariance/shift derivative asks for that nonexistent trace. The uniform actual-parameter conclusion above does not make that extra assertion.

For completeness an individual additive summand has an explicit trace without second derivatives of g_i. Put $t = -z$, $a = u_i^T v$, $s = \sqrt{1 - a^2}$, $e = (u_i - av)/s$ when $s > 0$, and $\xi = (tv, 1)$, $\bar{e} = (e, 0)$, $\bar{P} = \mathrm{diag}(I - vv^T, 0)$. With an independent standard normal G, define

$$M_0 = \mathbb{E}[g_i(at + sG)], \quad M_1 = \mathbb{E}[Gg_i(at + sG)], \quad M_2 = \mathbb{E}[(G^2 - 1)g_i(at + sG)].$$

Then its conditional matrix in (HK9) equals

$$M_0(\bar{P} + \xi\xi^T) + M_1(\bar{e}\xi^T + \xi\bar{e}^T) + M_2\bar{e}\bar{e}^T. \tag{HK18}$$

For $s > 0$, weak integration by parts also gives $M_1 = s\mathbb{E}[g_i'(at+sG)]$ and $M_2 = s\mathbb{E}[Gg_i'(at+sG)]$. At $s = 0$ the matrix is $g_i(at)(\bar{P} + \xi\xi^T)$ using the continuous one-dimensional H^1 representative. Summing (HK18) with coefficients c_i gives the full additive trace, including means and all chaos orders.

F.5 Interfaces for coupled Euclidean population GD

Use actual raw coordinates $x = (q_j, \theta_j)_{j=1}^m$ and the fixed normalization

$$f_x = \kappa \sum_{j=1}^{m} q_j \sigma(\theta_j^T \widetilde{X}), \quad \ell(x) = \tfrac{1}{2}\mathbb{E}[(y - f_x)^2], \quad x^+ = x - \eta\nabla\ell(x). \tag{HK19}$$

The full MSE is $L = 2\ell$ and its corresponding raw rate is $\eta/2$. No coordinates or row interactions are removed. Let $R = \|x\|$, $F_x = \mathbb{E}[yf_x]$, $V_x = \|f_x\|_2^2$, $Q = F_x/R^2$. For $A = 1$, exactly as follows from (HK6) and Euler's identity,

$$|F_x| \le \kappa R^2/2, \quad \|\nabla F_x\| \le \kappa R, \quad \|f_x\|_2 \le \kappa R^2/2, \quad V_x \le \kappa^2 R^4/4, \quad \|\nabla V_x\| \le \kappa^2 R^3, \quad \|\nabla Q\| \le \kappa/R. \tag{HK20}$$

At a point with every nonzero hidden row satisfying $|q_j| \le \beta\|\theta_j\|$, the row-block Hessian and quotient rule give

$$\|D^2 F_x\| \le \kappa(1 + \beta B_y), \qquad \|D^2 Q\| \le \kappa(8 + \beta B_y)/R^2. \tag{HK21}$$

Indeed each Hessian block of F_x is $\kappa \begin{pmatrix} 0 & \nabla C_y^T \\ \nabla C_y & q_j D^2 C_y \end{pmatrix}$; its norm is at most $\kappa(1 + \beta B_y)$. Differentiating R^{-2} contributes at most $4\kappa/R^2 + 3\kappa/R^2$ to the quotient Hessian. In particular, the $\beta = 2$ interpolation segments have quotient constant $8 + 2B_y$. These are the curvature estimates used in Lemma M.2; its entry conditions and complete same-step continuation proof appear in Appendix M.

The whole student Gram also has sufficient regularity. If $G(\theta, \psi) = \mathbb{E}[\sigma(\theta^T\widetilde{X})\sigma(\psi^T\widetilde{X})]$, then away from zero rows

$$\|D_\theta^2 G\| \le 4\sqrt{2}(1 - \alpha)\|\psi\|/\|\theta\|, \qquad \|D_\theta D_\psi G\| \le 1. \tag{HK22}$$

The first bound uses (HK7) with the Sobolev teacher $\sigma(\psi^T\widetilde{X})$, whose H^1 norm is at most $\sqrt{2}\|\psi\|$. The second uses the exact mixed matrix $\mathbb{E}[\sigma'(\theta^T\widetilde{X})\sigma'(\psi^T\widetilde{X})\widetilde{X}\widetilde{X}^T]$ and Cauchy–Schwarz on unit directions. These derivatives are jointly continuous, including coincident nonzero hyperplanes: for the pure second derivative use (HK7) and continuity of the feature in H^1; for the mixed derivative use almost-everywhere convergence and Gaussian domination.

For example, on the same balanced region,

$$\|D^2 V_x\| \le \{3 + 4\sqrt{2}\beta(1 - \alpha)\}\kappa^2 R^2. \tag{HK23}$$

To verify this, write the second variation as $2\|Df_x[h]\|_2^2 + 2\langle f_x, D^2 f_x[h, h]\rangle$ using the Gram derivatives. The first term is at most $2\kappa^2 R^2\|h\|^2$. The head/hidden cross term is at most $2\kappa\|f_x\|_2\|h\|^2 \le \kappa^2 R^2\|h\|^2$. The remaining hidden Hessians are at most $2\kappa\beta\{4(1-\alpha)\|f_x\|_{H^1}\}\|h\|^2$; here $\|f_x\|_{H^1} \le \kappa R^2/\sqrt{2}$. This proves (HK23) without declaring the feature itself twice differentiable as an L^2-valued map.

Rows with $(q_j, \theta_j) = (0, 0)$ stay zero under the conventional actual update and may be omitted. No C^2 claim at $\theta = 0$ with nonzero head is made. The usual discrete row balance $\|\theta_j\|^2 - q_j^2 \ge 0$, when transported by the actual algorithm, excludes that singular case; Gaussian initialization gives nonzero hidden rows almost surely. Neither regularity nor a positive reached Q implies acquisition of all teacher directions.

G ORIGINAL GAUSSIAN INITIALIZATION AND PREDICTION BASELINES

G.1 POPULATION MODEL AND GENUINE INITIAL EIGENDIRECTIONS

Function norms and gate expectations use the displayed Gaussian input; expectations of initial alignment statistics use the parameter draw. Write $P_U = UU^T$ for the teacher-space projector when U is an orthonormal axis matrix, and $h_k = \mathrm{He}_k/\sqrt{k!}$ for normalized probabilists' Hermite polynomials. Let $d \geq 2$, $m \geq 1$, $0 \leq \alpha < 1$, $J = 1 - \alpha$, and

$$f(X) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma_\alpha(w_j^T X + b_j), \qquad \sigma_\alpha(t) = \alpha t + J t_+, \qquad X \sim N(0, I_d). \tag{IA1}$$

Use spatial rows $W \in \mathbb{R}^{m\times d}$ and define

$$\begin{gathered} d_j(X) = \sigma'_\alpha(w_j^T X + b_j), \quad K_{jl} = \mathbb{E}[d_j d_l], \quad D_a = \mathrm{diag}(a_j), \\ G_s = \mathbb{E}[\nabla_X f \nabla_X f^T] = \frac{1}{m^2} W^T D_a K D_a W. \end{gathered} \tag{IA2}$$

The initialization has IID spatial rows $w_j \sim N(0, s_w^2 I_d)$ with one common $s_w > 0$, independent Gaussian biases, and independent nondegenerate Gaussian heads. The positive spatial, bias and head block scales may differ; the spatial scale is common to all rows. All heads, biases and rows are those of the original draw; no teacher-based screening or initialization rejection is used.

Lemma G.1 (Initial population spectral rank and simplicity)**.** *With probability one, $G_s(0)$ has rank $p = \min(m, d)$ and its p positive eigenvalues are pairwise distinct. Consequently, for any $1 \leq q \leq p$ with $q < d$, its leading rank-q projector $\widehat{P}_{q,0}$ is uniquely defined and has a positive boundary gap. This is qualitative almost-sure nondegeneracy, not a lower bound on the size of that gap.*

Proof. Almost surely W has rank p, each head is nonzero, each spatial row is nonzero, and its affine kink hyperplanes are pairwise distinct. For such a hidden bank the gate Gram K is positive definite. Indeed, if $v^T K v = 0$, the piecewise constant function $\sum_j v_j d_j(X)$ is zero almost everywhere. For each hyperplane choose a point not on any other hyperplane. On a sufficiently small ball about that point, the two open sides differ in just the corresponding gate. Both have positive Gaussian measure, so their constants must both be zero. Their difference is Jv_j, hence $v_j = 0$ for every j. Equation IA2 then gives rank $G_s = p$.

Condition on this entire hidden/bias bank, and choose an orthonormal basis of its p-dimensional spatial span. The restriction of G_s to that span is a symmetric $p \times p$ matrix polynomial in the head coordinates. The discriminant of its characteristic polynomial is therefore a polynomial in those coordinates. It is not identically zero, as the following explicit existence argument shows.

Choose p linearly independent spatial rows, indexed by I, and temporarily set the other heads to zero only to evaluate that polynomial. In the chosen span basis let S be the invertible $p \times p$ selected spatial matrix. The restricted AGOP is $m^{-2} S^T D K_{II} D S$. There are fixed $0 < c_- \leq c_+ < \infty$ such that its ordered eigenvalues satisfy

$$c_- a_{(i)}^2 \leq \lambda_i(m^{-2} S^T D K_{II} D S) \leq c_+ a_{(i)}^2 \quad (1 \leq i \leq p), \tag{IA3}$$

where the selected head magnitudes are sorted decreasingly. For example, use

$$\begin{aligned} c_- &= m^{-2}\lambda_{\min}(K_{II}) s_{\min}(S)^2, \\ c_+ &= m^{-2}\lambda_{\max}(K_{II}) s_{\max}(S)^2. \end{aligned}$$

These inequalities follow from matrix order and the singular-value bounds for multiplication by the fixed invertible S. Choose successive squared head magnitudes with ratio greater than $2c_+/c_-$. The intervals in IA3 are disjoint, so all eigenvalues are positive and distinct at that polynomial evaluation.

A nonzero real polynomial vanishes on a Lebesgue-null set. The conditional Gaussian head law has a density, so repeated positive eigenvalues have probability zero. The temporary zero-head choice was a polynomial witness, not an initialization used by the theorem. Integrating the conditional statement proves the result. For $q < p$ the boundary eigenvalues are distinct and positive; for $q = p < d$ the boundary is the positive last eigenvalue versus zero. □

G.2 Haar law, mean/min statistics and initial headroom

For an orthogonal R, the transformation $W \mapsto WR^T$ leaves the original spatial Gaussian law unchanged. The gate Gram is unchanged, and IA2 transforms as $G_s \mapsto RG_sR^T$. Thus the initial AGOP law is rotationally invariant. Its ordered eigenvalues are unchanged by this transformation, while its unique leading projector transforms equivariantly.

It follows that $\widehat{P}_{q,0}$ is a Haar rank-q projector, including conditionally on its eigenvalues. To justify uniqueness of this invariant law directly, average any bounded function of a rank-q projector over an independent uniform orthogonal rotation. That average is the same for every starting rank-q projector. Rotational invariance makes the original expectation equal to this average. The identical argument with an arbitrary bounded function of the eigenvalues proves the conditional statement.

Fix the teacher's orthonormal basis $U \in \mathbb{R}^{d\times r}$, with $1 \le r \le m$ and $r < d$, and use $q = r$ for the direction metric. Set

$$B_0 = U^T \widehat{P}_{r,0} U, \qquad A_{\min,0} = \lambda_{\min}(B_0), \qquad A_{\mathrm{mean},0} = \mathrm{tr}(B_0)/r. \tag{IA4}$$

Then the exact initial mean and variance are

$$\mathbb{E}[A_{\mathrm{mean},0}] = r/d, \qquad \mathrm{Var}(A_{\mathrm{mean},0}) = \frac{2(d-r)^2}{d^2(d-1)(d+2)}, \qquad 0 \le A_{\min,0} \le A_{\mathrm{mean},0}. \tag{IA5}$$

For completeness, in coordinates adapted to U, each diagonal entry of the Haar projector has beta law with parameters $r/2, (d-r)/2$, hence variance $2r(d-r)/[d^2(d+2)]$. Since its diagonal sum is exactly r, exchangeability makes the covariance of two different diagonal entries equal to minus that variance divided by $d-1$. Summing over the first r coordinates and dividing by r^2 proves IA5.

For any $0 < \delta_0 < 1$ and $a > r/d$, Chebyshev therefore gives

$$\Pr\{A_{\mathrm{mean},0} > a\} \le \frac{2(d-r)^2}{d^2(d-1)(d+2)(a-r/d)^2}. \tag{IA6}$$

In particular, the explicit size regime

$$m \ge r, \qquad d \ge \max\{8r, \sqrt{128/\delta_0}\} \tag{IA7}$$

ensures both $A_{\mathrm{mean},0} \le 1/4$ and $A_{\min,0} \le 1/4$ with probability at least $1-\delta_0$. This gives initial geometric headroom for a half-unit improvement; it does not prove that such an improvement is attained. No upper restriction on width occurs in IA5–7.

The complete weight-energy statistic remains distinct:

$$A_{\mathrm{sub}}(W_0) = \frac{\|W_0 P_U\|_F^2}{\|W_0\|_F^2} \sim \mathrm{Beta}(mr/2, m(d-r)/2). \tag{IA8}$$

This follows by adding the independent squared spatial Gaussian coordinates inside and outside U. Its mean is also r/d, but its distribution and later dynamics are not those of IA4. Any desired simultaneous initial score event must be intersected explicitly; no independence between IA4 and IA8 is asserted.

G.3 A one-feature Hermite estimate

Let $X \sim N(0, I_d)$ and $U \in \mathbb{R}^{d\times r}$ have orthonormal columns, $1 \le r < d$. Suppose $F = F(U^TX)$, $\|F\|_2 = 1$, and F is orthogonal to every Gaussian polynomial of total degree below an integer $k_0 \ge 1$. Let $y = F + e$, $\|y\|_2 = 1$, $\|e\|_2 \le \zeta$. Only L^2 regularity is needed here. Put $\sigma_\alpha(z) = \alpha z + (1-\alpha)z_+$, $0 \le \alpha < 1$.

Lemma G.2. *For $w \ne 0$, $n = w/\|w\|$, $t = \|U^Tn\|$, and any finite b, define*

$$\phi_{w,b} = \frac{\sigma_\alpha(w^TX + b)}{\sqrt{\|w\|^2 + b^2}}.$$

Then

$$\|\phi_{w,b}\|_2 \le 1, \qquad |\mathbb{E}[y\phi_{w,b}]| \le t^{k_0} + \zeta. \tag{BW1}$$

For a pure-bias feature the same bound holds with $t = 0$; a zero augmented vector is assigned zero feature.

Proof. The norm bound follows from $|\sigma_\alpha(z)| \le |z|$. For $0 < t \le 1$, let $v = UU^T n/t$, $Z = v^T X$, and $H(Z) = \mathbb{E}[F \mid Z]$. The variable $n^T X = tZ + \sqrt{1-t^2}G$ uses a Gaussian G independent of the complete teacher input $U^T X$, with the evident convention at $t = 1$. Consequently $\mathbb{E}[F \mid n^T X] = \mathbb{E}[H(Z) \mid n^T X]$. Expand $H = \sum_{\ell \ge 0} a_\ell h_\ell$ in normalized Hermites. Its coefficients below k_0 vanish because $a_\ell = \mathbb{E}[F h_\ell(v^T X)]$, and $\sum a_\ell^2 \le 1$.

For a Gaussian pair of correlation t, $\mathbb{E}[h_\ell(Z) \mid n^T X] = t^\ell h_\ell(n^T X)$: condition the generating function $\exp(uZ - u^2/2)$ and match polynomial coefficients. Finite sums and L^2 contraction justify the identity for the infinite expansion. Hence

$$\|\mathbb{E}[F \mid n^T X]\|_2^2 = \sum_{\ell \ge k_0} a_\ell^2 t^{2\ell} \le t^{2k_0}. \tag{BW2}$$

Cauchy–Schwarz with the feature proves its central correlation bound; the full residual costs at most ζ. For $t = 0$, independence and $\mathbb{E}[F] = 0$ give zero central correlation. Pure-bias and zero features satisfy the same assertion. □

G.4 The original Gaussian bank and the declared readout

Let $m \ge 1$ be arbitrary, and let the original spatial rows be independent $w_j \sim N(0, s_w^2 I_d)$ with common public $s_w > 0$. Biases and training heads retain their separately declared original laws; this argument imposes no further condition on them. Fix a public $B_1 > 0$ and the common feasible set $\mathcal{V} \subseteq \{v : \|v\|_1 \le B_1\}$. Define

$$\mathcal{R}(0) = \inf_{v \in \mathcal{V}} \left\| y - \sum_j v_j \phi_{w_j, b_j} \right\|_2^2. \tag{BW3}$$

Any existing ℓ^2 cap is kept; no intercept or projection is added. For $0 < \delta_I < 1$, set

$$\ell_I = \log(2m/\delta_I), \quad Q_I = r + 2\sqrt{r\ell_I} + 2\ell_I, \qquad d \ge 16\ell_I. \tag{BW4}$$

Lemma G.3. *With probability at least $1 - \delta_I$ under the original draw,*

$$\max_j \|U^T(w_j/\|w_j\|)\|^2 \le 2Q_I/d, \qquad \mathcal{R}(0) \ge 1 - 2B_1\{(2Q_I/d)^{k_0/2} + \zeta\}. \tag{BW5}$$

There is no condition $m < d$. The bound is simultaneous for all coefficients in the unchanged class, including adaptive choices.

Proof. For a standardized row g, the numerator $\|U^T g\|^2$ and denominator $\|g\|^2$ have chi-square laws of degrees r and d. The usual bounds, proved directly below, are

$$\Pr\{Z_k > k + 2\sqrt{kx} + 2x\} \le e^{-x}, \quad \Pr\{Z_k < k - 2\sqrt{kx}\} \le e^{-x}. \tag{BW6}$$

Indeed the Gaussian MGF gives $\log \mathbb{E}[e^{u(Z_k - k)}] \le ku^2/(1-2u)$ for $0 < u < 1/2$ and $\log \mathbb{E}[e^{-u(Z_k - k)}] \le ku^2$ for $u > 0$. Exponential Markov, with respectively $u = \sqrt{x}/(\sqrt{k} + 2\sqrt{x})$ and $u = \sqrt{x/k}$, gives BW6. The denominator is at least $d/2$ when $x = \ell_I$. A union over both tails and all rows has failure $2me^{-\ell_I} = \delta_I$; no independence of numerator and denominator is asserted. On this event BW1 applies to every feature. Finally, writing $\phi_j = \phi_{w_j, b_j}$, for every feasible v,

$$\left\| y - \sum_j v_j \phi_j \right\|_2^2 = 1 - 2\sum_j v_j \mathbb{E}[y\phi_j] + \left\| \sum_j v_j \phi_j \right\|_2^2 \ge 1 - 2B_1 \max_j |\mathbb{E}[y\phi_j]|. \tag{BW7}$$

Taking the infimum proves the claim with no feature-independence or Gram-conditioning premise. □

H MIXTURE PROFILE ACQUISITION

All public quantities are those of Appendix B.1. Here $X \sim N(0, I_d)$, $P_U = UU^T$, $P_{U^\perp} = I_d - P_U$, φ_v is the $N(0, v)$ density, $\varphi = \varphi_1$, and Φ is the standard normal CDF. The h_k below are normalized probabilists' Hermite polynomials. The equation families BM, BA and BAJ all refer to calculations printed in this section. Write $\mathcal{V}_n = \sum_{j=1}^m V_{j,n}$ for the complete normalized row energy. Within the base acquisition calculation write $\varepsilon = \varepsilon_{\mathrm{BM}}$ and $\kappa = \kappa_{\mathrm{BM}}$. Its base endpoint $n_0 + n_1$ is N_*, and its base envelope is $M_0 = 2\sqrt{S_0}$. In the continuation calculation the auxiliary notation means $M_1 = M$, $T_{\mathrm{pub}} = T$, $S_* = S_0$ and $T_{\mathrm{suf}} = T_G + 1$, $S_{\mathrm{tube}} = S$, and $N_A = N$. Every step and force bound is imposed with the larger (M, T); the same step is used on the entire trajectory.

H.1 FULL LINKS AND UNCHANGED ACTUAL DYNAMICS

Let $J = 1 - \alpha > 0$, $0 \le \alpha < 1$, and $\sigma_\alpha(t) = \alpha t + Jt_+$. For probability measures μ_i on $[1/3, 1]$ define

$$g_i(t)\varphi(t) = -\int \varphi_v'''(t) d\mu_i(v), \quad N_i = \|g_i\|_2, \quad \bar{g}_i = g_i/N_i,$$
$$y_c = \sum_{i=1}^r \lambda_i \bar{g}_i(u_i^T X), \quad U^T U = I_r, \quad \lambda_i > 0, \quad \sum_i \lambda_i^2 = 1. \tag{BM1}$$

Signs can be absorbed into the axes because these full links are odd. The exact H1-convergent series is

$$g_i = \sum_{j \ge 0} \frac{(-1)^j \sqrt{(2j+3)!}}{2^j j!} \left[\int (1-v)^j d\mu_i(v)\right] h_{2j+3},$$
$$\sqrt{6} \le N_i < 9, \quad \|y_c\|_2 = 1, \quad \|y_c\|_{H^1} < 13. \tag{BM2}$$

Indeed integration by parts gives the generating function $t^3 \int e^{-(1-v)t^2/2} d\mu_i(v)$; the squared coefficients and their degree weights are dominated by the summable sequence at $v = 1/3$. For $a = 1 - v$, the squared L2 and derivative norms before normalization are $3(2 + 3a^2)/(1 - a^2)^{7/2} < 80$ and $(18 + 69a^2 + 18a^4)/(1 - a^2)^{9/2} < 800$. The coefficient of h_3 is $\sqrt{6}$, proving BM2. If $\mu_i(v < 1) > 0$, every displayed coefficient is nonzero. At $\mu_i = \delta_{1/3}$ the normalized link is at L2 distance greater than one from h_3. Thus this is not a tiny cubic neighborhood.

The full teacher may be any function of $U^T X$ satisfying

$$y = y_c + e, \qquad \|y\|_2 = 1, \qquad \|e\|_{H^1} \le \zeta. \tag{BM3}$$

This optional small residual retains its actual mean, linear part and every tail coefficient. The central links intrinsically have zero mean and first moment; no centering operation is performed. This partial branch does not cover arbitrary order-one mean or linear components. Train the original averaged network and full MSE by simultaneous raw GD:

$$f = m^{-1} \sum_j a_j \sigma_\alpha(w_j^T X + b_j), \qquad L = \mathbb{E}[(y - f)^2], \qquad \eta = mh/2. \tag{BM4}$$

Every initial raw entry is independent $N(0, s^2/d)$. For analysis only, write $(a, w, b) = s(q, \theta)$, $\theta = (w, b)$, suppressing decorations; put $\rho = \|w\|$, $V = q^2 + \|\theta\|^2$. The exact recurrence is

$$q^+ = q + h(C + \theta^T E), \quad \theta^+ = \theta + hq(\nabla C + E), \quad C = \mathbb{E}[y_c \sigma_\alpha(w^T X + b)], \tag{BM5}$$

$$E_j = \mathbb{E}[(e - s^2 \widehat{f})\sigma_j'(X, 1)], \quad \widehat{f} = m^{-1} \sum_l q_l \sigma_\alpha(\theta_l^T(X, 1)), \quad \|E_j\| \le \zeta + \frac{s^2}{2m} \sum_l V_l. \tag{BM6}$$

Here $\sigma_j' = \sigma_\alpha'(\theta_j^T(X, 1))$ is a scalar gate; its following factor $(X, 1)$ is the augmented input vector. The same adaptive vector generates both errors. Gaussian Cauchy–Schwarz gives the last bound. The full uncentered H1 kernel and BM2 give

$$|C| \le \|\theta\|, \quad \|\nabla C\| \le 1, \quad \|\theta\| \|D^2 C\| \le 52, \quad \theta^T \nabla C = C. \tag{BM7}$$

This holds away from the augmented origin, including pure-bias states. The regularity and compatible-update bounds are proved in Appendix F.

Let B_i^* be the unique solution

$$B_i^*(1+B_i^*)m_i(B_i^*)=1, \quad m_i(B)=\frac{\int v^{-5/2}e^{-B/(2v)}d\mu_i(v)}{\int v^{-3/2}e^{-B/(2v)}d\mu_i(v)}, \quad \frac{\sqrt{7/3}-1}{2}\le B_i^*\le\frac{\sqrt{5}-1}{2}. \tag{BM16}$$

Put $b_i^*=\sqrt{B_i^*}$ and take the actual integer checkpoint

$$n_0=\lceil 4R/(hk)\rceil, \quad n_1=\lceil T_1/h\rceil, \quad N_*=n_0+n_1, \qquad N_*h\le T_0. \tag{BM17}$$

H.2 Exact selected-cone dynamics, including candidate exits

For $x_i=u_i^Tw$, $\rho>0$, put $D_{iv}=\rho^2-(1-v)x_i^2$, $B=b^2/\rho^2$, $B_{iv}=b^2/D_{iv}$. Three integrations by parts give the full uncentered correlation

$$C=\sum_i\int C_{iv}d\mu_i(v), \qquad C_{iv}=-\frac{J\lambda_i}{N_i}bx_i^3D_{iv}^{-3/2}\varphi(b/\sqrt{D_{iv}}). \tag{BM23}$$

Write $H_{iv}=qC_{iv}$, $H=qC$. On $B\le 3/4$ one has $\rho^2/3\le D_{iv}\le\rho^2$ and $B_{iv}\le 9/4$. For a selected axis i fix $\epsilon=-\operatorname{sign}(qb)$ and set

$$x=\epsilon x_i>0, \quad u^2=\sum_{l\ne i}x_l^2, \quad O=\|P_{U^\perp}w\|, \quad u\le\vartheta x,$$
$$H_i=\int H_{iv}d\mu_i>0, \quad H_{\text{other}}=\sum_{l\ne i}\int|H_{lv}|d\mu_l. \tag{BM24}$$

Persistence of the head and bias signs is proved below. Uniformly,

$$\sum_l\int|C_{lv}|d\mu_l\le J|b|, \quad |C_i|\ge\frac{J\lambda_i|b|x^3}{72\rho^3}, \quad H_{\text{other}}\le\eta_iH_i, \qquad \eta_i=\frac{72\vartheta^3}{\lambda_i}<1/100. \tag{BM25}$$

For the upper bound use $N_l\ge\sqrt{6}$, $D_{lv}\ge\rho^2/3$, $3\sqrt{3}\varphi(0)/\sqrt{6}<1$, and $\sum_l\lambda_l|x_l|^3\le\rho^3$. For the lower bound use $N_i<9$, $\varphi(3/2)>1/8$. The same upper estimate on the other coordinates has numerator u^3.

The pure teacher spatial candidate has a common radial multiplier and nonnegative quadratic boosts in $y_l=\epsilon x_l$:

$$y_l^t=Ay_l+d_ly_l^2, \quad P_{U^\perp}w^t=AP_{U^\perp}w, \quad A=1+\sum_l\int\frac{h(B_{lv}-3)H_{lv}}{D_{lv}}d\mu_l, \tag{BM26}$$

$$d_l=\frac{hpJ|b|\lambda_l}{N_l}\int D_{lv}^{-3/2}\varphi(b/\sqrt{D_{lv}})\left[3+\frac{(1-v)x_l^2(3-B_{lv})}{D_{lv}}\right]d\mu_l, \qquad p=|q|.$$

The bracket is between 3 and 9. As $p\le\|\theta\|\le\sqrt{7/4}\rho$, BM25 gives

$$|A-1|\le 11h, \qquad d_i\ge\frac{hpJ|b|\lambda_i}{24\rho^3}, \qquad \|(d_ly_l^2)_{l\ne i}\|\le\frac{9hpJ|b|u^2}{\rho^3}. \tag{BM27}$$

Thus $A>0$, and $u\le\vartheta x$ implies the candidate inequality

$$u^t-\vartheta x^t\le-\frac{\vartheta hpJ|b|\lambda_ix^2}{48\rho^3}. \tag{BM28}$$

The force displacement changes its left side by at most $2hp\nu$. Since $|C|\ge k$ implies $J|b|\ge k$, (31) pays that displacement whenever $x\ge x_0/2$, $\rho\le M$. This proves the actual candidate cone constraint, not just an inward derivative at its boundary.

The own-component contribution to the oriented x increment is

$$\int hH_{iv}\left[\frac{3}{x}-\frac{vx(3-B_{iv})}{D_{iv}}\right]d\mu_i=\int hH_{iv}\frac{3(u^2+O^2)+B_{iv}vx^2}{xD_{iv}}d\mu_i\ge\frac{3hH_iu^2}{x\rho^2}. \tag{BM29}$$

The adverse radial contribution of the other axes has magnitude at most $9hH_{\text{other}}x/\rho^2$. Relative to the displayed lower bound, it is at most $216\vartheta/\lambda_i < 1$ when $u > 0$; when $u = 0$ it vanishes. Hence $x^t \geq x$. The cumulative actual decrease is at most $\nu MT \leq x_0/4$. Starting with $x \geq x_0$ therefore closes $x \geq x_0/2$ at every candidate state.

For the upper bias barrier let $F = b^2 - (3/4)\rho^2$. Its derivative is

$$DF[q\nabla C] = 2\sum_l \int [1 - (7/4)B_{lv}]H_{lv}d\mu_l. \tag{BM30}$$

If $2/3 \leq B \leq 3/4$, this is at most $2[-1/6 + 3\eta_i]H_i \leq -H_i/4 \leq -pk/8$. The pure quadratic remainder and full force displacement together are at most $hpM(4h + 2\nu) < hpk/8$ by (30)–(31). If $B < 2/3$, the initial gap is $F \leq -\rho^2/12$, whereas the absolute actual one-step change is at most $8h\rho^2$. Thus both cases, and every transition between them, give $B^+ \leq 3/4$.

The radial multiplier further satisfies

$$A \leq 1 - \frac{hH}{3\rho^2}, \qquad O^+ \leq (1 - t/3)O + tM^2\nu/k, \qquad t = hH/\rho^2 \geq 0. \tag{BM31}$$

Indeed its own term is at most $-(3/4)hH_i/\rho^2$ and its adverse terms sum to at most $9h\eta_i H_i/\rho^2$. Use $H \leq (1 + \eta_i)H_i$, $A > 0$, and $|C| \geq k$. Then $\nu \leq 2k/(3M^2)$ preserves $O \leq 2$. No analogous cone or favorable sign has been assumed for other rows.

H.3 Favorable growth and a lower bias barrier

We record the generic full-H1 quotient estimate explicitly:

$$Q = \frac{qC}{V}, \qquad Q^+ \geq Q - h\nu^2, \qquad |q| \leq \|\theta\|, \quad \|E\| \leq \nu \leq 1, \quad h \leq 2^{-12}. \tag{BM32}$$

To verify it, put $z = (q, \theta)$, $F_H = \nabla(qC)$, $p_E = (\theta^T E, qE)$, $a = z^T p_E/V$, $p_\perp = p_E - az$, $G = F_H - 2Qz$. Homogeneity makes the new quotient $Q(z+\tau(G+p_\perp))$, where $\tau = h/[1+h(2Q+a)]$. This is a tangent retraction. The joining segment has norm at least $\sqrt{V}$ and head/hidden ratio at most two. Differentiation with BM7 gives $\|D^2Q\| \leq 128/V$ there: the Hessian-of-H term contributes at most $105/V$ and the remaining quotient terms at most $9/V$. Also $\|p_\perp\| \leq \nu\sqrt{V}$ and $\gamma = \|G\|/\sqrt{V} \leq 1$. Taylor's theorem yields

$$Q^+ - Q \geq \tau(\gamma^2 - \nu\gamma) - 64\tau^2(\gamma + \nu)^2 \geq \tau(\gamma^2/4 - 3\nu^2/4) \geq -h\nu^2.$$

Balance persists exactly, since homogeneity also gives

$$\|\theta^+\|^2 - (q^+)^2 \geq (1 - h^2\|\nabla C + E\|^2)(\|\theta\|^2 - q^2), \qquad V^+ \leq (1 + 2h)^2 V.$$

These estimates apply to every original row, independent of signs.

For a selected seed with $p_0 \geq 1/(2\sqrt{d})$, $|C_0| \geq k$, $H_0 > 0$, $V_0 \leq 5$, BM32 gives $Q_n \geq Q_* - T\nu^2 \geq 1599Q_*/1600$. Orient C by the initial head sign and call it K. Euler's identity gives $K/\|\theta\| \geq 2Q_n \geq Q_*$. Thus $\nu \leq Q_*/4$ is at most one quarter of both $K/\|\theta\|$ and $\|\nabla K\|$. The head sign persists and $p^+ \geq p + 3hK/4$. Taylor's theorem with BM7 gives for every $0 \leq u \leq 1$

$$K(\theta + uhp(\nabla K + \widetilde{E})) \geq K(\theta) + (u/2)hp\|\nabla K\|^2, \qquad p^+ \geq p + 3hk/4. \tag{BM33}$$

Here $\widetilde{E}$ includes the fixed head orientation. The first-order inner product is at least $(3/4)uhp\|\nabla K\|^2$, while the curvature remainder is at most $82u^2h^2p\|\nabla K\|^2$. Both $\|\theta\|^2$ and V increase. Since BM23 vanishes at $b = 0$ or $w = 0$, BM33 preserves the original bias sign along each entire hidden step. BM28–29 preserve the chosen spatial orientation. This supplies the signs used above without assuming a future phase. The argument is a simultaneous induction on signs, $K \geq k$, the cone and the two barriers: every inequality uses only its pre-state conditions and has just been established for its complete candidate.

Set $W = (\rho^2 - 8b^2)_+$. When $B \leq 1/6$, the first-order change of the untruncated quadratic is $2h\sum_l \int (9B_{lv} - 8)H_{lv}d\mu_l \leq 0$: the own coefficient is at most $-7/2$, the other absolute coefficients at most 8. When $B > 1/6$, its initial negative gap is at least $\rho^2/3$; the possible positive first-order

change is at most $30h\rho^2$. Indeed $\sum_l \int |H_{lv}| \le pJ|b| \le (6/5)\rho^2$ and $|9B_{lv} - 8| \le 49/4$. The pure quadratic remainder is at most h^2M^2, and the force cost is at most $20h\nu M^2$ (the quadratic matrix has norm eight). The positive part therefore obeys, including all branch crossings,

$$W_n \le W_0 + TM^2(h + 20\nu) < 5. \tag{BM34}$$

Also $\mathcal{D} = \|\theta\|^2 - p^2$ satisfies $\mathcal{D}^+ - \mathcal{D} \le 4h^2M^2$, so $\mathcal{D}_n \le 6$. Consequently, at every state with $\rho \ge 8$,

$$1/16 \le B \le 3/4, \qquad p/\rho \ge \sqrt{1 - 6/\rho^2} > 0.95. \tag{BM35}$$

No monotone spatial radius has been assumed. At n_0 BM33 gives $p > 3R$; at all later states balance and the upper bias barrier give $\rho \ge p/\sqrt{7/4} > 2R$. This is a persistent late-radius bound.

H.4 Actual convergence to the link-dependent bias

In BM16, $m_i(B) = \mathbb{E}_B[1/v]$ under probability weights proportional to $v^{-3/2}e^{-B/(2v)}d\mu_i$. Thus $1 \le m_i \le 3$ and $m_i' = -\mathrm{Var}_B(1/v)/2 \in [-1/2, 0]$. On $0 \le B \le 3/4$, the derivative of $B(1 + B)m_i(B)$ is between one and nine. At the endpoints stated in BM16 the function is, respectively, at most and at least one. This proves existence, uniqueness and the stated bounds for B_i^*.

At a late selected row write $e_a = 1 - x^2/\rho^2 \le \vartheta^2 + 4/R^2$. The own dimensionless variance in BM23 is $d_v = v + (1 - v)e_a$ rather than v. Let $m_{i,e_a}(B)$ be the resulting weighted mean of $1/d_v$. Differentiation with respect to e_a gives a direct term of magnitude at most nine, and a covariance term at most 48: use $1/d_v \le 3$ and $|\partial_{e_a} \log(d_v^{-3/2}e^{-B/(2d_v)})| \le 8$. Hence, uniformly on the corridor,

$$|m_{i,e_a}(B) - m_i(B)| \le 60e_a. \tag{BM36}$$

With $t_i = hH_i/\rho^2$, BM25 and BM35 give

$$\kappa h \le t_i \le (6/5)h. \tag{BM37}$$

For the lower bound use $p/\rho > 0.95$, $|b|/\rho \ge 1/4$ and $x/\rho \ge 0.9$ in BM25; the coefficient is greater than $J\lambda_i/1024$. For the upper bound use $H_i \le \sum_l \int |H_{lv}| \le pJ|b|$.

Expand the actual quotient $B^+ = (b^+)^2/\|w^+\|^2$:

$$B^+ - B = \frac{2h}{\rho^2} \sum_l \int [1 - (1 + B)B_{lv}]H_{lv}d\mu_l + \mathcal{E}, \qquad |\mathcal{E}| \le 200h^2 + 20h\nu. \tag{BM38}$$

Here is an explicit remainder check. Both $\|\Delta w\|/\rho$ and $|\Delta b|/\rho$ are at most $3h$. The denominator divided by ρ^2 differs from one by at most $9h$. In the numerator for $B^+ - B$, the first-order teacher term has absolute value at most $8h$, the force term at most $5h\nu$, and the quadratic term at most $13h^2$. Division by $1 - 9h$ proves the conservative remainder in BM38. BM25 and BM36 change the teacher term from $2t_i[1 - (1 + B)Bm_i(B)]$ by at most $2t_i(80e_a + 3\eta_i) \le 400he_a$. The ideal bracket equals $-a(B)(B - B_i^*)$ with $1 \le a(B) \le 9$. Its candidate coefficient has no sign reversal, since $2t_ia(B) < 1$. Therefore

$$|B^+ - B_i^*| \le (1 - \kappa h)|B - B_i^*| + h[400(\vartheta^2 + 4/R^2) + 200h + 20\nu]. \tag{BM39}$$

By (26)–(31) the last bracket is at most $\kappa\varepsilon/4$. After $n_1h \ge T_1$, the initial error, at most one, is at most $\varepsilon/4$, and the accumulated error at most $\varepsilon/4$. Thus $|B_{N_*} - B_i^*| \le \varepsilon$. The original bias sign persists, and $|\sqrt{B} - \sqrt{B_i^*}| \le 2|B - B_i^*|$. Also $\|w/\rho - \epsilon u_i\| \le \sqrt{2}\vartheta + 3/R$. These inequalities and (26) prove the calibrated profile assertion with room to spare. This proof includes the actual integer endpoint, all intermediate bias transitions, and a radius bound acquired before the late phase.

H.5 Original Gaussian event and full coupled-force closure

Write the normalized initial row as $(q, w, b) = (Z, G, B_0)/\sqrt{d}$. For each (i, ϵ, τ) use the event

$$\epsilon G_i \in [1, 2], \quad |G_l| \le \vartheta/(4\sqrt{r - 1})\ (l \ne i,\ l \le r), \quad \tau B_0 \in [1/2, 1], \quad -\epsilon\tau Z \in [1/2, 1], \tag{BM40}$$

omitting the other-axis restriction for $r = 1$. It has probability exactly p_{rect} and places no restriction on the remaining $d - r$ coordinates. The union bound over $4r$ categories shows that every category

occurs among the original m rows except with probability $\delta/4$. Distinct categories need distinct rows. This is a property of the unscreened draw; no row is filtered from the actual training network. Separately, Gaussian and chi-square tails give, for all original rows,

$$1/2 \le \rho_{j,0} \le 2, \qquad |Z_j|, |B_{0,j}| \le \sqrt{2\ell}, \tag{BM41}$$

except with probability at most $m(4e^{-\ell} + 2e^{-d/8}) \le 6me^{-\ell} < \delta/8$. On BM41, $V_{j,0} < 5$, $q_{j,0}^2 \le \|\theta_{j,0}\|^2$ and $O_{j,0} \le 2$. Each designated row additionally has $x \ge x_0$, $u \le \vartheta x_0/4$, $B \le 4/d \le 1/16$, $p_0 \ge 1/(2\sqrt{d})$, and $|C_0| \ge J\lambda_i/(2048d^2) \ge k$ by BM25. Its signs in BM40 give $H_0 > 0$.

At every pre-state with force at most ν, BM7 and BM5 give $V^+ \le (1+2h)^2 V$ for every original row. Starting from $\sum V_0 \le 5m$, induction through all candidate states with $nh \le T$ yields $\sum V_n \le 5me^{4T} = M^2/4$. BM6 and (34)–(35) then close $\|E_j\| \le \nu$ at each such state, including the force recomputed at the final endpoint. This is a bootstrap on a public bound, not an assumed trajectory. BM28–39 consequently apply simultaneously to all designated rows of the same actually interacting run. Furthermore $\|f_n\|_2 \le s^2 M^2/(8m) \le \epsilon_L/160$; since $\|y\|_2 = 1$, this proves the original-loss bound. The raw head bound follows from $\sum q_j^2 \le M^2/4$. Each of the $4r$ distinct designated rows has raw initial spatial norm at most $2s$ and final norm exceeding $2sR$, giving the stated spatial-motion bound by the reverse triangle inequality.

H.6 Full actual tube and extension of the selected profiles

Work first on the event in (BM40)–(BM41), of failure at most $3\delta/8$. Its initial all-row bound is $V_{j,0} < 5$ and its initial balance holds. The actual force obeys

$$\|E_j\| \le \zeta + \frac{s^2}{2m}\mathcal{V}_n. \tag{BA18}$$

Under a pre-state force cap, exact compatibility preserves balance and gives $V_j^+ \le (1+2h)^2 V_j$. The choices in (30)–(35) meet all bounds used in (BM25)–(BM41), so the preceding induction supplies $\mathcal{V}_n \le S_*$ through N_* and its same acquired rows. For every subsequent candidate with $(n - N_*)h \le T_{\mathrm{suf}}$,

$$\mathcal{V}_n \le S_*(1+2h)^{2(n-N_*)} \le S_* e^{4(n-N_*)h} \le S_{\mathrm{tube}}. \tag{BA19}$$

Equation BA18 and (35) then prove the force cap recomputed at that candidate. This closes a full coupled induction through N_A, not an assumed future force condition. The pre-acquisition event is unchanged when these cuts are strengthened.

For clarity, here is why the same selected profiles persist over this much longer interval, instead of merely reusing the calibrated profile assertion at a new time. Every one-step selected-cone, head/bias sign, upper-bias, correlation and outside-radius assertion in BM25–33 uses only a current norm envelope, a force cap, and cumulative time. Their proofs therefore apply with the larger current envelope M_1 and cumulative time T_{pub} in (30)–(31). Specifically:

- The uniform quotient debit is $Q^+ \ge Q - h\nu^2$. The selected initial floor Q_* remains above $Q_* - T_{\mathrm{pub}}\nu^2 \ge 1599Q_*/1600$, because $\nu \le Q_*/4$ and $\nu \le 1/(400T_{\mathrm{pub}})$. The oriented correlation $|C| \ge k$ and the head/bias signs then persist by BM33's candidate-segment argument.
- BM28's cone-force payment uses $\nu \le \vartheta\lambda_* k x_0^2/(2048M_1^3)$. BM29's possible accumulated loss of the own-axis coordinate is at most $\nu M_1 T_{\mathrm{pub}} \le x_0/4$. Thus every candidate preserves the selected cone and own coordinate $x \ge x_0/2$.
- BM30's upper-bias barrier uses $h, \nu \le k/(512M_1)$. BM31's selected outside multiplier and its force payment use $\nu \le 2k/(3M_1^2)$, keeping $O \le 2$. These statements cover candidate exits and repeated bias transitions.

The two cumulative quadratics are now paid over the entire extended window:

$$\begin{aligned} W_n = (\rho_n^2 - 8b_n^2)_+ &\le W_0 + T_{\mathrm{pub}}M_1^2(h + 20\nu) < 5, \\ 0 \le \|\theta_n\|^2 - q_n^2 &\le \|\theta_0\|^2 - q_0^2 + 4hT_{\mathrm{pub}}M_1^2 \le 6. \end{aligned} \tag{BA20}$$

The first inequality uses BM34's exact positive-part calculation; the second is the compatible balance identity. (30)–(31) pay the displayed debits by $1/4 + 1/4$ and 1, respectively. The initial quadratics are below 5 (indeed $W_0 \le 4$).

The selected head grows as $p^+ \ge p + 3hk/4$ throughout. At $n_0 = \lceil 4R/(hk) \rceil$ it has $p > 3R$, and subsequently balance and the upper bias barrier give $\rho > 2R$. Thus BA20 gives $p/\rho \ge \sqrt{1 - 6/\rho^2} > 19/20$, and the lower bias barrier $B \ge 1/16$, at every later state. None of these steps assumes that the spatial radius itself is monotone.

Finally BM36–39's exact bias-ratio calculation gives, at every $n \ge n_0$ up to the current candidate,

$$|B^+ - B_i^*| \le (1 - \kappa_{\mathrm{BM}} h)|B - B_i^*| + h\{400(\vartheta^2 + 4/R^2) + 200h + 20\nu\}. \tag{BA21}$$

The expression in braces is at most $\kappa_{\mathrm{BM}}\varepsilon/4$ by the unchanged R, ϑ choices and the smaller h, ν. At every $n \ge N_*$ the initial error has decayed below $\varepsilon/4$, while the accumulated forcing contributes at most $\varepsilon/4$. Hence $|B_n - B_i^*| \le \varepsilon/2 \le \varepsilon$ for the whole suffix. The persistent angular error is at most $\sqrt{2}\vartheta + 3/R$; the bias-ratio error is at most 2ε. Since $\vartheta \le e_*/16$, $R \ge 16/e_*$ and $\varepsilon = e_*/8$, their sum is below e_*. This proves the same calibrated profile assertion at each later checkpoint, with the same original rows and signs. The enlarged cumulative cuts are essential to this extension.

H.7 Actual tube and growth of every acquired axis

Work on the event in (BM40)–(BM41), of failure at most $3\delta/8$. Every row starts with $V_j < 5$ and nonnegative balance difference. Under a pre-state force cap exact compatibility gives

$$V_j^+ \le (1 + 2h)^2 V_j, \quad \|E_j\| \le \zeta + \frac{s^2}{2m} \sum_l V_l, \quad \sum_j V_{j,n} \le S \quad (n \le N). \tag{BAJ17}$$

The base bound through N_* is S_0; remaining duration is at most $T_G + 1$, paid in S. (34)–(35) recomputes the cap at each candidate, closing a full-bank induction with every signed student interaction.

Selected profiles persist by the proof BA18–21 with (M, T). Specifically, cone error is paid by $\nu \le \vartheta\lambda_* k x_0^2/(2048M^3)$, coordinate drift by $\nu MT \le x_0/4$, outside control by $\nu \le 2k/(3M^2)$, and the upper-bias barrier by $h, \nu \le k/(512M)$. The cumulative quadratic debits satisfy $TM^2(h + 20\nu) < 1$ and $4hTM^2 \le 1$. The unchanged $R, \vartheta, \varepsilon$ retain the contraction around each B_i^* over all $N_* \le n \le N$. Thus all four original categories per axis retain their signs and satisfy

$$\begin{aligned} &\rho_j > 2R, \quad |q_j|/\rho_j > 19/20, \quad 1/16 \le b_j^2/\rho_j^2 \le 3/4, \\ &\|(w_j/\rho_j, b_j/\rho_j) - (\epsilon u_i, \tau\sqrt{B_i^*})\| \le e_*, \quad \mathrm{sign}(q_j) = -\epsilon\tau. \end{aligned} \tag{BAJ18}$$

The compatible balance identity holds for every row. Its upper debit is at most $4h^2 V_j$ per step, so the same bounds imply

$$0 \le D_j := \|\theta_j\|^2 - q_j^2 \le 6 \quad (n \le N). \tag{BAJ19}$$

No acquired balance assumption is imposed on an unselected row.

I Control of every signed row for mixture teachers

I.1 Full model and the sharp static ceiling

Use $X \sim N(0, I_d)$, $\sigma_\alpha(t) = \alpha t + (1 - \alpha)t_+$, $(A_j, W_j, B_j) = s(q_j, w_j, b_j)$ and $\theta_j = (w_j, b_j)$. For a normalized row, $C(\theta) = \mathbb{E}[y_c \sigma_\alpha(w^T X + b)]$; $P_U = UU^T$ and $P_{U^\perp} = I_d - P_U$. The densities φ_v and φ are those of $N(0, v)$ and $N(0, 1)$, respectively.

Use precisely the full teacher, network, original Gaussian law and simultaneous update (24) and BM4:

$$\begin{aligned} g_v(t)\varphi(t) = -\varphi_v'''(t), \quad g_i = \int_{[1/3,1]} g_v \, d\mu_i(v), \quad N_i = \|g_i\|_2, \\ y_c = \sum_i \frac{\lambda_i}{N_i} g_i(u_i^T X), \quad y = y_c + e, \quad \|y\|_2 = 1. \end{aligned} \tag{SB1}$$

Here each μ_i is an arbitrary probability measure, $\lambda_i > 0$, $\sum_i \lambda_i^2 = 1$, and $U^T U = I_r$. No Hermite coefficient is removed. In particular $\sqrt{6} \le N_i < 9$ and $\|y_c\|_{H^1} < 13$. The full H^1 residual retains its mean, linear and interaction terms. For $J = 1 - \alpha > 0$, let

$$K_i(B) = \frac{J\lambda_i}{2N_i}\sqrt{\frac{B}{1+B}}\int v^{-3/2}\varphi(\sqrt{B/v})\,d\mu_i(v), \qquad K_i^* = \max_{B\ge 0} K_i(B), \qquad K_* = \max_i K_i^*. \tag{SB2}$$

For each fixed axis i, $\mathbb{E}_B[\cdot]$ denotes expectation under the probability measure proportional to $v^{-3/2}e^{-B/(2v)}d\mu_i(v)$; the axis index is suppressed. The maxima occur at the unique mixture roots $B_i^*(1+B_i^*)\mathbb{E}_{B_i^*}[1/v] = 1$; thus $B_i^* \in [(\sqrt{7/3}-1)/2, (\sqrt{5}-1)/2]$. The proof following (BM35) in Appendix H establishes uniqueness in this interval. There is no other global maximum: the logarithmic derivative of the axis objective is $1/[2B(1+B)] - \mathbb{E}_B[1/v]/2 < 0$ for $B > \beta$; the objective vanishes at zero and infinity. In particular $J/(1000\sqrt{r}) \le K_* \le 1/2$.

Write $p = |q|$, $\rho = \|w\|$, $V = p^2 + \rho^2 + b^2$, $H = qC$, $Q = H/V$, and $B = b^2/\rho^2$. At zero spatial radius $C = 0$, including pure-bias rows. Put

$$\beta = \frac{\sqrt{5}-1}{2}, \qquad \Theta(a) = \left\{\frac{a/(1+a)}{\beta/(1+\beta)}e^{-(a-\beta)}\right\}^{1/2}. \tag{SB3}$$

For every $a \ge \beta$, every row, every head ratio and every collection of measures and coefficients,

$$B \ge a \implies |Q| \le \Theta(a)K_*. \tag{SB4}$$

This ceiling is sharp uniformly over the family: equality holds for a single cubic axis $\mu = \delta_1$, $B = a$, and $p^2 = (1+a)\rho^2$.

Here is a full proof that includes signed spatial coordinates. Let $s_i = (u_i^T w/\rho)^2$, $d = 1 - (1-v)s_i$. The head factor is at most $1/2$. The squared ratio of the absolute i, v contribution to s_i times its axis contribution at β is

$$\frac{B/(1+B)}{\beta/(1+\beta)}s_i(v/d)^3\exp(-B/d + \beta/v). \tag{SB5}$$

Since $v \le d$, the function $3\log v + \beta/v$ is increasing on $[1/3, 1]$. Consequently the last two factors are at most $\exp[-(B-\beta)/d] \le \exp[-(B-\beta)]$. Sum the resulting unsquared positive bounds, using $\sum_i s_i \le 1$ and $K_i(\beta) \le K_i^*$. Finally Θ is decreasing for $B \ge \beta$, because $(\log\Theta^2)' = 1/[B(1+B)] - 1$. This proves SB4 without a parity cancellation or a favorable row assumption.

In particular

$$\begin{gathered}\Theta(3/4) = 0.991615\ldots, \qquad \Theta(29/20) = 0.821163\ldots < 411/500,\\ \Theta(3/2) = 0.806393\ldots.\end{gathered} \tag{SB6}$$

Only the strict rational bound in the middle is used below. It can be certified without numerical optimization: first write $\beta/(1+\beta) = 1 - \beta$, enclose β between $618033/10^6$ and $618035/10^6$, substitute the upper endpoint in both increasing factors of SB3, and lower bound $e^{831965/10^6}$ by its Taylor polynomial of degree eight. All remaining comparisons are rational.

I.2 A BOUNDED BIAS WEIGHT FOR EVERY SIGNED COMPONENT

For $B \ge 0$, define the continuous function

$$c(B) = \begin{cases} 1/5, & 0 \le B \le 1,\\ \dfrac{(1+B)^{-1} - 9/2 + 6B}{2\{3B(1+B)-1\}}, & 1 \le B \le 3/2,\\ c(3/2), & B \ge 3/2,\end{cases} \qquad F(B) = \exp\left(\int_0^B c(u)\,du\right). \tag{SB7}$$

On $[0, 3/2]$, $1/5 \le c \le 1/4$, c is Lipschitz with constant two, and

$$1 \le F(B) \le e^{3/8} < 3/2. \tag{SB8}$$

For example, the numerator of $c - 1/5$, after multiplication by the positive denominator and $10(1+B)$, is $(B-1)(-12B^2 + 24B + 31)$. For $c \le 1/4$, the relevant numerator is $(1+B)^{-1} - 4 +$

$(9/2)B - (3/2)B^2 \le 1/2 - 5/8 < 0$. Differentiating the displayed quotient bounds $|c'|$ by two; its numerator is at most $49/10$, its derivative at most six, its denominator at least ten, and the latter's derivative at most 24. The constant continuation preserves this Lipschitz bound.

For $qb \neq 0$ and $\rho > 0$, let

$$\tau = -\operatorname{sign}(qb), \quad E = \|P_{U^\perp} w\|^2 + \sum_i \min\{\tau u_i^T w, 0\}^2, \quad X = p\sqrt{E}, \qquad Z = F(B)X. \tag{SB9}$$

Orientation is analytical; no parameter is reflected. The unweighted X alone is used for pure-bias or unmarked rows; Z is needed only on marked states, where $Q > 0$ already ensures $\rho > 0$. Fix $11/10 \le a < 3/2$, put $\Delta = 3/2 - a$, and assume

$$B \le a, \qquad \frac{1}{1+B} \le k := \frac{\rho^2}{p^2} \le \frac{1+\varepsilon}{1+B}, \qquad 0 < \varepsilon \le \min\{1/15, \Delta\}. \tag{SB10}$$

For each full kernel component put $t = 1/d \in [1, 3]$. The two decisive coefficients are

$$\begin{aligned} L &= 3t - Bt^2 - k - 2c(B) + 2c(B)B(1+B)t, \\ R &= k + Bt + 2c(B) - 2c(B)B(1+B)t. \end{aligned} \tag{SB11}$$

They satisfy, uniformly over all mixtures,

$$L \ge \Delta, \qquad R \ge \Delta, \qquad L + R = (3+B)t - Bt^2. \tag{SB12}$$

For $0 \le B \le 1$, L is concave in t and its two endpoint values exceed one; $R \ge k + 2/5 \ge 9/10$. Explicitly those endpoint values of L are $13/5 - 3B/5 + 2B^2/5 - k$ and $43/5 - 39B/5 + 6B^2/5 - k$. The first exceeds one by $k \le 16/15$; the second decreases to a value at least $2 - 8/15$. For $1 \le B \le a$, put $k_0 = (1+B)^{-1}$. The choice SB7 gives at $t = 3$

$$L = (9 - 6B)/2 - (k - k_0) \ge (5/2)\Delta, \qquad R = (9 - 6B)/2 + (k - k_0) \ge 3\Delta. \tag{SB13}$$

At $t = 1$, $L > 1$ and $R \ge 2/5 + 2/5 - 3/8 = 17/40 > \Delta$. Concavity of L and affinity of R finish SB12. This uses $\Delta \le 2/5$, the reason for the harmless lower restriction $a \ge 11/10$.

To explain the new weight, use the exact components of BM5

$$H_{iv} = -\frac{J\lambda_i q}{N_i} c_i^3 b D_{iv}^{-3/2} \varphi(b/\sqrt{D_{iv}}), \quad D_{iv} = \rho^2 - (1-v)c_i^2, \quad c_i = u_i^T w. \tag{SB14}$$

Their signs equal $\operatorname{sign}(\tau c_i)$. The common radial spatial coefficient is $-\sum \int H_{iv}(3t - Bt^2)/\rho^2$. The coordinate boost magnitude is

$$k_{iv}^{\mathrm{boost}} = \frac{|H_{iv}|}{|c_i|}\{(3+B)t - Bt^2\}, \qquad \dot{B} = \frac{2}{\rho^2} \sum_i \int H_{iv}\{1 - B(1+B)t\}\, d\mu_i. \tag{SB15}$$

Dots denote the vector field of the exact simultaneous teacher Euler candidate, not a replacement flow. Form a vector z of $F(B)pw_\perp$ and the negative coordinates $F(B)p\tau c_i$. It has norm Z. For a positive component the first-order vector field is $-H_{iv}Lz/\rho^2$. For a negative component, writing $A_{iv} = |H_{iv}|$, its field is

$$A_{iv}Lz/\rho^2 + F(B)pk_{iv}^{\mathrm{boost}} e_i, \qquad -\langle z, \dot{z}_{iv}\rangle = A_{iv}F(B)^2 p^2\{R + L(1 - E/\rho^2)\} \ge \Delta\, A_{iv} Z^2/\rho^2. \tag{SB16}$$

The negative coordinate itself pays its potentially expanding radial contribution. Summing all signed components proves

$$-\langle z, \dot{z}\rangle \ge \Delta H Z^2/\rho^2 \quad \text{when } H > 0. \tag{SB17}$$

I.3 The actual candidate, including every error and crossing

Here E_{force} is the augmented compatible vector E_j of (BM6); it is distinct from the scalar signed-spatial energy E in (SB9).

Assume SB10, $Q \ge \kappa > 0$, $\|E_{\mathrm{force}}\| \le \nu \le \min\{1, \kappa\}$, and $0 < h \le 10^{-4}$. In the actual compatible simultaneous update BM4, the candidate keeps the nonzero signs of q, b, and

$$Z^+ \le (1 - h\Delta H/\rho^2)Z + (5000h^2 + 20h\nu)V. \tag{SB18}$$

Only the pre-state must be in SB10; the candidate can cross its bias or energy boundary. The function in SB7 is defined beyond the corridor precisely for this assertion.

Here are explicit finite-step remainder bounds. At $B \le 3/2$, upper balance gives $\sum_i \int |H_{iv}| \le 2\rho^2$, $|3t - Bt^2| \le 9$, and $0 \le (3+B)t - Bt^2 \le 9$. Also $|H_{iv}|/c_i^2 \le 2$. Thus the common radial coefficient has magnitude at most 18, each coordinate boost divided by its absolute coordinate is at most 18, and all pure coordinate multipliers are at least $1 - 36h > 0$. Zero coordinates stay zero. The pure head grows, since $H > 0$. The pure bias multiplier is at least $1 - 6h$, by its exact form $1 + hH/b^2 - h\sum \int H_{iv}/D_{iv}$. The full kernel bounds $|C| \le |b|$, $|C| \le \|\theta\|$ give $|b| \ge \kappa p$. The actual bias error is at most $h\nu p$; the head error is at most $h\nu\|\theta\|$. Consequently neither sign changes, including at the candidate.

Parameterize the straight pure hidden segment by $u \in [0, h]$. Because $p/\rho < 2$ and $\|\nabla C\| \le 1$, its spatial radius is at least $(1 - 2h)\rho$ and its squared bias ratio is below two. Direct quotient differentiation yields

$$|B'(u)| \le 16, \qquad |B''(u)| \le 200. \tag{SB19}$$

For detail, put $N = (b + u\dot{b})^2$ and $D = \|w + u\dot{w}\|^2$. The bounds

$$N''/D, D''/D \le 9, \quad |D'/D| \le 5, \quad N/D \le 2, \quad |N'/D| \le 6$$

give $|B'| \le 16$ and $|B''| \le 9 + 60 + 18 + 100 < 200$. Since $|c'| \le 2$, the scalar

$$s(u) = (1 + uC_{\text{oriented}}/p)\, F(B(u))/F(B(0)) \tag{SB20}$$

has $|s'| \le 10$ and $|s''| \le 1300$ on this segment. Indeed

$$C_{\text{oriented}}/p \le \sqrt{1+\varepsilon} < 2, \qquad F(B(u))/F(B(0)) \le e^{4h} < 1.001;$$

the differentiated bounds give $|s'| < 7$ and $|s''| < 1300$. Here the deliberately looser 10 is used. The bounds also hold across $B = 1, 3/2$ in the Lipschitz derivative sense.

Because pure coordinate signs persist, the exact weighted defect vector equals

$$s(h)\{z + hF(B)p\dot{w}_{\text{defect}}\}.$$

Its difference from $z + h\dot{z}$ is at most $1000h^2V$. Moreover $\|\dot{z}\| \le 42Z$: the head, bias-weight, radial and negative-coordinate terms are bounded respectively by $2Z, 4Z, 18Z, 18Z$. Using SB17 and $\sqrt{1+x} \le 1 + x/2$ therefore gives the pure part of SB18 with an error below $2000h^2V$. At $Z = 0$, the pure defect stays zero, so no division by zero is involved in that argument.

Finally connect the pure and actual candidates by a straight segment. Throughout it $\rho > \rho_{\text{old}}/2$, $B < 2$, and $p \le 2p_{\text{old}}$. The map $F(B)p\ \text{dist}(w, \text{cone}_\tau)$ is locally Lipschitz. Its head derivative is at most 2ρ; its hidden derivative norm is at most

$$p\{F + \rho|F'|\|\nabla B\|\} \le 5p, \qquad \|\nabla B\| = 2\sqrt{B + B^2}/\rho. \tag{SB21}$$

The compatible displacements $h\nu\|\theta_{\text{old}}\|$ and $h\nu p_{\text{old}}$ thus cost less than $20h\nu V$. This proves SB18 for all coordinate sign crossings caused by the actual force. The teacher and full student/residual terms have not been separated into different dynamics.

We now prove the acquisition and switching bounds used by the full-bank estimate. The actual candidate induction in Appendix H, with the public M, T, gives all four original categories per axis throughout $N_* \le n \le N$, with their same calibrated errors and $\rho > 2R$. Every row obeys $\sum_j V_{j,n} \le S$, the complete compatible force cap $\|E_{j,n}\| \le \nu$, and

$$0 \le D_j := \|\theta_j\|^2 - q_j^2 \le 6. \tag{SB27}$$

These estimates are (BAJ17)–(BAJ19) in that appendix; their proofs do not use the comparable-signal condition (25). They include each candidate and recomputed endpoint force.

At a selected profile, balance and BAJ19 give $t^2 := q^2/\|\theta\|^2 \ge 1 - 6/(4R^2) \ge 125/128$, hence $2t/(1+t^2) \ge 1 - 10^{-4}$. Normalizing the actual and ideal augmented hidden profiles costs at most $2e_*$; the central correlation on the unit augmented sphere is 1-Lipschitz by HK6. The balanced ideal profile has normalized signal exactly K_i^*. It follows that

$$Q_{j,N_*} \ge (1 - 10^{-4})K_i^* - e_* \ge \{.9999(5/6) - 1/40960\}K_* > .833K_*. \tag{SB28}$$

This follows from $q^2/\|\theta\|^2 \geq 125/128$ and $2t/(1+t^2) \geq 1-10^{-4}$, not an assumed balanced head. The exact quotient debit $Q^+ \geq Q - h\nu^2$ and (30)–(31) keep every selected $Q_j > .832K_*$. The exact energy lower multiplier $V^+ \geq [1+h(4Q-2\nu)]V$ then gives

$$V_j(N_* + l) \geq v_* e^{3.32K_* lh}. \tag{SB29}$$

Indeed its multiplier is at least $1 + 3.3278hK_*$, and $hK_* \leq 1/40000$ pays the logarithmic quadratic debit.

Mark each original row at its first suffix checkpoint $Q_j \geq .826K_*$. Labels do not affect training. Thereafter $Q_j \geq .825K_*$, so SB4–6 imply $B_j < 29/20$. Its head and bias signs persist: upper balance alone, the component bounds used in SB18, and $\nu \leq \kappa = .825K_*$ give the same sign margins. Before marking the exact energy upper bound gives, including the first marking candidate,

$$\log(V_j^+/V_j) \leq 4hQ_j + 2h\nu + 4h^2 < 3.31hK_*. \tag{SB30}$$

Every marked row has increasing V_j. Once $V_j \geq 480$, SB27 implies

$$\frac{q_j^2}{\|\theta_j\|^2} = \frac{V_j - D_j}{V_j + D_j} \geq \frac{474}{486} > \frac{20}{21}. \tag{SB31}$$

Hence SB10 holds with $a = 29/20$ and $\varepsilon = \Delta = 1/20$. SB18 applies at every ensuing pre-state, including candidate exits, and its optional contraction can be dropped. If marking occurs below 480, the first candidate above 480 has $V^+ < 481$ and $Z^+ < 400$. If marking occurs above 480, its entry defect is at most $3V/4$. Before any such threshold crossing, $Z < 360$. These cases, never-marked rows and zero rows exhaust the bank. Using $X \leq Z \leq 3X/2$ on marked states gives

$$\begin{aligned} \Xi(N_* + l) &:= \sum_j X_j(N_* + l) \\ &\leq (3S_0/4 + 400m)e^{3.31K_* lh} + (5000h + 20\nu)TS, \\ \Xi_N/v_G &\leq \Gamma e^{-.01K_* t_G} + (5000h + 20\nu)TS/v_* \leq c_G. \end{aligned} \tag{SB32}$$

Before marking X_j uses its instantaneous orientation; always $X_j \leq V_j/2$. Thus no initial or later sign event has been assumed for an adverse row.

J THE COMPLETE SYMMETRIC DERIVATIVE FRAME

Here $X \sim N(0, I_d)$, $U = (u_1, \ldots, u_r)$, $P_U = UU^T$, $P_{U^\perp} = I_d - P_U$, $J = 1 - \alpha$, and $\sigma_\alpha(t) = \alpha t + Jt_+$. Use normalized rows $(A_j, W_j, B_j) = s(q_j, w_j, b_j)$, $\rho_j = \|w_j\|$, and the complete student $f = (s^2/m)\sum_j q_j \sigma_\alpha(w_j^T X + b_j)$. The density φ and scalar G below are standard normal; $\nabla_x f$ differentiates the input.

For any orientation $\tau \in \{-1, 1\}$ define $\mathcal{E}_\tau(w) = \|P_{U^\perp} w\|^2 + \sum_i \min\{\tau u_i^T w, 0\}^2$ and $\Xi = \sum_j |q_j| \sqrt{\mathcal{E}_{\tau_j}(w_j)}$, where $\tau_j = -\operatorname{sign}(q_j b_j)$ when $q_j b_j \neq 0$ and either orientation otherwise. This is the same defect as SB9.

J.1 A SYMMETRIC SIGNED FRAME FROM FIXED QUADRATIC TESTS

This subsection is deterministic at any one actual bank, with no corridor hypothesis. Put $Z = U^T X$, $e = r^{-1/2}(1, \ldots, 1)^T$, and use the r centered Gaussian quadratic tests

$$\xi_i(X) = (e^T Z)Z_i - (e)_i, \qquad T := \mathbb{E}[\xi\xi^T] = I_r + ee^T, \qquad I_r \preceq T \preceq 2I_r. \tag{BF21}$$

In the profile comparisons below e_i denotes the ith coordinate basis vector, whereas $(e)_i$ is the scalar component in (BF21). For each row with $\rho_j > 0$ define

$$x_j = \tau_j U^T w_j/\rho_j, \quad z_j = b_j/\rho_j, \qquad \omega_j = \frac{Js^2}{m}|q_j \rho_j z_j|\varphi(z_j) \geq 0. \tag{BF22}$$

If $q_j b_j = 0$, set $\omega_j = 0$ and choose either orientation. If $\rho_j = 0$, the spatial gradient contribution is zero and all formulas below assign the row zero contribution. The *complete* current-head frame is exactly symmetric:

$$D := \mathbb{E}[(U^T \nabla_x f)\xi^T] = \sum_{j=1}^m \omega_j (e^T x_j) x_j x_j^T. \tag{BF23}$$

Indeed Gaussian regression onto $w_j^T X/\rho_j$ and $\mathbb{E}[\mathbf{1}_{G>-z}(G^2-1)] = -z\varphi(z)$ give

$$\mathbb{E}[\sigma'_\alpha(w_j^T X + b_j)\xi_i] = -Jz_j\varphi(z_j)(U^T w_j/\rho_j)_i e^T(U^T w_j/\rho_j).$$

Multiplication by the actual gradient row $(s^2/m)q_j w_j$ proves BF23; the constant derivative α has zero correlation with each centered quadratic test. It remains present in the full AGOP and in its outside estimate below.

Split BF23 by the sign of $e^T x_j$ as $D = D_+ - D_-$, where both D_+, D_- are positive semidefinite. Every signed original row is in exactly one sum, with zero summands harmless. Then

$$\|D_-\|_{\mathrm{op}} \le \frac{J\varphi(1)s^2}{m}\,\Xi, \qquad \Xi = \sum_j |q_j|\sqrt{\mathcal{E}_{\tau_j}(w_j)}. \tag{BF24}$$

To see this, $(-e^T x)_+ \le \|x_-\|$, $\|x\| \le 1$, and $|z|\varphi(z) \le \varphi(1)$. Thus each negative summand has operator norm at most $(J\varphi(1)s^2/m)|q_j|\rho_j\|(x_j)_-\|$, which is bounded by its BF24 debit. No angular separation, common head sign or axis assignment of the other rows is needed.

For an explicit chosen-profile interface, suppose there are g disjoint groups of r actual rows, one row for every i in each group, obeying

$$\|x_{\ell i} - e_i\| \le \epsilon, \quad 0 \le \epsilon < r^{-1/2}, \qquad \omega_{\ell i} \ge \omega_0 > 0, \quad 1 \le \ell \le g. \tag{BF25}$$

If the scalar on the right is positive, then

$$\lambda_{\min}(D) \ge \Lambda := g\omega_0(r^{-1/2} - \epsilon)(1 - \sqrt{r}\epsilon)^2 - \frac{J\varphi(1)s^2}{m}\Xi > 0. \tag{BF26}$$

For each group the column matrix of its $x_{\ell i}$ differs from I_r in operator norm by at most $\sqrt{r}\epsilon$. All its $e^T x_{\ell i}$ are at least $r^{-1/2} - \epsilon$. Consequently its positive semidefinite contribution has least eigenvalue at least $\omega_0(r^{-1/2} - \epsilon)(1 - \sqrt{r}\epsilon)^2$. All other positive summands only increase D_+, and BF24 pays all negative ones. This proves BF26 for the full bank.

J.2 Full AGOP, outside gradient, and a conditional spectral join

Let $\mathsf{G} = \mathbb{E}[\nabla f \nabla f^T]$ for the actual current heads. The exact outside derivative is

$$P_{U^\perp}\nabla f = \frac{s^2}{m}\sum_j q_j P_{U^\perp} w_j \sigma'_\alpha(w_j^T X + b_j).$$

The L^2 triangle inequality, retaining all cross terms, gives

$$\mathrm{tr}(P_{U^\perp}\mathsf{G}P_{U^\perp}) \le \beta := \frac{s^4}{m^2}\Xi^2. \tag{BF27}$$

This is a full outside-gradient estimate, not just a spatial-bank score. Orthogonalizing the fixed tests in BF21 yields, by Bessel,

$$U^T\mathsf{G}U \succeq DT^{-1}D^T \succeq \tfrac{1}{2}DD^T. \tag{BF28}$$

Hence BF26 implies $\lambda_r(\mathsf{G}) \ge a := \Lambda^2/2$. If $r < d$, minmax and BF27 give $\lambda_{r+1}(\mathsf{G}) \le \beta$. When $\beta < a$, the genuine top-r projector P is unique and

$$\lambda_r - \lambda_{r+1} \ge a - \beta > 0, \qquad \lambda_{\min}(U^T P U) \ge 1 - \beta/a, \qquad r^{-1}\,\mathrm{tr}(U^T P U) \ge 1 - \beta/(ra). \tag{BF29}$$

Indeed $P \preceq \mathsf{G}/a$, so outside trace and operator norm of P are bounded by β/a; the equal-rank principal-angle identities give BF29. All cross blocks remain in the positive semidefinite matrix inequality. For $r = d$ the projector is I_d and outside angle assertions are automatic; no omitted eigenvalue is required.

K A BOUNDED READOUT FOR THE FULL MIXTURE LINKS

K.1 THE FULL MIXTURE AND ITS EXACT SCALAR PROJECTION

Let φ_v be the $N(0,v)$ density and $\varphi=\varphi_1$. Here $Z\sim N(0,1)$, Φ is its CDF, $h_k=\mathrm{He}_k/\sqrt{k!}$, and scalar inner products and norms are in $L^2(N(0,1))$. The activation is $\sigma_\alpha(t)=\alpha t+(1-\alpha)t_+$, $0\le\alpha<1$. For any probability measure μ on $[1/3,1]$, define

$$g_v\varphi=-\varphi_v''', \qquad g_\mu=\int g_v\,d\mu(v), \quad N_\mu=\|g_\mu\|_2, \quad y_\mu=g_\mu/N_\mu,$$
$$g_v(t)=v^{-7/2}(t^3-3vt)e^{-(v^{-1}-1)t^2/2}. \tag{MP1}$$

These are the full links in (24). They are odd, have zero first Gaussian moment, and belong to Gaussian H^1. For completeness their coefficients and exact norm kernel are

$$\langle g_\mu,h_{2j+3}\rangle=\frac{(-1)^j\sqrt{(2j+3)!}}{2^j\,j!}\int(1-v)^j\,d\mu(v),$$
$$K(t)=\frac{3(2+3t)}{(1-t)^{7/2}}, \qquad N_\mu^2=\iint K((1-v)(1-w))\,d\mu(v)d\mu(w),$$
$$N_v^2=\frac{3\{2+3(1-v)^2\}}{[v(2-v)]^{7/2}}. \tag{MP2}$$

The generating function $\mathbb{E}[g_v(Z)e^{tZ-t^2/2}]=t^3e^{-(1-v)t^2/2}$ gives the coefficients. Squaring and summing the absolutely convergent coefficient series gives MP2. The same series weighted by Hermite degree is summable uniformly on this interval, so the full H^1 claim and integration in μ are justified. No truncation is used.

For $B=b^2>0$, set

$$p_b=2\Phi(b)-1, \qquad q_b(z)=\mathrm{clip}(z,-b,b)-p_bz,$$
$$V_b=\|q_b\|_2^2=p_b-2b\varphi(b)+B(1-p_b)-p_b^2,$$
$$L_\mu(b)=2b\int v^{-3/2}\varphi(b/\sqrt{v})\,d\mu(v)>0. \tag{MP3}$$

The function q_b is odd and orthogonal to Z. Integrating MP1 by parts against the clipped linear function gives the exact full-link identity

$$\langle g_\mu,q_b\rangle=-L_\mu(b), \qquad E_b(y_\mu)=\frac{L_\mu(b)^2}{N_\mu^2V_b}. \tag{MP4}$$

Here E_b is the full nonconstant projection energy in the four-profile span $\mathrm{span}\{\sigma_\alpha(\epsilon z+\tau b):\epsilon,\tau\in\{-1,1\}\}$. Directly, that span is generated by $1,Z,q_b$ and the centered even function $\max(|Z|,b)-\mathbb{E}[\max(|Z|,b)]$. Oddness and zero first moment leave only q_b for MP1. In particular the unexplained infinite tail has exactly squared norm $1-E_b(y_\mu)$; it is not dropped.

Uniform full-link prediction lemma. For every such μ and every

$$B\in[B_-,B_+], \qquad B_-=(\sqrt{7/3}-1)/2, \quad B_+=(\sqrt{5}-1)/2, \tag{MP5}$$

one has $E_b(y_\mu)>2/5$. This includes every link-dependent stationary bias supplied by the mixture acquisition interface, without using its stationarity equation.

Proof. Write $L_v=2bv^{-3/2}\varphi(b/\sqrt{v})$. Minkowski's inequality and positivity imply

$$\frac{\int L_v\,d\mu}{N_\mu}\ge\frac{\int L_v\,d\mu}{\int N_v\,d\mu}\ge\inf_{v\in[1/3,1]}\frac{L_v}{N_v}. \tag{MP6}$$

For fixed B, the logarithm of L_v^2/N_v^2, up to a constant, is

$$f_B(v)=\tfrac12\log v+\tfrac72\log(2-v)-B/v-\log\{2+3(1-v)^2\}.$$

Its second derivative is

$$f''_B(v) = -\frac{1}{2v^2} - \frac{7}{2(2-v)^2} - \frac{2B}{v^3} + \frac{-12 + 18(1-v)^2}{\{2 + 3(1-v)^2\}^2} < 0. \tag{MP7}$$

Thus its minimum over the interval is at one of the two endpoints. The explicit rational scalar certificate proved below gives

$$\begin{gathered} 39/1000 < V_b < 13/250, \qquad p_b < 57/100, \\ Be^{-B} > 1/5, \qquad Be^{-3B} > 967/10000, \qquad N_{1/3}^2 < 79, \qquad \pi < 22/7. \end{gathered} \tag{MP8}$$

Since $N_1^2 = 6$, the two endpoint ratios obey

$$\begin{aligned} \frac{L_1^2}{N_1^2 V_b} &= \frac{Be^{-B}}{3\pi V_b} > \frac{175}{429} > \frac{2}{5}, \\ \frac{L_{1/3}^2}{N_{1/3}^2 V_b} &= \frac{54Be^{-3B}}{\pi N_{1/3}^2 V_b} > \frac{182763}{451880} > \frac{2}{5}. \end{aligned} \tag{MP9}$$

Equations MP6–9 prove the claim for every positive measure, including non-atomic measures. They use the full norm MP2 and full correlation MP4.

K.2 A FINITE EXACT-RATIONAL CERTIFICATE FOR MP8

The following gives a finite rational certificate. For $n \geq 0$ define

$$T_n(x) = \sum_{j=0}^{n} \frac{(-x)^j}{j!}, \qquad I_n(b) = \sum_{j=0}^{n} \frac{(-1)^j b^{2j+1}}{2^j j!(2j+1)}.$$

Taylor's theorem, followed by integration, gives $T_{2k+1}(x) \leq e^{-x} \leq T_{2k}(x)$ for $x \geq 0$ and $I_{2k+1}(b) \leq \int_0^b e^{-t^2/2} dt \leq I_{2k}(b)$. Put

$$c_- = 3989422804/10^{10}, \qquad c_+ = 3989422805/10^{10}.$$

The alternating series for $\pi = 16\arctan(1/5) - 4\arctan(1/239)$, through the first sixteen terms and with the next term as remainder bound, verifies $c_- < (2\pi)^{-1/2} < c_+$ by squaring positive rational endpoints. Define rational enclosures

$$P_-(b) = 2c_- I_9(b), \quad P_+(b) = 2c_+ I_{10}(b), \quad F_-(b) = c_- T_9(b^2/2), \quad F_+(b) = c_+ T_{10}(b^2/2).$$

They bound p_b and $\varphi(b)$ on $[.51, .79]$. Since p_b increases and both p_b/b and $\varphi(b)$ decrease,

$$\frac{dV_b}{dB} = 1 - p_b - 2(p_b/b)\varphi(b) \geq 1 - P_+((k+1)/100) - \frac{200}{k} P_+(k/100) F_+(k/100) > 3/1000 \tag{MP10}$$

whenever $b \in [k/100, (k+1)/100]$, for each of the twenty-eight integers $k = 51, \ldots, 78$. The last inequality is checked by substitution of the displayed rational polynomials. Therefore V_b is increasing on $[.51, .79]$. At a rational $b < 1$ use

$$\begin{aligned} b^2 + (1-b^2)P_-(b) - P_+(b)^2 - 2bF_+(b) &\leq V_b, \\ V_b &\leq b^2 + (1-b^2)P_+(b) - P_-(b)^2 - 2bF_-(b). \end{aligned} \tag{MP11}$$

At $b = .51$ the lower expression exceeds $39/1000$, and at $b = .787$ the upper expression is less than $13/250$; also $P_+(.787) < 57/100$. These are three further rational checks.

The exact MP5 interval is contained in $[B_l, B_u] = [2637/10000, 6181/10000]$, itself contained in $(.51^2, .787^2)$. This follows by substituting the endpoints into the increasing polynomial $B + B^2$. On this interval Be^{-B} is increasing and Be^{-3B} has only one critical point, a maximum. The three rational checks

$$B_l T_{13}(B_l) > 1/5, \qquad B_l T_{13}(3B_l) > 967/10000, \qquad B_u T_{13}(3B_u) > 967/10000$$

prove the exponential parts of MP8. Finally $N_{1/3}^2 = 10(9/5)^{7/2} < 79$ follows from $100(9/5)^7 < 79^2$. This proves every scalar inequality used above by a finite, reproducible exact calculation.

K.3 General additive rank and scale-independent paid budgets

Let $X \sim N(0, I_d)$ and $U = (u_1, \dots, u_r)$ have orthonormal columns, choose arbitrary probability measures μ_i on $[1/3, 1]$, and define the full central teacher

$$y_0(X) = \sum_{i=1}^{r} a_i y_{\mu_i}(u_i^T X), \qquad \sum_i a_i^2 = 1. \tag{MP12}$$

Signs are unrestricted. This teacher is unit norm, centered, and has Hermite rank at least three. If every $a_i \neq 0$, its true minimal linear index space is $\mathrm{span}(U)$: the nonzero degree-three tensor has contraction range equal to that space. Zero coefficients may instead be omitted from the definition of rank. Allow a full actual unit target

$$y = y_0 + e_y, \qquad \|y\|_2 = 1, \quad \|e_y\|_{H^1(\gamma_d)} \leq \zeta. \tag{MP13}$$

Every mean, lower Hermite term and interaction in e_y is retained and paid below. No stronger index identification for the perturbed target is inferred without an additional same-index assumption.

Choose any b_i^2 in MP5 and put

$$t_i = -\frac{L_{\mu_i}(b_i)}{N_{\mu_i} V_{b_i}}, \qquad F_0(X) = \sum_i a_i t_i q_{b_i}(u_i^T X),$$
$$E = \|F_0\|_2^2 = \sum_i a_i^2 E_{b_i}(y_{\mu_i}) > 2/5, \qquad \|y_0 - F_0\|_2^2 = 1 - E < 3/5. \tag{MP14}$$

Independence of the Gaussian coordinates makes the scalar projection energies add exactly. F_0 uses the four-profile coefficients

$$d_{i,\epsilon,\tau} = \frac{a_i t_i \epsilon}{2}\left(\frac{\tau}{J} - \frac{p_{b_i}}{1+\alpha}\right), \qquad J = 1 - \alpha > 0. \tag{MP15}$$

To verify the representation for every $b > 0$, set $D_\tau(z) = \sigma_\alpha(z + \tau b) - \sigma_\alpha(-z + \tau b)$. The two elementary identities are

$$D_+(z) + D_-(z) = 2(1+\alpha)z, \qquad D_+(z) - D_-(z) = 2J\,\mathrm{clip}(z, -b, b).$$

Substituting them into (MP15) gives $\sum_{\epsilon,\tau} d_{i,\epsilon,\tau}\sigma_\alpha(\epsilon z + \tau b_i) = a_i t_i q_{b_i}(z)$. Set $C_2 = \|d\|_2$, $C_1 = \|d\|_1$. Cauchy–Schwarz gives $|t_i| \leq V_{b_i}^{-1/2}$, so MP8 proves

$$C_2^2 = \sum_i a_i^2 t_i^2 \left(J^{-2} + \frac{p_{b_i}^2}{(1+\alpha)^2}\right) \leq \frac{1000}{39J^2}\left(1 + \frac{57^2}{100^2}\right) < \frac{36}{J^2},$$
$$C_1 \leq 2\sqrt{r}C_2 < 12\sqrt{r}/J. \tag{MP16}$$

Thus rank and leak are explicitly charged; there is no uniform bound as $\alpha \uparrow 1$.

Suppose at one actual common checkpoint N there are $4r$ distinct original rows with $\rho_j = \|W_{j,N}\| > 0$ and

$$\left\| \left(\frac{W_{j,N}}{\rho_j}, \frac{B_{j,N}}{\rho_j}\right) - (\epsilon u_i, \tau b_i) \right\| \leq \varepsilon, \qquad 0 \leq \varepsilon \leq 1/100. \tag{MP17}$$

The assertion MP17 must come from the full actual coupled trajectory; it is not implied by initial category counts or by aggregate alignment. Define, at both endpoints and using every original row,

$$A_{j,n} = \sqrt{\|W_{j,n}\|^2 + B_{j,n}^2}, \qquad \phi_{j,n}(X) = \frac{\sigma_\alpha(W_{j,n}^T X + B_{j,n})}{A_{j,n}},$$
$$\mathcal{R}(n) = \inf_{\substack{\|v\|_2 \leq 32/J \\ \|v\|_1 \leq 64\sqrt{r}/J}} \left\| y - \sum_{j=1}^{m} v_j \phi_{j,n} \right\|_2^2. \tag{MP18}$$

A zero augmented row is assigned the zero feature. The normalization, identity spatial projection, row set, and both budgets agree exactly at $n = 0, N$. Actual training heads remain untouched.

On the selected rows take $v_j = d_{i,\epsilon,\tau}\sqrt{1+(B_{j,N}/\rho_j)^2}$, and put zero diagnostic coefficients on other rows. Since $|B_{j,N}/\rho_j| \le .787 + .01$ and $\sqrt{1+(.787+.01)^2} < 1.3$, MP16 shows that both MP18 budgets hold. Positive homogeneity gives the exact radius-canceling identity

$$v_j\phi_{j,N} = d_{i,\epsilon,\tau}\sigma_\alpha((W_{j,N}/\rho_j)^T X + B_{j,N}/\rho_j).$$

The activation is one-Lipschitz. Gaussian affine second moments therefore show that its sum F_N obeys

$$\|F_N - F_0\|_2 \le C_1\varepsilon \le \frac{12\sqrt{r}}{J}\varepsilon =: \Delta, \qquad \mathcal{R}(N) \le (\sqrt{1-E}+\zeta+\Delta)^2. \tag{MP19}$$

This is an original-loss diagnostic retaining the full target MP13. The entire MP18 class has L^2 norm, Lipschitz norm and affine variation cost at most $64\sqrt{r}/J$. The witness has $\|F_N\|_2 \le \sqrt{E}+\Delta$ and Lipschitz norm at most C_1. None of these budgets depend on initialization scale or learned radii. No feature-Gram conditioning claim is required or implied.

L Acquisition of full-bank signal before loss release

Use the teacher in (24), writing $a_i = \lambda_i$. Its full central kernel C is defined in (BM5) and evaluated in (BM23) of Appendix H, and its normalized H^1 norm is less than 13. The residual is the full $e = y - y_c$. This section proves a deterministic implication from one actual signal row. In Theorem D.1 that premise is supplied by the preceding acquisition proof, with $(\gamma, U, v) = (K_*/2, S, v_*)$.

L.1 Actual simultaneous update and two complete-row inequalities

Here $X \sim N(0, I_d)$, $\theta_j = (w_j, b_j)$, $\sigma_\alpha(t) = \alpha t + (1-\alpha)t_+$, and $\sigma'_{\alpha,j} = \sigma'_\alpha(\theta_j^T(X,1))$ is the scalar gate. The scalar U used in the energy premise (MR11) is a public energy ceiling; teacher axes remain u_i.

Train the averaged network with the one unchanged full-MSE step

$$\begin{gathered} f_n = m^{-1}\sum_j A_{j,n}\sigma_\alpha(W_{j,n}^T X + B_{j,n}), \\ (A_j, W_j, B_j) = s(q_j, w_j, b_j), \quad L_n = \|y - f_n\|_2^2, \quad \eta = mh/2, \quad 0 < h \le 1/1024. \end{gathered} \tag{MR6}$$

These are analytic normalized coordinates only. The exact row update is

$$q^+ = q + h(C + \theta^T E), \quad \theta^+ = \theta + hq(T+E), \quad T = \nabla C, \quad E_j = \mathbb{E}[(e - f_n)\sigma'_{\alpha,j}(X,1)]. \tag{MR7}$$

The gate in the last expression multiplies the augmented vector. Write $z = (q, \theta)$, $H = qC$, $V = \|z\|^2$, $Q = H/V$, $F = \nabla H = (C, qT)$ and $p = (\theta^T E, qE)$, so $z^+ = z + h(F+p)$. The full-H1 kernel bounds give

$$|C| \le \|\theta\|, \quad \|T\| \le 1, \quad \|D^2 C\| \le 52/\|\theta\|, \quad |Q| \le \tfrac{1}{2}, \quad \|F\| \le \sqrt{V}. \tag{MR8}$$

The mean/linear-free central target has $C = T = 0$ at $w = 0, b \ne 0$; regularity there is included in Appendix F. Identically zero joint rows stay zero and can have Q defined as zero. They are still present in the network and all sums.

At any balanced nonzero row, $|q| \le \|\theta\|$, with $\|E\| \le \nu \le 1$, the exact update preserves the balance and nonzero hidden state. In fact, for $D = \|\theta\|^2 - q^2$,

$$D^+ \ge (1 - h^2\|T+E\|^2)D, \quad \|\theta^+ - \theta\| \le 2h\|\theta\|, \quad V^+ \le (1+2h)^2 V. \tag{MR9}$$

The two useful inequalities, valid with every possible row sign, are

$$Q^+ \ge Q - h\nu^2, \qquad \Delta(V - 4hH) \le 4hH + 3h\nu V. \tag{MR10}$$

Here is a proof independent of rank-one geometric invariants. One has $\|p\| \le \nu\sqrt{V}$ and $|z^T p| \le \nu V$. Along the actual joining segment the head/hidden ratio is less than two, so $\|D^2 H\| \le 105$. Taylor's formula and $F^T p \ge -\|F\|^2/4 - \nu^2 V$ yield

$$\Delta H \ge h(3\|F\|^2/4 - \nu^2 V) - 105h^2(\|F\|^2 + \nu^2 V) \ge h\|F\|^2/2 - 2h\nu^2 V.$$

The exact energy expansion has upper bound $\Delta V \le 4hH + 2h\nu V + 2h^2\|F\|^2 + 2h^2\nu^2 V$. Subtracting $4h\Delta H$ leaves at most $4hH + 2h\nu V + 10h^2\nu^2 V \le 4hH + 3h\nu V$.

For the normalized inequality put $a = z^T p/V$, $F_\perp = F - 2Qz$, $p_\perp = p - az$, $\tau = h/[1+h(2Q+a)]$. Homogeneity gives $Q(z^+) = Q(z + \tau(F_\perp + p_\perp))$ and $\tau \le h/(1-2h) \le 4h/3$. The latter segment is tangent to z, has norm at least $\sqrt{V}$, and its head/hidden ratio is less than two: each block moves by at most $4\tau\|\theta\|$. The quotient-Hessian bound from the full-H1 calculation is $\|D^2 Q\| \le 112/V$ along this segment. With $\beta = \|F_\perp\|/\sqrt{V} \le 1$,

$$\Delta Q \ge \tau(\beta^2 - \nu\beta) - 56\tau^2(\beta+\nu)^2 \ge \tau(7\beta^2/8 - \nu\beta - \nu^2/8) \ge -5\tau\nu^2/8 \ge -h\nu^2.$$

Here $112\tau \le 112/1022 < 1/8$ and $\nu\beta \le (\beta^2+\nu^2)/2$ suffice. No lower correlation detector or spatial radius appears in MR10.

L.2 A PUBLIC SUFFIX THAT PAYS ALL UNMARKED ENERGY

The precise acquisition obligation is this: on an explicitly specified event of the original draw, an actual integer N obeys

$$\begin{gathered}\max_{0\le n\le N} \sum_j V_{j,n} \le U, \qquad |q_{j,N}| \le \|\theta_{j,N}\| \quad (j \le m),\\ V_{j_*,N} \ge v > 0, \quad Q_{j_*,N} \ge \gamma, \quad 0 < \gamma \le \tfrac{1}{2},\end{gathered} \tag{MR11}$$

where $U \ge v$ and γ are public constants chosen before the draw. Nonzero hidden states are also part of entry, or follow from preceding balanced actual dynamics. The original independent Gaussian law is $(A_{j,0}, W_{j,0}, B_{j,0}) \sim N(0, s^2/d)$ entrywise; no row is screened, replaced or sign-flipped. The deterministic implication here assumes (MR11); Appendix H supplies the original-law acquisition event used in Theorem D.1.

Choose all following parameters before that original draw:

$$\begin{gathered}k = \left\lceil \frac{2}{h\gamma} \log \frac{32U}{\gamma v} \right\rceil, \quad T = h + \frac{2}{\gamma}\log\frac{32U}{\gamma v}, \quad \overline{U} = Ue^{4T},\\ 0 < \nu \le \min\left\{\frac{\gamma}{16}, \sqrt{\frac{\gamma}{8T}}\right\}, \quad \zeta \le \nu/2, \quad s^2 \le \frac{m\nu}{\overline{U}}.\end{gathered} \tag{MR12}$$

This adds k actual steps; $kh \le T$. The full force cap $\|E_j\| \le \nu$ holds at every suffix checkpoint, including its endpoint, by induction. Indeed

$$\|E_j\| \le \zeta + \|f\|_2 \le \zeta + \frac{s^2}{2m}\sum_l V_l, \qquad \sum_j V_{j,N+l} \le U(1+2h)^{2l} \le Ue^{4lh} \le \overline{U}. \tag{MR13}$$

The initial force follows from U and MR12; MR9 then proves each candidate energy bound, and recomputing the force closes the induction. Every head and hidden row remains trained and balanced.

Mark a row at its first suffix checkpoint with $Q \ge \gamma/4$, including the entry checkpoint. This is only an analytic label. By MR10 and $\nu^2 kh \le \gamma/8$, every marked endpoint has $Q \ge \gamma/8$. The original signal row obeys throughout the suffix $Q_{j_*} \ge 7\gamma/8 \ge 3\gamma/4$. Its exact energy increment gives

$$V_{j_*}^+ \ge [1 + h(4Q_{j_*} - 2\nu)]V_{j_*} \ge (1 + 23h\gamma/8)V_{j_*}, \qquad V_{j_*,N+l} \ge ve^{2\gamma lh}. \tag{MR14}$$

For the last inequality, if $x = h\gamma \le 1/2048$, then $\log(1+23x/8) \ge (23x/8)/(1+23x/8) \ge 2x$.

A row still unmarked at the final checkpoint has $Q < \gamma/4$ at every preceding suffix checkpoint, regardless of its sign. Let $Z = V - 4hH = (1-4hQ)V$. MR10 gives

$$Z^+ \le \left(1 + h\frac{\gamma+3\nu}{1-h\gamma}\right) Z, \qquad V_{j,N+l} \le \frac{1+2h}{1-h\gamma}\exp\left(\frac{\gamma+3\nu}{1-h\gamma} lh\right) V_{j,N} \le 2e^{3\gamma lh/2} V_{j,N}. \tag{MR15}$$

Both scalar constants use $h \le 1/1024$, $\gamma \le 1/2$, $\nu \le \gamma/16$. Summing every still-unmarked row and comparing with the genuinely growing signal row proves, at $M = N + k$,

$$p_{\mathrm{un}} := \frac{\sum_{j \text{ unmarked}} V_{j,M}}{\sum_j V_{j,M}} \le \frac{2U}{v} e^{-\gamma kh/2} \le \frac{\gamma}{16}. \tag{MR16}$$

Rows crossing the threshold join the positive marked set; MR15 is needed only for those that never cross. No assertion that the selected row dominates other marked rows is made or needed.

All signs now have an explicit payment. Since every unmarked row has $Q \geq -1/2$, while every marked row has $Q \geq \gamma/8$,

$$\frac{\sum_j H_{j,M}}{\sum_j V_{j,M}} \geq \frac{\gamma}{8}(1 - p_{\mathrm{un}}) - \frac{1}{2}p_{\mathrm{un}} \geq \frac{\gamma}{16}. \tag{MR17}$$

The last inequality follows from $\gamma \leq 1/2$ and MR16; indeed the preceding lower bound is at least $23\gamma/256$. This is the complete original bank's energy-weighted central signal.

L.3 Full original loss, unchanged rate, and explicit gain

Write $\widehat{f} = m^{-1}\sum_j q_j \sigma_\alpha(\theta_j^T(X,1))$, so $f = s^2\widehat{f}$. The compatibility in MR7 implies exactly

$$\mathbb{E}[e\widehat{f}] = m^{-1}\sum_j q_j \theta_j^T E_j + s^2\|\widehat{f}\|_2^2. \tag{MR18}$$

Thus no retained teacher mean, linear term or other tail is dropped. At M, MR17 and the complete force cap give, for the raw joint radius $\mathfrak{R}_M = s\sqrt{\sum_j V_{j,M}}$,

$$Q_{\mathrm{raw}}(M) := \frac{\mathbb{E}[yf_M]}{\mathfrak{R}_M^2} \geq \frac{1}{m}\left(\frac{\gamma}{16} - \frac{\nu}{2}\right) \geq \frac{\gamma}{32m}. \tag{MR19}$$

Choose publicly

$$a_{\mathrm{sig}} = \frac{\gamma}{64m}, \quad R_{\mathrm{sig}}^2 = a_{\mathrm{sig}}m^2 = \frac{\gamma m}{64}, \quad G_{\mathrm{sig}} = \frac{a_{\mathrm{sig}}R_{\mathrm{sig}}^2}{2} = \frac{\gamma^2}{8192},$$
$$s^2 \leq \min\left\{\frac{R_{\mathrm{sig}}^2}{4\overline{U}}, \frac{mG_{\mathrm{sig}}}{\overline{U}}\right\}. \tag{MR20}$$

These are further original scale restrictions; no modification occurs at N or M. They give $\mathfrak{R}_M \leq R_{\mathrm{sig}}/2$ and

$$\|f_n\|_2 \leq G_{\mathrm{sig}}/2, \qquad L_n \geq 1 - G_{\mathrm{sig}} \qquad (0 \leq n \leq M). \tag{MR21}$$

The same radius bound covers the prior acquisition window via MR11. The full normalized teacher has $\|y\|_{H^1} < 14$. In Lemma M.2, take network factor $\kappa = 1/m$, half-MSE step $\eta_{1/2} = mh = 2\eta$, and parameter $a = a_{\mathrm{sig}}$. Its entry and step conditions hold because

$$Q_{\mathrm{raw}}(M) \geq 2a_{\mathrm{sig}}, \quad a_{\mathrm{sig}} \leq 1/(4m), \quad R_{\mathrm{sig}}^2 = a_{\mathrm{sig}}/\kappa^2, \quad h \leq 1/1024 < [32(1 + J\|y\|_{H^1})]^{-1}. \tag{MR22}$$

The same actual coupled GD therefore has a finite first later radius hit $K > M$, without resetting or refitting anything, such that

$$L_K \leq 1 - 2G_{\mathrm{sig}}, \qquad L_n - L_K \geq G_{\mathrm{sig}} \quad (0 \leq n \leq M). \tag{MR23}$$

Its additional full-MSE physical clock, after the explicitly paid suffix $\eta k = mkh/2$, is

$$\frac{m}{4}\log\frac{R_{\mathrm{sig}}}{\mathfrak{R}_M} \leq \eta(K - M) \leq \eta + a_{\mathrm{sig}}^{-1}\log\frac{R_{\mathrm{sig}}}{\mathfrak{R}_M}. \tag{MR24}$$

Also $R_{\mathrm{sig}} \leq \mathfrak{R}_K < \sqrt{2}R_{\mathrm{sig}}$, the joint displacement from M is at least $R_{\mathrm{sig}}/2$, and

$$\frac{\gamma}{64}R_{\mathrm{sig}} \leq \|(A_{j,K})_j\|_2 < R_{\mathrm{sig}}. \tag{MR25}$$

No retention of a spectral subspace, alignment or refit accuracy is asserted during either continuation.

M Same-step release of the original population risk

Throughout this section $X \sim \gamma_d = N(0, I_d)$, $y \in H^1(\gamma_d)$, $\|y\|_2 = 1$, $\mathcal{H} = \|y\|_{H^1}$, $J = 1 - \alpha > 0$, and $\sigma(t) = \alpha t + Jt_+$. Function norms are Gaussian, with $\|g\|_{H^1}^2 = \mathbb{E}[g(X)^2 + \|\nabla g(X)\|^2]$; φ is the standard normal density. Parameter derivatives below act on the explicitly displayed raw coordinates. The full teacher is retained in every expectation.

M.1 Raw coordinates, full teacher, and exact row balance

For any finite m and fixed network factor $\kappa > 0$ use

$$f_x(X) = \kappa \sum_{j=1}^{m} q_j \sigma(\theta_j^T \widetilde{X}), \quad \theta_j = (w_j, b_j), \quad \widetilde{X} = (X, 1), \quad x = (q_j, \theta_j)_{j=1}^m. \tag{48}$$

Every displayed coordinate takes the same raw Euclidean step

$$x^+ = x - \eta \nabla \ell(x), \qquad \ell(x) = \tfrac{1}{2}\mathbb{E}[(y - f_x)^2], \qquad L = 2\ell. \tag{49}$$

The averaged model has $\kappa = 1/m$; this factor is not a change of coordinates. The identical update has full-MSE raw rate $\eta/2$, and its full-MSE physical clock is $\eta n/2$.

Set $K(\theta) = \mathbb{E}[y\sigma(\theta^T \widetilde{X})]$ and

$$R = \|x\|, \quad F = \mathbb{E}[y f_x], \quad V = \|f_x\|_2^2, \quad Q = F/R^2 \ (R > 0), \quad L = 1 - 2F + V. \tag{50}$$

F and V have homogeneous degrees two and four respectively; V is the energy of the whole interacting bank. With

$$g_j = \kappa \mathbb{E}[(f_x - y)\sigma'(\theta_j^T \widetilde{X})\widetilde{X}], \qquad D_j = \|\theta_j\|^2 - q_j^2,$$

the exact coupled update and balance identity are

$$\begin{aligned} q_j^+ &= q_j - \eta \theta_j^T g_j, & \theta_j^+ &= \theta_j - \eta q_j g_j, \\ D_j^+ &= D_j + \eta^2\{q_j^2 \|g_j\|^2 - (\theta_j^T g_j)^2\} \geq (1 - \eta^2 \|g_j\|^2) D_j, & \|g_j\| &\leq \kappa\sqrt{L}. \end{aligned} \tag{51}$$

These follow by Euler homogeneity and Cauchy–Schwarz, since $\mathbb{E}[(h^T \widetilde{X})^2] = 1$ for a unit augmented vector h. Thus $D_j \geq 0$ is transported through any preceding actual loss band with $\eta\kappa\sqrt{L_n} \leq 1$.

For ordinary independent blocks $\theta_{j,0} \sim N(0, s^2 I_{d+1})$, $q_{j,0} \sim N(0, s_q^2)$, $s_q \leq s$, all initial balances hold with probability at least $1 - 3me^{-(d+1)/8}$. Indeed each failure is contained in $\{|q_j| > s\sqrt{d+1}/2\} \cup \{\|\theta_j\| < s\sqrt{d+1}/2\}$; the two costs are at most $2e^{-(d+1)/8}$ and $\exp[-(d+1)(\log 4 - 3/4)/2]$. This is a sufficient event of the original law, not a rejection or conditioning instruction. More general public scales may use a direct balance check on their proved Gaussian event.

M.2 Dimension-free H1 trace and normalized-correlation bounds

Lemma M.1 (Full-teacher kink regularity)**.** *The function K is C^2 on $\theta \neq 0$ and satisfies*

$$|K(\theta)| \leq \|\theta\|, \qquad \|\nabla K(\theta)\| \leq 1, \qquad \|\theta\|\,\|D^2 K(\theta)\| \leq B, \quad B := 4J\mathcal{H}. \tag{52}$$

Consequently, globally where derivatives are evaluated,

$$|F| \leq \kappa R^2/2, \quad \|\nabla F\| \leq \kappa R, \quad \|f_x\|_2 \leq \kappa R^2/2, \quad V \leq \kappa^2 R^4/4, \quad \|\nabla V\| \leq \kappa^2 R^3, \tag{53}$$

$$\|\nabla Q\| \leq \kappa/R, \qquad \|D^2 Q\| \leq M\kappa/R^2, \quad M := 8 + 2B, \tag{54}$$

where the Hessian assertion applies on any region with $|q_j| \leq 2\|\theta_j\|$ for every nonzero joint row. An identically zero joint row is omitted and stays zero in GD.

Proof. Here is a direct proof of the H1 trace interface, without a growth substitute. For a Hilbert-valued function $g \in H^1(\gamma_1)$ put $u(t) = \sqrt{\varphi(t)}g(t)$. For smooth functions, integration by parts and then density give

$$\|u'\|_{L^2(dt)}^2 + \tfrac{1}{4}\|tu\|_{L^2(dt)}^2 = \|g'\|_{L^2(\gamma_1)}^2 + \tfrac{1}{2}\|g\|_{L^2(\gamma_1)}^2.$$

The one-dimensional Sobolev inequality $\sup_t \|u(t)\|^2 \leq \|u\|_2 \|u'\|_2$ yields

$$\|g(t)\| \leq 2^{-1/4}\varphi(t)^{-1/2}\|g\|_{H^1(\gamma_1)}.$$

Apply this with Hilbert space $L^2(\gamma_{d-1})$ and the slicing $g(t) = y(tv + X_\perp)$ for a unit v. For $w = \rho v \neq 0$, $z = b/\rho$, Gaussian hyperplane differentiation gives

$$D^2K(\theta) = J\frac{\varphi(z)}{\rho}\mathbb{E}[y\widetilde{X}\widetilde{X}^T \mid v^TX = -z].$$

For unit augmented h, the conditional fourth moment of $h^T\widetilde{X}$ is at most $3(1+z^2)^2$. Hence

$$\|\theta\|\|D^2K(\theta)\| \leq J\mathcal{H}\sqrt{3}\,2^{-1/4}(1+z^2)^{3/2}\sqrt{\varphi(z)} < 4J\mathcal{H}.$$

The maximum in the last line occurs at $z^2 = 5$ and is $\sqrt{3}\,2^{-1/4}(2\pi)^{-1/4}6^{3/2}e^{-5/4} < 4$. The first two estimates follow from testing against affine Gaussian forms and Cauchy–Schwarz.

For completeness, approximate y in Gaussian H1 by Hermite polynomials. Their correlations are C^2 away from $\theta = 0$, including $w = 0, b \neq 0$ by Gaussian exponential decay. Applied to differences, the same three bounds give uniform convergence of functions, gradients and Hessians on compact sets away from zero. This proves the asserted C^2 regularity of the full H1 correlation. If $\mu = \mathbb{E}[y]$ and $v_1 = \mathbb{E}[yX]$, the extension is exactly

$$K(0,b) = \mu\sigma(b), \qquad \nabla K(0,b) = \sigma'(b)(v_1,\mu), \qquad D^2K(0,b) = 0 \quad (b \neq 0).$$

Thus the teacher mean and the kink trace are both retained.

The bounds on F, f, V follow from $\sum_j |q_j|\|\theta_j\| \leq R^2/2$ and $\|Df_x\|_{\mathrm{op}} \leq \kappa R$. Euler's identity gives $x^T\nabla F = 2F$, hence $R^4\|\nabla Q\|^2 = \|\nabla F\|^2 - 4F^2/R^2$. The joint-row Hessian blocks of F are

$$\kappa\begin{pmatrix} 0 & \nabla K^T \\ \nabla K & qD^2K \end{pmatrix},$$

with norm at most $(1+2B)\kappa$ on the stated region. Differentiating $Q = FR^{-2}$ adds at most $4\kappa/R^2 + 3\kappa/R^2$, proving $M = 8 + 2B$. □

M.3 A STRONGER SAME-STEP DETERMINISTIC RELEASE

Lemma M.2 (Positive balanced H1 entry releases original risk)**.** *At an actual integer checkpoint suppose*

$$R_0 > 0, \qquad |q_{j,0}| \leq \|\theta_{j,0}\| \ (j \leq m), \qquad Q(x_0) \geq 2a > 0.$$

Choose public constants, before reducing the original hidden scale, such that

$$0 < a \leq \kappa/4, \qquad 0 < \eta\kappa \leq \frac{1}{4M} = \frac{1}{32(1+J\mathcal{H})}, \qquad 0 < R_*^2 \leq \frac{a}{\kappa^2}, \qquad R_0 < R_*. \tag{55}$$

Continue the identical simultaneous raw GD. At its finite first crossing N of R_*,

$$R_* \leq R_N < \sqrt{2}R_*, \qquad Q(x_n) \geq 29a/16 \quad (0 \leq n \leq N), \qquad |q_{j,n}| \leq \|\theta_{j,n}\|, \tag{56}$$

$$L_N \leq 1 - aR_*^2, \qquad 1 - 2\kappa R_*^2 \leq L_n \leq 1 \quad (0 \leq n \leq N), \qquad \frac{\max_{n\leq N} L_n}{\min_{n\leq N} L_n} \leq \frac{1}{1-2\kappa R_*^2}, \tag{57}$$

$$\frac{1}{2\kappa}\log\frac{R_*}{R_0} \leq \eta N \leq \eta + \frac{2}{a}\log\frac{R_*}{R_0}. \tag{58}$$

If a preceding checkpoint t of this same run has $L_t \geq 1 - \epsilon$ with $\epsilon \leq aR_*^2/2$, then

$$L_t - L_N \geq aR_*^2/2. \tag{59}$$

At the maximal permitted radius this gain is $a^2/(2\kappa^2)$. There is no head refit, reset, selected sub-bank, or new random event.

Proof. Use a first-exit induction with $R < R_*$, $Q \geq a$, and the row balances. Since $\kappa R^2 \leq a/\kappa \leq 1/4$,

$$\|g_j\| \leq \kappa(1 + \kappa R^2/2) < 2\kappa, \quad R^+ \leq (1+2\eta\kappa)R.$$

The balances propagate by (51); no nonzero hidden row can hit zero. Moreover Euler's identities give exactly

$$(R^+)^2 - R^2 = 4\eta(F - V) + \eta^2\|\nabla F - \tfrac{1}{2}\nabla V\|^2 \geq 2\eta a R^2, \tag{60}$$

because $V/R^2 \le \kappa^2 R^2/4 \le a/4$.

The improved estimate preserves the positive teacher angular work. Let $P = I - xx^T/R^2$, $G = \|\nabla Q\|$, and $v = \frac{1}{2}P\nabla V$, so $\|v\| \le \kappa^2 R^3/2$. Set

$$\lambda = 1 + 2\eta(Q - V/R^2) \ge 1, \qquad h = \eta/\lambda, \qquad e = h(R^2\nabla Q - v).$$

The actual step is $x^+ = \lambda(x+e)$ and $e \perp x$, so $Q(x^+) = Q(x+e)$ and the joining tangent segment has norm at least R. Rowwise, for either a head or a hidden block, the movement of this segment is at most

$$\eta(2\kappa + \kappa^2 R^2)\|\theta_j\| \le \tfrac{9}{4}\eta\kappa\|\theta_j\| \le \tfrac{9}{128}\|\theta_j\|.$$

Indeed the teacher part costs $\kappa\|\theta_j\|$, the student part costs $\kappa^2 R^2\|\theta_j\|/2$, and the subtracted radial part costs at most $(\kappa + \kappa^2 R^2/2)\|\theta_j\|$. Thus every nonzero hidden row stays nonzero and its head/hidden ratio is below $(1 + 9/128)/(1 - 9/128) < 2$. The Hessian bound in (54) applies along the entire segment.

Taylor's formula, $(u+v)^2 \le 2u^2 + 2v^2$, and $M\kappa h \le 1/4$ give

$$\begin{aligned} Q(x^+) - Q(x) &\ge h\{R^2G^2 - G\|v\|\} - \frac{M\kappa h^2}{2R^2}(R^2G + \|v\|)^2 \\ &\ge \tfrac{3}{4}hR^2G^2 - hG\|v\| - \tfrac{1}{4}h\|v\|^2/R^2 \\ &\ge \tfrac{1}{4}hR^2G^2 - \tfrac{3}{4}h\|v\|^2/R^2 \ge -\tfrac{3}{16}\eta\kappa^4R^4. \end{aligned}$$

The penultimate line uses $G\|v\| \le R^2G^2/2 + \|v\|^2/(2R^2)$. This fourth-power debit does not accumulate a constant Euler error over a logarithmically long release.

By (60), $(R_{k+1}^4 - R_k^4) \ge 4\eta a R_k^4$. Every stopped checkpoint, including a possible first exit, has $R_n^2 < 2R_*^2$, because $\eta\kappa \le 1/32$ and the upper radial multiplier is below $\sqrt{2}$. Hence

$$\sum_{k<n} \eta R_k^4 \le \frac{R_n^4 - R_0^4}{4a} < \frac{R_*^4}{a}, \qquad Q(x_n) \ge 2a - \frac{3\kappa^4 R_*^4}{16a} \ge \frac{29a}{16} > a.$$

Thus angular exit is impossible and the balance induction closes. Geometric radial growth proves finite crossing. The upper and lower radial multipliers give (58), using $\log(1 + 2\eta a) \ge \eta a$ and $\log(1 + 2\eta\kappa) \le 2\eta\kappa$.

At every checkpoint up to crossing, $V/R^2 \le \kappa^2 R_*^2/2 \le a/2$ and $Q \ge a$. Therefore $1 - L = 2F - V \ge aR^2$, and at crossing this is at least aR_*^2. Conversely $L \ge 1 - \kappa R^2 \ge 1 - 2\kappa R_*^2$, proving the stated whole-integer loss band and checkpoint gain. □

Joint motion, head cost, and the nonzero-mean distinction. The release certifies $\|x_N - x_0\| \ge R_* - R_0$. Writing $q = (q_j)$, Cauchy–Schwarz gives $F \le \kappa\|q\|\|\Theta\|_F$, hence

$$\|q_N\| \ge \frac{aR_*}{\kappa}, \qquad \|q_N\| \le R_N/\sqrt{2} < R_*.$$

Thus the actual head cost is explicit; it is not an oracle budget. There is no general lower bound on spatial motion in this release. For example a nonzero constant target can produce positive Q with $W = 0$ and improve solely through head/bias growth. Such a teacher is an illustration of the release interface, not an alignment theorem.

If additionally $\mu = \mathbb{E}[y] = 0$, the exact Lipschitz bound $|K(w,b) - K(0,b)| \le \|w\|$ gives $|K(w,b)| \le \|w\|$, and then

$$\|W_N\|_F \ge \frac{aR_*}{\kappa}, \qquad \|W_N - W_0\|_F \ge \frac{aR_*}{\kappa} - \|W_0\|_F.$$

For $\mu \ne 0$, the retained inequality is only $|K(w,b)| \le \|w\| + |\mu||b|$; it cannot be simplified by centering the target or discarding the bias contribution. Neither form proves a half-unit complete-score gain or all-direction coverage. The earlier entry window must supply those separately.

N CUBIC ROW GROWTH AND SELECTED-CONE INVARIANCE

Use the teacher and public constants of Appendix B.2: $X \sim N(0, I_d)$, $y_c = \sum_i a_i h_3(u_i^T X)$, $h_3(t) = (t^3 - 3t)/\sqrt{6}$, and $\sigma_\alpha(t) = \alpha t + (1-\alpha)t_+$. In normalized coordinates $(A, W, B) = s(q, w, b)$, put $\theta = (w, b)$, $\rho = \|w\|$, $n = w/\rho$, $z = b/\rho$ when $\rho > 0$, $P(n) = \sum_i a_i (u_i^T n)^3$, and $C(\theta) = \mathbb{E}[y_c \sigma_\alpha(w^T X + b)]$. Write $P_U = UU^T$, $P_{U^\perp} = I_d - P_U$ and let φ denote the standard normal density.

This is a deterministic argument for one category representative from AC13–16. Put $Z_X = (X, 1)$, $v = y - y_c$, $\varsigma = \operatorname{sign} C_0$, $p = \varsigma q$, $K = \varsigma C$, $B = b^2/\rho^2$ and $c = (1-\alpha)/\sqrt{6}$. Its initial conditions are $|C_0| \geq k$, $q_0 C_0 > 0$, $\rho_0 \in [1/2, 2]$, $|q_0| \leq \|\theta_0\|$, $B_0 \leq 1/16$ and initial joint norm below 3. The temporary radius cap denoted R in the following prefix estimates is the public envelope M of AJ8; the prescribed force in AJ9 pays every displayed inequality. The category milestone remains the radius R in AJ7. In the selected-cone calculation write $R_B = M$, $k_E = k$, $a_E = a_0$, $b_2 = b_*^2$, and $\overline{S} = 4\log(8M)$. Let $x = s_i u_i^T w$, $u = (\sum_{l \neq i} (u_l^T w)^2)^{1/2}$ and s_i be the category's original axis sign.

N.1 DETERMINISTIC ACTUAL COMPATIBLE-FORCE THEOREM

Let $T_c = \nabla C$. The cubic correlation bounds are

$$|C(\theta)| \leq \|\theta\|, \qquad \|T_c(\theta)\| \leq 1, \qquad \|D^2 C(\theta)\| \leq 8/\|\theta\|. \tag{DC11}$$

The exact normalized original coupled recurrence is

$$\begin{aligned} q_j^+ &= q_j + h(C_j + \theta_j^T E_j), \qquad \theta_j^+ = \theta_j + h q_j (T_{c,j} + E_j), \\ E_j &= \mathbb{E}[(v - s^2 \widehat{f}) \sigma_\alpha'(\theta_j^T Z_X) Z_X], \qquad \widehat{f} = m^{-1} \sum_l q_l \sigma_\alpha(\theta_l^T Z_X). \end{aligned} \tag{DC12}$$

The head force is exactly $\theta_j^T E_j$. Below E_j may be any adaptive compatible force of norm at most ν; no derivative of that force or independence of its rows is assumed.

N.2 DIRECT PHASE, BIAS, AND OUTSIDE-ENERGY PROOF

Use $g = \varsigma T_c$, $e = \varsigma E$, so $p^+ = p + h(K + \theta^T e)$ and $\theta^+ = \theta + hp(g + e)$. Initial balance and exact compatibility imply

$$D^+ \geq (1 - h^2 \|g + e\|^2) D \geq 0, \qquad D = \|\theta\|^2 - p^2. \tag{DC17}$$

Stop provisionally at the first failure of $p > 0, K \geq k, B = z^2 \leq 1$ or $\rho > 0$, and at the radius/common stop. At each pre-step point $\rho < R$, and including its candidate step one has $\rho \leq (1+2h)R < 2R$, since $\|\theta^+ - \theta\| \leq 2h\|\theta\|$ and, for $B \leq 1$, the sharper bound $\|w^+ - w\| \leq 2h\rho$ follows from $p \leq \sqrt{2}\rho$, $\|T_c\| \leq 1$, and $\nu \ll 1$. In this enlarged $2R$ tube,

$$\nu \leq \|g\|/4, \qquad |\theta^T e| \leq K/4, \tag{DC18}$$

by Euler's identity $\|\theta\|\|g\| \geq K \geq k$ and AJ9. The hidden segment stays above $\|\theta\|/2$. Taylor's formula and DC11 give

$$K^+ - K \geq hp\|g\|^2(3/4 - 25h/2) \geq \tfrac{1}{2} hp\|g\|^2, \qquad p^+ - p \geq 3hK/4. \tag{DC19}$$

Thus p, K remain positive with their original common sign. Also $\|\theta^+\|^2 - \|\theta\|^2 \geq 3hpK/2 \geq 0$.

For the exact teacher intermediate spatial step put $t = hpK/\rho^2$, $v_P = \nabla P(n)/P(n)$ and $a = 1 - (3 - B)t$. Nonzero K ensures $zP \neq 0$. Then

$$\begin{aligned} \overline{w}/\rho = an + t v_P, \quad \overline{b}/\rho = z + t(1-B)/z, \quad 0 < t \leq h\sqrt{B} \leq h, \quad 1 - 3h \leq a \leq 1, \\ D_w := \|\overline{w}\|^2/\rho^2 = (1 + Bt)^2 + (\|v_P\|^2 - 9)t^2 \geq 1, \\ N_b := \overline{b}^2/\rho^2 = B + 2(1-B)t + (1-B)^2 t^2/B. \end{aligned} \tag{DC20}$$

This is an identity inside the simultaneous step, not a splitting of the algorithm. The actual added hidden displacement is hpe.

If $B \le .75$, then $N_b \le .75 + 2h + h^2 < .755$ and $D_w \ge 1$. The added displacement has norm at most $2h\rho\nu$, so the actual bias-to-radius ratio remains below one. If $.75 \le B \le 1$,

$$D_w - N_b = 1 - B + (4B - 2)t + (B^2 - B + 2 - B^{-1} + \|v_P\|^2 - 9)t^2 \ge 1 - B + t. \quad \text{(DC21)}$$

The quadratic coefficient is nonnegative in this interval. Since $\|\overline{\theta}\| \le 2\rho$, the force changes $\|\overline{w}\|^2 - \overline{b}^2$ by at most $4\rho hp\nu + (hp\nu)^2$ in the unfavorable direction. Divided by $\rho^2 t$, this is at most $16R\nu/k + 8hR\nu^2/k < .01$. This proves the candidate upper-bias barrier. The new spatial radius is positive since $\rho^+ \ge \rho - hp\nu \ge (1 - 2h\nu)\rho > 0$. The nondecreasing augmented norm then gives $\rho \ge \|\theta_0\|/\sqrt{2} \ge 1/(2\sqrt{2}) > 1/3$. All provisional failures have been strictly excluded.

Writing $O = \|P_{U^\perp} w\|$, the exact teacher outside update gives

$$O^+ \le aO + hp\nu \le (1 - 2t)O + \frac{4R^2\nu}{k}t. \quad \text{(DC22)}$$

The coefficients are nonnegative, and its scalar equilibrium is $2R^2\nu/k \le 1/5000$. Since $O_0 \le 2$, this proves $O_n \le 2$. It controls the reliable row only; other original rows have not been discarded from any asserted complete-bank score.

N.3 Intrinsic force sums

For this actual row put $S_n = \sum_{l<n} t_l$. The augmented-energy increase above, together with $B \le 1$, gives

$$\|\theta^+\|^2 \ge \|\theta\|^2(1 + 3t/4), \qquad S_n \le \frac{20}{7}\log(4R) < \overline{S}, \quad \sum_{l<n} hp_l \le \frac{4R^2}{k}\overline{S}, \quad \sum_{l<n} hp_l K_l \le \frac{4}{3}\rho_n^2. \quad \text{(DC23)}$$

Here $\|\theta_0\| \ge .5$ and $\|\theta_n\| \le \sqrt{2}(1 + 2h)R < 2R$; $\log(1 + 3t/4) \ge 7t/10$ proves the clock bound. Every sum includes the candidate exit step. The pure bias step retains its sign because $bC_b = C(1 - B)$ and $B \le 1$. Also $|b| \ge |C| \ge k$, whereas the actual bias perturbation is at most $hp\nu \le hM\nu < k$. Thus the original bias sign persists, so the cone orientation below is the category's original orientation.

N.4 Direct invariant projected cone

Orient all teacher coordinates by the selected sign s_i. Inside the same step the selected coordinate and the other-coordinate norm satisfy exactly or respectively

$$\overline{x} = ax + d_c a_i x^2, \qquad \overline{u} \le au + d_c u^2, \qquad d_c = \frac{3hpc|z|\varphi(z)}{\rho^2} = \frac{3t}{\rho|P(n)|} > 0. \quad \text{(DX27)}$$

The norm inequality uses $(\sum_{l\ne i} a_l^2 (u_l^T w)^4)^{1/2} \le u^2$. The compatible force changes x in absolute value by at most $hp\nu$ and increases u by at most $hp\nu$.

Stop provisionally before failure of $x \ge 1/(2\sqrt{d})$ or $u \le \vartheta x$. In that cone the selected cubic term dominates the others because $a_i x^3 > u^3$, and $a_i x$ is the largest positive weighted coordinate, so $|P(n)| \le (a_i x/\rho)\|P_U n\|^2 \le a_i x/\rho$. Thus $\overline{x} \ge x(1 + Bt) \ge x$. Initial $x_0 \ge 1/\sqrt{d}$ and DC23 give at every candidate state

$$x_n \ge x_0 - \nu \sum_{l<n} hp_l \ge \frac{1}{\sqrt{d}} - \frac{4R_B^2 \nu \overline{S}}{k_E} \ge \frac{15}{16\sqrt{d}} > \frac{1}{2\sqrt{d}}. \quad \text{(DX28)}$$

For the cone boundary, monotonicity of $au + d_c u^2$ for $u \ge 0$ gives on the entire closed cone

$$u^+ - \vartheta x^+ \le -d_c \vartheta(a_i - \vartheta)x^2 + (1 + \vartheta)hp\nu. \quad \text{(DX29)}$$

Here $a_i - \vartheta \ge 7a_E/8$, $x^2 \ge 1/(4d)$ and $c|z|\varphi(z) = K/(\rho|P|) \ge k_E/\rho$. At the candidate step the pre-step spatial radius is below R_B; using the weaker $2R_B$ bound gives

$$d_c \vartheta(a_i - \vartheta)x^2 \ge \frac{21}{256}\frac{hpk_E \vartheta a_E}{R_B^3 d} > (1 + \vartheta)hp\nu. \quad \text{(DX30)}$$

The strict final inequality follows from AJ9. Both candidate failures are excluded. This proves the direct cone retention the selected-cone assertion for every original cone representative through the declared horizon N. It is not an assumed favorable sign or a relabeled reference basin.

N.5 Direct bias calibration and retention after a milestone

Let

$$W = b^2 - b_2\rho^2, \qquad b_2 + b_2^2 = 1.$$

The exact teacher intermediate gives

$$\overline{W} = (1 - 2(1 + b_2)t)W + h^2p^2\{T_b^2 - b_2\|T_w\|^2\}, \qquad T = \nabla C. \tag{DX31}$$

Indeed $bT_b = C(1 - B)$ and $w^TT_w = BC$ for a cubic target. The coefficient in front of W is in $[0, 1]$ since $t \le h$. For the full actual force displacement hpe, the quadratic W changes by at most $4\rho hp\nu + h^2p^2\nu^2$, using $\|\overline{\theta}\| \le 2\rho$. Correlation ascent, monotonicity of p, K and balance give

$$\sum_{l<n} h^2p_l^2\|T_l\|^2 \le 2h\sum_{l<n} p_l(K_{l+1} - K_l) \le 2hp_nK_n \le 4h\rho_n^2. \tag{DX32}$$

For the force sum use $\rho_l/K_l \le 2R_B/k_E$ and DC23:

$$\sum_{l<n} 4\rho_l hp_l\nu \le \frac{32R_B\nu}{3k_E}\rho_n^2.$$

The sum of $h^2p_l^2\nu^2$ is at most $8hR_B\nu^2\rho_n^2/(3k_E)$, absorbed by the stated weaker bound. Since $|W_0| < 5$, iterating the absolute-value inequality in DX31 proves the bias bound, with slack in both constants.

At $\rho \ge R_D$, the oriented direction error obeys

$$\|w/\rho - s_iu_i\| \le \sqrt{2}\sqrt{\vartheta^2 + 4/R_D^2} < \xi/4. \tag{DX33}$$

Also AJ3, AJ9 and $b_* > 1/2$ give

$$||b|/\rho - b_*| \le 2\{5/R_D^2 + 8h + 100R_B\nu/k_E\} < \xi/4. \tag{DX34}$$

The sign of b is the original s_b, so the actual full affine profile error is below $\xi/2$, hence below ξ. By DC20 each pure teacher spatial step has nondecreasing radius; force can decrease it by at most $hp\nu$. On any subinterval of the prefix the total decrease is at most the whole DC23 budget, which is at most $1/(16\sqrt{d})$. Thus a row reaching $2R_D$ retains radius at least $2R_D - 1/(16\sqrt{d}) > R_D$ through N, and its cone and bias bounds continue to apply. This is actual post-milestone retention, not a rule freezing that row.

O Cubic normalized correlation and complete-row geometry

O.1 Actual model and compatible force

In the first calculation the teacher coefficients and axes are (λ_i, v_i); in the later OE and BH calculations the axes are called u_i. These correspond to (a_i, u_i) in AJ1 when the interfaces are applied to that theorem. Function norms are Gaussian and $\theta_j = (w_j, b_j)$ denotes a normalized augmented hidden row. Let $X \sim N(0, I_d)$, $Z = (X, 1)$, and fix a normalized additive cubic

$$y_c = \sum_{i=1}^{r} \lambda_i h_3(v_i^TX), \qquad \sum_i \lambda_i^2 = 1, \quad v_i^Tv_l = \mathbf{1}_{i=l}, \qquad h_3(t) = (t^3 - 3t)/\sqrt{6}. \tag{NC1}$$

The coefficient magnitudes need not be equal. For $0 \le \alpha < 1$ let $\sigma(t) = \alpha t + (1 - \alpha)t_+$. Retain the full target $y = y_c + v$, including all its lower terms and tails, and the original averaged network

$$f = m^{-1}\sum_j a_j\sigma(w_j^TX + b_j), \qquad L = \mathbb{E}[(y - f)^2], \qquad (a_j, w_j, b_j) = s(q_j, \theta_j). \tag{NC2}$$

All raw coordinates train simultaneously at the same full-MSE step $\eta = mh/2$, where throughout $0 < h \le 1/512$. Define

$$C(\theta) = \mathbb{E}[y_c\sigma(\theta^TZ)], \qquad H(x) = qC(\theta), \quad x = (q, \theta), \quad F = \nabla H,$$

$$V = \|x\|^2, \qquad Q = H/V, \qquad \widehat{f} = m^{-1} \sum_l q_l \sigma(\theta_l^T Z). \tag{NC3}$$

The letter V here denotes a single row's joint energy; complete-bank energy below is denoted $\mathcal{V}$. The exact actual update is

$$x_j^+ = x_j + h\{F(x_j) + p_j\}, \qquad p_j = (\theta_j^T E_j, q_j E_j),$$
$$E_j = \mathbb{E}[(v - s^2 \widehat{f}) \sigma_j' Z]. \tag{NC4}$$

Here $\sigma_j' = \sigma'(\theta_j^T Z)$ is the scalar gate. Every original row, bias and student cross term is present. The analysis scaling by s changes no optimizer or metric.

The previously Gaussian-H1 bounds for the unit cubic give

$$|C| \le \|\theta\|, \quad \|\nabla C\| \le 1, \quad \|D^2 C\| \le 8/\|\theta\|, \quad F = (C, q\nabla C), \quad \|F(x)\| \le \|x\|,$$
$$H(tx) = t^2 H(x), \quad F(tx) = tF(x) \quad (t > 0), \qquad \|DF(q,\theta)\| \le 1 + 8|q|/\|\theta\|. \tag{NC5}$$

These bounds use the augmented hidden norm, so zero spatial weight with nonzero bias is allowed. In particular $|Q| \le 1/2$.

Assume each original initial row has $0 < \|\theta_{j,0}\|$ and $|q_{j,0}| \le \|\theta_{j,0}\|$. No head sign or teacher detector lower bound is assumed. If the compatible force at an update obeys $\|E_j\| \le \nu \le 1$, then balance and nonzero augmented hidden state persist through that update. Indeed, with $G = \nabla C + E$ and $D = \|\theta\|^2 - q^2$,

$$D^+ = D + h^2\{q^2\|G\|^2 - (\theta^T G)^2\} \ge (1 - h^2\|G\|^2) D \ge 0,$$
$$\|\theta^+\| \ge (1 - 2h)\|\theta\| > 0, \qquad \|x^+\| \le (1 + 2h)\|x\|. \tag{NC6}$$

This exact compatible balance is the only row-ratio hypothesis used.

O.2 Normalized correlation ascends up to a linear force debit

Put $u = x/\sqrt{V}$ and define its teacher tangent field

$$g(u) = F(u) - 2Qu, \qquad \langle u, g\rangle = 0, \qquad \|g\| \le 1, \qquad F(u) = 2Qu + g. \tag{NC7}$$

Euler's identity for degree two gives the factor $2Q$. For every actual update satisfying the preceding force and balance conditions,

$$Q^+ \ge Q + \frac{h}{2}\|g(u)\|^2 - 2h\nu. \tag{NC8}$$

Consequently $Q_n + 2\nu h n$ is nondecreasing on every prefix having that force bound. With zero force, Q itself is nondecreasing at the original fixed discrete step.

Here is a full one-step proof. First take a pure teacher step from the current actual row, solely as an algebraic intermediate. Its unnormalized unit-row update is

$$u + hF(u) = (1 + 2hQ)(u + tg), \qquad t = \frac{h}{1 + 2hQ}, \qquad \frac{h}{1+h} \le t \le \frac{h}{1-h} \le \frac{1}{511}.$$

Since the starting unit row is balanced, the segment $u + stg$, $0 \le s \le 1$, has augmented hidden norm greater than .7 and head/hidden ratio less than two. Thus $\|D^2 H\| = \|DF\| \le 17$ on the entire Taylor segment. Exact homogeneity and orthogonality give

$$Q^+_{\text{teach}} = \frac{H(u + tg)}{1 + t^2\|g\|^2}, \qquad H(u + tg) \ge Q + t\|g\|^2 - \frac{17}{2}t^2\|g\|^2.$$

It follows that

$$Q^+_{\text{teach}} - Q \ge \frac{t - 9t^2}{1 + t^2\|g\|^2}\|g\|^2 \ge \frac{h}{2}\|g\|^2. \tag{NC9}$$

For the last coarse inequality, use $t \ge 3h/4$, $1 - 9t \ge 3/4$ and $1 + t^2\|g\|^2 \le 9/8$, all implied by $h \le 1/512$.

The normalized compatible perturbation is $r = p/\sqrt{V}$, with $\|r\| \le \nu$. For nonzero a, b the exact identity

$$\|a - b\|^2 = (\|a\| - \|b\|)^2 + \|a\|\|b\|\|a/\|a\| - b/\|b\|\|^2$$

applied to $a = u + hF(u)$ and $b = a + hr$ bounds the difference of their normalized directions by $2h\nu$: their norms are at least $1 - h$ and $1 - 2h$. Both normalized endpoints are balanced by NC6. Their close chord has nonzero hidden component and norm at most one; there $\|\nabla H\| = \|F\| \le 1$. The resulting change in $Q = H(u)$ is at most $2h\nu$, proving NC8. No comparison trajectory or Euler remainder is being charged.

O.3 A COMPENSATED ENERGY INEQUALITY VALID AT NEGATIVE CORRELATION

The same actual update satisfies the stronger potential bound

$$\Delta\{\log V - 4hQ\} \le 4hQ + 5h\nu. \tag{NC10}$$

Unlike a bound that replaces Q by its positive part, NC10 retains the energy decrease of adversely correlated rows. To prove it, the exact pure-teacher radial/tangent decomposition is

$$\log(V^+_{\mathrm{teach}}/V) = 2\log(1+2hQ) + \log(1+t^2\|g\|^2) \le 4hQ + 2h^2\|g\|^2.$$

This uses $1 + 2hQ > 0$, $\log(1+a) \le a$ for $a > -1$, and $t^2 \le 2h^2$. Along the additional force segment the joint norm is at least $1 - 2h$, so the change in log joint norm is at most $2h\nu$, and the change in log energy is at most $4h\nu$. By NC8,

$$2h^2\|g\|^2 \le 4h(Q^+ - Q) + 8h^2\nu.$$

Combining the inequalities and using $4 + 8h \le 5$ proves NC10.

On any interval $a \le n$ of such updates, put $t = (n-a)h$. Telescoping NC10 gives

$$\log(V_n/V_a) \le 4h\sum_{l=a}^{n-1} Q_l + 5\nu t + 4h(Q_n - Q_a). \tag{NC11}$$

If $Q_l \le K$ for every pre-update index $a \le l < n$, with $K \ge 0$, this yields

$$V_n \le V_a \exp\{(4K+5\nu)t + 4h(Q_n - Q_a)\} \le V_a \exp\{(4K+5\nu)t + 4h\}. \tag{NC12}$$

There is no requirement on the terminal Q_n. In particular NC12 includes the candidate step which first crosses the threshold.

O.4 PUBLIC-HORIZON IRREVERSIBLE CLASSIFICATION, INCLUDING ALL CROSSINGS

Fix public $T \ge 0$, $\overline{T} = T + 1$, and $N_{\max} = \lceil T/h \rceil$, so $N_{\max}h \le \overline{T}$. Use the entirely polynomial constants

$$\kappa = \frac{1}{100\overline{T}}, \qquad 0 \le \nu \le \frac{1}{400\overline{T}^2} = \frac{\kappa}{4\overline{T}}. \tag{NC13}$$

All following assertions hold on any force-controlled prefix through an integer $N \le N_{\max}$. A first candidate energy exit may be that terminal integer, as formalized below. For each original row define its permanent marking time

$$\tau_j = \inf\{0 \le n \le N : Q_{j,n} \ge 2\kappa\}, \qquad \inf \varnothing = +\infty. \tag{NC14}$$

Then:

1. Every low-phase and threshold-crossing state has

$$V_{j,n} < \frac{8}{7}V_{j,0} \quad (0 \le n \le \min\{\tau_j, N\}). \tag{NC15}$$

   For an initially marked row this statement concerns just $n = 0$. For an unmarked row it covers the entire available horizon.

2. Once marked, the row cannot return to $Q \le \kappa$ on that horizon; more precisely,

$$Q_{j,n} \ge \frac{3}{2}\kappa \quad (\tau_j \le n \le N). \tag{NC16}$$

   The actual values can cross 2κ again. The permanent label, not exact monotonicity of the forced Q, is the irreversible object.

3. At every marked state the actual teacher head and correlation have the same nonzero sign, and

$$\frac{|C_j|}{\|\theta_j\|} \ge 2Q_{j,n} \ge 3\kappa, \qquad \frac{|q_j|}{\|\theta_j\|} \ge Q_{j,n} \ge \frac{3}{2}\kappa. \tag{NC17}$$

   Their common sign persists and $|q_j|$ increases on subsequent controlled updates. The relative head-force and hidden-gradient force are bounded by

$$\frac{|\theta_j^T E_j|}{|C_j|}, \quad \frac{\|E_j\|}{\|\nabla C_j\|} \le \frac{\nu}{3\kappa} \le \frac{1}{12\overline{T}}. \tag{NC18}$$

To prove NC15 for a positive marking time, all pre-crossing $Q_{j,l} < 2\kappa$. Apply NC12 with $K = 2\kappa$, including the crossing candidate. The exponent is at most

$$(8\kappa + 5\nu)\overline{T} + 4h \le \frac{8}{100} + \frac{5}{400} + \frac{4}{512} < \frac{1}{8}.$$

Since $e^{1/8} < 8/7$, NC15 follows. NC8 gives at every later state, including a proposed first exit below κ,

$$Q_{j,n} \ge Q_{j,\tau_j} - 2\nu(n - \tau_j)h \ge 2\kappa - 2\nu\overline{T} \ge \frac{3}{2}\kappa.$$

This excludes the candidate exit and proves NC16 without an iteration over repeated threshold crossings.

For NC17 put $z = |q|/\|\theta\| \in (0, 1]$ at a marked state. The identity $Q = qC/(q^2 + \|\theta\|^2)$ gives $|C|/\|\theta\| = Q(z + z^{-1}) \ge 2Q$ and, since $|C| \le \|\theta\|$, gives $Q \le z/(1 + z^2) \le z$. Euler's identity $\theta^T \nabla C = C$ then gives $\|\nabla C\| \ge |C|/\|\theta\|$, proving NC18. Finally the head update has increment with its current sign because $|C| - |\theta^T E| \ge (3\kappa - \nu)\|\theta\| > 0$. Together with NC16 at the next state, this preserves also the correlation sign. There was no initial detector quantile for any of these classified rows.

All original rows stay in the bank. For example, at any common time the entire energy of the still-unmarked rows is at most $(8/7)\sum_j V_{j,0}$. Marked rows are not presumed aligned: NC16–18 only provide the positive-correlation and relative-force interface needed for a separate direct spatial argument.

O.5 Conditional full-bank correlation from actual energy growth

There is also a useful endpoint consequence that avoids classifying which rows supplied the energy. For any interval of $n \ge 1$ actual steps starting at zero, set $t = nh$ and retain the force bound. NC8 implies $Q_l \le Q_n + 2\nu(n - l)h$. Substitution in NC11 yields the rowwise inequality

$$Q_n \ge \frac{\log(V_n/V_0) + 4hQ_0 - \nu\{4t(t + h) + 5t\}}{4(t + h)}. \tag{NC19}$$

The sum $h\sum_{l=0}^{n-1} 2\nu(n - l)h = \nu t(t + h)$ is exact, so no unreported time-step error occurs.

Define the complete actual quantities

$$\mathcal{V}_n = \sum_j V_{j,n}, \qquad \mathcal{H}_n = \sum_j H_{j,n}, \qquad \overline{Q}_n = \mathcal{H}_n/\mathcal{V}_n.$$

Multiplying NC19 by $V_{j,n}/\mathcal{V}_n$, summing every original row, and using $Q_{j,0} \ge -1/2$ gives

$$\overline{Q}_n \ge \frac{\log(\mathcal{V}_n/\mathcal{V}_0) - 2h - \nu\{4t(t + h) + 5t\}}{4(t + h)}. \tag{NC20}$$

For completeness, the log-sum step is

$$\sum_j \frac{V_{j,n}}{\mathcal{V}_n} \log \frac{V_{j,n}}{V_{j,0}} = \log \frac{\mathcal{V}_n}{\mathcal{V}_0} + \sum_j p_j \log(p_j/p_{j,0}) \ge \log \frac{\mathcal{V}_n}{\mathcal{V}_0},$$

where $p_j = V_{j,n}/\mathcal{V}_n$ and $p_{j,0} = V_{j,0}/\mathcal{V}_0$. All these weights are positive by NC6. This is an energy-weighted identity for the complete bank, not a selected-row estimate.

Under NC13, if an actual integer $1 \le N \le N_{\max}$ has $\mathcal{V}_N \ge 2\mathcal{V}_0$, then

$$\overline{Q}_N \ge \frac{1}{8\overline{T}}. \tag{NC21}$$

Indeed $t \le \overline{T}$ and $h \le 1/512$ bound the debit $2h + \nu\{4t(t + h) + 5t\}$ by $1/32$. The numerator of NC20 is then greater than $5/8$, since $\log 2 \ge 2/3$, and its denominator is at most $5\overline{T}$. This proves NC21. The energy-doubling hypothesis is an actual checkpoint requirement still to be acquired, not a proved exit time or a replacement definition of subspace learning.

For the full target rather than its central cubic, exact compatibility gives at the endpoint

$$\mathbb{E}[v\widehat{f}] = m^{-1}\sum_j q_j\theta_j^T E_j + s^2\|\widehat{f}\|_2^2. \tag{NC22}$$

If the endpoint force itself also obeys $\|E_{j,N}\| \le \nu$, then with complete raw joint norm $\mathfrak{R}_N^2 = s^2\mathcal{V}_N$,

$$\frac{\mathbb{E}[yf_N]}{\mathfrak{R}_N^2} \ge \frac{\overline{Q}_N - \nu/2}{m} \ge \frac{1}{16m\overline{T}}. \tag{NC23}$$

Here the nonnegative student term in NC22 is retained, and $\sum_j |q_j|\|\theta_j\| \le \mathcal{V}_N/2$ bounds the debit. The last inequality uses NC13 and NC21. The endpoint-force condition is explicitly included because a first energy-crossing state need not lie inside an unpadded pre-update tube.

O.6 Full actual recurrence and the cubic kernel

Let $X \sim N(0, I_d)$, $Z = (X, 1)$, and

$$y_c = \sum_{i=1}^r \lambda_i h_3(u_i^T X), \qquad h_3(t) = (t^3 - 3t)/\sqrt{6}, \qquad u_i^T u_l = \mathbf{1}_{i=l}, \qquad \sum_i \lambda_i^2 = 1. \tag{OE1}$$

Write $U = \mathrm{span}\{u_i\}$ and let P_U be its orthogonal projector. Coefficient signs and magnitudes are unrestricted. For $0 \le \alpha < 1$ set $\sigma(t) = \alpha t + (1-\alpha)t_+$ and retain the full teacher $y = y_c + v$. The original averaged network and full-MSE risk are

$$f = m^{-1}\sum_j a_j\sigma(w_j^T X + b_j), \qquad L_{\mathrm{risk}} = \mathbb{E}[(y-f)^2].$$

Normalize only for analysis by $(a_j, w_j, b_j) = s(q_j, \theta_j)$, $\theta_j = (\widetilde{w}_j, \widetilde{b}_j)$. Below we suppress the tildes. With the original simultaneous raw Euclidean step $\eta = mh/2$, the exact actual recurrence is

$$q^+ = q + h(C + \theta^T E), \qquad \theta^+ = \theta + hq(T + E), \qquad C(\theta) = \mathbb{E}[y_c\sigma(\theta^T Z)], \quad T = \nabla C, \tag{OE2}$$

$$E_j = \mathbb{E}[(v - s^2\widehat{f})\sigma'(\theta_j^T Z)Z], \qquad \widehat{f} = m^{-1}\sum_l q_l\sigma(\theta_l^T Z). \tag{OE3}$$

The same vector E generates both errors in OE2. It may depend on every row and the whole history; no independence or derivative of E is assumed. A choice of σ' at zero affects no nondegenerate Gaussian row, and the identically zero joint row is absorbing.

For $\rho = \|w\| > 0$, put $n = w/\rho$, $z = b/\rho$, $B = z^2$, and $P(n) = \sum_i \lambda_i(u_i^T n)^3$. The exact cubic correlation is

$$C(\theta) = -c\,b\varphi(z)P(n), \qquad c = (1-\alpha)/\sqrt{6}, \qquad \varphi(z) = (2\pi)^{-1/2}e^{-z^2/2}. \tag{OE4}$$

It extends to a C^2 function on the punctured augmented space; $C = T = 0$ at $w = 0, b \ne 0$. The Gaussian-H1 theorem gives

$$|C| \le \|\theta\|, \quad \|T\| \le 1, \quad \theta^T T = C, \quad \|D^2 C\| \le 8/\|\theta\|. \tag{OE5}$$

For trace and differentiability, see Appendix F. This is the full augmented-parameter bound, including pure-bias rows; no unmentioned spatial-radius lower bound is used.

For a row define

$$V = q^2 + \|\theta\|^2, \quad H = qC, \quad Q = H/V, \quad \mathcal{B} = C^2 + q^2\|T\|^2,$$

$$I^2 = \|P_U w\|^2, \quad O^2 = \|(I - P_U)w\|^2, \quad A = I^2/\rho^2 \quad (\rho > 0), \qquad \mathcal{M}(\theta) = \max\{O^2, b^2 - I^2\}. \tag{OE6}$$

Thus $\mathcal{M} = O^2 + (b^2 - \rho^2)_+ \ge 0$ controls outside energy and excess squared bias together. Always $O^2 \le \mathcal{M}$ and $b^2 \le \rho^2 + \mathcal{M}$.

O.7 SIGN-FREE ENERGY AND NORMALIZED CORRELATION

Throughout this subsection suppose $\|E\| \le \nu \le 1$, $0 < h \le 1/128$, $\theta \neq 0$, and $|q| \le \|\theta\|$. Balance and nonzero augmented state persist through every actual step. Indeed, setting $G = T + E$ and $D = \|\theta\|^2 - q^2$, homogeneity gives

$$D^+ = D + h^2\{q^2\|G\|^2 - (\theta^T G)^2\} \ge (1 - h^2\|G\|^2)D, \qquad \|\theta^+ - \theta\| \le 2h\|\theta\|. \tag{OE7}$$

In particular the joining hidden segment stays away from zero. The following two sign-free estimates hold:

$$H^+ - H \ge \frac{h}{2}\mathcal{B} - 3h\nu^2\|\theta\|^2, \tag{OE8}$$

$$(V - 4hH)^+ - (V - 4hH) \le 4hH + 3h\nu V. \tag{OE9}$$

For completeness their short derivation follows. Put $a = C + \theta^T E$. Taylor's theorem and OE5 give

$$C^+ = C + hqT^T G + \mathcal{E}, \qquad |\mathcal{E}| \le 8h^2q^2\|G\|^2/\|\theta\|.$$

The first-order terms of $H^+ - H = hCa + hq^2T^TG + h^2qaT^TG + (q + ha)\mathcal{E}$ are at least $h(3\mathcal{B}/4 - 2\nu^2\|\theta\|^2)$ by Young's inequality. The absolute values of the last two terms sum to at most $18h^2(\mathcal{B} + \nu^2\|\theta\|^2)$: respectively use $|qaT^TG| \le 3(\mathcal{B} + \nu^2\|\theta\|^2)/2$ and $|(q + ha)\mathcal{E}| \le 16(1 + 2h)h^2(\mathcal{B} + \nu^2\|\theta\|^2)$. This proves OE8. The exact expansion

$$V^+ - V = 4hH + 4hq\theta^T E + h^2(a^2 + q^2\|G\|^2) \le 4hH + 2h\nu V + 2h^2\mathcal{B} + 2h^2\nu^2 V$$

minus $4h$ times OE8 proves OE9, since $14h\nu \le 1$.

We also need the following detector-free debit:

$$Q^+ \ge Q - 2h\nu. \tag{OE10}$$

Here is a direct proof, independent of a prescribed bias corridor. Write $x = (q, \theta)$, $F = \nabla H = (C, qT)$, and $g = \nabla Q$. Homogeneity gives $x^T F = 2H$, $x^T g = 0$, and

$$\|F\| \le \sqrt{V}, \qquad g = (F - 2Qx)/V, \qquad \|g\| \le V^{-1/2}. \tag{OE11}$$

The final bound follows by squaring the numerator and using $\|F - 2Qx\|^2 = \|F\|^2 - 4Q^2V \le V$. For $|q| \le 2\|\theta\|$, direct differentiation of H/V yields

$$\|D^2 Q\| \le 32/V. \tag{OE12}$$

Indeed $\|D^2 H\| \le 17$; the three remaining quotient terms have bounds $4/V$, $1/V$ and $4/V$, so even $26/V$ suffices. For the pure teacher candidate let $\lambda = 1 + 2hQ \ge 1 - h > 0$. By degree-zero homogeneity,

$$Q(x + hF) = Q\left(x + \frac{hV}{\lambda}g\right).$$

The tangent joining segment has joint norm at least $\sqrt{V}$. Its hidden and head displacements have magnitude at most $2h\|\theta\|/(1 - h)$, so its head/hidden norm ratio is below two. Taylor's theorem, OE12 and $16h/\lambda < 1$ give $Q(x + hF) \ge Q(x)$. The compatible joint perturbation has norm $h\|(\theta^T E, qE)\| \le h\nu\sqrt{V}$. Its joining segment from $x + hF$ has joint norm at least $(1 - 2h)\sqrt{V}$ and nonzero hidden part. Integrating OE11 along that segment gives loss at most $h\nu/(1 - 2h) \le 2h\nu$, proving OE10. This proof also covers zero correlation and zero head.

O.8 A MAXIMUM OF QUADRATICS HANDLES EVERY BIAS TRANSITION

For every balanced row, the specific cubic formula improves the generic bound $|H| \le V/2$ to

$$|H| \le \rho^2/4. \tag{OE13}$$

For $\rho > 0$, $|P| \le (\sum_i |u_i^T n|^6)^{1/2} \le 1$ and balance give

$$\frac{|H|}{\rho^2} \le c\varphi(\sqrt{B})\sqrt{B(1 + B)} \le c(1 + B)\varphi(\sqrt{B}) \le 2c\varphi(1) < \frac{1}{4}.$$

The last strict inequality follows from $2c\varphi(1) \le 1/\sqrt{3\pi e} < 1/4$. The maximum of $(1 + B)e^{-B/2}$ occurs at $B = 1$. At $\rho = 0$, $C = 0$, so OE13 holds there as well.

Favorable one-step theorem. If $H \geq 0$, then every actual candidate step satisfies

$$\mathcal{M}^+ - \mathcal{M} \leq h^2\mathcal{B} + 2h\nu V, \tag{OE14}$$

$$(\mathcal{M} - 2hH)^+ - (\mathcal{M} - 2hH) \leq 3h\nu V. \tag{OE15}$$

There is no condition on B and no condition on the next sign of H.

To prove OE14 first take the pure hidden candidate $\theta^{\mathrm{t}} = \theta + hqT$. Let $F_1 = O^2$ and $F_2 = b^2 - I^2$; each is a quadratic form with matrix operator norm at most one. Differentiating OE4, with no division by P or C, gives

$$DF_1(\theta)[qT] = -2(3-B)H(1-A),$$
$$DF_2(\theta)[qT] = 2H\{(3-B)A - 2 - B\}. \tag{OE16}$$

For example, $w^T T_w = BC$, $bT_b = (1-B)C$, and $w_\perp^T T_{w,\perp} = -(3-B)C(1-A)$ imply both identities. At $\rho > 0$ put $t = hH/\rho^2$. By OE13, $0 \leq t \leq h/4 \leq h$. We claim

$$F_i(\theta) + hDF_i(\theta)[qT] \leq \mathcal{M}(\theta) \quad (i = 1, 2). \tag{OE17}$$

For F_1, the derivative is nonpositive if $B \leq 3$. If $B > 3$, the other branch exceeds it by $(B-1)\rho^2$, while its positive first-order change is at most $2t(B-3)\rho^2$; the gap absorbs this for $h \leq 1/2$. For F_2, the bracket in OE16 is nonpositive when $B \geq 1$: its maximum over $A \in [0,1]$ is $1-2B$ for $B \leq 3$ and $-2-B$ for $B \geq 3$. If $B < 1$, the gap from F_1 is $(1-B)\rho^2$ and the bracket is at most $1-2B$. It is nonpositive for $B \geq 1/2$; for $B < 1/2$, $2t(1-2B) \leq 1-B$ when $h \leq 1/2$. This proves OE17. The exact quadratic second-order term of each branch is at most $h^2q^2\|T\|^2 \leq h^2\mathcal{B}$. Taking their maximum yields

$$\mathcal{M}(\theta^{\mathrm{t}}) \leq \mathcal{M}(\theta) + h^2\mathcal{B}. \tag{OE18}$$

All possible branch crossings are already included. At $\rho = 0$, $C = T = 0$ and OE18 is immediate, so no limiting ratio is needed.

For the force displacement $\delta = hqE$, either quadratic obeys

$$F_i(\theta^{\mathrm{t}} + \delta) - F_i(\theta^{\mathrm{t}}) \leq 2\|\theta^{\mathrm{t}}\|\|\delta\| + \|\delta\|^2 \leq 2h\nu|q|\|\theta\| + (2h^2\nu + h^2\nu^2)q^2 \leq (1+3h)h\nu V \leq 2h\nu V.$$

Taking maxima proves OE14. Subtracting $2h$ times OE8 gives

$$(\mathcal{M} - 2hH)^+ - (\mathcal{M} - 2hH) \leq (2h\nu + 6h^2\nu^2)V \leq 3h\nu V,$$

which is OE15. Consequently, on any favorable prefix starting at an actual state J, including its last candidate step,

$$\mathcal{M}_n \leq \mathcal{M}_J + 2h(H_n - H_J) + 3\nu\sum_{l=J}^{n-1} hV_l \leq \mathcal{M}_J + \frac{h}{2}\rho_n^2 + 3\nu\sum_{l=J}^{n-1} hV_l. \tag{OE19}$$

This is a bias-energy barrier valid even when the normalized bias crosses one or three repeatedly, or when a spatial row vanishes.

O.9 A PUBLIC POLYNOMIAL BUDGET JOINS ALL ORIGINAL ROWS

Fix a public horizon $T_0 \geq 1$ and parameters

$$0 < h \leq 1/512, \qquad \kappa = \frac{1}{100T_0}, \qquad 0 \leq \nu \leq \frac{1}{400T_0^2}. \tag{OE20}$$

Suppose every pre-step force up to a considered integer state n obeys OE2 with $\|E\| \leq \nu$, and $nh \leq T_0$. Start at any nonzero augmented row with $|q_0| \leq \|\theta_0\|$. Let J be the first integer state with $Q_J \geq 2\kappa$, or $+\infty$ if none. Then

$$V_l < \frac{8}{7}V_0 \quad (0 \leq l \leq \min\{J, n\}), \qquad Q_l \geq \frac{3}{2}\kappa > 0 \quad (J \leq l \leq n). \tag{OE21}$$

The second statement is vacuous for $J = +\infty$.

To check the low phase including its crossing step, put $Z_V = V - 4hH$. Before J, $Z_V \geq (1-8h\kappa)V$. By OE9,

$$Z_V^+ \leq \left(1 + \frac{h(8\kappa + 3\nu)}{1 - 8h\kappa}\right) Z_V.$$

Globally $(1-2h)V \le Z_V \le (1+2h)V$, so every low or crossing state has

$$V_l \le \frac{1+2h}{1-2h}\exp\left(\frac{(8\kappa+3\nu)lh}{1-8h\kappa}\right)V_0 < \frac{8}{7}V_0. \tag{OE22}$$

For explicit arithmetic, the prefactor is at most $257/255$, the exponent is below .09, and $e^{.09} \le 100/91$; their product is less than $8/7$. If $J = 0$, the energy claim is immediate. Summing OE10 after J gives $Q_l \ge 2\kappa - 2\nu T_0 \ge 3\kappa/2$. Thus there is one persistent marked phase, with no uncontrolled later adverse episode. These constants also follow from NC13–18.

Apply OE19 to the marked phase and use $\mathcal{M}_J \le V_J < 8V_0/7$. Before marking, simply use $\mathcal{M}_l \le V_l < 8V_0/7$. Every original row therefore obeys

$$\mathcal{M}_n \le \frac{8}{7}V_0 + \frac{h}{2}\rho_n^2 + 3\nu\sum_{l=0}^{n-1} hV_l. \tag{OE23}$$

In particular

$$O_n^2 \le \frac{8}{7}V_0 + \frac{h}{2}\rho_n^2 + 3\nu\sum_{l<n} hV_l, \qquad b_n^2 \le (1+h/2)\rho_n^2 + \frac{8}{7}V_0 + 3\nu\sum_{l<n} hV_l, \tag{OE24}$$

$$V_n \le (4+h)\rho_n^2 + \frac{16}{7}V_0 + 6\nu\sum_{l<n} hV_l. \tag{OE25}$$

The last inequality uses balance and $V \le 2(\rho^2 + b^2)$. These conclusions require no row's initial detector to be nonzero; they retain every initial row, including either initial head sign. An identically zero initial joint row contributes zero throughout and can be included by continuity or directly.

O.10 Exact model and starting interfaces

Let $X \sim N(0, I_d)$, $Z = (X, 1)$, and retain the full teacher $y = y_c + v$, where

$$y_c = \sum_{i=1}^{r} \lambda_i h_3(u_i^T X), \quad h_3(t) = (t^3 - 3t)/\sqrt{6}, \quad u_i^T u_l = \mathbf{1}_{i=l}, \quad \sum_i \lambda_i^2 = 1. \tag{BH1}$$

Put $U = \operatorname{span}\{u_i\}$ and $\sigma(t) = \alpha t + (1-\alpha)t_+$, $0 \le \alpha < 1$. In the original averaged network $f = m^{-1}\sum_j a_j\sigma(w_j^T X + b_j)$ with full-MSE risk, write the analysis normalization $(a_j, w_j, b_j) = s(q_j, \theta_j)$. The unchanged simultaneous raw Euclidean step $\eta = mh/2$ is exactly

$$q^+ = q + h(C + \theta^T E), \qquad \theta^+ = \theta + hq(T + E), \qquad C = \mathbb{E}[y_c\sigma(\theta^T Z)], \quad T = \nabla C, \tag{BH2}$$

$$E_j = \mathbb{E}[(v - s^2\widehat{f})\sigma'(\theta_j^T Z)Z], \qquad \widehat{f} = m^{-1}\sum_l q_l\sigma(\theta_l^T Z). \tag{BH3}$$

All rows and student cross terms remain in this recurrence. The force E is arbitrary and history-dependent subject to its norm bound; no derivative or independence of E is assumed.

Suppressing the normalization marks on $\theta = (w, b)$, let

$$\rho = \|w\|, \quad n = w/\rho, \quad B = b^2/\rho^2, \quad P(n) = \sum_i \lambda_i(u_i^T n)^3, \quad c = (1-\alpha)/\sqrt{6}.$$

For $\rho > 0$ the exact kernel is $C = -cb\varphi(b/\rho)P(n)$, with $\varphi(t) = (2\pi)^{-1/2}e^{-t^2/2}$. It is C^2 off the augmented origin and has $C = T = 0$ at $w = 0, b \ne 0$. For the quantities

$$\begin{gathered} V = q^2 + \|\theta\|^2, \quad H = qC, \quad Q = H/V, \quad \mathcal{B} = C^2 + q^2\|T\|^2, \\ I^2 = \|P_U w\|^2, \quad O^2 = \|(I - P_U)w\|^2, \quad A = I^2/\rho^2, \\ M = \max\{O^2, b^2 - I^2\}, \quad D = \|\theta\|^2 - q^2, \end{gathered} \tag{BH4}$$

Appendix O gives:

$$|C| \le \|\theta\|, \quad \|T\| \le 1, \quad \theta^T T = C, \quad \|D^2 C\| \le 8/\|\theta\|, \qquad |H| \le \rho^2/4 \quad \text{when } D \ge 0; \tag{BH5}$$

$$H^+ - H \geq \frac{h}{2}\mathcal{B} - 3h\nu^2\|\theta\|^2, \qquad Q^+ \geq Q - 2h\nu; \tag{BH6}$$

$$(M - 2hH)^+ - (M - 2hH) \leq 3h\nu V \quad \text{if } H \geq 0. \tag{BH7}$$

These hold for $0 < h \leq 1/128$, $\|E\| \leq \nu \leq 1$, initial $D \geq 0$, and nonzero augmented hidden state. Balance and nonzero augmented state persist. The proofs, including every bias transition and zero-spatial case, are OE4–19 above. Only BH5–7, the quotient-Hessian bound, and the sign-free compensated-energy inequality are used here. The latter inequality is

$$(V - 4hH)^+ - (V - 4hH) \leq 4hH + 3h\nu V. \tag{BH7a}$$

For a public horizon $T_0 \geq 1$, take

$$0 < h \leq 1/512, \qquad \kappa = \frac{1}{100T_0}, \qquad 0 \leq \nu \leq \frac{1}{400T_0}. \tag{BH8}$$

The linear-in-horizon force budget follows from a sharper actual normalized-correlation estimate than the second inequality of BH6:

$$Q^+ \geq Q - h\nu^2 \qquad (0 < h \leq 1/512). \tag{BH8a}$$

Here is a self-contained reduction using the kernel bounds. Write $\xi = (q, \theta)$, $F = \nabla H$, and $p = (\theta^T E, qE)$. Then $\|F\| \leq \sqrt{V}$, $\|p\| \leq \nu\sqrt{V}$, and $\xi^T F = 2H$. Set

$$a = \xi^T p/V, \quad p_\perp = p - a\xi, \quad d = F - 2Q\xi + p_\perp, \quad \lambda = 1 + h(2Q + a), \quad \tau = h/\lambda.$$

We have $|a| \leq \nu$, $\|p_\perp\| \leq \nu\sqrt{V}$, $\xi^T d = 0$, and $\lambda \geq 1 - 2h > 0$. Degree-zero homogeneity gives $Q(\xi + h(F + p)) = Q(\xi + \tau d)$. Put $\gamma = \|F - 2Q\xi\|/\sqrt{V} \leq 1$; the last inequality follows from $\|F - 2Q\xi\|^2 = \|F\|^2 - 4Q^2V \leq V$. The joining segment has norm at least $\sqrt{V}$, and each of its head and hidden displacements has norm at most $4\tau\|\theta\|$. Since $\tau \leq 1/510$, the head/hidden ratio along it is less than two. On this segment direct differentiation of $Q = H/V$ gives $\|D^2Q\| \leq 32/V$: its four terms have bounds $17/V$, $4/V$, $1/V$ and $4/V$, respectively, using BH5. Taylor's theorem therefore gives

$$Q^+ - Q \geq \tau(\gamma^2 - \nu\gamma) - 16\tau^2(\gamma + \nu)^2 \geq \tau\{(1 - 32\tau)\gamma^2 - \nu\gamma - 32\tau\nu^2\}.$$

Now $\tau \leq 1/128$, $\tau \leq 4h/3$, and $\nu\gamma \leq (\gamma^2 + \nu^2)/2$ imply $Q^+ - Q \geq \tau(\gamma^2/4 - 3\nu^2/4) \geq -h\nu^2$. No detector, head sign or bias-ratio bound is used.

At integer states with $nh \leq T_0$, define J as the first state with $Q_J \geq 2\kappa$, or infinity if no such state exists. The direct estimates imply, including the crossing candidate,

$$V_l < \frac{8}{7}V_0 \quad (l \leq \min\{J, n\}), \qquad Q_l \geq \frac{3}{2}\kappa \quad (J \leq l \leq n). \tag{BH9}$$

The second inequality follows from BH8a and $\nu^2 T_0 \leq \kappa/1600 < \kappa/2$. For the first, while $Q < 2\kappa$ one has $V - 4hH \geq (1 - 8h\kappa)V$. Iterating BH7a and using the global $|H| \leq V/2$ gives the sign-free compensated-energy bound

$$V_l \leq \frac{1 + 2h}{1 - 2h} \exp\left(\frac{(8\kappa + 3\nu)lh}{1 - 8h\kappa}\right) V_0 < \frac{8}{7}V_0.$$

The prefactor is at most $257/255$, and the exponent is below $.09$; $e^{.09} \leq 100/91$ verifies the final constant. No initial head sign or nonzero detector is required.

O.11 The positive-part barrier through every bias transition

Define the two scalar functions

$$W(\theta) = \rho^2 - 4b^2, \qquad Y(\theta) = W(\theta)_+. \tag{BH10}$$

Suppose $H \geq 0$, balance holds, $\|E\| \leq \nu \leq 1$, and $0 < h \leq 1/128$. Then the full actual candidate satisfies

$$Y^+ - Y \leq h^2\mathcal{B} + 5h\nu V, \tag{BH11}$$

$$(Y - 2hH)^+ - (Y - 2hH) \leq 6h\nu V. \tag{BH12}$$

In expressions such as $(Y - 2hH)^+$ the superscript means the next iterate; the positive part is used only in the explicit definition $Y = W_+$. This convention avoids confusing the two operations.

For the proof, the exact kernel derivatives give $w^T T_w = BC$ and $bT_b = (1 - B)C$. Thus

$$DW(\theta)[qT] = 2H(5B - 4). \tag{BH13}$$

If $B \le 4/5$, this derivative is nonpositive. If $B > 4/5$, then $W = -(4B - 1)\rho^2 < 0$. With $t = hH/\rho^2$ and BH5, $0 \le t \le h/4$, so

$$2t(5B - 4) \le \frac{h}{2}(5B - 4) \le \frac{5}{2}hB \le \frac{11}{4}B \le 4B - 1. \tag{BH14}$$

Here $h \le 1$ suffices for the third inequality, and $B > 4/5$ gives the last one. The original negative gap therefore absorbs the entire positive first-order change. In all cases $W + hDW[qT] \le Y$.

For the pure hidden candidate $\theta^{\mathrm{t}} = \theta + hqT$, the exact quadratic second-order term is $h^2q^2(\|T_w\|^2 - 4T_b^2) \le h^2\mathcal{B}$. Consequently $W(\theta^{\mathrm{t}}) \le Y(\theta) + h^2\mathcal{B}$; the right side is nonnegative, so taking positive parts yields

$$Y(\theta^{\mathrm{t}}) \le Y(\theta) + h^2\mathcal{B}. \tag{BH15}$$

This treats all candidate crossings of $W = 0$, including crossings from a large normalized bias. If $\rho = 0$, then $C = T = 0$ and the pure candidate is unchanged, so the same conclusion is immediate.

The matrix of W has operator norm four. For the force displacement $\delta = hqE$, its quadratic expansion gives

$$\begin{aligned} W(\theta^{\mathrm{t}} + \delta) - W(\theta^{\mathrm{t}}) &\le 8\|\theta^{\mathrm{t}}\|\|\delta\| + 4\|\delta\|^2 \\ &\le 8h\nu|q|\|\theta\| + (8h^2\nu + 4h^2\nu^2)q^2 \le (4 + 12h)h\nu V \le 5h\nu V. \end{aligned} \tag{BH16}$$

The scalar inequality $(a + d)_+ \le a_+ + d$ for $d \ge 0$ turns this into the same upper increment for Y. Combining with BH15 proves BH11. Subtracting $2h$ times BH6 proves BH12, since $5h\nu + 6h^2\nu^2 \le 6h\nu$.

O.12 Balance-defect growth and the row's own energy budget

The balance defect has the exact recurrence

$$D^+ - D = h^2\{q^2\|T + E\|^2 - (\theta^T(T + E))^2\} \le 2h^2\mathcal{B} + 2h^2\nu^2 V. \tag{BH17}$$

Subtracting $4h$ times BH6 gives the sign-free upper bound

$$(D - 4hH)^+ - (D - 4hH) \le 14h^2\nu^2 V. \tag{BH18}$$

Balance remains nonnegative by the lower inequality $D^+ \ge (1 - h^2\|T + E\|^2)D$.

On a marked phase $J \le l < n$ satisfying BH9, exact joint-energy expansion yields

$$V_{l+1} - V_l = 4hH_l + 4hq_l\theta_l^T E_l + h^2\{(C_l + \theta_l^T E_l)^2 + q_l^2\|T_l + E_l\|^2\} \ge h(6\kappa - 2\nu)V_l > 0. \tag{BH19}$$

The strict positivity uses $\nu \le \kappa/4$ and $V_l > 0$. Therefore for $t = (n - J)h$ the marked row's own cumulative budget is

$$\mathcal{I}_{J,n} := \sum_{l=J}^{n-1} hV_l \le tV_n \le T_0 V_n, \qquad \mathcal{I}_{J,n} \le \frac{V_n - V_J}{6\kappa - 2\nu}. \tag{BH20}$$

The inequalities include the last candidate step. No bank energy or width enters this bound, and no terminal spatial radius is assumed in proving it.

Write $E_* = V_J$ and $x = \rho_n^2$. Summing BH7, BH12 and BH18, using $H_J \ge 0$, $M_J, Y_J, D_J \le E_*$ and BH5, gives

$$M_n \le E_* + \frac{h}{2}x + 3\nu\mathcal{I}_{J,n}, \qquad Y_n \le E_* + \frac{h}{2}x + 6\nu\mathcal{I}_{J,n}, \tag{BH21}$$

$$0 \le D_n \le E_* + hx + 14h\nu^2\mathcal{I}_{J,n}. \tag{BH22}$$

In particular no uncontrolled sum is assigned to a row that merely happens to be present in a large bank.

Since $b_n^2 \le x + M_n$ and balance gives $V_n \le 2(x + b_n^2)$, BH20–21 imply, with $\epsilon = \nu T_0$,

$$V_n \le \frac{(4+h)x + 2E_*}{1-6\epsilon}, \qquad M_n \le \frac{E_* + (h/2 + 12\epsilon)x}{1-6\epsilon}. \tag{BH23}$$

The denominator is positive under BH8. For $x > 0$, put $e = E_*/x$ and

$$R(e) = \frac{4 + h + 2e}{1 - 6\epsilon}.$$

The simultaneous direct bounds can then be stated as

$$\frac{V_n}{x} \le R(e), \quad \frac{M_n}{x} \le e + h/2 + 3\epsilon R(e), \quad \frac{Y_n}{x} \le e + h/2 + 6\epsilon R(e), \tag{BH24}$$

$$\frac{D_n}{x} \le e + h + 14h\nu\epsilon R(e). \tag{BH25}$$

These estimates are valid at arbitrary bias transitions and every candidate exit within the public horizon. No spatial nonvanishing assumption was needed before the endpoint division by x.

O.13 Fixed geometry for every sufficiently large original row

Under BH8, any marked endpoint with

$$\rho_n^2 \ge 16 V_J \tag{BH26}$$

has the fixed geometry

$$A_n \ge \frac{9}{10}, \qquad \frac{1}{5} \le \frac{b_n^2}{\rho_n^2} \le \frac{9}{8}, \qquad \frac{9}{8} \le \frac{q_n^2}{\rho_n^2} \le \frac{17}{8}. \tag{BH27}$$

Here $A_n = \|P_U w_n\|^2/\rho_n^2$ refers to the complete represented cubic subspace, not to a particular favored direction.

To verify the constants, $e \le 1/16$, $h \le 1/512$ and $\epsilon, \nu \le 1/400$ give

$$R(e) < 17/4, \qquad M_n/x < 1/10, \qquad Y_n/x < 1/5, \qquad D_n/x < 3/40. \tag{BH28}$$

For example substitute $17/4$ in the last three expressions of BH24–25; they are at most $1/16 + 1/1024 + 51/1600$, $1/16 + 1/1024 + 51/800$, and $1/16 + 1/512 + 238/(512 \cdot 160000)$, respectively. Now $O_n^2 \le M_n$ proves $A_n \ge 9/10$. From $x - 4b_n^2 \le Y_n$ we obtain $b_n^2/x \ge 1/5$, while $b_n^2 \le x + M_n$ gives the upper bound $11/10 < 9/8$. Finally

$$q_n^2/x = 1 + b_n^2/x - D_n/x \ge 1 + 1/5 - 3/40 = 9/8,$$

and balance gives $q_n^2/x \le 1 + b_n^2/x \le 17/8$.

This is an every-large-row statement. In particular, suppose every original initial row is balanced and $V_{j,0} \le 5$. Every actual row at every considered state with

$$\rho_{j,n}^2 \ge 96 \tag{BH29}$$

must already be marked, since unmarked or crossing energies are less than $40/7$. Its marked entry has $V_{j,J} < 40/7$, and $96 > 16(40/7)$, so BH27 follows. Rows that never become large remain in the bank; they are not discarded. If a different initial energy cap $V_{j,0} \le v_*$ is available, the deterministic sufficient threshold is $\rho_{j,n}^2 \ge 128 v_*/7$; a strict larger threshold can be used to avoid endpoint equality conventions.

For a general-rank prediction interface define $\bar{n} = P_U n/\sqrt{A}$ at an endpoint obeying BH27. Its marked sign gives the exact identity

$$H_n = c\rho_n^2 \sqrt{q_n^2/\rho_n^2}\sqrt{B_n}\varphi(\sqrt{B_n}) A_n^{3/2} |P(\bar{n})| \ge \frac{c}{16}\rho_n^2 |P(\bar{n})|. \tag{BH30}$$

For the coarse constant use $\sqrt{q_n^2/\rho_n^2} \ge 1$, $\sqrt{B_n} > 2/5$, $A_n^{3/2} > 4/5$, and $\varphi(\sqrt{B_n}) > 1/5$. The last bound follows from $B_n \le 9/8$ and $2\pi e^{9/8} < (44/7)(32/23)^4 < 25$, using $e^{9/8} \le (1 - 9/32)^{-4}$. The product $8/125$ exceeds $1/16$. In rank one $|P(\bar{n})| = 1$. At higher rank it may be small because of cancellation; this calculation introduces no new lower-detector assumption. It leaves acquired control of that factor to the general-rank prediction analysis.

P SIMULTANEOUS MATURATION OF THE CUBIC CATEGORIES

Use Appendix B.2. Here $(A_j, W_j, B_j) = s(q_j, w_j, b_j)$, $\theta_j = (w_j, b_j)$, $\rho_j = \|w_j\|$, and $V_n = \sum_j (q_{j,n}^2 + \|\theta_{j,n}\|^2)$ is the complete normalized joint energy. Write $P_U = UU^T$, $P_{U^\perp} = I_d - P_U$ and $A_{\mathrm{sub},n} = \sum_j \|P_U w_{j,n}\|^2 / \sum_j \|w_{j,n}\|^2$. The activation is $\sigma_\alpha(t) = \alpha t + (1-\alpha)t_+$, the cubic constant is $c = (1-\alpha)/\sqrt{6}$, and φ is the standard normal density. In the category and Gaussian-tail calculation here, δ denotes its $\delta_0 = \delta_{\mathrm{theorem}}/2$. Write $R_M = R$ and define J to be the last first crossing of $2R_M$ among the $4r$ category representatives. The symbol J in this section is a checkpoint, while the leak factor is written $1-\alpha$.

Original Gaussian event without a conditional-law substitution. Write each initial normalized spatial, bias and head row as $(G, B, Z)/\sqrt{d}$, where $G \sim N(0, I_d)$ and $B, Z \sim N(0,1)$ are independent; different rows are independent as well. For every i and $(s_i, s_b) \in \{-1, 1\}^2$ require a row with

$$s_i u_i^T G \in [1,2], \quad |u_l^T G| \le \vartheta/(4\sqrt{r-1})\ (l \ne i), \quad s_b B \in [1/2, 2], \quad -s_i s_b Z \in [1/2, 2]. \tag{AC13}$$

The other-coordinate restrictions are empty at rank one. Each specified rectangle has probability exactly p_C under the original independent Gaussian law. Its failure of occupancy is at most e^{-mp_C}, so all $4r$ categories occur except with probability $\delta/4$. Distinct categories have disjoint rectangles, hence there are at least $4r$ distinct rows.

The elementary Gaussian tails give, except with probability $7\delta/96$, simultaneously for every row

$$\rho_{j,0} \in [1/2, 2], \quad \|P_U G_j\|^2 \le H, \quad |B_j|, |Z_j| \le \sqrt{2\ell}, \quad |q_{j,0}| \le \|\theta_{j,0}\|, \quad V_0 < 5m. \tag{AC14}$$

Here is the tail accounting explicitly. For $Z_k \sim \chi_k^2$, the Gaussian moment-generating function yields

$$\log \mathbb{E}[e^{u(Z_k - k)}] \le \frac{ku^2}{1-2u} \quad (0 < u < 1/2), \qquad \log \mathbb{E}[e^{-u(Z_k - k)}] \le ku^2 \quad (u > 0).$$

Exponential Markov with $u = \sqrt{\ell}/(\sqrt{k} + 2\sqrt{\ell})$ and $u = \sqrt{\ell/k}$, respectively, therefore gives

$$\Pr\{Z_k > k + 2\sqrt{k\ell} + 2\ell\} \le e^{-\ell}, \qquad \Pr\{Z_k < k - 2\sqrt{k\ell}\} \le e^{-\ell}.$$

These are also the bounds BW6 in Appendix G. For a standard normal scalar, its moment-generating function $\mathbb{E}[e^{uG}] = e^{u^2/2}$ similarly gives $\Pr\{|G| > \sqrt{2\ell}\} \le 2e^{-\ell}$. Apply the two chi-square tails to $\|G_j\|^2$, the upper tail to $\|P_U G_j\|^2$, and the scalar bound to each of B_j, Z_j. Their per-row costs sum to $(2+1+2+2)e^{-\ell}$, so the union over all rows costs $7me^{-\ell} = 7\delta/96$. Since $d \ge 128\ell$, these events give $1/4 < \|G_j\|^2/d < 4$ and $q_{j,0}^2, b_{j,0}^2 \le 2\ell/d \le 1/64$. Thus balance holds and each joint row energy is below $4 + 1/32 < 5$, proving every assertion in AC14. They also give $A_{\mathrm{sub},0} \le 4H/d \le 1/4$.

For each AC13 row the exact cubic correlation is

$$C(\theta) = -cb\varphi(b/\rho)P(w/\rho), \qquad P(n) = \sum_i a_i (u_i^T n)^3, \quad \rho = \|w\|. \tag{AC16}$$

The selected signed cubic contribution in Gaussian coordinates is at least a_0, whereas the magnitude of the remainder is at most $(\vartheta/4)^3 < a_0/2$. Using AC14 gives $|C_0| \ge k$, $q_0 C_0 > 0$, $|q_0|, |b_0| \in [1/(2\sqrt{d}), 2/\sqrt{d}]$, $(b_0/\rho_0)^2 \le 1/16$, selected coordinate at least $1/\sqrt{d}$, and initial other-coordinate cone ratio at most $\vartheta/4$. All signs and seeds are acquired properties of these original rectangles.

A complete actual force induction through the public time. Put $v = y - y_c$ and $\widehat{f} = m^{-1} \sum_l q_l \sigma_\alpha(\theta_l^T(X,1))$. The exact actual recurrence is

$$q_j^+ = q_j + h(C_j + \theta_j^T E_j), \qquad \theta_j^+ = \theta_j + hq_j(\nabla C_j + E_j), \qquad E_j = \mathbb{E}[(v - s^2 \widehat{f})\sigma_j'(X,1)]. \tag{AC17}$$

Here $\sigma_j' = \sigma_\alpha'(\theta_j^T(X,1))$ is a scalar; the following $(X,1)$ in AC17 multiplies it as an augmented vector. Thus the last expression is a vector. Every self/cross interaction and every retained teacher term occurs in it. Duality and homogeneity give

$$\|\nabla C_j\| \le 1, \quad \|\widehat{f}\|_2 \le V/(2m), \quad \|E_j\| \le \|v\|_2 + s^2 V/(2m). \tag{AC18}$$

Assume inductively the force is controlled up to a candidate step. Exact compatibility gives

$$\|\theta^+\|^2-(q^+)^2\ge(1-h^2\|\nabla C+E\|^2)(\|\theta\|^2-q^2)\ge 0,\qquad V^+\le(1+2h)^2V. \tag{AC19}$$

Nonzero augmented hidden states persist since $\|\theta^+\|\ge(1-2h)\|\theta\|$. Starting from AC14, every candidate $n\le N$ consequently has

$$V_n\le(1+2h)^{2n}V_0<5me^{4nh}\le 5me^{4\overline{T}}=M^2/4. \tag{AC20}$$

At every such candidate AC18 and AJ10–11 give $\|E_j\|\le\nu/2+\nu/8<\nu$. The initial step satisfies the same inequality, so induction closes the full force and radius tube through N, including its endpoint. This is not a stopped proof ending at the old $512m$ energy hit.

Direct phase, cone and bias estimates on the longer actual prefix. The DC phase proof and DX cone/bias proof are local compatible-update statements. Their cap may be any public radius, not specifically $\sqrt{512m}$. Apply their proofs with radius cap M, seed k, lower coefficient a_0, and horizon N. AC20 keeps every row strictly inside that cap, while every corresponding force budget is included in AJ9. For completeness their direct inequalities are recorded here. For a category row orient its head by $\varsigma=\operatorname{sign}C_0$ and set $p=\varsigma q$, $D=\varsigma C$, $B=b^2/\rho^2$. Throughout the prefix, including candidate steps,

$$\begin{gathered}p>0,\quad D\ge k,\quad p\le\|\theta\|,\quad 0<B\le 1,\quad \rho>1/3,\\ D^+-D\ge\tfrac12 hp\|\nabla C\|^2,\quad p^+-p\ge 3hk/4.\end{gathered} \tag{AC21}$$

Taylor's theorem uses $\|D^2C\|\le 8/\|\theta\|$ and $\nu\le k/(10^4M^2)$; it excludes a phase crossing. The exact cubic bias barrier uses $bC_b=C(1-B)$ and $w^T\nabla_w C=BC$ and gives $B\le 1$, without a reference path. With $t=hpD/\rho^2\le h$ and $\bar a=1-(3-B)t$, the teacher intermediate has outside contraction and nondecreasing spatial radius. Compatibility supplies

$$O\le 2,\qquad \sum_{l<n}hp_l\le\frac{4M^2}{k}\overline{S},\qquad \sum_{l<n}hp_lD_l\le\frac43\rho_n^2,\quad O=\|P_{U^\perp}w\|. \tag{AC22}$$

The outside recurrence is $O^+\le(1-2t)O+(4M^2\nu/k)t$, with equilibrium at most $1/5000$. The intrinsic sums follow by telescoping augmented energy, exactly as DC23; their cap is now M.

Write $x=s_iu_i^Tw$, $u=\|P_{U\cap u_i^\perp}w\|$. The teacher intermediate and actual force obey

$$\overline{x}=\bar a x+d_ca_ix^2,\quad \overline{u}\le\bar a u+d_cu^2,\quad d_c=\frac{3hpc|b/\rho|\varphi(b/\rho)}{\rho^2},\qquad |\Delta_{\text{force}}x|,\Delta_{\text{force}}u\le hp\nu. \tag{AC23}$$

Inside $u\le\vartheta x$, the selected weighted coordinate dominates, so $\overline{x}\ge x$. Thus AJ9 and AC22 give $x_n\ge 1/\sqrt{d}-4M^2\nu\overline{S}/k\ge 15/(16\sqrt{d})>1/(2\sqrt{d})$. At the cone boundary the strict teacher advantage is at least $21hpk\vartheta a_0/(256M^3d)$, whereas the force debit is at most $(1+\vartheta)hp\nu$. AJ9 strictly pays it. This closes the cone by candidate induction, with the original bias sign retained.

The exact signed quadratic $b^2-b_*^2\rho^2$ satisfies the DX31 recursion. Its initial absolute value is below five, and correlation ascent plus AC22 pay the discrete and force remainders, yielding

$$|B-b_*^2|\le 5/\rho^2+8h+100M\nu/k. \tag{AC24}$$

The force is allowed to be adaptive and coupled throughout. At $\rho\ge R_M\ge R_D$, AC23–24 imply

$$\|w/\rho-s_iu_i\|\le\sqrt{2}\sqrt{\vartheta^2+4/R_D^2}<\xi/4,\qquad \big||b|/\rho-b_*\big|\le 2(5/R_D^2+8h+100M\nu/k)<\xi/4. \tag{AC25}$$

This proves AC21–25 whenever the row has that radius.

The teacher spatial intermediate never decreases its radius, and the cumulative actual radius debit on *any* subinterval is at most the full $\nu\sum hp\le 1/(16\sqrt{d})$. Thus a row reaching $2R_M$ retains radius greater than R_M through N and retains its profile there. If a category row had not reached $2R_M$ by N, AC21 would give $p_N\ge 3Nhk/4>6R_M$, while balance and $B\le 1$ would give $p_N<2\sqrt{2}R_M$. This contradiction proves the acquired all-category hitting claim. The larger force envelope, rather than an assumption that the former bank stop occurs late, is what removes the DX alternative.

Every original row in the endpoint score. The normalized-correlation marking proof applies on any force-controlled horizon $\overline{T}$, without initial signal lower bounds for unseeded rows. It costs $\nu \le 1/(400\overline{T}^2)$, supplied by AJ9. Every unmarked or marking state has row energy at most $8/7$ times its initial energy; after marking the head/correlation product stays positive. The max-quadratic outside proof then gives, at *every* integer on this prefix,

$$O_n^{\mathrm{bank}} \le \frac{8}{7}V_0 + \frac{h}{2}S_n + 3I_n, \qquad V_n \le (4+h)S_n + \frac{16}{7}V_0 + 6I_n,$$
$$S_n = \sum_j \|w_{j,n}\|^2, \quad O_n^{\mathrm{bank}} = \sum_j \|P_{U^\perp} w_{j,n}\|^2, \quad I_n = \nu \sum_{l<n} hV_l \le \nu\overline{T}M^2 \le m/1024. \tag{AC26}$$

These are the sum of OE24–25 with the complete force integral retained; they do not require n to be the first energy crossing. For $n \ge J$, the $4r$ distinct retained category representatives give $V_n \ge S_n \ge 4rR_M^2 \ge 1024m$. Using $V_0 < 5m$ in AC26 gives $S_n > .24V_n$ and

$$1 - A_{\mathrm{sub},n} \le h/2 + \frac{(40/7 + 3/1024)m}{.24V_n} < .025 < .1. \tag{AC27}$$

This proves the complete spatial-energy conclusion for the full original denominator. It also gives $\|W_n - W_0\|_F > 8s\sqrt{m}$ and $\|W_n - W_0\|_F/\|W_0\|_F > 4$ at those integers. The energy inequality AC19 and $V_J \ge 1024m > V_0$ imply the lower physical clock in the stated loss band and physical clock, including the first candidate hit.

Q Complete head-weighted outside control for cubic teachers

Q.1 Actual run and the stronger force integral already paid by AC

Use AC's unchanged model, raw step and original draw. Here $X \sim N(0, I_d)$, $U = (u_1, \dots, u_r)$, $P_U = UU^T$, $P_{U^\perp} = I_d - P_U$, and $(A_j, W_j, B_j) = s(q_j, w_j, b_j)$. The public $R_M = R$, M, k, $\overline{T}$ and ν are those of AJ7–9; $a_0 = \min_i a_i$. In its normalized coordinates the predictor and exact simultaneous recurrence are

$$f_n(X) = \frac{s^2}{m}\sum_j q_{j,n}\sigma_\alpha(w_{j,n}^T X + b_{j,n}), \quad \sigma_\alpha(t) = \alpha t + (1-\alpha)t_+, \quad 0 \le \alpha < 1, \quad \eta = mh/2,$$
$$q_j^+ = q_j + h(C_j + \theta_j^T E_j), \qquad \theta_j^+ = \theta_j + hq_j(T_j + E_j), \quad \theta_j = (w_j, b_j), \quad T_j = \nabla C_j,$$
$$C_j = \mathbb{E}[y_c\sigma_\alpha(\theta_j^T(X,1))], \qquad E_j = \mathbb{E}[(y - y_c - f_n)\sigma'_\alpha(\theta_j^T(X,1))(X,1)]. \tag{CW1}$$

The central teacher is the complete additive cubic in AJ1. All retained original means, linear terms and tails, and every student interaction, are in E_j. Nothing is centered or reset. Let

$$V_{j,n} = q_{j,n}^2 + \|\theta_{j,n}\|^2, \quad \mathcal{V}_n = \sum_j V_{j,n}, \quad \rho_{j,n} = \|w_{j,n}\|, \quad O_{j,n} = \|P_{U^\perp} w_{j,n}\|,$$
$$H_{j,n} = q_{j,n}C_{j,n}, \quad Q_{j,n} = H_{j,n}/V_{j,n}, \quad I_{j,n} = \nu\sum_{l<n} hV_{j,l}, \qquad I_n = \sum_j I_{j,n}. \tag{CW2}$$

The notation $O_{j,n}$ is a norm, not a squared energy.

On AC's event of original probability at least $1 - \delta$, at all integers $0 \le n \le N$ one already has

$$V_{j,0} < 5, \quad |q_{j,n}| \le \|\theta_{j,n}\|, \quad \|T_{j,n}\| \le 1, \quad \|E_{j,n}\| \le \nu,$$
$$\mathcal{V}_n < M^2/4, \quad Nh \le \overline{T}, \quad h \le 1/512, \quad \nu \le 1/64. \tag{CW3}$$

The individual initial bound follows from AC14: $\rho_{j,0} \le 2$ and $q_{j,0}^2, b_{j,0}^2 \le 2\ell/d \le 1/64$. In particular $V_{j,0} \le 4 + 1/32 < 5$; no bound on a typical row is substituted for a simultaneous bound.

AC's existing cone-force cutoff is much stronger than the coarse $I_n \le m/1024$ previously used. Its constants satisfy

$$\overline{T} = 2 + 8R_M/k, \quad M > 4R_M, \quad k \le R_M, \quad \vartheta \le a_0/8, \quad a_0 \le 1, \quad d \ge 256,$$
$$\nu \le \frac{k\vartheta a_0}{1024M^3 d}. \tag{CW4}$$

Here $k \le R_M$ follows directly from AC's displayed $k < 1$ and $R_M \ge 16\sqrt{m}$. Hence, without a new restriction,

$$I_n \le \nu\overline{T}M^2 \le \frac{(2k+8R_M)\vartheta a_0}{1024Md} < \frac{5\vartheta a_0}{2048d} \le \frac{5}{16384d} \le \frac{5}{2^{22}} < 2^{-19} =: \epsilon_I. \tag{CW5}$$

The actual radius bound in CW3 would improve this by a further factor four, which is not needed. Each $I_{j,n}$ is nonnegative and at most I_n.

Q.2 Permanent marking and a new balance-defect upper bound

Put $\kappa_0 = 1/(100\overline{T})$ and permanently mark a row at its first actual state with $Q_j \ge 2\kappa_0$. The NC13–18/OE20–23 estimates apply because AC includes $\nu \le 1/(1024\overline{T}^2) < 1/(400\overline{T}^2)$. Every unmarked and marking state has

$$V_{j,n} < \frac{8}{7}V_{j,0} < \frac{40}{7}. \quad \text{After marking,} \quad Q_{j,n} \ge \frac{3}{2}\kappa_0 > 0. \tag{CW6}$$

The latter assertion holds at all later actual states, not just while the row stays above the original marking threshold. In particular its head and cubic correlation have the same nonzero sign. NC obtains this from the exact debit $Q^+ \ge Q - 2h\nu$ and a compensated low-phase energy estimate that includes the crossing step. No post-marking Gaussianity or minimum initial detector is used.

We also use the sign-free OE8 inequality and exact cubic bound OE13, with $\mathcal{B}_j = C_j^2 + q_j^2\|T_j\|^2$:

$$H_j^+ - H_j \ge \frac{h}{2}\mathcal{B}_j - 3h\nu^2\|\theta_j\|^2, \qquad |H_j| \le \rho_j^2/4. \tag{CW7}$$

The latter remains valid at zero spatial weight. It follows from $C = -cb\varphi(b/\rho)P(w/\rho)$, $|P| \le 1$ and balance, since $c\varphi(\sqrt{B})\sqrt{B(1+B)} < 1/4$ for $B \ge 0$. The upper bound on the *balance defect* $D_j = \|\theta_j\|^2 - q_j^2 \ge 0$ is new here. Exact compatibility, followed by CW7, gives

$$\begin{aligned} D_j^+ - D_j &= h^2\{q_j^2\|T_j + E_j\|^2 - (C_j + \theta_j^T E_j)^2\} \\ &\le 2h^2\mathcal{B}_j + 2h^2\nu^2 V_j \le 4h(H_j^+ - H_j) + 14h^2\nu^2 V_j. \end{aligned} \tag{CW8}$$

Let τ_j be its finite marking time. Sum CW8 from τ_j through the current state, use $H_{j,\tau_j} > 0$ and CW6–7, and enlarge the nonnegative force integral to start at zero. Every marked state therefore satisfies

$$D_{j,n} \le \frac{8}{7}V_{j,0} + h\rho_{j,n}^2 + 14h\nu I_{j,n}. \tag{CW9}$$

It includes the marking state itself. This upper bound is separate from the preserved nonnegative balance lower bound in CW3.

OE's all-bias maximum-of-quadratics estimate supplies, for every row at every state, including unmarked rows and every branch crossing,

$$\begin{aligned} b_{j,n}^2 &\le (1 + h/2)\rho_{j,n}^2 + \frac{8}{7}V_{j,0} + 3I_{j,n}, \\ V_{j,n} &\le (4+h)\rho_{j,n}^2 + \frac{16}{7}V_{j,0} + 6I_{j,n}. \end{aligned} \tag{CW10}$$

These are precisely OE24–25 with the full actual row force integral. They require no axis cone or initial head sign.

Every state with $\rho_{j,n} \ge 4$ is already permanently marked, because its energy is at least $16 > 40/7$. At such a state, CW9–10 and CW5 imply

$$\begin{aligned} \frac{q_j^2}{\rho_j^2} &\ge 1 - h - \frac{40/7 + 14h\nu\epsilon_I}{16} > \frac{16}{25}, \\ B_j := \frac{b_j^2}{\rho_j^2} &\le 1 + h/2 + \frac{40/7 + 3\epsilon_I}{16} < \frac{34}{25}. \end{aligned} \tag{CW11}$$

For explicit conservative arithmetic, $14h\nu\epsilon_I < 1/1024$ makes the first lower bound exceed $1 - 1/512 - 5/14 - 1/16384 > 16/25$. The second upper bound is below $1 + 1/1024 + 5/14 + 3/(16\,2^{19}) < 34/25$. These are acquired all-row bounds at a fixed absolute entry radius; they are not assumptions about the unselected rows.

Q.3 THE EXACT HEAD–OUTSIDE PRODUCT CONTRACTS ABOVE THAT RADIUS

At an actual state with $\rho \geq 4$, omit the row subscript and put $p = |q|$, $O = \|P_{U^\perp} w\|$, and $t = hH/\rho^2$. By CW6–7, $0 < t \leq h/4$. Use one pure-teacher step only as an algebraic intermediate from the current actual state. The exact cubic identities give

$$p^{\mathrm{t}} = p + h|C| = p\left(1 + t\frac{\rho^2}{q^2}\right), \qquad w_\perp^{\mathrm{t}} = [1 - (3 - B)t]w_\perp. \tag{CW12}$$

The outside multiplier is in $[0, 1]$, because $B < 34/25$ and $t \leq h/4$. Set $a = \rho^2/q^2$ and $b = 3 - B$. CW11 implies $a \leq 25/16$, $b \geq 41/25$, hence $b - a \geq 31/400$. Their product, with its favorable quadratic term retained, satisfies

$$p^{\mathrm{t}} O^{\mathrm{t}} = (1 + at)(1 - bt)pO \leq \left(1 - \frac{31}{400}t\right) pO. \tag{CW13}$$

No selected-axis geometry appears in this exact outside- *teacher-subspace* calculation.

The full actual step still includes both compatible errors: $|q^+ - q^{\mathrm{t}}| \leq h\nu\|\theta\|$ and $\|w^+ - w^{\mathrm{t}}\| \leq hp\nu$. Since $O^{\mathrm{t}} \leq O \leq \|\theta\|$ and $p^{\mathrm{t}} \leq p + h\|\theta\|$, expanding their product gives

$$\begin{aligned} |q^+|O^+ &\leq p^{\mathrm{t}} O^{\mathrm{t}} + h\nu\{\|\theta\| O^{\mathrm{t}} + pp^{\mathrm{t}}\} + h^2\nu^2 p\|\theta\| \\ &\leq \left(1 - \frac{31}{400}t\right) pO + h\nu(1 + h/2 + h\nu/2)V \\ &\leq \left(1 - \frac{31}{400}t\right) pO + 2h\nu V. \end{aligned} \tag{CW14}$$

This estimate is valid for an arbitrary adaptive coupled E_j of the actual run. It charges both the hidden force and the head force; dropping the latter would not justify the weighted estimate.

Q.4 ALL LOW STATES, CANDIDATE CROSSINGS, AND EVERY ORIGINAL ROW

At a pre-step state with $\rho_j < 4$, CW10 gives

$$V_j < (4 + 1/512)16 + 80/7 + 6\epsilon_I < 76. \tag{CW15}$$

The exact compatible update has $V_j^+ \leq (1 + 2h)^2 V_j$, by balance and $\|T_j + E_j\| \leq 2$. Thus its next candidate state, whether or not it crosses radius four or the marking threshold, has $V_j^+ < (257/256)^2\, 76 < 77$. In particular $|q_j^+|O_j^+ \leq V_j^+/2 < 39$. An unmarked or marking state already has $V_j < 40/7$ by CW6, so it cannot be an omitted high-radius adverse state.

Starting from $|q_{j,0}|O_{j,0} \leq V_{j,0}/2 < 5/2$, combine CW14 at high-radius pre-step states with CW15 at every other pre-step state. Induction proves, for every original row and every integer $0 \leq n \leq N$,

$$|q_{j,n}|O_{j,n} \leq 39 + 2I_{j,n} < 40. \tag{CW16}$$

There is no need to assume a single radius crossing. Every return below radius four is handled by CW15, and every new exit candidate has the same bound. Permanent marking only supplies the positive phase at high states; no unmarked row is discarded. The rowwise integrals sum exactly, giving the stronger full-bank statements

$$\begin{aligned} \sum_j |q_{j,n}|O_{j,n} &\leq 39m + 2I_n, \\ \sum_j q_{j,n}^2 O_{j,n}^2 &\leq 1521m + 156I_n + 4I_n^2 < 1522m. \end{aligned} \tag{CW17}$$

The last inequality uses $I_n < 2^{-19}$ and $m \geq 1$. This is a complete-bank head-weighted outside energy estimate, not an estimate on the $4r$ category representatives alone.

Q.5 Exact full outside AGOP and the remaining inside obligation

Let $\mathsf{G}_n = \mathbb{E}[\nabla_x f_n \nabla_x f_n^T]$ be the actual raw student AGOP. Positive homogeneity with $s > 0$ gives the exact derivative

$$P_{U^\perp} \nabla_x f_n = \frac{s^2}{m} \sum_j q_{j,n} P_{U^\perp} w_{j,n} \sigma'_\alpha(\theta_{j,n}^T(X,1)). \tag{CW18}$$

The L^2 triangle inequality and $|\sigma'_\alpha| \le 1$ retain all cross-row terms and yield

$$\operatorname{tr}(P_{U^\perp} \mathsf{G}_n P_{U^\perp}) \le \frac{s^4}{m^2}(39m + 2I_n)^2 < 1522 s^4, \qquad \|P_{U^\perp} \mathsf{G}_n P_{U^\perp}\|_{\mathrm{op}} < 1522 s^4. \tag{CW19}$$

The operator bound follows because this is positive semidefinite. It is uniform through AC's entire actual horizon, independent of M and later row-radius growth. No cancellation was assumed in obtaining it, and the linear part of leaky ReLU was not removed.

R Original-coordinate anchors and the cubic signed frame

R.1 The exact map and ordering without a head-sign assumption

Here $X \sim N(0, I_d)$, $h_3(t) = (t^3 - 3t)/\sqrt{6}$, $\sigma_\alpha(t) = \alpha t + (1-\alpha)t_+$, and φ is the standard normal density. With $U = (u_1, \dots, u_r)$, write $P_U = UU^T$ and $P_{U^\perp} = I_d - P_U$; later uses of U as a span refer to the same teacher subspace. The normalized student is $\widehat{f} = m^{-1} \sum_j q_j \sigma_\alpha(w_j^T X + b_j)$, $\sigma'_j = \sigma'_{\alpha,j} = \sigma'_\alpha(\theta_j^T(X,1))$ denotes a scalar gate, and E_w is the spatial component of E. Absorb coefficient signs into the orthonormal teacher axes and write

$$y_c = \sum_{i=1}^r a_i h_3(u_i^T X), \quad a_i > 0, \quad \sum_i a_i^2 = 1, \quad c = (1-\alpha)/\sqrt{6}, \quad a_{\max} = \max_i a_i. \tag{WG1}$$

Use the unchanged compatible recurrence for an original row,

$$\begin{gathered} q^+ = q + h(C + \theta^T E), \quad \theta^+ = \theta + hq(\nabla C + E), \\ \theta = (w, b), \quad |q| \le \|\theta\|, \quad \|E\| \le \nu \le 1, \quad 0 < h \le 1/512. \end{gathered} \tag{WG2}$$

In the original raw model $(A, W, B) = s(q, w, b)$ and full-MSE step $\eta = mh/2$, the actual force is $E_j = \mathbb{E}[(y - y_c - s^2 \widehat{f}) \sigma'_j(X,1)]$. All full-teacher and student terms remain in it. The proofs allow adaptive forces but do not assert that every allowed force is generated by one fixed teacher. Let $\rho = \|w\| > 0$, $z = b/\rho$, and define

$$\begin{gathered} F(z) = -cz\varphi(z), \quad S(w) = \sum_i a_i (u_i^T w)^3, \quad C = F(z)S(w)/\rho^2, \\ V = q^2 + \rho^2 + b^2, \quad H = qC, \quad Q = H/V. \end{gathered} \tag{WG3}$$

The weighted spatial coordinates $x_i = a_i u_i^T w$ obey exactly

$$\begin{aligned} &x_i^+ = Ax_i + Dx_i^2 + \eta_i, \qquad \eta_i = hqa_i u_i^T E_w, \\ &A = 1 + (z^2 - 3)hH/\rho^2, \qquad D = 3hqF(z)/\rho^2, \qquad P_{U^\perp} w^+ = AP_{U^\perp} w + hqP_{U^\perp} E_w. \end{aligned} \tag{WG4}$$

Indeed differentiation of $F(z)S(w)/\rho^2$, using $zF'(z)/F(z) = 1 - z^2$ away from zero and continuity at zero, gives the displayed radial and coordinate terms.

The following bounds hold with either sign of H:

$$|A-1| \le h, \qquad |D| \max_i |x_i| \le h, \qquad A + Dx_i \ge 1 - 2h, \qquad A + D(x_i + x_l) \ge 1 - 3h. \tag{WG5}$$

For details, $|S|/\rho^3 \le 1$, $|q|/\rho \le \sqrt{1+z^2}$, and $|x_i|/\rho \le a_{\max} \le 1$. With $B = z^2$,

$$\frac{|A-1|}{h} \le c(B+3)\sqrt{B(1+B)}\varphi(\sqrt{B}) \le \frac{(8 + 4\sqrt{5})e^{-\sqrt{5}/2}}{\sqrt{12\pi}} < 1,$$

$$\frac{|D|\max_i|x_i|}{h} \le 3c\sqrt{B(1+B)}\varphi(\sqrt{B}) \le 6c\varphi(1) < 1. \tag{WG6}$$

The first maximum follows by differentiating $(B+1)(B+3)e^{-B/2}$; its maximum is at $B=\sqrt{5}$. The second uses the maximum of $(B+1)\varphi(\sqrt{B})$ at $B=1$.

Consequently the pure map ($E=0$) preserves every coordinate sign and every strict weighted-coordinate order, even in an adverse head phase. It preserves both original extrema, the weighted maximum and minimum. For a forced trajectory there is the exact identity

$$\Delta_{il,n} = G_{il,n}\left\{\Delta_{il,0} + \sum_{v=0}^{n-1}\frac{\eta_{i,v}-\eta_{l,v}}{G_{il,v+1}}\right\}, \quad G_{il,n} = \prod_{v<n}[A_v + D_v(x_{i,v}+x_{l,v})] > 0, \tag{WG7}$$

where $\Delta_{il} = x_i - x_l$. Thus the absolute sum of the displayed force debits being less than $|\Delta_{il,0}|$ preserves its sign. The same formula for a single coordinate uses factors $A_v + D_v x_{i,v}$ and forces $\eta_{i,v}$. These are actual-history identities, including the final candidate step, not reference-trajectory approximations. At a zero spatial row the pure cubic spatial gradient is zero; the ratio statements below begin only after $\rho > 0$.

R.2 Permanent marking acquires a fixed orientation

Suppose on a considered phase the actual row has $Q \ge q_0 > 0$ and $\nu \le q_0$. Then q, b and C are nonzero. The sign of q persists through its candidate update: since $|C| \le \|\theta\|$, $H \ge q_0 V$ implies $|q| \ge q_0\|\theta\|$, and the head force has relative magnitude at most $h\nu/q_0 \le h$.

The exact pure bias candidate is

$$b^{\mathrm{t}} = b\left[1 + \frac{hH}{\rho^2}\frac{1-z^2}{z^2}\right]. \tag{WG8}$$

For $z^2 \le 1$ its multiplier is at least one; for $z^2 > 1$ it is at least $1 - h/4$, using the cubic balanced bound $0 \le H \le \rho^2/4$. Also

$$H \le c\varphi(0)|qb|, \qquad \frac{|q|}{|b|} \le \frac{c\varphi(0)}{q_0}. \tag{WG9}$$

The bias-force magnitude divided by $|b|$ is therefore at most $hc\varphi(0)\nu/q_0 < h/4$, so b keeps its sign as well. It follows that the orientation $\epsilon = \operatorname{sign}(qF) = -\operatorname{sign}(qb)$ is fixed throughout any such phase, including candidate endpoints. Put

$$y_i = \epsilon x_i, \qquad D_+ = |D| \ge 0. \quad \text{Then} \quad y_i^+ = Ay_i + D_+ y_i^2 + \epsilon\eta_i, \qquad \sum_i y_i (u_i^T w)^2 > 0. \tag{WG10}$$

Thus at least one oriented coordinate is positive. The phase does not require an initial favorable sign for the row.

R.3 An exact winner clock, negative coordinates and force tolerance

Fix a marked starting integer J with unique oriented maximum $R_J = y_{*,J} > 0$. At that state define

$$r_{i,J} = y_{i,J}/R_J, \qquad \gamma = 1 - \max\{0, \max_{i\ne *} r_{i,J}\} > 0, \qquad M_0 = \max\{1, a_{\max}\rho_J/R_J\}. \tag{WG11}$$

For $r = 1$ take the maximum over competitors to be zero, so $\gamma = 1$. Negative competitors can have arbitrarily large magnitude; they are included in M_0. Define, along the actual candidate prefix,

$$R_n = y_{*,n}, \quad \beta_n = D_{+,n}R_n, \quad \sigma_n = \frac{\beta_n}{A_n + 2\beta_n}, \quad e_n = \frac{a_{\max}h|q_n|\|E_{w,n}\|}{R_n}. \tag{WG12}$$

The conditions below are interpreted by candidate induction while $R_n > 0$. It is sufficient to establish the public or actual-history budget

$$\sum_{n=J}^{N-1} e_n \le \frac{\gamma}{4(M_0+2)}. \tag{WG13}$$

Lemma R.1 (Forced winner and odds contraction)**.** *Under WG2, the fixed marked orientation, and WG11–13, $R_n > 0$, the same coordinate is the unique winner, and every competitor has $r_{i,n} \le 1 - \gamma/2$ through N. Its sign need not be fixed. Let $v_n = P_{u_*^\perp} w_n/(u_*^T w_n)$ and define a transformed transverse vector by leaving its $U^\perp$ components unchanged and replacing its component $v_i = (a_*/a_i)r_i$ by*

$$\mathcal{T}_i(v_i) = \begin{cases} v_i/(1-r_i), & r_i \ge 0, \\ v_i, & r_i < 0. \end{cases} \qquad \Psi_n = \|\mathcal{T}(v_n)\|. \tag{WG14}$$

Then, for $K_\gamma = 8(M_0+2)/\gamma^2$,

$$\Psi_{n+1} \le (1-\sigma_n)\Psi_n + K_\gamma e_n, \qquad \Psi_J \le M_0/\gamma, \tag{WG15}$$

$$\Psi_N \le \Psi_J e^{-\sum_{n=J}^{N-1}\sigma_n} + K_\gamma \sum_{n=J}^{N-1} e_n e^{-\sum_{v=n+1}^{N-1}\sigma_v}, \quad \|w_N/\rho_N - \epsilon u_*\| \le \sqrt{2}\,\Psi_N. \tag{WG16}$$

For a pure phase this gives a $1/\gamma$ prefactor and the same clock rate for positive and negative competitors; it does not cost $\exp(1/\gamma)$. If the clock diverges, the pure spatial direction converges to the winning signed axis. No convergence is asserted when the clock stays finite.

Proof. Write $L = A + \beta$. The pure ratio map at the actual pre-state is

$$r_i^{\mathrm{t}} = \frac{r_i(A+\beta r_i)}{A+\beta}. \tag{WG17}$$

By WG5 all numerator multipliers are positive. When $r_i < 0$ its absolute value contracts by at least $A/(A+\beta)$; when $0 \le r_i \le 1$ it remains between zero and r_i. The outside spatial ratio contracts by exactly $A/(A+\beta)$. The actual ratio differs from WG17 by at most $2(M_0+2)e_n$ in each weighted-coordinate ratio and in the Euclidean norm of v, as long as their current magnitudes are at most $M_0 + 1$. Indeed the denominator is $L + \epsilon\eta_*/R$, at least $1 - h - e_n > 1/2$, and quotient subtraction gives the numerator "force minus pure ratio times winner force." For the vector v, the numerator-force norm is at most e_n, since $a_* \le a_{\max}$. Budget WG13 therefore bounds their total possible increase by $\gamma/2 \le 1/2$. Induction closes the asserted magnitude bounds, $r_i \le \max(0, r_{i,J}) + \gamma/2 \le 1 - \gamma/2$, and positivity of the winner denominator at every candidate step.

For $0 \le r_i < 1$, the exact positive odds satisfy

$$\frac{r_i^{\mathrm{t}}}{1-r_i^{\mathrm{t}}} = \frac{r_i}{1-r_i}\frac{A+\beta r_i}{A+\beta(1+r_i)} \le (1-\sigma)\frac{r_i}{1-r_i}. \tag{WG18}$$

Negative components and outside components contract at least this fast. On the invariant domain $r_i \le 1 - \gamma/2$, the map $\mathcal{T}$ has Euclidean Lipschitz constant at most $4/\gamma^2$; the derivative of $v_i/(1 - (a_i/a_*)v_i)$ is $(1-r_i)^{-2}$, and the negative-side derivative is one. It is continuous at zero, so force-induced sign changes cause no exception. Applying this bound to the preceding quotient-force estimate proves WG15. Iteration proves WG16. Finally $\|v\| \le \Psi$ and elementary normalization give the signed-axis distance bound. □

R.4 Head-weighted transverse control with a rowwise force sum

Suppose additionally, at every pre-state of this phase,

$$V_n \ge \frac{9}{4}\rho_n^2. \tag{WG19}$$

This is an acquired large-row interface, not a preferred-axis cone. BH23 and BH27 supply it for every row after its first joint-energy crossing $V \ge 512$, provided initially $V_0 < 5$ and the marking budgets above hold. Indeed marked entry energy is below $40/7$, $V \le [(4+h)\rho^2 + 80/7]/(1 - 6\nu T_0)$, and $V \ge 512$ forces $\rho^2 > 96$. Marked energy is increasing (BH19), so this continues to hold. BH27 then gives $V/\rho^2 \ge 1 + 1/5 + 9/8 = 93/40 > 9/4$. The first candidate crossing has $V_J < 517$, since $\|\nabla H + p\| \le 2\sqrt{V}$ implies $V^+ \le (1+2h)^2 V$. Rows that never cross retain $|q|\|P_{u^\perp} w\| \le V/2 < 256$ for any axis u.

Put $Z_n = V_n - 4hH_n = (1 - 4hQ_n)V_n$. The exact compensated energy bound BH7a gives

$$Z_{n+1} - Z_n \le 4hH_n + 3h\nu V_n. \tag{WG20}$$

Because every oriented coordinate is at most R_n,

$$hH_n = \frac{\beta_n}{3R_n} \sum_i y_{i,n}(u_i^T w_n)^2 \le \frac{\beta_n \rho_n^2}{3}. \tag{WG21}$$

Thus $4hQ_n \le 16\beta_n/27$ under WG19, and

$$\log(Z_{n+1}/Z_n) \le \frac{4hQ_n + 3h\nu}{1 - 4hQ_n} \le \frac{3}{5}\beta_n + 4h\nu. \tag{WG22}$$

Here $0 < Q_n \le 1/2$ and $h \le 1/512$; in particular $(16/27)/(1 - 2h) < 3/5$. Also $\beta_n \le h$, $A_n \le 1 + h$, so $\sigma_n \ge \beta_n/(1 + 3h)$ and $1/(1 + 3h) - 3/5 > 3/8$. Combining WG15 and WG22 gives

$$Z_{n+1}\Psi_{n+1} \le e^{-3\beta_n/8 + 4h\nu} Z_n \Psi_n + K_\gamma Z_{n+1} e_n. \tag{WG23}$$

Since $|q|\|P_{u_*^\perp} w\| \le V\Psi/2$ and $(1 - 2h)V \le Z \le V$, iteration proves the explicit estimate

$$\begin{aligned} |q_N|\|P_{u_*^\perp} w_N\| \le \frac{e^{4\nu(N-J)h}}{2(1-2h)} \Big\{ & V_J \Psi_J e^{-\frac{3}{8}\sum_{v=J}^{N-1} \beta_v} \\ & + K_\gamma \sum_{n=J}^{N-1} V_{n+1} e_n e^{-\frac{3}{8}\sum_{v=n+1}^{N-1} \beta_v} \Big\}. \end{aligned} \tag{WG24}$$

In a pure phase this is uniformly at most $V_J M_0/[2(1 - 2h)\gamma]$ and decreases with the winner clock. The forced result keeps each row's actual energy and force; it has not substituted the full-bank M^2 envelope. Obtaining a uniform small numerical bound from that convolution still requires paying the displayed force terms. It is not free from WG2 alone.

To connect to the complete signed frame the cubic derivative frame, write its normalized row coefficient as $t = -(1 - \alpha)s^2 q\rho z\varphi(z)/(\sqrt{2}m)$. For a marked winning orientation, $t\epsilon = |t|$. Since $\sup_z |z|\varphi(z) = \varphi(1)$ and the positive winning hemisphere gives $\|w/\rho - \epsilon u_*\| \le \sqrt{2}\|P_{u_*^\perp} w\|/\rho$, rank-one expansion gives

$$\left\| t\, v(v^{\odot 2})^T - |t| e_* e_*^T \right\|_{\mathrm{op}} \le 3|t|\|w/\rho - \epsilon u_*\| \le \frac{3(1-\alpha)\varphi(1)s^2}{m} |q|\|P_{u_*^\perp} w\|, \quad v = U^T w/\rho. \tag{WG25}$$

Thus WG24 can be summed over every row for which its gap transport and force budget are proved. No original row is removed from the complete signed frame (FA25). Unmarked and low-energy rows remain explicitly bounded by their energy or retained exactly.

Inherited actual run and explicit additional costs. Use the complete normalized cubic teacher, full retained target, initial law and simultaneous raw GD of Appendix B.2:

$$\begin{gathered} y_c = \sum_{i=1}^r a_i h_3(u_i^T X), \quad a_i > 0, \quad \sum_i a_i^2 = 1, \\ a_0 = \min_i a_i, \quad a_{\max} = \max_i a_i, \quad U = \mathrm{span}\{u_i\}. \end{gathered} \tag{FA1}$$

All original raw coordinates, including heads and biases, remain independent $N(0, s^2/d)$, and $(A_j, W_j, B_j) = s(q_j, w_j, b_j)$. The step is one fixed common raw step $\eta = mh/2$ throughout. For each original row the exact actual recurrence is

$$\begin{gathered} q^+ = q + h(C + \theta^T E), \qquad \theta^+ = \theta + hq(\nabla C + E), \qquad \theta = (w, b), \\ C = \mathbb{E}[y_c \sigma_\alpha(\theta^T(X, 1))], \qquad E_j = \mathbb{E}[(y - y_c - f)\sigma'_{\alpha,j}(X, 1)]. \end{gathered} \tag{FA2}$$

Here $f = s^2 m^{-1} \sum_j q_j \sigma_\alpha(w_j^T X + b_j)$ is the complete current student. The force includes every student self/cross interaction and all retained teacher terms. It is not a frozen-Gram, teacher-only or reference update.

Fix AC's original confidence parameter δ_{AC} and an additional $0 < \zeta < 1$. Its existing Gaussian quantities H, d, m and coefficients define

$$\tau = \frac{\zeta a_0}{4mr^2}, \qquad g_0 = \frac{\tau}{\sqrt{d}}, \qquad b_0 = a_{\max}\sqrt{H/d}, \qquad D_a = \sum_i a_i^{-2}, \qquad L_a = (a_{\max}^2 D_a)^{1/3},$$
$$Z_* = 40 + 640\sqrt{D_a}\left\{\frac{b_0^2}{g_0} + (1+L_a)b_0\right\}. \tag{FA3}$$

These are public deterministic constants. In particular $g_0 \le 1/4$, since $d, m, r \ge 1$ and $a_0, \zeta \le 1$. One may first enlarge the public milestone R_M to any value at least $\max(R_D, 16\sqrt{m})$, then define AC's $T, \overline{T}, M$ and all its subsequent budgets from that chosen value. The AC and CW proofs use only this lower bound on R_M, and their displayed definitions of $T, \overline{T}, M$; their conclusions therefore continue to hold with this consistent replacement. In addition to all the AC cutoffs require

$$0 < h \le \frac{1}{16\overline{T}}, \qquad 0 < \nu \le \frac{g_0 e^{-4\overline{T}}}{10^5 \overline{T} M}. \tag{FA4}$$

Here ν is decreased before specifying the retained neighborhood $\|y - y_c\|_{H^1} \le \nu/2$ and the original scale s in AJ10–11. The actual AC force closure still gives $\|E_j\| \le \nu$ on $0 \le n \le N = \lceil T/h \rceil$; no force is removed from FA2. The complete run satisfies

$$Nh \le \overline{T}, \quad h \le 1/512, \quad \nu \le 1, \quad |q_j| \le \|\theta_j\|, \quad \mathcal{V}_n := \sum_j (q_j^2 + \|\theta_j\|^2) < M^2/4,$$
$$|q_{j,n}|\, O_{j,n} < 40, \qquad O_{j,n} := \|P_{U^\perp} w_{j,n}\|. \tag{FA5}$$

The last inequality is the complete-row conclusion CW16. Decreasing h and ν preserves all its hypotheses.

The simultaneous original-coordinate event. Write the original projected Gaussian coordinates as $u_i^T w_{j,0} = G_{ji}/\sqrt{d}$ and set $x_{ji} = a_i u_i^T w_j$. For any one coordinate and distinct pair, Gaussian density bounds give

$$\mathbb{P}(|x_{ji,0}| \le g_0) \le \tau/a_0, \qquad \mathbb{P}(|x_{ji,0} - x_{jl,0}| \le g_0) \le \tau/(a_0\sqrt{\pi}). \tag{FA6}$$

Indeed the second Gaussian has standard deviation $\sqrt{a_i^2 + a_l^2}/\sqrt{d} \ge \sqrt{2}a_0/\sqrt{d}$. Union bounding the mr coordinates and the $mr(r-1)/2$ unordered gaps shows that their total failure probability is at most

$$\frac{\zeta}{4r} + \frac{\zeta(r-1)}{8r\sqrt{\pi}} < \frac{\zeta}{2}. \tag{FA7}$$

Intersect this event with the unchanged AC event. On that intersection, simultaneously for all rows,

$$|x_{ji,0}| > g_0, \quad |x_{ji,0} - x_{jl,0}| > g_0 \ (i \ne l), \quad |x_{ji,0}| \le b_0, \quad O_{j,0}^2 \ge \frac{1}{4} - \frac{H}{d} \ge \frac{3}{16}. \tag{FA8}$$

The upper bound and outside lower bound use AC's existing $\|P_U w_{j,0}\|^2 \le H/d$, $\rho_{j,0} \ge 1/2$ and $d \ge 16H$. No later conditional Gaussian distribution is used. The intersection has original probability at least $1 - \delta_{\mathrm{AC}} - \zeta/2$; for total confidence $1-\delta$, one can take $\delta_{\mathrm{AC}} = \delta/2$ and $\zeta = \delta$. These are events in the original law, not sampling filters.

Exact scalar dynamics through every adverse phase. Fix one original row and omit its row subscript. Put $\rho = \|w\|$, $z = b/\rho$, $F(z) = -(1-\alpha)z\varphi(z)/\sqrt{6}$, and $\mathsf{H} = qC$. As long as $\rho > 0$, WG4–5 give the exact simultaneous coordinate and outside recurrences

$$x_i^+ = Ax_i + Dx_i^2 + \eta_i, \qquad \eta_i = hqa_i u_i^T E_w, \qquad |\eta_i| \le hM\nu,$$
$$A = 1 + (z^2-3)h\mathsf{H}/\rho^2, \qquad D = 3hqF(z)/\rho^2, \qquad w_\perp^+ = Aw_\perp + hqE_\perp, \tag{FA9}$$
$$|A-1| \le h, \quad |D| \max_i |x_i| \le h, \quad A + Dx_i \ge 1 - 2h, \quad A + D(x_i + x_l) \ge 1 - 3h.$$

These bounds are sign-free. In particular they hold before any favorable phase or permanent marking. All scalar multipliers in FA9 are positive. Define the actual-history linear anchor

$$\Lambda_0 = 1, \qquad \Lambda_n = \prod_{v<n} A_v. \quad \text{Then} \quad \Lambda_n \ge e^{-2nh} \ge e^{-2\overline{T}}. \tag{FA10}$$

The bound follows from $\log(1-h) \geq -h/(1-h) \geq -2h$. Similarly every product of coordinate or pair-gap factors in FA9 is at least e^{-4nh}, since $\log(1-3h) \geq -3h/(1-3h) \geq -4h$.

For a coordinate or difference, write the exact variation formula along the actual history, as in WG7:

$$\begin{aligned} x_{i,n} &= P_{i,n}\left(x_{i,0} + \sum_{v<n} \frac{\eta_{i,v}}{P_{i,v+1}}\right), \\ P_{i,n} &= \prod_{v<n}(A_v + D_v x_{i,v}), \\ x_{i,n} - x_{l,n} &= P_{il,n}\left(x_{i,0} - x_{l,0} + \sum_{v<n} \frac{\eta_{i,v} - \eta_{l,v}}{P_{il,v+1}}\right), \\ P_{il,n} &= \prod_{v<n}[A_v + D_v(x_{i,v} + x_{l,v})]. \end{aligned} \tag{FA11}$$

Each total absolute debit inside parentheses is at most $2\overline{T}M\nu e^{4\overline{T}} \leq 2g_0/10^5 < g_0/2$. Consequently every original sign and strict order is preserved through every candidate endpoint, with the uniform bounds

$$|x_{i,n}| \geq \frac{g_0}{2}e^{-4\overline{T}}, \qquad |x_{i,n} - x_{l,n}| \geq \frac{g_0}{2}e^{-4\overline{T}}. \tag{FA12}$$

No transported favorable-entry gap is assumed.

For completeness these statements are not circular at $\rho = 0$. Let $e_0 = w_{\perp,0}/O_0$, which exists by FA8. Up to any first candidate exit one has the exact projection identity

$$\langle w_{\perp,n}, e_0\rangle = \Lambda_n\left\{O_0 + \sum_{v<n} \frac{hq_v\langle E_{\perp,v}, e_0\rangle}{\Lambda_{v+1}}\right\}. \tag{FA13}$$

The absolute sum is at most $\overline{T}M\nu e^{2\overline{T}} \leq g_0 e^{-2\overline{T}}/10^5 < 1/8 < O_0/2$. Thus induction closes $\rho_n \geq O_n > 0$ and proves

$$O_n \geq \frac{O_0}{2}\Lambda_n, \qquad |q_n|\Lambda_n < \frac{80}{O_0} < 320. \tag{FA14}$$

The second bound uses the full actual weighted outside estimate in FA5, including its compatible head-force payment.

A discrete pair anchor with only a second-order pure debit. For any distinct coordinates define the positive quantity

$$K_{il,n} = \frac{|x_{i,n}x_{l,n}|}{|x_{i,n} - x_{l,n}|\Lambda_n}. \tag{FA15}$$

Its denominators never vanish by FA10–12. From the current actual pre-state form just the algebraic teacher intermediate $x_i^{\mathrm{t}} = x_i(A + Dx_i)$ and $\Lambda^+ = A\Lambda$; this does not define another trajectory. Factoring the difference gives exactly

$$\frac{K_{il}^{\mathrm{t}}}{K_{il}} = \frac{(A + Dx_i)(A + Dx_l)}{A[A + D(x_i + x_l)]} = 1 + \frac{D^2 x_i x_l}{A[A + D(x_i + x_l)]}. \tag{FA16}$$

All factors are positive. If the pair has opposite signs, this ratio is at most one. If it has the same sign, it is at most $1 + 3h^2$: the numerator of the last fraction is at most h^2 and its denominator is at least $(1-h)(1-3h) > 1/3$. Both statements allow either sign of D, so include all adverse head/bias phases and their reversals.

The additive force debit has no hidden second trajectory or uncontrolled Hessian term. At both endpoints of the segment from $(x_i^{\mathrm{t}}, x_l^{\mathrm{t}})$ to (x_i^+, x_l^+), every coordinate and their difference has the same sign as originally. By FA9 and FA12 their absolute values throughout this straight segment are at least $(g_0/3)e^{-4\overline{T}}$. Along the segment,

$$d\log\frac{|x_i x_l|}{|x_i - x_l|} = \frac{dx_i}{x_i} + \frac{dx_l}{x_l} - \frac{dx_i - dx_l}{x_i - x_l}. \tag{FA17}$$

Integrating this exact derivative, and using $|\eta_i|, |\eta_l| \le hM\nu$, yields

$$|\log K_{il}^{+} - \log K_{il}^{\mathrm{t}}| \le \frac{12hM\nu e^{4\overline{T}}}{g_0}. \tag{FA18}$$

The anchor is the same $A\Lambda$ for the intermediate and actual endpoint, so it creates no extra term here. Iteration of FA16–18 proves

$$\begin{aligned} K_{il,n} &\le K_{il,0} \exp\left(3nh^2 + \frac{12\overline{T}M\nu e^{4\overline{T}}}{g_0}\right) < 2K_{il,0}, \\ K_{il,n} &\le K_{il,0} \exp\left(\frac{12\overline{T}M\nu e^{4\overline{T}}}{g_0}\right) < 2K_{il,0} \qquad (x_{i,0}x_{l,0} < 0). \end{aligned} \tag{FA19}$$

Indeed $3nh^2 \le 3/16$ and the remaining exponent is at most $12/10^5$; their sum is below $\log 2$. For an initial same-sign pair $K_{il,0} \le b_0^2/g_0$. For an initial opposite-sign pair, $K_{il,0} = |x_{i,0}x_{l,0}|/(|x_{i,0}| + |x_{l,0}|) \le b_0$. The force-free result is exact up to the favorable $O(h^2)$ discrete defect; the exponential cost appears solely in paying the actual force against original absolute coordinate/gap floors.

Any favorable actual row has an original extremal axis. Consider any one time $n \le N$ at which $\mathsf{H}_n = q_nC_n > 0$. There is no assumption on the row's past sign history. Define $\epsilon = \mathrm{sign}(q_nF(z_n))$ and $y_i = \epsilon x_{i,n}$. Because $C = F\sum_i x_i^3/(a_i^2\rho^2)$,

$$\sum_i \frac{y_i^3}{a_i^2} > 0. \tag{FA20}$$

In particular the oriented maximum $p = y_* > 0$ exists and is unique. Its label is the original maximum of $\epsilon x_{i,0}$ by FA12. The orientation ϵ is only an analysis choice at this actual endpoint; all pair bounds already hold for both orientations and every time, so it need not have been fixed earlier.

If $0 < y_i < p$ is a positive competitor, then

$$y_i \le \frac{py_i}{p - y_i} = \Lambda_n K_{*i,n} \le \frac{2b_0^2}{g_0}\Lambda_n. \tag{FA21}$$

If there are negative coordinates, let $-\ell < 0$ be the most negative one, with axis label l. Positivity in FA20 implies

$$\frac{\ell^3}{a_l^2} < \sum_{y_i>0} \frac{y_i^3}{a_i^2} \le p^3 D_a, \qquad \ell \le L_a p. \tag{FA22}$$

The opposite-sign anchor consequently gives

$$\ell = \left(1 + \frac{\ell}{p}\right)\frac{p\ell}{p+\ell} \le (1+L_a)\Lambda_n K_{*l,n} \le 2(1+L_a)b_0\Lambda_n. \tag{FA23}$$

Every negative competitor has magnitude at most ℓ. This argument covers an arbitrary mixed-sign history. If all oriented coordinates are positive only FA21 is needed; an all-negative orientation is excluded by FA20. If $r = 1$ there are no competitors and the inside transverse norm is zero.

Combining FA21–23, summing squared spatial components and using FA14 proves the uniform complete-row conclusion

$$\begin{aligned} |q_n|\|P_{U\cap u_*^{\perp}} w_n\| &\le 640\sqrt{D_a}\left\{\frac{b_0^2}{g_0} + (1+L_a)b_0\right\}, \\ |q_n|\|P_{u_*^{\perp}} w_n\| &\le Z_* \quad (q_nC_n > 0,\ 0 \le n \le N). \end{aligned} \tag{FA24}$$

The outside term is at most 40 by FA5 and is combined by the triangle inequality. The result is uniform over all original rows, without selecting a favorable subbank. Its deterministic constant is independent of the radius envelope M and of the clock T. In particular $b_0^2/g_0 = a_{\max}^2 H/(\tau\sqrt{d})$ is polynomial in the original Gaussian parameters; no later relative gap occurs.

Every signed frame contribution is retained. Let $v_j = U^T w_j/\rho_j$ and put

$$t_j = -\frac{(1-\alpha)s^2}{\sqrt{2}m} q_j \rho_j z_j \varphi(z_j), \qquad \mathsf{D}_n = \sum_j t_j v_j (v_j^{\odot 2})^T. \tag{FA25}$$

This is the complete signed inside gradient-Hermite frame in the cubic derivative frame, not a Gram matrix made from selected rows. For each favorable row, choose its endpoint winner from FA20 and note $t_j\epsilon_j = |t_j|$ and $\epsilon_j u_*^T w_j > 0$. The hemisphere normalization inequality and a rank-one telescoping expansion give

$$\begin{aligned} \|w_j/\rho_j - \epsilon_j u_*\| &\le \sqrt{2}\,\|P_{u_*^\perp} w_j\|/\rho_j, \\ \|t_j v_j (v_j^{\odot 2})^T - |t_j| e_* e_*^T\|_{\mathrm{op}} &\le 3|t_j|\|w_j/\rho_j - \epsilon_j u_*\| \\ &\le \frac{3(1-\alpha)\varphi(1)s^2}{m} Z_*. \end{aligned} \tag{FA26}$$

Here $\|v_j\| \le 1$ and the squared-coordinate map is 2-Lipschitz on the unit ball; the last line uses $\sup_z |z|\varphi(z) = \varphi(1)$. Thus every favorable row contributes a nonnegative diagonal mass plus an explicitly bounded error. No sign of a head has been chosen or changed by the algorithm.

For any row with $\mathsf{H}_j \le 0$, CW6 implies that it is unmarked, so $V_j < 40/7$. The coarser uniform bound $V_j < 512$ therefore suffices. For each such row,

$$\|t_j v_j (v_j^{\odot 2})^T\|_{\mathrm{op}} \le |t_j| \le \frac{(1-\alpha)\varphi(1)s^2}{\sqrt{2}m}|q_j|\rho_j < \frac{256(1-\alpha)\varphi(1)s^2}{\sqrt{2}m}. \tag{FA27}$$

This also covers $\mathsf{H}_j = 0$: such a row may have a nonzero frame, and it has not been silently discarded. Summing all m row bounds gives the exact full-bank decomposition

$$\begin{aligned} &\mathsf{D}_n = \mathsf{D}_{+,n} + \mathsf{E}_n, \qquad \mathsf{D}_{+,n} = \sum_{\mathsf{H}_{j,n}>0} |t_{j,n}| e_{*,j} e_{*,j}^T \succeq 0, \\ &\|\mathsf{E}_n\|_{\mathrm{op}} \le (1-\alpha)\varphi(1)s^2(3Z_* + 182), \qquad 0 \le n \le N. \end{aligned} \tag{FA28}$$

The diagonal may have repeated winner axes; coverage of every axis must come from the separately acquired AC categories. Together with their positive diagonal lower bounds and CW's complete outside AGOP estimate, FA28 is a paid interface for a full spectral theorem. Such a theorem must still state its public radius choice, actual eigenvalue floor, gap and angle comparison; none is inferred merely from a count of winners here.

S THE IDENTICAL PROJECTED-BANK DIAGNOSTIC

S.1 EXACT INPUTS AND THE DIAGNOSTIC BEING COMPARED

Use the normalized additive cubic and full target of AC. Here $X \sim N(0, I_d)$, $h_3(t) = (t^3 - 3t)/\sqrt{6}$, $\sigma_\alpha(t) = \alpha t + (1-\alpha)t_+$, and φ, Φ are the standard normal density and CDF. The imported $R_M = R$, ξ, b_* and ν are the public constants in AJ3 and AJ7–9; $\|g\|_2$ denotes its Gaussian L^2 norm. Thus

$$y_c = \sum_{i=1}^r a_i h_3(u_i^T X), \quad a_i > 0, \quad \sum_i a_i^2 = 1, \quad U^T U = I_r, \quad \|y\|_2 = 1, \quad \|y - y_c\|_{H^1} \le \nu/2. \tag{PJ1}$$

Here U denotes the matrix of teacher axes and $P_U = UU^T$. All nonzero coefficient signs may be absorbed into those axes. The target includes its actual mean, linear part and all higher tails, as in AC; the current result does not extend AC's acquired dictionary to an unrestricted H^1 teacher. At the same fixed public integer N in AJ8, choose for analysis one of the acquired representatives for each $(i, \epsilon, \tau) \in \{1, \dots, r\} \times \{-1, 1\}^2$. There are $4r$ distinct original rows. For each selected row,

$$\begin{aligned} (A_j, W_j, B_j) = s(q_j, w_j, b_j), \quad \rho_j = \|w_j\| \ge R_M > 7, \quad n_j = w_j/\rho_j, \quad z_j = b_j/\rho_j, \\ \|(n_j, z_j) - (\epsilon u_i, \tau b_*)\| \le \xi. \end{aligned} \tag{PJ2}$$

These are acquired conclusions of AC21–25; all unchosen rows remain in training, in the diagnostic bank and in the AGOP.

Let $G_t = \mathbb{E}[\nabla f_t \nabla f_t^T]$ be the AGOP of the actual full original student at $t = 0, N$. Assume $r < d$ and

$$\lambda_r(G_N) > \lambda_{r+1}(G_N) \geq 0, \qquad \lambda_{\min}(U^T P_N U) \geq \mu > 0,$$
$$P_N = \text{the leading rank-}r \text{ projector of } G_N. \tag{PJ3}$$

Thus the r modes are nonzero and the boundary is genuine. At initialization the same projector P_0 exists and is unique almost surely by IA, for every $m \geq r$, including $m \geq d$.

Fix one public common multiplier $\kappa > 0$ and define at both times

$$\psi_{j,t}^P(X) = \kappa \sigma_\alpha((P_t W_{j,t})^T X + B_{j,t}), \qquad B_\kappa = \frac{C_\alpha}{7\kappa s},$$
$$\mathcal{R}_{B_\kappa}^P(t) = \inf_{\|v\|_2 \leq B_\kappa} \left\| y - \sum_{j=1}^m v_j \psi_{j,t}^P \right\|_2^2, \qquad \mathcal{R}_\infty^P(t) = \inf_{v \in \mathbb{R}^m} \left\| y - \sum_j v_j \psi_{j,t}^P \right\|_2^2. \tag{PJ4}$$

There is no re-normalization of $P_t W_{j,t}$ after projection. The choice $\kappa = 1$ is AC's raw-feature convention. No diagnostic coefficient changes the trained heads.

S.2 The four exact coefficients and cancellation of large radii

Recall the exact AJ3 constants

$$b_*^2 = (\sqrt{5} - 1)/2, \quad p = 2\Phi(b_*) - 1, \quad g(z) = \mathrm{clip}(z, -b_*, b_*) - pz,$$
$$v_g = p - 2b_*\varphi(b_*) + b_*^2(1 - p) - p^2, \quad t_g = -2b_*\varphi(b_*)/\sqrt{6}, \quad K_* = t_g^2/v_g > 2/3,$$
$$D_\alpha = \sqrt{(1-\alpha)^{-2} + p^2/(1+\alpha)^2}, \qquad C_\alpha = |t_g| D_\alpha / v_g. \tag{PJ5}$$

For $\epsilon, \tau \in \{-1, 1\}$ set

$$c_{\epsilon,\tau} = \frac{\epsilon}{2}\left(\frac{\tau}{1-\alpha} - \frac{p}{1+\alpha}\right), \qquad d_{i,\epsilon,\tau} = (t_g/v_g) a_i c_{\epsilon,\tau}. \tag{PJ6}$$

The identities

$$\sum_{\epsilon,\tau} c_{\epsilon,\tau} \sigma_\alpha(\epsilon z + \tau b_*) = g(z), \qquad \sum c_{\epsilon,\tau}^2 = D_\alpha^2,$$
$$\|d\|_2 = C_\alpha, \qquad \|d\|_1 = \frac{2|t_g|}{v_g(1-\alpha)} \sum_i a_i \leq 2C_\alpha \sqrt{r} \tag{PJ7}$$

follow by taking the odd differences of the two biased activations: their sum is $2(1+\alpha)z$, and their difference is $2(1-\alpha)\,\mathrm{clip}(z, -b_*, b_*)$. The ideal readout is therefore

$$F(X) = \sum_{i,\epsilon,\tau} d_{i,\epsilon,\tau} \sigma_\alpha(\epsilon u_i^T X + \tau b_*) = \frac{t_g}{v_g} \sum_i a_i g(u_i^T X), \qquad \|y_c - F\|_2^2 = 1 - K_*. \tag{PJ8}$$

The last equality uses that g is odd, its variance is v_g, and $\langle h_3, g \rangle = t_g$. Independence across teacher axes and $\sum a_i^2 = 1$ then prove it for the full cubic.

On each selected row use the coefficient

$$v_j = \frac{d_{i,\epsilon,\tau}}{\kappa s \rho_j}, \quad v_j = 0 \text{ on all other rows}, \qquad \|v\|_2 \leq \frac{C_\alpha}{\kappa s R_M} < B_\kappa. \tag{PJ9}$$

Positive homogeneity gives the exact equality

$$v_j \psi_{j,N}^P(X) = d_{i,\epsilon,\tau} \sigma_\alpha((P_N n_j)^T X + z_j). \tag{PJ10}$$

Thus neither the largest raw radius nor an M-sized full-bank norm enters the error or the budget. The denominator is the original radius ρ_j, not the radius after projection. Since the activation is 1-Lipschitz and P_N is an orthogonal projector, PJ2, PJ7 and the Gaussian identity $\mathbb{E}[(v^T X + b)^2] = \|v\|^2 + b^2$ imply

$$\left\| \sum_j v_j \psi_{j,N}^P - F(P_N X) \right\|_2 \leq \|d\|_1 \xi \leq 2C_\alpha \sqrt{r}\, \xi \leq 1/64. \tag{PJ11}$$

Biases are unchanged throughout this calculation. Zeros on unchosen diagnostic coefficients are a legal witness in the original full bank, not a training subbank or a different before-time comparator.

An elementary projection estimate, useful without the structure of g, is

$$\|F(P_N X) - F(X)\|_2 \le 2C_\alpha \sqrt{r}\sqrt{1-\mu}. \tag{PJ12}$$

Indeed $\|(I-P_N)u_i\|^2 \le 1-\mu$ in every teacher direction, and one can sum the individual activation errors using PJ7. It already suffices to impose $1-\mu \le (1024 C_\alpha^2 r)^{-1}$ to make PJ12 at most $1/16$.

S.3 A DIMENSION-FREE DISTORTION BOUND FOR THE STRUCTURED WITNESS

The following refinement avoids the rank cost in PJ12. Put

$$L_g = \max\{p, 1-p\}, \qquad D_g = |t_g| L_g / v_g, \qquad Q = U^T P_N U, \quad \mu I \preceq Q \preceq I. \tag{PJ13}$$

Here L_g is a global Lipschitz constant of g, and D_g is independent of r and α. Define independent Gaussian vectors $Z = U^T P_N X$, $W = U^T(I - P_N)X$, with covariances Q and $I - Q$. For $H(z) = \sum_i a_i g(z_i)$, conditional Gaussian Poincare yields

$$\mathbb{E}[\mathrm{Var}(H(Z+W) \mid Z)] \le \mathbb{E}[\nabla H(Z+W)^T (I-Q) \nabla H(Z+W)] \le (1-\mu) L_g^2, \tag{PJ14}$$

because $\|\nabla H\|^2 \le L_g^2 \sum_i a_i^2 = L_g^2$. This includes singular $I - Q$ by restriction to its Gaussian support. The Gaussian Poincare inequality itself follows by expanding a function of standard Gaussian coordinates into Hermites: its variance sums the nonconstant squared coefficients, whereas its derivative energy weights each such coefficient by its degree.

To bound the conditional mean error, let $q_i = Q_{ii} \ge \mu$ and

$$h_i(z) = \mathbb{E}_T[g(z + \sqrt{1-q_i}\, T)] - g(z), \quad T \sim N(0,1), \quad \|h_i(Z_i)\|_2^2 \le L_g^2 (1-q_i). \tag{PJ15}$$

The last bound is Jensen followed by the Lipschitz inequality. Each h_i is odd, so $\mathbb{E}[h_i(Z_i)] = 0$. For $V_i = Z_i/\sqrt{q_i}$, the correlation matrix of the standardized coordinates of Z is $R = \mathrm{diag}(q_i)^{-1/2} Q\, \mathrm{diag}(q_i)^{-1/2}$, with $\|R\| \le 1/\mu$. For any centered square-integrable functions of these coordinates, their Hermite expansions give

$$\mathbb{E}\left[\left(\sum_i a_i h_i(Z_i)\right)^2\right] \le \|R\| \sum_i a_i^2 \|h_i(Z_i)\|_2^2 \le \frac{1-\mu}{\mu} L_g^2. \tag{PJ16}$$

Here is the matrix detail. At degree $k \ge 1$, the covariance is the quadratic form of the coefficient vector against $R^{\odot k}$, the entrywise kth power of R. The identity for two coordinates follows by comparing coefficients in $\mathbb{E}[e^{tV_i - t^2/2} e^{zV_j - z^2/2}] = e^{R_{ij} tz}$. For a correlation matrix R, Schur multiplication preserves order and sends the identity to itself. Thus $\|R^{\odot k}\| \le \|R\|$ by induction from $R^{\odot(k-1)} \preceq \|R\| I$. Sum the degree-wise bound using Parseval. Centering is valid here because these *analytical error functions* are odd; no target or data-centering operation is imposed.

The noise about the conditional mean is orthogonal to its conditional mean error, and $\mathbb{E}[H(Z+W) \mid Z] - H(Z) = \sum_i a_i h_i(Z_i)$. Combining PJ14–16 therefore gives the exact sufficient bound

$$\|F(X) - F(P_N X)\|_2 \le D_g \sqrt{(1-\mu)(1+1/\mu)}. \tag{PJ17}$$

The argument holds for the observed P_N, however it depends on the full trained bank and teacher. No later Haar law or independence between P_N and the learned rows is used.

Let

$$\Delta(\mu) = \min\{2C_\alpha \sqrt{r(1-\mu)}, D_g \sqrt{(1-\mu)(1+1/\mu)}\}. \tag{PJ18}$$

The full endpoint risks then satisfy

$$\mathcal{R}^P_{B_\kappa}(N),\ \mathcal{R}^P_\infty(N) \le \left(\sqrt{1-K_*} + 3/128 + \Delta(\mu)\right)^2. \tag{PJ19}$$

The costs are the ideal cubic error, PJ11, the full original perturbation $\|y - y_c\|_2 \le \nu/2 \le 1/128$, and PJ17 or PJ12. No selected-head replacement of the actual network is used to define P_N.

S.4 Initial top-r baseline without a width ceiling

For ordinary unscreened IID Gaussian raw spatial rows, independent nondegenerate Gaussian heads and biases, IA proves that P_0 is a Haar rank-r projector whenever $m \geq r$ and $r < d$. Its genuine nonzero boundary is almost surely simple, also when $m \geq d$. For the fixed normalized cubic in PJ1,

$$\mathbb{E}_{\text{init}}\left[\|\mathbb{E}_X[y_c \mid P_0X]\|_2^2\right] = q_{r,d} := \frac{r(r+2)(r+4)}{d(d+2)(d+4)}. \tag{PJ20}$$

For a fixed projector the projected Hermite inner product is $(u_i^T P_0 u_j)^3$; off-diagonal Haar averages vanish by a coordinate reflection, and diagonal averages are the third beta moment. This proves PJ20 without independence of the initialized projector and the initial feature bank. Fix $0 < \delta_I < 1$ and write $e = y - y_c$, $m_y = \mathbb{E}[y]$, and $e^\circ = e - \mathbb{E}[e]$. By Markov, full conditional-expectation contraction, and the identity $\|e\|_2^2 = (\mathbb{E}[e])^2 + \|e - \mathbb{E}[e]\|_2^2$, with probability at least $1 - \delta_I$ under the original draw,

$$\begin{aligned}\|\mathbb{E}[y \mid P_0X]\|_2^2 &\leq m_y^2 + \left(\sqrt{q_{r,d}/\delta_I} + \|e^\circ\|_2\right)^2 \leq \left(\sqrt{q_{r,d}/\delta_I} + \nu/2\right)^2, \\ \mathcal{R}^P_{B_\kappa}(0),\ \mathcal{R}^P_\infty(0) &\geq 1 - \left(\sqrt{q_{r,d}/\delta_I} + \nu/2\right)^2.\end{aligned} \tag{PJ21}$$

Every initial projected feature, with its retained bias, is measurable in P_0X; the lower bound applies even to the larger class of all measurable functions of that input. The same full target and the same complete projected bank occur at both times. The full teacher mean was charged, not deleted. The rank in PJ20 is r, not the width m.

Consequently if $q_{r,d} \leq \delta_I/32$ and $\nu/2 \leq 1/128$, the initial risk is at least $1 - 9/256$. Either explicit condition

$$\mu \geq 1 - \frac{1}{1024 C_\alpha^2 r}, \qquad \text{or} \qquad \mu \geq \mu_* := \max\left\{\frac{1}{2}, 1 - \frac{1}{768 D_g^2}\right\} \tag{PJ22}$$

makes $\Delta(\mu) \leq 1/16$. The second sufficient minimum is independent of rank, width, dimension and leak parameter. The two same-budget before/after gains, including the unrestricted one, therefore obey

$$\begin{aligned}\mathcal{R}^P_{B_\kappa}(0) - \mathcal{R}^P_{B_\kappa}(N) \geq G_{\text{proj}} > 1/2, \qquad \mathcal{R}^P_\infty(0) - \mathcal{R}^P_\infty(N) \geq G_{\text{proj}} > 1/2, \\ G_{\text{proj}} := 1 - 9/256 - (1/\sqrt{3} + 11/128)^2.\end{aligned} \tag{PJ23}$$

For an entirely rational check of positivity, replace $1/\sqrt{3}$ by $7/12$ in this lower bound; the result still exceeds $1/2$. Approximate scalar values are $D_g = 2.066824$, $\mu_* = 0.99969519$, and $G_{\text{proj}} = 0.52489308$. These evaluations are of the displayed constants, not training experiments. The certified conditions are PJ5 and PJ22–23. A simple numerical sufficient condition is $\mu \geq 0.9997$. For a finite check without relying on rounded evaluations, use $0.78615 < b_* < 0.78616$ and $0.39894 < \varphi(0) < 0.39895$. The alternating Taylor sums of orders seven and eight for e^{-x} and for $\int_0^b e^{-t^2/2}\,dt$ then give $0.56821 < p < 0.56824$ and $0.29288 < \varphi(b_*) < 0.29291$. Substitute these rational intervals in PJ5 and use $\sqrt{6} > 2.44948$: they give $v_g > 0.05160647$ and $D_g < 2.08$. Finally $3(2.08)^2(0.0003) = 0.00389376 < 1/256$, which verifies PJ22's second sufficient bound at 0.9997.

The same rational intervals certify the strict inequality $K_* > 2/3$ used in AJ3 and PJ5. Indeed, writing their decimal endpoints as exact rationals gives

$$\begin{aligned}0 < v_g &< 0.56824 - 2(0.78615)(0.29288) \\ &\quad + (0.78616)^2(1 - 0.56821) - (0.56821)^2 < \frac{5177}{100000}, \\ t_g^2 = \frac{2}{3} b_*^2 \varphi(b_*)^2 &> \frac{2}{3}(0.78615)^2(0.29288)^2 > \frac{3532}{100000}.\end{aligned}$$

Positivity of v_g also follows directly because it is the variance of the nonconstant function g. Since $3 \cdot 3532 = 10596 > 10354 = 2 \cdot 5177$, these bounds imply $3t_g^2 > 2v_g$ and hence $K_* > 2/3$. This argument uses only the displayed rational bounds.

T SwiGLU STUDENTS: LEADING-DIRECTION LEARNING DURING A PLATEAU

This appendix gives a complementary result for a gated student with trainable inner weights and trainable output heads. Its geometric conclusion concerns the leading population-AGOP direction. For a teacher subspace of dimension larger than one, this is weaker than the minimum top-r alignment used in the main theorems. The results below do not extend those theorems to all Gaussian Sobolev teachers or replace their initialization restrictions.

The useful distinction is between three statements: leading-direction alignment on a long, nearly constant-loss interval; approximation by an unrestricted refitted readout; and a later decrease of the original trained loss. The first two are treated separately below. The numerical examples also exhibit later loss decreases, but a general theorem joining that release to the small-initialization trajectory remains open.

T.1 WHICH TEACHER LINKS ARE COVERED?

The alignment theorem requires a nonzero Hermite component of degree one, two, or three. The examples below make this condition concrete. Each displayed link is centered and divided by its standard deviation when used as a teacher; this preserves the listed nonzero coefficients and the value of k. Here Z, S, T are independent standard Gaussians, $h_1(z) = z$, $h_2(z) = (z^2-1)/\sqrt{2}$, and $h_3(z) = (z^3-3z)/\sqrt{6}$. Coefficient values in the table are for the displayed link before normalization. In general $h_j = \mathrm{He}_j/\sqrt{j!}$ is the normalized probabilists' Hermite polynomial, and $H_{\geq 4}$ denotes a Gaussian L^2 function supported in Hermite degrees at least four. Expectations in this table are over the displayed Gaussian arguments. The index k is the degree of the leading homogeneous student–teacher interaction: $k = 3$ if degree one or two is present, and $k = 4$ if degree three is the first nonzero degree, as formalized in (65).

Teacher link F	A nonzero Hermite coefficient	k
$h_1(z)$ or $h_2(z)$	$\mathbb{E}[F(Z)h_i(Z)] = 1$ for $F = h_i$	3
$h_3(z)$	$\mathbb{E}[F(Z)h_3(Z)] = 1$; lower degrees vanish	4
ReLU or leaky ReLU, $F(z) = \max(z, \lambda z)$, $0 \leq \lambda \leq 1$	$\mathbb{E}[F(Z)h_1(Z)] = (1+\lambda)/2$	3
$\lvert z\rvert$	$\mathbb{E}[F(Z)h_2(Z)] = 1/\sqrt{\pi}$	3
Softplus, SiLU, or GELU	$\mathbb{E}[F(Z)h_1(Z)] = 1/2$	3
$\sin z$ or $\tanh z$	$\mathbb{E}[Z \sin Z] = e^{-1/2}$; $\mathbb{E}[Z \tanh Z] > 0$	3
$ah_3(z) + H_{\geq 4}(z)$, $a \neq 0$, $H_{\geq 4} \in L^2$	Cubic coefficient a; lower degrees vanish	4
Pure $h_j(z)$, $j \geq 4$	No coefficient in degrees one through three	Excluded
$\mathrm{SiLU}(s)t$ or $\mathrm{ReLU}(s)t$	$\mathbb{E}[F(S,T)ST] = 1/2$	3
st or $\tanh(s)t$	$\mathbb{E}[F(S,T)ST] = 1$ or $\mathbb{E}[S \tanh S] > 0$	3
$h_2(s)t$	$\mathbb{E}[F(S,T)h_2(S)T] = 1$; lower degrees vanish	4

These checks use Hermite orthogonality, independence, parity, and the Gaussian characteristic function. Softplus $\log(1+e^z)$, SiLU $z/(1+e^{-z})$, and GELU $z\,\mathbb{P}(Z \leq z)$ all have odd part $z/2$. Arbitrary square-integrable additions supported in degrees at least four leave the classification unchanged. A surviving degree-one or degree-two component always sets $k = 3$, even when a cubic component is also present.

The rank-one refit theorem is narrower: it requires a scalar link with zero degree-one and degree-two coefficients and a nonzero cubic coefficient. Thus it applies to $ah_3 + H_{\geq 4}$, but does not by itself apply to the bivariate $h_2(s)t$ example. For mixtures and interaction teachers alike, the alignment theorem concerns the leading AGOP direction in the teacher subspace; it does not assert recovery of every teacher direction.

T.2 SETUP AND LEADING-DIRECTION THEOREM

This result concerns a SwiGLU *student* whose head, gates, values, and inner biases all follow the same population gradient update. Only the output intercept is minimized exactly. At every fixed

width, a small initialization produces a long interval with vanishing loss variation, on which the leading AGOP direction enters the teacher subspace. The conclusion concerns every vector in the leading eigenspace; it does not assert recovery of all teacher directions when their number exceeds one.

Setting and clocks. Let $x \sim N(0, I_d)$, $y = F(U^\top x)$, $U^\top U = I_r$, $1 \le r < d$, $F \in L^2(\gamma_r)$, and $\mathrm{Var}(y) = 1$. Put $\widetilde{x} = (x, 1)$ and

$$\widetilde{f}_\theta(x) = \frac{\kappa}{m} \sum_{j=1}^{m} a_j S(p_j^\top \widetilde{x})(v_j^\top \widetilde{x}), \qquad S(s) = \frac{s}{1 + e^{-s}}, \qquad \kappa > 0. \tag{61}$$

Write $\theta_j = (a_j, p_j, v_j)$, $p_j = (w_j, b_j)$, $v_j = (z_j, c_j)$, and $g_j(x) = S(p_j^\top \widetilde{x})(v_j^\top \widetilde{x})$. Here d is the input dimension, m the student width, r the teacher rank, and $U \in \mathbb{R}^{d \times r}$ an orthonormal basis of the teacher subspace. The measure $\gamma_s = N(0, I_s)$ is standard Gaussian measure in dimension s. The scalar a_j is the output head, $w_j, z_j \in \mathbb{R}^d$ are the gate and value weights, and b_j, c_j their biases. The fixed positive κ is the SwiGLU output scale, distinct from the leaky-ReLU slope in the main text. Expectations, covariances, and variances in this subsection are over x unless specified; probabilities are over the Gaussian initialization. Vector norms are Euclidean, matrix norms are operator norms unless marked F for Frobenius norm, and function norms $\|\cdot\|_2$ use Gaussian L^2. The fitted intercept is $b_0(\theta) = \mathbb{E}[y - \widetilde{f}_\theta]$, so that

$$L(\theta) = \mathrm{Var}(y - \widetilde{f}_\theta), \qquad \mathcal{F}_j(\theta) = -m \nabla_{\theta_j} L = 2\kappa \,\mathrm{Cov}\left(y - \widetilde{f}_\theta, \nabla_{\theta_j}(a_j g_j)\right). \tag{62}$$

We consider either $\dot{\theta}_j = \mathcal{F}_j(\theta)$ or $\theta_j^{n+1} = \theta_j^n + h\mathcal{F}_j(\theta^n)$, with a fixed $h > 0$. Thus a raw GD learning rate η corresponds to $h = \eta/m$ in this mean-field clock. No step size is changed as the initialization scale shrinks. Initialize $\theta_j(0) = \varepsilon \vartheta_j$, where independently

$$\vartheta_j \sim N(0, \beta_0^2) \otimes N(0, s_0^2 I_{d+1}) \otimes N(0, s_0^2 I_{d+1}), \qquad \beta_0, s_0 > 0. \tag{63}$$

All dimensions, scales, and the teacher are fixed before $\varepsilon \downarrow 0$. The Gaussian draw $\vartheta = (\vartheta_1, \dots, \vartheta_m)$ is coupled across initialization scales. Set $P = UU^\top$ and

$$M(\theta) = \mathbb{E}[\nabla_x \widetilde{f}_\theta \nabla_x \widetilde{f}_\theta^\top], \qquad A_{\mathrm{top}}(\theta) = \min_{\substack{\|e\|=1 \\ M(\theta)e = \lambda_{\max}(M(\theta))e}} \|Pe\|^2. \tag{64}$$

The minimum resolves any multiplicity of the largest eigenvalue.

The required teacher signal. Define

$$g_1 = \mathbb{E}[yx], \quad H_2 = \mathbb{E}[y(xx^\top - I_d)], \quad T_3 = \mathbb{E}[y\,\mathrm{He}_3(x)],$$

where $(\mathrm{He}_3(x))_{ijk} = x_i x_j x_k - x_i\delta_{jk} - x_j\delta_{ik} - x_k\delta_{ij}$. Here δ_{ij} is the Kronecker delta, and tensor contractions mean $T_3[a, b, c] = \sum_{i,j,\ell} (T_3)_{ij\ell} a_i b_j c_\ell$. We use the Frobenius norm for tensors when bounding these contractions. These expectations exist by Cauchy–Schwarz. Gaussian independence in $U \oplus U^\perp$ shows that g_1, H_2, T_3 are supported on U. Assume that at least one of these three tensors is nonzero, and define

signal	k	$T_{\mathrm{lead}}(p, v)$
$(g_1, H_2) \neq (0, 0)$	3	$\frac{1}{2}\{w^\top H_2 z + g_1^\top (cw + bz)\}$
$g_1 = H_2 = 0,\ T_3 \neq 0$	4	$\frac{1}{4} T_3[w, w, z]$.

(65)

There is no smallness condition on higher Hermite components of F. In particular, the theorem is not restricted to a cubic polynomial teacher, but it does not cover a teacher having no signal in degrees one through three.

A decoupled comparator. For one neuron put $\Psi(a, p, v) = 2\kappa a T_{\mathrm{lead}}(p, v)$. This is a nonzero homogeneous polynomial of degree k, odd in a. Let $X_j(\tau)$ solve

$$X_j' = \nabla\Psi(X_j), \qquad X_j(0) = \vartheta_j. \tag{66}$$

Let τ_j be its maximal forward existence time and $\tau_* = \min_j \tau_j$. A finite τ_j means that the comparator norm diverges there. Define the escape event $\mathcal{E} = \{\tau_* < \infty\}$.

Theorem 5.1 (comparator formulation). Fix the setting above and a GD mean-field step $h > 0$. Then $\mathbb{P}(\mathcal{E}) \geq 1 - 2^{-m}$, and almost surely on $\mathcal{E}$ the earliest escaping neuron W is unique. On this same probability-one subset, for every $\delta \in (0, 1)$ there are random constants

$$0 < \tau_1 = \tau_1(\vartheta, \delta) < \tau_*, \qquad C = C(\vartheta, \delta) < \infty, \qquad \varepsilon_0 = \varepsilon_0(\vartheta, \delta, h) > 0$$

such that, for every $0 < \varepsilon < \varepsilon_0$, the following hold for both gradient flow and this fixed-step GD. For flow take $t_1 = \tau_1 \varepsilon^{-(k-2)}$; for GD take $n_1 = \lfloor \tau_1/(h\varepsilon^{k-2}) \rfloor$.

1. The loss stays on a plateau:

$$\sup_{0 \leq t \leq t_1} |L(\theta(t)) - L(\theta(0))| \leq C\varepsilon^k, \qquad \max_{0 \leq n \leq n_1} |L(\theta^n) - L(\theta^0)| \leq C\varepsilon^k.$$

   Its initial value has the signed expansion

$$L(\theta(0)) = 1 - \frac{\varepsilon^k}{m} \sum_j \Psi(\vartheta_j) + O_\vartheta(\varepsilon^{k+1}) = 1 + O_\vartheta(\varepsilon^k). \tag{67}$$

2. At the respective endpoints, $A_{\mathrm{top}}(\theta(t_1)) \geq 1 - \delta$ and $A_{\mathrm{top}}(\theta^{n_1}) \geq 1 - \delta$.

3. At either endpoint the winner obeys

$$\big(\|(I - P)w_W\|^2 + \|(I - P)z_W\|^2\big)^{1/2} \leq \delta\big(\|Pw_W\|^2 + \|Pz_W\|^2\big)^{1/2}.$$

   When $k = 4$, the stronger separate inequalities $\|(I - P)w_W\| \leq \delta\|Pw_W\|$ and $\|(I - P)z_W\| \leq \delta\|Pz_W\|$ hold.

The time τ_1 can be chosen arbitrarily close to τ_*, but is fixed before taking $\varepsilon \downarrow 0$. The cutoff ε_0 is initialization dependent; no explicit finite-scale confidence bound is asserted.

For $k = 4$ the plateau has duration proportional to ε^{-2} and loss variation $O(\varepsilon^4)$. For $k = 3$ these orders are ε^{-1} and $O(\varepsilon^3)$. The statement gives an interval inside the initial plateau, not its actual exit time or a later fixed-size loss decrease. It also makes no claim about stability of a fixed GD step after the small-parameter phase.

Lemma T.1 (Escape and the unique first winner)**.** *For a homogeneous gradient flow $X' = \nabla\Psi(X)$ of degree $k > 2$, let τ_b denote its maximal forward existence time. If $\Psi(X(0)) > 0$, then*

$$\tau_b \leq \frac{\|X(0)\|^2}{k(k-2)\Psi(X(0))}, \qquad \Psi(X(\tau)) \geq \Psi(X(0)) \left(\frac{\|X(\tau)\|}{\|X(0)\|}\right)^k. \tag{68}$$

In the present initialization, $\mathbb{P}(\Psi(\vartheta_j) > 0) = 1/2$. The law of τ_j has no finite *atoms; it may have an atom at ∞. Consequently $\mathbb{P}(\mathcal{E}) \geq 1 - 2^{-m}$ and its finite minimum is almost surely unique. Moreover, whenever the set on the right is nonempty,*

$$\tau_* \leq \min_{j:\Psi(\vartheta_j)>0} \frac{\|\vartheta_j\|^2}{k(k-2)\Psi(\vartheta_j)}.$$

Proof. Writing $N = \|X\|^2$, Euler's identity and Cauchy–Schwarz give $N' = 2k\Psi$ and $\Psi' = \|\nabla\Psi\|^2 \geq k^2\Psi^2/N$. While $\Psi > 0$, $(\log\Psi - \frac{k}{2}\log N)' \geq 0$. Integrating $N' \geq 2k\Psi(X(0))N^{k/2}/N(0)^{k/2}$ proves (68) and forces finite escape by the polynomial ODE's continuation criterion. The zero set of the nonzero polynomial Ψ is Lebesgue null; changing the sign of a proves its equal sign probabilities.

Homogeneity gives $\tau_b(s\vartheta) = s^{-(k-2)}\tau_b(\vartheta)$ for $s > 0$. For a fixed $t \in (0, \infty)$, the level set $\{\tau_b = t\}$ therefore meets each radial ray in at most one point. Polar coordinates, Fubini, and absolute continuity of the Gaussian law show that this set has probability zero. Escape times are measurable, for example by expressing the maximal existence time as the increasing limit of exit times from balls. Independence then rules out equality between any pair of finite escape times. A finite minimum exists whenever one neuron starts with positive potential, giving the stated probability. □

Lemma T.2 (Uniform field, loss, and AGOP expansions)**.** *On every fixed ball of radius R in the rescaled parameters $Z = (Z_1, \dots, Z_m)$, uniformly for $0 < \varepsilon \le 1$,*

$$\varepsilon^{1-k}\mathcal{F}_j(\varepsilon Z) = \nabla\Psi(Z_j) + O_R(\varepsilon), \tag{69}$$

$$L(\varepsilon Z) = 1 - \frac{\varepsilon^k}{m}\sum_j \Psi(Z_j) + O_R(\varepsilon^{k+1}), \tag{70}$$

$$M(\varepsilon Z) = \varepsilon^6\left(\frac{\kappa}{2m}\right)^2\{B(Z)^2 + \ell(Z)\ell(Z)^\top\} + O_R(\varepsilon^7), \tag{71}$$

where the last error is in operator norm and, writing the coordinates of Z_j as $(a_j, w_j, b_j, z_j, c_j)$,

$$B(Z) = \sum_j a_j(w_j z_j^\top + z_j w_j^\top), \qquad \ell(Z) = \sum_j a_j(c_j w_j + b_j z_j).$$

For $k = 4$, the error in (69) is in fact $O_R(\varepsilon^2)$. The subscript on O_R allows its constant to depend on this fixed radius (and the fixed model parameters); O_ϑ and O_τ similarly allow dependence on the realized initialization and fixed comparator time, respectively. The vector $\ell(Z)$ here is an AGOP expansion coefficient, not the normalized experimental loss.

Proof. The identity $S(s) = s/2 + (s/2)\tanh(s/2)$ gives

$$S(s) = s/2 + s^2/4 + R(s), \qquad |R(s)| \le |s|^4/48, \qquad |R'(s)| \le |s|^3/12. \tag{72}$$

Indeed $|\tanh u - u| \le |u|^3/3$ follows by integrating $|\tanh'(u) - 1| = \tanh^2 u \le u^2$; differentiation gives the second bound. We also have $|S(s)| \le |s|$, $|S'(s)| \le 2$, $|S(s) - s/2| \le s^2/4$, and $|S'(s) - 1/2| \le |s|/2$. These are analytic bounds, not numerical estimates.

For affine Gaussian forms, every fixed-order moment satisfies $\|p^\top \widetilde{x}\|_{L^q} \le C_q\|p\|$; the same bound with additional fixed powers of $\|\widetilde{x}\|$ follows from Hölder's inequality. Thus $F \in L^2$ suffices for all ensuing covariance bounds and parameter derivatives, by Cauchy–Schwarz. Derivatives up to order two of the features have polynomial envelopes on bounded parameter sets, so L is twice continuously differentiable and its vector field is locally Lipschitz.

Put $P_0 = w^\top x + b$ and $V_0 = z^\top x + c$. Hermite decomposition yields

$$\begin{aligned}\mathrm{Cov}(y, P_0V_0) &= w^\top H_2 z + g_1^\top(cw + bz),\\ \mathrm{Cov}(y, P_0^2V_0) &= T_3[w,w,z] + \|w\|^2 g_1^\top z + 2(w^\top z)g_1^\top w\\ &\quad + 2b\, w^\top H_2 z + c\, w^\top H_2 w + b^2 g_1^\top z + 2bc\, g_1^\top w.\end{aligned}$$

Consequently $T(p, v) = \mathrm{Cov}(y, S(P_0)V_0)$ has the leading term in (65). Its next contribution to aT is degree four when $k = 3$, and is of order six when $k = 4$. The differentiated remainder obeys the corresponding bounds by (72). The student feedback in (62) is $O_R(\varepsilon^5)$, since $\|\widetilde{f}_{\varepsilon Z}\|_2 = O_R(\varepsilon^3)$ and $\|\nabla_{\theta_j}(a_j g_j)\|_2 = O_R(\varepsilon^2)$. This proves the field expansion, including its sharper cubic-signal version. Expanding $L = 1 - 2\,\mathrm{Cov}(y, \widetilde{f}) + \mathrm{Var}(\widetilde{f})$, with the last term $O_R(\varepsilon^6)$, proves the loss formula.

Finally, $\nabla_x g_j = S'(P_0)V_0 w_j + S(P_0)z_j$. The elementary bounds above give, in $L^2(\gamma_d; \mathbb{R}^d)$,

$$\nabla_x \widetilde{f}_{\varepsilon Z} = \frac{\kappa\varepsilon^3}{2m}\{B(Z)x + \ell(Z)\} + O_R(\varepsilon^4).$$

Taking its outer product and expectation proves (71), because B is symmetric and $\mathbb{E}[xx^\top] = I_d$, $\mathbb{E}[x] = 0$. □

Lemma T.3 (Tracking up to a fixed comparator time)**.** *Fix a realization with $0 < T < \tau_*$, and let $X = (X_j)_j$ be its comparator. For flow define $Z_\varepsilon(\tau) = \varepsilon^{-1}\theta(\tau\varepsilon^{-(k-2)})$. For GD define $Z_\varepsilon^n = \varepsilon^{-1}\theta^n$ and $\Delta = h\varepsilon^{k-2}$. Then, for sufficiently small ε,*

$$\sup_{0\le\tau\le T}\|Z_\varepsilon(\tau) - X(\tau)\| \le C_T\varepsilon,$$

$$\max_{0\le n\Delta\le T}\|Z_\varepsilon^n - X(n\Delta)\| \le C_T(\varepsilon + \Delta).$$

All these rescaled states lie in one fixed ball. For $k = 4$, the ε terms can be replaced by ε^2.

Proof. Choose a ball with radius $R > \sup_{\tau \le T} \|X(\tau)\| + 1$. On this ball $G(Z) = (\nabla\Psi(Z_j))_j$ is Lipschitz, say with constant L_0, and the rescaled actual field is $G(Z) + r_\varepsilon(Z)$ with $\|r_\varepsilon\| \le C_0\varepsilon$. Stopping at the ball's first exit, Grönwall bounds the flow error by $C_0\varepsilon T e^{L_0 T}$. Taking it smaller than half the margin closes the stopping argument and proves existence through T.

For GD, $Z^{n+1} = Z^n + \Delta\{G(Z^n) + r_\varepsilon(Z^n)\}$. Along the bounded comparator, the one-step Taylor defect is at most $C_1\Delta^2$: this follows directly by integrating $G(X(s)) - G(X(n\Delta))$, using a bound for $\|X'\|$. Writing $e_n = \|Z^n - X(n\Delta)\|$, we therefore obtain

$$e_{n+1} \le (1 + L_0\Delta)e_n + C_0\varepsilon\Delta + C_1\Delta^2, \qquad e_0 = 0.$$

Summing this recursion gives $e_n \le T e^{L_0 T}(C_0\varepsilon + C_1\Delta)$. Induction closes the same radius bootstrap. Since $|T - \lfloor T/\Delta \rfloor \Delta| < \Delta$ and X' is bounded, the rounded GD endpoint also converges to $X(T)$. The sharper bound follows by replacing $C_0\varepsilon$ by $C_0\varepsilon^2$. □

Proof of Theorem 5.1. Lemma T.1 gives the escape probability and uniqueness. Fix a realization in this event with unique winner. Every nonwinner extends across τ_* and stays bounded there. Write $K(\tau) = \|X_W(\tau)\|$. Since $(K^2)' = 2k\Psi(X_W)$ and $\Psi' = \|\nabla\Psi\|^2 \ge 0$, finite escape forces $\Psi(X_W)$ to become positive: otherwise the norm is bounded. In fact it becomes unbounded, since a bounded potential would keep K^2 bounded on a finite time interval. Applying (68) from a positive-potential time gives, for all sufficiently late $\tau < \tau_*$,

$$\Psi(X_W(\tau)) \ge cK(\tau)^k, \qquad K(\tau) \longrightarrow \infty \quad (\tau \uparrow \tau_*), \quad c > 0. \tag{73}$$

This argument permits a winner whose initial potential was negative.

The comparator depends on the spatial weights only through Pw_j, Pz_j. Thus $(I - P)w_j$ and $(I - P)z_j$ are constant along it; for $k = 4$, b_j, c_j are also constant. For the winner, set

$$B_W = a_W(w_W z_W^\top + z_W w_W^\top), \quad D = PB_W P,$$
$$\ell_W = a_W(c_W w_W + b_W z_W), \quad v = P\ell_W.$$

Every term removed by these projections contains a frozen transverse factor. Since all other factors are bounded by K,

$$\|B_W - D\| + \|\ell_W - v\| = O(K^2). \tag{74}$$

Here and below constants may depend on the fixed initialization.

For $k = 4$, combine (73) with

$$|\Psi| \le (\kappa/2)\|T_3\|\,|a_W|\|Pw_W\|^2\|Pz_W\|.$$

Each of $|a_W|, \|Pw_W\|, \|Pz_W\|$ is then at least a positive constant times K. Using $\|uv^\top + vu^\top\| = \|u\|\|v\| + |u^\top v|$, we obtain $\|D\| \ge c_1 K^3$. For $k = 3$, instead use the exact identity

$$\Psi(X_W) = \kappa\{\tfrac{1}{2}\operatorname{tr}(H_2 D) + g_1^\top v\}.$$

It implies $\|D\|_{\mathrm{F}} + \|v\| \ge c_2 K^3$. In both cases the positive semidefinite matrix $Q = D^2 + vv^\top$ has range contained in U and satisfies

$$\lambda_{\max}(Q) \ge c_3 K^6, \tag{75}$$

using $\|D\|^2 \ge \|D\|_{\mathrm{F}}^2/r$ in the second case. The bounded nonwinners and (74) give

$$M_c(\tau) := B(X(\tau))^2 + \ell(X(\tau))\ell(X(\tau))^\top = Q + E, \qquad \|E\| = O(K^5). \tag{76}$$

For *any* unit top eigenvector e of $Q + E$,

$$\lambda_{\max}(Q) - \|E\| \le e^\top(Q + E)e \le \lambda_{\max}(Q)\|Pe\|^2 + \|E\|.$$

Hence $\|Pe\|^2 \ge 1 - 2\|E\|/\lambda_{\max}(Q) = 1 - O(K^{-1})$. This controls the entire top eigenspace without assuming an eigengap.

The hidden-weight claims follow as well. For $k = 4$ the individual projected norms are of order K, whereas the transverse norms are fixed. For $k = 3$, writing $s = (\|Pw_W\|^2 + \|Pz_W\|^2)^{1/2} \le K$, the potential formula yields $|\Psi(X_W)| \le C_2 K^2 s$. Together with (73) this gives $s \ge c_4 K$. Thus the combined spatial norm dominates the fixed transverse norm; no separate lower bound on the two projected norms is needed in this case.

Choose $\tau_1 < \tau_*$ sufficiently late that all comparator alignment bounds hold with strict margin relative to δ. By Lemmas T.2 and T.3, for both the flow and the rounded GD endpoint,

$$\varepsilon^{-6}\left(\frac{2m}{\kappa}\right)^2 M \longrightarrow M_c(\tau_1) = Q(\tau_1) + E(\tau_1).$$

The same quadratic-form inequality remains valid after adding the operator-norm error $o(1)$ to E. It proves the claimed uniform top-eigenspace alignment, even at multiple top eigenvalues. Parameter convergence transfers the hidden-weight inequalities, since their projected denominators are nonzero at this fixed time. Uniform boundedness on $[0, \tau_1]$ and (70) prove the entire-prefix loss bounds and (67). Shrinking ε_0 makes the flow and GD conclusions simultaneous. Since all the comparator estimates improve as $K \to \infty$, τ_1 can be chosen arbitrarily close to τ_*. Rational choices of these times and countable choices of cutoffs can be used throughout, so the random endpoints and thresholds may be chosen measurably. □

Corollary T.4 (Unconditional initial headroom and a same-run gain)**.** *For every $\varepsilon > 0$, $A_{\mathrm{top}}(\theta(0))$ is stochastically dominated by $B \sim \mathrm{Beta}(r/2, (d-r)/2)$; in particular its expectation is at most r/d. Fix $\delta \in (0,1)$ and $q \in (0, 1-\delta)$. Using the random endpoints of Theorem 5.1,*

$$\liminf_{\varepsilon \downarrow 0} \mathbb{P}\left(\begin{array}{c}\textit{the theorem's plateau conclusions hold, and}\\ A_{\mathrm{top}}(\textit{endpoint}) - A_{\mathrm{top}}(0) \geq 1-\delta-q\end{array}\right) \geq \mathbb{P}(B \leq q) - 2^{-m}. \tag{77}$$

The endpoint event is understood to fail outside $\mathcal{E}$. The same bound applies to the event that both the flow and the fixed-step GD conclusions hold. Equivalently, for any $\zeta > 0$ there is a deterministic sufficiently small cutoff for which the right side minus ζ holds at each initialization scale below it.

Proof. At initialization the joint spatial-weight law is rotationally invariant, and rotating these weights by $O \in O(d)$ conjugates M to $OMO^\top$. Conditionally on its top eigenspace choose a uniform unit vector e in that space using auxiliary randomness. Its unconditional distribution is uniform on S^{d-1}, and $A_{\mathrm{top}}(0) \leq \|Pe\|^2$ pointwise. The Gaussian representation of a uniform sphere vector shows that $\|Pe\|^2$ has the displayed Beta distribution.

Let $G_\varepsilon = \mathcal{E} \cap \{\varepsilon < \varepsilon_0\}$, with the measurable common cutoff in the theorem. Then $\mathbb{P}(G_\varepsilon) \to \mathbb{P}(\mathcal{E}) \geq 1 - 2^{-m}$. The union bound and initial stochastic domination give

$$\mathbb{P}(G_\varepsilon \cap \{A_{\mathrm{top}}(0) \leq q\}) \geq \mathbb{P}(G_\varepsilon) - \mathbb{P}(B > q).$$

On this intersection, both endpoint alignments are at least $1 - \delta$, which proves the claim. No independence, or conditional Haar law given escape, is used. □

Scope for several teacher directions. For any choice P_r of a projector onto r top AGOP eigenvectors, the endpoint satisfies $\mathrm{tr}(PP_r)/r \geq (1-\delta)/r$, because its range contains a top eigenvector. This is an average-alignment lower bound, not a bound on the smallest principal-angle cosine between the two r-dimensional spaces. A single winner can contribute rank two, so one cannot attribute all remaining eigenvalues to nonwinners. More generally, $\mathrm{range}(M) \subseteq \mathrm{span}\{w_j, z_j : 1 \leq j \leq m\}$ and $\mathrm{rank}(M) \leq 2m$. Full-subspace recovery and the later loss-release regime require additional arguments.

T.3 Unrestricted population refitting of small SwiGLU features

This section proves an *attained refit-risk bound* on the initial plateau. It first proves the rank-one bound without asserting a comparison with initialization. A separate proposition establishes such a comparison at the boundary width only. All readout coefficients and the intercept are unrestricted; the constructed coefficients may diverge as initialization vanishes. These conclusions concern population least squares, not a bounded readout, ridge regression, or a finite-sample estimator.

Setting. Let $x \sim N(0, I_d)$, $\widetilde{x} = (x, 1)$, and

$$g_j(x) = S(p_j^\top \widetilde{x})(v_j^\top \widetilde{x}), \qquad S(t) = \frac{t}{1+e^{-t}}, \qquad \widetilde{f} = \frac{\kappa}{m}\sum_{j=1}^m a_j g_j, \qquad \kappa > 0.$$

Write $p_j = (w_j, b_j)$ and $v_j = (z_j, c_j)$. Here d is the input dimension, m the student width, $\kappa > 0$ the output scale, a_j the trained head, $w_j, z_j \in \mathbb{R}^d$ the gate and value weights, and b_j, c_j their biases. Write $\gamma_s = N(0, I_s)$; function inner products and norms are $\langle f, g\rangle = \mathbb{E}[f(x)g(x)]$ and $\|f\|_2^2 = \mathbb{E}[f(x)^2]$ under the appropriate Gaussian measure. Unless explicitly subscripted, expectations, variances, and covariances are over the Gaussian input, whereas probabilities are over initialization. Vector norms are Euclidean, matrix norms are operator norms, and $\|\cdot\|_{\mathrm{F}}$ denotes Frobenius norm. The training loss is $L = \mathrm{Var}(y - \widetilde{f})$: its intercept is profiled out exactly. The neuron parameters $\theta_j = (a_j, p_j, v_j)$ follow $\dot\theta_j = -m\nabla_{\theta_j} L$, or $\theta_j^{n+1} = \theta_j^n - hm\nabla_{\theta_j} L$ with fixed $h > 0$. Thus the raw learning rate is mh. Initialization is $\theta_j(0) = \varepsilon\vartheta_j$, where the ϑ_j are independent with the same nondegenerate centered Gaussian law, invariant under $a \mapsto -a$. The head and inner-block variances may differ and are fixed independently of ε.

The teacher is $y = F(U^\top x)$ with $U^\top U = I_r$ and $F \in L^2(\gamma_r)$, $\mathrm{Var}(y) = 1$. In this section it is of type three:

$$\mathbb{E}[(y - \mathbb{E}[y])x] = 0, \qquad \mathbb{E}[(y - \mathbb{E}[y])(xx^\top - I_d)] = 0, \qquad T_3 := \mathbb{E}[y\,\mathrm{He}_3(x)] \neq 0.$$

Here $U \in \mathbb{R}^{d\times r}$ spans the rank-r teacher subspace, $(\mathrm{He}_3(x))_{ijk} = x_ix_jx_k - x_i\delta_{jk} - x_j\delta_{ik} - x_k\delta_{ij}$, and δ_{ij} is the Kronecker delta. The notation $T_3[a, b, c]$ denotes contraction with three vectors; the tensor norm used to bound this contraction is the Frobenius norm. Let $\mathcal{Q}$ be the real polynomials of degree at most two, as a subspace of $L^2(\gamma_d)$, with orthogonal projector $\Pi_{\mathcal{Q}}$ and dimension

$$D = \dim \mathcal{Q} = 1 + d + \frac{d(d+1)}{2} = \frac{(d+1)(d+2)}{2}. \tag{78}$$

Then $y - \mathbb{E}[y] \perp \mathcal{Q}$. The refit risk at a frozen parameter state is

$$\mathcal{R}(\theta) = \inf_{\eta_0\in\mathbb{R},\, \eta\in\mathbb{R}^m} \mathbb{E}\left[\left(y - \eta_0 - \sum_{j=1}^m \eta_j g_j\right)^2\right]. \tag{79}$$

This infimum is attained because the feature span is finite dimensional; its coefficients need not be unique. The trained a_j are not constraints on the refitted coefficients.

Comparator and evaluation times. Let $X_j(\tau)$ solve the independent polynomial flows

$$X_j' = \nabla\Psi(X_j), \qquad X_j(0) = \vartheta_j, \qquad \Psi(a, p, v) = \frac{\kappa}{2} a T_3[w, w, z]. \tag{80}$$

Let $T_j \in (0, \infty]$ be their maximal forward existence times and $\tau_* = \min_j T_j$. The event $E_{\mathrm{esc}} = \{\tau_* < \infty\}$ has a unique winner W almost surely. At fixed comparator time τ write

$$t_\varepsilon(\tau) = \tau\varepsilon^{-2} \quad \text{for flow}, \qquad t_\varepsilon(\tau) = h\left\lfloor \frac{\tau}{h\varepsilon^2}\right\rfloor \quad \text{for GD}.$$

All limits below first keep $\tau < \tau_*$ fixed and then send $\varepsilon \downarrow 0$. The selected τ may depend on the initialization, but never on ε. For GD, $\theta(t_\varepsilon(\tau))$ means the iterate with index $\lfloor\tau/(h\varepsilon^2)\rfloor$. Below $w_\perp = (I - UU^\top)w$ and $z_\perp = (I - UU^\top)z$ denote transverse spatial weights. Subscripts on O_τ allow constants to depend on the fixed comparator time; O_{L^2} specifies the norm in which the remainder is bounded.

Lemma T.5 (Expansion, tracking, and escape facts). Write $Z = (Z_1, \ldots, Z_m)$ for the rescaled parameter tuple and $P = p^\top\widetilde{x}$, $V = v^\top\widetilde{x}$ for one neuron's affine Gaussian forms. For tuples, $\nabla\Psi(Z)$ denotes $(\nabla\Psi(Z_j))_{j=1}^m$. On every compact set of rescaled parameters, uniformly on that set,

$$S(\varepsilon P)\varepsilon V = \varepsilon^2 q + \varepsilon^3\chi + O_{L^2}(\varepsilon^5), \qquad q = \tfrac{1}{2}PV, \quad \chi = \tfrac{1}{4}P^2V, \tag{81}$$

$$\varepsilon^{-3}[-m\nabla L](\varepsilon Z) = \nabla\Psi(Z) + O(\varepsilon^2), \tag{82}$$

$$L(\varepsilon Z) = 1 - \frac{\varepsilon^4}{m}\sum_j \Psi(Z_j) + O(\varepsilon^6). \tag{83}$$

Here q_j and χ_j denote the quadratic and cubic terms of neuron j; χ_j is distinct from its value bias c_j. Consequently, for every realization and fixed $\tau < \tau_*$, the rescaled flow, and fixed-step GD at its rounded times, satisfy

$$\sup_{0\le s\le\tau}\max_j \left\|\varepsilon^{-1}\theta_j(t_\varepsilon(s)) - X_j(s)\right\| = O_\tau((1+h)\varepsilon^2), \tag{84}$$

where $h = 0$ denotes flow. Moreover $\Pr(E_{\mathrm{esc}}) \geq 1 - 2^{-m}$, there are no finite atoms in the law of T_j, and, almost surely on E_{esc}, the following hold as $\tau \uparrow \tau_*$:

1. all nonwinners extend smoothly to τ_*;
2. $K(\tau) := \|X_W(\tau)\| \to \infty$ and $\Psi(X_W(\tau)) \geq cK(\tau)^4$ for some random $c > 0$;
3. the winner's a, $U^\top w$, and $U^\top z$ each have norm at least $c'K$, for some $c' > 0$, while $w_\perp, z_\perp, b, c$ remain constant.

Proof. The identity $S(t) = t/2 + (t/2)\tanh(t/2)$ gives

$$|S(t) - t/2 - t^2/4| \leq |t|^4/48, \qquad |S'(t) - 1/2 - t/2| \leq |t|^3/12.$$

Indeed $|\tanh u - u| \leq |u|^3/3$ follows by integrating $\tanh'(u) - 1 = -\tanh^2 u$, and differentiating the remainder gives the second bound. Gaussian moments of affine functions prove (81) and its parameter derivative bounds. Type three cancels the degree-two teacher correlation, and $\mathrm{Cov}(y, P^2V) = T_3[w, w, z]$. The next teacher contribution to the parameter field is $O(\varepsilon^5)$; the student feedback is also $O(\varepsilon^5)$ since $\|\widetilde{f}\|_2 = O(\varepsilon^3)$ and $\|\nabla_\theta(ag)\|_2 = O(\varepsilon^2)$. This proves (82). Expanding $1 - 2\,\mathrm{Cov}(y, \widetilde{f}) + \mathrm{Var}(\widetilde{f})$ proves (83). Cauchy–Schwarz uses only $y \in L^2$; no differentiability of F is needed. On a compact comparator segment, the polynomial vector field is Lipschitz. Gronwall, stopped in a slightly larger compact set, bounds the flow error by $O_\tau(\varepsilon^2)$ and closes the stopping argument. Euler's global error in rescaled time adds $O_\tau(h\varepsilon^2)$, including the rounded endpoint. This proves (84) without asserting tracking up to blow-up.

For completeness, put $N = \|X\|^2$. Homogeneity gives $N' = 8\Psi$ and $\Psi' = \|\nabla\Psi\|^2 \geq 16\Psi^2/N$. Where $\Psi > 0$, these imply that Ψ/N^2 is nondecreasing and $N' \geq 8cN^2$, hence finite-time escape. The nonzero polynomial Ψ is odd in a; its zero set has Lebesgue measure zero, so $\Pr(\Psi(\vartheta_j) > 0) = 1/2$. Independence gives the claimed escape probability. Homogeneity gives $T(\rho\omega) = \rho^{-2}T(\omega)$. Each finite level set meets every ray in at most one point, hence has Lebesgue measure zero by polar coordinates. An atom at infinity is allowed. Independence now gives uniqueness of the earliest finite escape and $T_j > \tau_*$ for $j \neq W$. For an escaping solution, unbounded N forces unbounded Ψ by $N' = 8\Psi$; monotonicity of Ψ then makes it positive eventually. The preceding differential inequalities give $\Psi \geq cK^4$ and $K \to \infty$. The upper bound

$$|\Psi| \leq (\kappa/2)\|T_3\|\,|a|\,\|U^\top w\|^2\|U^\top z\|$$

and the fact that each factor is at most K give the lower bounds on all three active blocks (with adjusted constants). Finally, Ψ depends only on those blocks; all other coordinates are frozen. □

Lemma T.6 (A nonwinner quadratic bank at the random escape endpoint). Suppose $m \geq D$. Almost surely on E_{esc}, the map

$$B(\tau) : \mathbb{R} \times \mathbb{R}^{m-1} \longrightarrow \mathcal{Q}, \qquad B(\tau)(\beta_0, \beta) = \beta_0 + \sum_{j \neq W} \beta_j q_j(\tau) \tag{85}$$

is surjective at $\tau = \tau_*$. Here q_j is built from the comparator parameters as in (81). There is a random $\tau_a < \tau_*$ such that its smallest positive singular value is bounded below on $[\tau_a, \tau_*]$. In particular, the right inverse $B^*(BB^*)^{-1}$ has uniformly bounded norm there, using the $L^2(\gamma_d)$ norm on $\mathcal{Q}$. Here B^* denotes the Hilbert-space adjoint of B.

Proof. Choose an orthonormal coordinate basis of $\mathcal{Q}$. For any $D - 1$ neurons, the determinant with columns $1, q_1, \ldots, q_{D-1}$ is a polynomial in their affine parameters and is not identically zero. To see this, realize the nonconstant monomials x_i, x_i^2, x_ix_j as products $PV/2$, one per neuron. Together with 1 they form a basis. The rank-deficient set is therefore Lebesgue null in the full neuron parameter space, including the unused heads.

The time τ_* is random, so fixed-time nonsingularity alone is insufficient. Fix a candidate winner i and condition on its entire initial value $\vartheta_i = \xi$, with finite $t = T(\xi)$. Conditional on i being the unique winner, the other initial values are independent copies of their Gaussian law restricted to

$$\mathcal{D}_t = \{\vartheta : T(\vartheta) > t\}.$$

This conditional law is absolutely continuous whenever its normalizing probability is positive; zero-probability cases contribute no winning realizations. The survival domain $\mathcal{D}_t$ is open. For a smooth autonomous vector field, its time-t flow map is a smooth diffeomorphism onto its open image, with inverse supplied by the backward flow along each surviving segment. In particular its Jacobian is nonsingular and it preserves absolute continuity. Thus the joint nonwinner endpoint law at this now fixed t is absolutely continuous. The determinant polynomial vanishes with conditional probability zero. Integrate over ξ and sum over the finitely many possible winners. This proves surjectivity at the selected random endpoint. Continuity of the surviving trajectories and of BB^* proves the final assertion. □

Lemma T.7 (Exact quadratic cancellation in the actual feature bank). On the event of Lemma T.6, fix $\tau \in [\tau_a, \tau_*)$. Form $q_j^\varepsilon, \chi_j^\varepsilon$ from the *actual rescaled parameters* $\varepsilon^{-1}\theta_j(t_\varepsilon(\tau))$. There are coefficients $\beta^\varepsilon \to \beta$, with $\beta_W^\varepsilon = \beta_W = 1$, such that

$$\beta_0^\varepsilon + \sum_j \beta_j^\varepsilon q_j^\varepsilon = 0, \qquad C_\varepsilon := \varepsilon^{-3}\left[\varepsilon^2\beta_0^\varepsilon + \sum_j \beta_j^\varepsilon g_j\right] \longrightarrow C := \sum_j \beta_j \chi_j \quad \text{in } L^2. \tag{86}$$

The limiting nonwinner/intercept coefficients obey $\|(\beta_0, (\beta_j)_{j\neq W})\| = O(K(\tau)^2)$, uniformly as $\tau \uparrow \tau_*$. For every $q \in \mathcal{Q}$ there is also $Q_\varepsilon(q)$ in the actual intercept-feature span such that $Q_\varepsilon(q) \to q$ in L^2. Its nonwinner quadratic-bank coefficients, before the ε^{-2} readout scaling, are bounded by a constant times $\|q\|_2$ once ε is sufficiently small for the chosen fixed τ; the bounding constant is uniform in these late comparator times, but the ε cutoff need not be.

Proof. Replace q_j by q_j^ε in B to obtain B_ε. Tracking gives $B_\varepsilon \to B$ for this fixed τ, so it is surjective for all small ε and its continuous right inverse $R_\varepsilon = B_\varepsilon^*(B_\varepsilon B_\varepsilon^*)^{-1}$ converges to R. Set the nonwinner/intercept vector to $-R_\varepsilon q_W^\varepsilon$ and the winner coefficient to one. This solves (86) exactly at every ε. Since $\|q_W\|_2 = O(K^2)$ and R is uniformly bounded, the coefficient estimate follows. Expanding the actual features now gives

$$C_\varepsilon = \sum_j \beta_j^\varepsilon \chi_j^\varepsilon + O_\tau(\varepsilon^2) \longrightarrow C.$$

The constant hidden in this error may depend on the fixed time τ and is not asserted uniform up to τ_*.

Similarly put $(\eta_0^\varepsilon, \eta^\varepsilon) = R_\varepsilon q$ and define

$$Q_\varepsilon(q) = \varepsilon^{-2}\left[\varepsilon^2\eta_0^\varepsilon + \sum_{j\neq W} \eta_j^\varepsilon g_j\right].$$

It equals $q + O_\tau(\varepsilon)$ in L^2. The factor $\varepsilon^2\beta_0^\varepsilon$ in the first witness is essential: omitting it leaves an uncanceled constant after division by ε^3. Cancellation of comparator quadratics alone would also be insufficient under mere $o(1)$ tracking; the present construction requires no such division of a tracking error. □

Theorem T.8 (Rank-one attained refit risk on the plateau). Suppose $r = 1$, $U = u$ is unit, and $m \geq D$. Write the orthogonal Hermite decomposition

$$y = \mathbb{E}[y] + Ah_3(u^\top x) + F_{\geq 4}(u^\top x), \qquad h_3(s) = \frac{s^3 - 3s}{\sqrt{6}}, \qquad A \neq 0, \qquad A^2 + \|F_{\geq 4}\|_2^2 = 1.$$

Here $A = \mathbb{E}[yh_3(u^\top x)]$ and $F_{\geq 4}$ contains only Hermite degrees at least four. Almost surely on E_{esc}, for every $\delta > 0$ there is a fixed $\tau_1 < \tau_*$, arbitrarily close to τ_* if desired, such that, for flow or any fixed MF GD step $h > 0$, all sufficiently small ε satisfy

$$\mathcal{R}\big(\theta(t_\varepsilon(\tau_1))\big) \leq \|F_{\geq 4}\|_2^2 + \delta, \qquad \sup_{0\leq t\leq t_\varepsilon(\tau_1)} |L(t) - L(0)| \leq C\varepsilon^4. \tag{87}$$

For GD the supremum is over its iterates. The constants and the smallness threshold may depend on the realized initialization, δ, and h. For pure h_3 the attained refit risk is at most δ. The escape event has probability at least $1 - 2^{-m}$; this is not a specified finite-ε confidence bound.

Proof. Put $s = u^\top x$, $\pi = u^\top w_W$, and $\nu = u^\top z_W$ in the comparator. Its winner has $P_W = \pi s + P_\perp$, $V_W = \nu s + V_\perp$, where the two affine remainders are fixed. Lemma T.5 gives $|\pi|, |\nu| \geq c'K$, so

$$\chi_W = \gamma s^3 + O_{L^2}(K^2), \qquad \gamma = \tfrac{1}{4}\pi^2\nu, \qquad |\gamma| \geq c''K^3.$$

The nonwinner χ_j stay bounded to τ_* and their cancellation coefficients are $O(K^2)$. Hence the limit C in Lemma T.7 satisfies $\|C/\gamma - s^3\|_2 = O(K^{-1})$. At this fixed comparator time use the actual feature-span witness

$$H_\varepsilon = \frac{1}{\sqrt{6}}\left(C_\varepsilon/\gamma - 3Q_\varepsilon(s)\right).$$

It converges to a function H_τ with $\|H_\tau - h_3(s)\|_2 = O(K^{-1})$. The predictor $\mathbb{E}[y] + AH_\varepsilon$ is an allowed refit and has limiting squared error at most

$$\|F_{\geq 4}\|_2^2 + 2|A|\|F_{\geq 4}\|_2\, O(K^{-1}) + A^2 O(K^{-2}).$$

Choose τ_1 sufficiently late that this bound is strictly below $\|F_{\geq 4}\|_2^2 + \delta$, then take ε small at this fixed τ_1. This order of choices proves the refit assertion. Tracking keeps the full comparator segment compact up to that time; (83) proves the plateau bound and $L(0) = 1 + O(\varepsilon^4)$. This also makes the result compatible with the same sufficiently late endpoint used for the leading-AGOP-direction conclusion in Theorem 5.1. □

Proposition T.9 (The lower-width, fixed-time obstruction). Suppose $m \leq D - 1$ and the teacher is of type three (any rank). For every *deterministic* $\tau \geq 0$, almost surely on $\{\tau < \tau_*\}$,

$$\lim_{\varepsilon \downarrow 0} \mathcal{R}\big(\theta(t_\varepsilon(\tau))\big) = 1. \tag{88}$$

In fact $0 \leq 1 - \mathcal{R} = O_\tau(\varepsilon^2)$. This includes initialization, but asserts neither a bound at all adaptive times nor a bound for $\tau = \tau(\varepsilon) \uparrow \tau_*$.

Proof. At the fixed τ, conditioning each initialization to survive past τ and applying its smooth flow diffeomorphism gives an absolutely continuous joint law. The $m + 1 \leq D$ functions $1, q_1(\tau), \ldots, q_m(\tau)$ are linearly independent almost surely: their Gram determinant is a nonzero polynomial, as witnessed by distinct members of the monomial construction in Lemma T.6. Define $A_0 : \mathbb{R}^{m+1} \to L^2$ to have these columns and A_ε to have columns $1, \varepsilon^{-2}g_1, \ldots, \varepsilon^{-2}g_m$ at the actual time. Expansion and tracking give $\|A_\varepsilon - A_0\| = O_\tau(\varepsilon)$ and $\sigma = \sigma_{\min}(A_0) > 0$. For $y_0 = y - \mathbb{E}[y] \perp \mathcal{Q}$ and every coefficient vector v,

$$|\langle y_0, A_\varepsilon v\rangle| \leq \|y_0\|_2 \|A_\varepsilon - A_0\|\, \|v\|, \qquad \|A_\varepsilon v\|_2 \geq (\sigma - \|A_\varepsilon - A_0\|)\|v\|.$$

The norm of the orthogonal projection of y_0 onto the actual feature span is therefore $O_\tau(\varepsilon)$, and its square is $1 - \mathcal{R}$. Countably many preselected times may be intersected, but this argument does not intersect an uncountable family of times. □

Proposition T.10 (An initial-risk statement at the boundary width). Assume $d \geq 2$, $r = 1$, and $m = D$. At initialization put $q_j = \frac{1}{2}P_jV_j$ and $\chi_j = \frac{1}{4}P_j^2V_j$ using the unscaled Gaussian parameters. Almost surely $1, q_1, \ldots, q_{D-1}$ is a basis of $\mathcal{Q}$. Let the unique coefficients satisfy

$$\beta_D = 1, \qquad \beta_0 + \sum_{j=1}^{D} \beta_j q_j = 0, \qquad H_0 = (I - \Pi_{\mathcal{Q}}) \sum_{j=1}^{D} \beta_j \chi_j.$$

Almost surely $H_0 \neq 0$ and H_0 is not a scalar multiple of $h_3(u^\top x)$. For the rank-one teacher in Theorem T.8,

$$R_0 := \lim_{\varepsilon \downarrow 0} \mathcal{R}(\theta(0)) = 1 - A^2 \frac{\langle h_3(u^\top x), H_0\rangle^2}{\|H_0\|_2^2} > \|F_{\geq 4}\|_2^2. \tag{89}$$

Thus, on E_{esc} and at this width, the endpoint theorem does imply a strictly positive, initialization-dependent asymptotic initial-to-endpoint improvement. No deterministic positive lower bound on that improvement is asserted.

Proof. The basis assertion follows from the same determinant polynomial. On its nonzero set, the β_j and the Hermite coefficients of H_0 are rational functions of the initial affine parameters. Both exceptional conditions $H_0 = 0$ and $H_0 \in \operatorname{span}\{h_3(u^\top x)\}$ are algebraic after multiplying by the basis determinant. They are proper conditions: rotate coordinates so that $u = e_1$ and realize $q_1, \dots, q_{D-1}$ as all nonconstant monomials, with $q = x_2^2$ realized by $P = x_2, V = 2x_2$. Realize the last feature by $P_D = 2x_2, V_D = x_2$, so $q_D = x_2^2$ also. Their cubic difference is $x_2^3/2$, hence $H_0 = \mathrm{He}_3(x_2)/2$, which is nonzero and orthogonal to $h_3(x_1)$. A nonzero polynomial's zero set is Lebesgue null, and the Gaussian initialization has a density.

The D actual functions $1, \varepsilon^{-2} g_1, \dots, \varepsilon^{-2} g_{D-1}$ converge to a basis of $\mathcal{Q}$. The extra function

$$\varepsilon^{-3}\left[\varepsilon^2 \beta_0 + \sum_{j=1}^{D} \beta_j g_j\right] \longrightarrow C_0 := \sum_{j=1}^{D} \beta_j \chi_j$$

converges in L^2. These functions span the entire actual refit space, since $\beta_D = 1$. Their limiting Gram matrix is positive definite because $(I - \Pi_{\mathcal{Q}})C_0 = H_0 \neq 0$. Inverting their Gram matrices shows convergence of the orthogonal projectors to the projector onto $\mathcal{Q} \oplus \operatorname{span}\{H_0\}$. Here H_0 lies in the third Hermite chaos. The teacher's higher Hermite remainder and its centered lower-degree part are therefore orthogonal to that space except for the displayed cubic projection. This proves the equality in (89); strict Cauchy–Schwarz and $A \neq 0$ prove the inequality.

For clarity, let $g = R_0 - \|F_{\geq 4}\|_2^2 > 0$ on a realized escape sample. Choose the endpoint theorem with $\delta = g/4$. Its endpoint risk is at most the tail plus $g/4$ for all small ε, while convergence of the initial risk makes it at least $R_0 - g/4$. Their difference is at least $g/2$. □

Corollary T.11 (A probability bound for the boundary-width improvement). In Proposition T.10, assume additionally that the initialization law is invariant under simultaneous spatial rotations of all w_j, z_j. This holds for independent isotropic Gaussian inner blocks, with any fixed positive head and inner variances. For every $q \in (0,1)$ and $\delta > 0$, there is an event of probability at least

$$1 - 2^{-D} - \frac{3}{q\, d(d+2)} \tag{90}$$

on which a fixed random $\tau_1 < \tau_*$ can be chosen so that, for flow or any fixed-step GD and all sufficiently small ε,

$$\mathcal{R}(\theta(0)) - \mathcal{R}\big(\theta(t_\varepsilon(\tau_1))\big) \geq A^2(1-q) - \delta, \qquad \sup_{t \leq t_\varepsilon(\tau_1)} |L(t) - L(0)| = O(\varepsilon^4). \tag{91}$$

The time and initialization cutoff are sample dependent, and the probability bound is informative only when its right-hand side is positive. It is an asymptotic statement at $m = D$, not a guarantee at a prescribed finite initialization scale.

Proof. The random unoriented line spanned by H_0 is rotation invariant: rotations act on the polynomial bank, its cancellation identity, and its Hermite projection equivariantly. Normalize it as $H_0/\|H_0\|_2 = \mathrm{He}_3[T]/\sqrt{6}$ for a symmetric tensor T with Frobenius norm one; its sign will not matter. Here $\mathrm{He}_3[T] = \sum_{ijk} T_{ijk}\, \mathrm{He}_3(x)_{ijk}$, and the Hermite isometry gives $\langle h_3(u^\top x), H_0/\|H_0\|_2 \rangle^2 = T[u,u,u]^2$. Write this squared correlation as Z. For every deterministic symmetric T with $\|T\|_F = 1$, the sixth sphere moment, or Gaussian Wick pairing followed by a radial integration, gives, with S^{d-1} the Euclidean unit sphere,

$$\mathbb{E}_{\omega \sim \mathrm{Unif}(S^{d-1})}\left[T[\omega,\omega,\omega]^2\right] = \frac{6 + 9\|\operatorname{tr} T\|^2}{d(d+2)(d+4)}.$$

There are six pairings connecting all three indices of one tensor to the other, giving $6\|T\|_F^2$, and nine pairings with an internal contraction in each tensor, giving $9\|\operatorname{tr} T\|^2$. The trace map $L : T \mapsto (\sum_j T_{ijj})_i$ has adjoint

$$(L^* v)_{ijk} = (v_i \delta_{jk} + v_j \delta_{ik} + v_k \delta_{ij})/3, \qquad LL^* = \frac{d+2}{3} I_d.$$

Therefore $\|\operatorname{tr} T\|^2 \leq (d+2)/3$ and the displayed moment is at most $3/[d(d+2)]$. Averaging an independent uniform rotation and using invariance of the unoriented line gives $\mathbb{E}_\vartheta[Z] \leq 3/[d(d+2)]$.

Markov's inequality yields $\Pr(Z > q) \le 3/[q\,d(d+2)]$. Intersect $\{Z \le q\}$ with E_{esc} by a union bound; no independence or rotation-invariance conditional on escape is claimed. On this intersection the limiting initial risk exceeds the Hermite tail by at least $A^2(1-q)$. Apply the endpoint theorem with tolerance $\delta/2$ and the initial convergence with tolerance $\delta/2$ to obtain (91). □

Proposition T.12 (A one-winner projection bound at general rank). Let $1 \le r \le d$, the teacher be of type three, and $m \ge D$. Almost surely on E_{esc}, for every sufficiently late fixed $\tau < \tau_*$ define unit vectors in the teacher subspace by

$$\mu = \frac{UU^\top w_W}{\|U^\top w_W\|}, \qquad \nu = \frac{UU^\top z_W}{\|U^\top z_W\|}, \qquad \rho = \mu^\top \nu.$$

There is a finite random constant C, independent of these late times, such that

$$\limsup_{\varepsilon \downarrow 0} \mathcal{R}\big(\theta(t_\varepsilon(\tau))\big) \le 1 - \frac{T_3[\mu,\mu,\nu]^2}{2+4\rho^2} + \frac{C}{K(\tau)}. \tag{92}$$

In particular the quotient in this formula is captured teacher energy, while the whole right-hand side is residual risk. For some random $c > 0$, $|T_3[\mu,\mu,\nu]| \ge 2c/\kappa$ at all sufficiently late times. This result asserts neither axis entry nor recovery of a full teacher subspace.

Proof. Replace the rank-one leading product in the cancellation proof by

$$\chi_W = \gamma(\mu^\top x)^2(\nu^\top x) + O_{L^2}(K^2), \qquad \gamma = \tfrac{1}{4}\|U^\top w_W\|^2\|U^\top z_W\| \ge c' K^3.$$

The same bounded nonwinner inverse and $O(K^2)$ coefficients give a feature-span limit within $O(K^{-1})$ of this cubic monomial. Use Q_ε to subtract its linear Hermite part and obtain a limit within $O(K^{-1})$ of

$$H_{\mu,\nu}(x) = (\mu^\top x)^2(\nu^\top x) - \nu^\top x - 2\rho\mu^\top x.$$

This function is in the third Hermite chaos. To compute its norm without numerical approximation, take independent $Z, W \sim N(0,1)$ and write $\mu^\top x = Z$, $\nu^\top x = \rho Z + \sqrt{1-\rho^2}W$. Then

$$H_{\mu,\nu} = \rho\,\mathrm{He}_3(Z) + \sqrt{1-\rho^2}(Z^2-1)W, \qquad \|H_{\mu,\nu}\|_2^2 = 6\rho^2 + 2(1-\rho^2) = 2+4\rho^2.$$

Its correlation with $y - \mathbb{E}[y]$ is exactly $T_3[\mu,\mu,\nu]$. Projecting onto its approximating single feature-span direction proves (92); the denominator is at least two, so the $O(K^{-1})$ perturbation is uniform over unit μ, ν. Finally

$$cK^4 \le \Psi(X_W) = \frac{\kappa}{2} a_W \|U^\top w_W\|^2 \|U^\top z_W\| T_3[\mu,\mu,\nu],$$

and the absolute product of the four parameter factors is at most K^4, giving the asserted lower bound. □

Remark T.13 (Scope and order of limits). The threshold $m = D$ counts the intercept: $m+1$ leading quadratic columns first admit a cancellation when $m = D$, and there are $m - D + 1$ such coefficient directions when the quadratic bank has full rank. They are combinations of neurons, not independently available residual neurons. At larger widths further cancellation layers can improve the initial refit, so Proposition T.10 is not a lower bound for arbitrary $m \ge D$. The witness readout coefficients can be $O_\tau(\varepsilon^{-3})$ and its intercept can be $O_\tau(\varepsilon^{-1})$. The limit first fixes a late pre-escape comparator time and then shrinks initialization; neither the lower-width proposition nor compact-time tracking controls an ε-dependent approach to escape. The high-Hermite remainder is retained in the rank-one bound and need not be small. No assertion about a subsequent fixed-size decrease of the *trained* loss is proved here; release remains an open transition, with only conditional routes available. One winner's AGOP contribution can have rank two, so this one-direction refit argument does not assign all remaining teacher directions to nonwinners.

T.4 A JOINT RANK-ONE CONSEQUENCE

The initial-risk calculation gives a concrete simultaneous consequence of the alignment and refit results. This consequence is restricted to the boundary width and retains the qualitative initialization cutoff. Here d is the input dimension, m the student width, $D = (d+1)(d+2)/2$ the dimension of the degree-at-most-two polynomial space, and ε the initialization scale. The scalar teacher has the form $y = F(u^\top x)$, $\|u\| = 1$, with unit variance, vanishing degree-one and degree-two Hermite coefficients, and nonzero cubic coefficient $a_3 = \mathbb{E}[F(Z)h_3(Z)]$, $Z \sim N(0,1)$. The statistic A_{top} is the minimum squared overlap with u among unit top AGOP eigenvectors, and $\mathcal{R}$ is the unrestricted population refit MSE from (79). Probabilities below refer to the coupled Gaussian initialization; $q_{\mathrm{A}}, q_{\mathrm{R}}$ are headroom thresholds and $\delta_{\mathrm{A}}, \delta_{\mathrm{R}}$ the endpoint tolerances.

Corollary T.14 (Simultaneous alignment and refit gains)**.** *Assume the rank-one, type-three setting, $d \geq 2$, $m = D = (d+1)(d+2)/2$, and the independent isotropic Gaussian initialization of* (63)*. Put $a_3 = \langle F, h_3 \rangle$ and let $B_d \sim \mathrm{Beta}(1/2, (d-1)/2)$. Fix $q_{\mathrm{A}}, q_{\mathrm{R}} \in (0,1)$ and positive tolerances $\delta_{\mathrm{A}}, \delta_{\mathrm{R}}$, with $q_{\mathrm{A}} + \delta_{\mathrm{A}} < 1$. For flow or any fixed-step GD, choose a common plateau endpoint as in the preceding theorems. The limit inferior, as $\varepsilon \downarrow 0$, of the probability that the plateau conclusions and both inequalities*

$$A_{\mathrm{top}}(\mathrm{endpoint}) - A_{\mathrm{top}}(0) \geq 1 - \delta_{\mathrm{A}} - q_{\mathrm{A}},$$
$$\mathcal{R}(0) - \mathcal{R}(\mathrm{endpoint}) \geq a_3^2(1 - q_{\mathrm{R}}) - \delta_{\mathrm{R}}$$

hold is at least

$$\Pr(B_d \leq q_{\mathrm{A}}) - 2^{-D} - \frac{3}{q_{\mathrm{R}} d(d+2)}. \tag{93}$$

The time is proportional to ε^{-2} and the entire-prefix loss variation is $O(\varepsilon^4)$, with sample-dependent constants.

Proof. Intersect the finite-escape event with the initial cubic-correlation event used in Corollary T.11. On this intersection choose a sufficiently late fixed comparator time so that the alignment and attained-refit estimates both have the desired strict margins. Their proofs give a common realization-dependent initialization cutoff. The probability of failing that cutoff tends to zero as $\varepsilon \downarrow 0$. At each initialization scale, the probability of failing the initial AGOP-headroom event is at most $\Pr(B_d > q_{\mathrm{A}})$ by Corollary T.4. The initial cubic-correlation failure probability is at most $3/[q_{\mathrm{R}} d(d+2)]$. A union bound, without any independence assumption or conditional rotation-invariance claim, proves (93). □

For example, with a pure h_3 teacher, $d = 16$, $m = 153$, $q_{\mathrm{A}} = 0.4$, $\delta_{\mathrm{A}} = 0.1$, $q_{\mathrm{R}} = 0.25$, and $\delta_{\mathrm{R}} = 0.05$, the two gains are at least 0.5 and 0.7, respectively. The exact probability expression in (93) is approximately 0.9519. This is an asymptotic probability lower bound: it supplies neither a practical numerical initialization scale nor a finite-sample guarantee. The refit is unrestricted, and the original trained loss still remains near one throughout the certified window. Its subsequent decrease is a separate question.

T.5 WHAT IS NOT ESTABLISHED BY THE PLATEAU THEOREM

For a normalized pure cubic teacher $y = h_3(u^T x)$, where $x \sim N(0, I_d)$, $\|u\| = 1$, and $h_3(s) = (s^3 - 3s)/\sqrt{6}$, the small-initialization comparison suggests that one neuron dominates the first departure from the plateau. A natural candidate description then follows the population gradient flow of one aligned SwiGLU neuron,

$$f(x) = b_0 + \frac{\kappa}{m} a \ \mathrm{SiLU}(\rho u^T x + b)(q u^T x + c),$$

where m is the original student width, $\kappa > 0$ its output scale, a the surviving head, ρ, q the scalar gate and value weights along u, and b, c their biases. The intercept b_0 is minimized out. This is a candidate mechanism for loss release, not a consequence of the preceding compact-time comparison.

A rigorous transfer requires control from a fixed pre-escape comparison time to an actual parameter norm independent of the initialization scale, as well as a characterization of the trajectory leaving the degenerate origin. The compact-time argument stops before that transition. Thus the long plateau

interval proved above should not be interpreted as a theorem identifying its endpoint with the first substantial loss decrease.

The candidate one-neuron release mechanism described here is restricted to pure h_3. Higher Hermite components of the teacher can affect the finite-amplitude vector field; there is no claim that the same curve applies to every teacher covered by the alignment theorem. Similarly, fixed-step GD at finite amplitude has a discretization error that does not vanish merely because initialization tends to zero. A continuous-flow release curve therefore does not establish the corresponding fixed-step GD limit without an additional argument. The numerical evidence is reported with these distinctions in place.

T.6 Population-quadrature experiments with a SwiGLU student

These experiments examine finite-initialization trajectories; they do not establish a finite-ϵ probability bound or a release theorem. We use the student and profiled intercept above, $x \sim \mathcal{N}(0, I_d)$, $\kappa = 1$, and the teacher basis $U = (e_1, \ldots, e_r)$, where e_i are coordinate unit vectors. Here d is the input dimension, m the student width, r the teacher rank, κ the output scale, and $\epsilon = \varepsilon$ the initialization scale. The independent Gaussian initializations have coordinate standard deviation $\epsilon/\sqrt{d+1}$ in each inner block and standard deviation ϵ in the output head. The intercept is minimized analytically at each state; the other three blocks share one learning rate. All expectations and variances in this protocol are population quantities over x; n counts optimizer updates and θ collects the trained neuron parameters.

Teacher normalization and clock. The reproduction engine uses $y = \gamma \sum_{i=1}^{r} c_i \phi(x_i)$ with $\mathbb{E}[y^2] = 1$, where ϕ is the listed teacher link, $c = (c_1, \ldots, c_r)$ its coefficient vector, and $\gamma > 0$ the normalizing scalar. We report $\ell = L/V_y$, where $L = \mathrm{Var}(y - \widetilde{f})$ is the profiled training MSE and $V_y = \mathrm{Var}(y)$, so the intercept-only baseline is one. This differs from rescaling the teacher itself to have unit variance: for rank-one ReLU, $V_y = 1 - 1/\pi$, whereas all centered teachers here have $V_y = 1$. The force is the gradient of the unnormalized L. Recorded time is the mean-field clock $t = \sum_n \Delta t_n$ for $\theta_{n+1} = \theta_n - m\Delta t_n \nabla_\theta L$. An ordinary parameter-space learning rate is therefore $\eta = m\Delta t$. Adaptive trajectories use explicit Euler steps with a maximum relative parameter change rule and are labeled as flow approximations. Separate fixed-Δt runs test ordinary gradient descent; adaptive trajectories are not evidence for a fixed-step guarantee.

Loss-selected plateau and seed policy. For $\delta = 10^{-3}$, define the measured plateau as the maximal initial prefix of recorded optimizer states on which $|\ell(t) - \ell(0)| \le \delta$. The entire update loss history is used, before examining alignment or refit. We also report the stricter $\delta = 10^{-4}$ prefix for comparison with the original source measurements. Alignment and refit at a prefix endpoint are evaluated at the last saved diagnostic checkpoint inside it; the audit retains the exact update-grid endpoint, diagnostic time and lag. Subsequent crossing times are first observed update times, without interpolation. Seed 0 is the fixed representative; all declared seeds and unsuccessful runs are retained in the tables and supplementary panels.

Population diagnostics. The AGOP $M = \mathbb{E}[\nabla_x f \nabla_x f^\top]$ includes all neuron cross terms; $f = b_0 + \widetilde{f}$ is the trained predictor with profiled intercept. We distinguish $A_{\mathrm{top}} = \|U^\top e_1(M)\|^2$ from $A_{\min} = \sigma^2_{\min}(U^\top Q_r(M))$. Here $e_1(M)$ is a unit leading eigenvector, $Q_r(M)$ has orthonormal columns spanning the selected top-r eigenspace, and $\sigma_{\min}$ is the smallest singular value. The first measures whether one leading direction lies in the teacher subspace; the second measures its weakest recovered direction. At rank one they agree. At higher rank we display both, and do not infer full-subspace recovery from A_{top}.

Integration and numerical refit. The population integrals are evaluated by deterministic quadrature, with no finite training sample. Two-projection feature pairs use tensor Gauss–Hermite rules, diagonal terms use a one-dimensional rule, and teacher terms condition on the relevant teacher coordinate. That coordinate is integrated on $[-10, 10]$ using piecewise Gauss–Legendre rules weighted by the Gaussian density, split at $-6, -4, -2, 0, 2, 4, 6$ and link kinks. These are numerical approximations, not certified exact integrals. The force uses Gaussian Stein identities before quadrature; at finite order it need not be the exact derivative of the quadrature-discretized loss scalar. Higher-order checks and Gram-matrix diagnostics are retained with the raw records.

For the frozen centered feature vector $\bar{g}$, with $\bar{g}_j = g_j - \mathbb{E}[g_j]$, write its Gram matrix as $G = \mathbb{E}[\bar{g}\bar{g}^\top]$, its teacher covariance as $b = \mathbb{E}[(y - \mathbb{E}[y])\bar{g}]$, and a candidate head as $w \in \mathbb{R}^m$. This w is a readout vector, distinct from the spatial gate weights w_j. Its normalized prediction risk is

$$R(w) = \frac{V_y - 2b^\top w + w^\top G w}{V_y}.$$

The unrestricted population refit is $\inf_w R(w)$, without a coefficient budget. Numerical ridge or eigencutoff solves approximate this infimum; their attained risks and cutoff sensitivity are reported explicitly. In particular, the source diagnostic $[V_y - b^\top (G + \lambda I)^{-1} b]/V_y$, where λ is the ridge parameter and $I = I_m$, includes the ridge penalty and is not the prediction risk. We retain it for source reproduction, but do not label it an exact unrestricted refit. This distinction is most relevant for the $d = 16, m = 256$ cancellation experiment.

Runtime amendments and censoring. The seed/configuration matrix was fixed before production, but the execution horizons were amended for runtime and are not presented as wholly preregistered. Before any fixed-step control launched, its nominal mean-field horizon was capped at 250; the original steps and seed were retained. The stopping rule is checked on the accumulated update clock, allowing at most one step of overshoot (the $\Delta t = .1$ run stops at $t = 250.1$), and actual endpoint times are retained. After a rank-two trajectory stalled with very small adaptive steps, a common rank-two budget of time 500 and 3000 updates was adopted. The balanced seed 2 run completed naturally at update3060; its raw tail is retained but the endpoint analysis stops at update3000. Only unbalanced seed 0 was interrupted and replayed with the same initialization and force, with the interrupted attempt and its log preserved. Other previously completed trajectories finished inside the amended bounds. The results distinguish final outcome slots from interrupted execution attempts and mark all budget-censored outcomes. These finite observation windows do not establish long-time convergence.

T.7 Finite-initialization outcomes

We retain the notation of Section T.6: d, m, r are the input dimension, student width, and teacher rank; ϵ is the initialization scale; c is the teacher coefficient vector; and t is mean-field time. The reported $\ell = L/\operatorname{Var}(y)$ is normalized trained loss, A_{top} and $A_{\min}$ are leading and weakest-direction AGOP alignments, and R is normalized attained numerical refit MSE. The polynomials are $h_2(s) = (s^2 - 1)/\sqrt{2}$ and $h_3(s) = (s^3 - 3s)/\sqrt{6}$. Subscripts or superscripts $0, P, D, f$ mark initialization, the last diagnostic inside a loss prefix, the last in-budget diagnostic, and the final in-budget update, respectively.

The predeclared matrix contains 24 trajectories: two source smoke checks, three rank-one cubic reproductions, nine rank-one illustrations with degree-one or degree-two teacher signal, six balanced/unbalanced rank-two illustrations, one width-256 source reproduction, and three fixed-step controls. Figures 11–17 retain all groups. Tables 3 and 4 report every seed, including the stricter loss-prefix comparison; Table 5 records final observations.

For rank-one h_3, initial leading-direction alignments are 0.128, 0.022, 0.000 (seeds 0, 1, 2). At the primary loss-selected endpoints they are 0.998, 0.994, 0.994, and at the stricter endpoints 0.986, 0.974, 0.977. The primary update-grid prefix durations are 226.4, 266.7, 312.1. The final observed losses are 0.4999, 0.4997, 0.5000. These subsequent decreases are measurements beyond the plateau, not consequences of the alignment theorem.

For $h_3 + .3h_2$, all three primary-prefix leading-direction alignments lie in [0.992,0.996], followed by final observed normalized losses in [0.6972,0.7000].

For ReLU, all three primary-prefix leading-direction alignments lie in [0.979,0.997], followed by final observed normalized losses in [0.6932,0.6972].

For tanh, all three primary-prefix leading-direction alignments lie in [0.956,0.983], followed by final observed normalized losses in [0.6926,0.6998].

For $c = (1, 1)$, the primary-prefix weakest-direction alignments are 0.976, 0.151, 0.178 (seeds 0, 1, 2), while the corresponding leading-direction alignments are 0.997, 0.997, 0.995.

For $c = (1, .6)$, the primary-prefix weakest-direction alignments are 0.008, 0.067, 0.004 (seeds 0, 1, 2), while the corresponding leading-direction alignments are 0.997, 0.994, 0.994.

Full-subspace alignment need not be monotone within the low-loss-movement regime: balanced rank-two seed 2 has $A_{\min} = .871$ at the stricter prefix checkpoint, .178 at the primary-prefix checkpoint and approximately .999 at its last in-budget diagnostic. These endpoints illustrate variability rather than persistent failure or monotone recovery.

For the width-256 reproduction, the numerical refit risk changes from 0.8183 to 0.0435 on the primary loss prefix, with normalized loss 1.000000 to 0.999037. At the source-comparison tolerance 1.2×10^{-5}, the legacy penalized diagnostic changes from 0.8183 to 0.3306, whereas the corrected eigencutoff prediction risk changes from 0.8183 to 0.3306. This additional threshold was fixed from the source statement before the new width-256 output arrived.

The source schedule additionally saved update 66, which lies between new diagnostics 52 and 69. An exact 14-update replay from saved state 52 recovers its $t = 48.834273$, $\ell = 0.999988250$ and legacy diagnostic 0.232107532. The attained eigencutoff prediction risk is 0.232107528 (all four cutoffs agree), changing by only 1.11×10^{-14} at doubled quadrature. Thus the source decrease from approximately .82 to .23 is reproduced at its original loss movement 1.17993×10^{-5}. This source-defined replay is a checkpoint validation, not an additional seed, a new primary endpoint or a selected minimum.

All 3830 valid saved-state refit diagnostics have finite normalized risks in $[0, 1]$ and positive equilibrated Gram eigenvalues (smallest eigenvalue ratio 2.88×10^{-4}). The four prescribed spectral cutoffs retain full rank and give the same risk to recorded precision. However, 955 saved states have Gram relative asymmetry above 10^{-8}, with maximum 1.07×10^{-4}; the solver uses the symmetric part. Positivity and cutoff stability therefore do not certify the underlying Gaussian integrals or the exact unrestricted infimum. Five post-failure diagnostics are explicitly omitted.

Doubling quadrature at the selected rank-one cubic seed-0 states changes refit risk by at most 1.08×10^{-6}; at the width-256 primary-prefix checkpoint the change is 3.36×10^{-8}, increasing to 4.36×10^{-4} at its final state. Across the six rank-two runs the largest checked change is 4.52×10^{-4} in refit risk and 7.28×10^{-4} in leading alignment; this includes the separately retained out-of-budget tail. These saved-state comparisons are neither integral-error bounds nor higher-order retraining.

The fixed $\Delta t = .1$ endpoint has a material quadrature sensitivity: doubled orders change its normalized loss from .650360 to .653321 and numerical refit risk from .412908 to .401575 (difference .011333). Its fine numerical endpoint values should not be interpreted as resolved to the displayed digits. The $\Delta t = .05$ endpoint is unchanged at recorded precision, but its trajectory is nonmonotone and returns to loss 1.000056 by the horizon.

The unchanged source validation script exits with code zero but prints four failed checks: two finite-difference gradient discrepancies (1.35×10^{-4} and 1.79×10^{-4}) and two Monte Carlo comparisons with largest standardized discrepancies 4.61 and 4.66. We retain these failures. The separate higher-order finite-difference diagnostic reduces its four gradient discrepancies to 2.12×10^{-9}, 3.86×10^{-7}, 4.88×10^{-8} and 7.01×10^{-7}, respectively; this convergence evidence does not replace the failed checks. The successful record-integrity audit is a distinct claim.

The fixed $\Delta t = 0.3$ control is a retained numerical failure: after exact fixed steps through update 762, the source relative-gradient safeguard activates and the numerical clock freezes near $t = 228.9005$. The recorded loss reaches approximately 6.11×10^{237}. Post-safeguard alignment/refit is omitted from the figure and no stability or release conclusion is drawn from that tail.

The fixed $\Delta t = 0.1$ control ends at its actual $t = 250.1000$ with normalized loss 0.6504 at the nominal observation horizon (censored).

The fixed $\Delta t = 0.05$ control ends at its actual $t = 250.0000$ with normalized loss 1.0001 at the nominal observation horizon (censored).

The rank-two panels show why leading-direction alignment and recovery of the whole teacher subspace must be separated: the dashed $A_{\min}$ curves can lag far behind the solid A_{top} curves. All three seeds are displayed for each coefficient vector. The width-256 and fixed-step experiments are interpreted separately, with the numerical limitations just described.

In all panels, circular markers identify the last diagnostic within the primary loss-selected prefix; a cross at a loss-curve endpoint marks a numerical failure or the common time/step analysis horizon reached before the loss stopping target. Any retained raw states beyond the common budget appear as faint dotted tails and are excluded from the endpoint tables. A single-run panel shades that prefix; a multi-run panel uses a separate vertical dotted boundary for each run. Refit curves use the predetermined 10^{-12} relative eigencutoff after diagonal equilibration. Shading around those curves spans the four cutoffs $10^{-8}, 10^{-10}, 10^{-12}, 10^{-14}$; it is a numerical sensitivity range, not a statistical confidence interval or a rigorous risk bound. Gray dotted refit curves, when visibly different, give the actual prediction risk of the source ridge coefficients. Hollow squares show doubled-quadrature evaluations of the same saved states, without retraining; the width-256 triangle marks the separately reproduced source checkpoint 66.

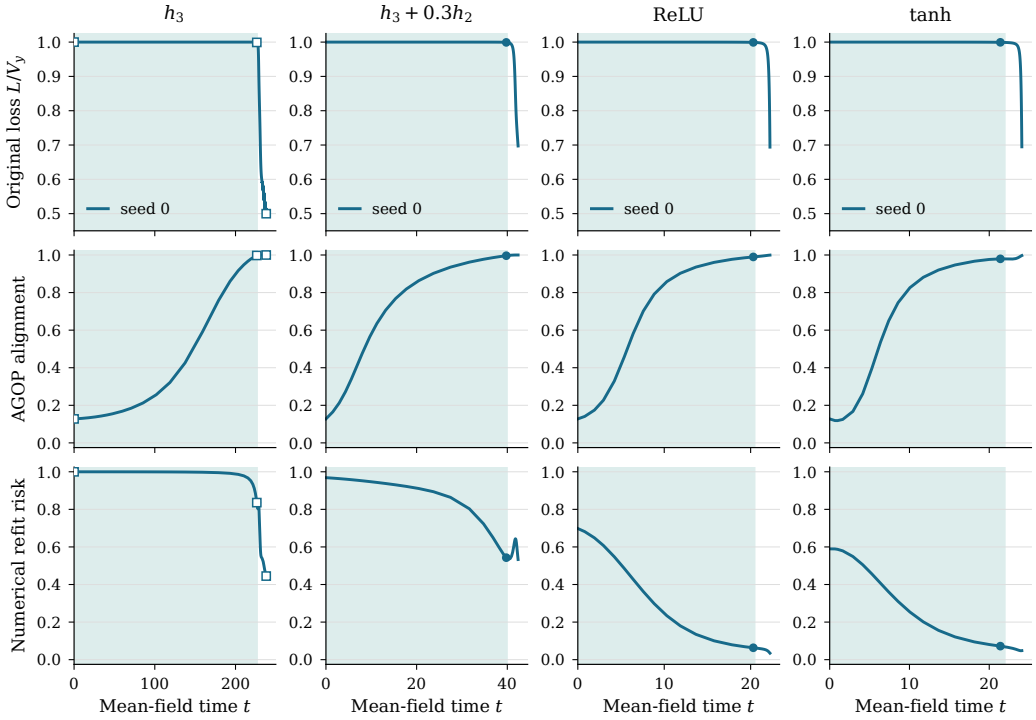


Figure 11: Fixed representative seed 0 for the four rank-one teacher links. Rows show original normalized loss, leading AGOP alignment and numerical frozen-feature refit risk. These are adaptive Euler flow approximations, with $d = 16, m = 64, \epsilon = .05$. The display includes the subsequent measured loss decrease.

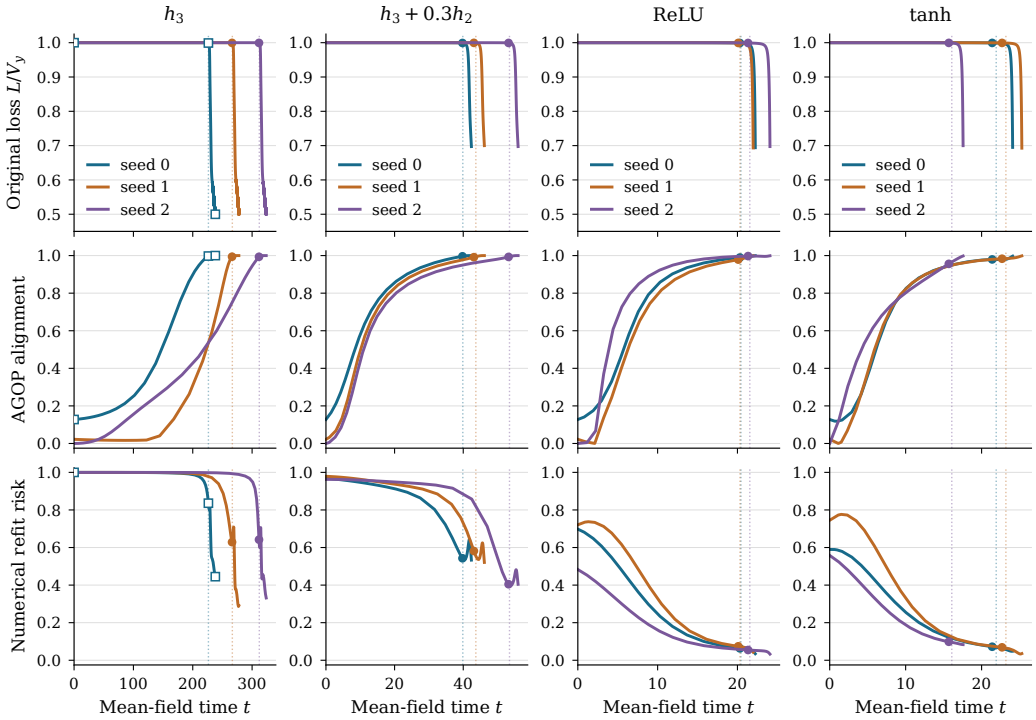


Figure 12: All three predeclared seeds for every rank-one link, using the same scales and diagnostics as Figure 11. Each curve is an individual run; no seed is selected by its outcome.

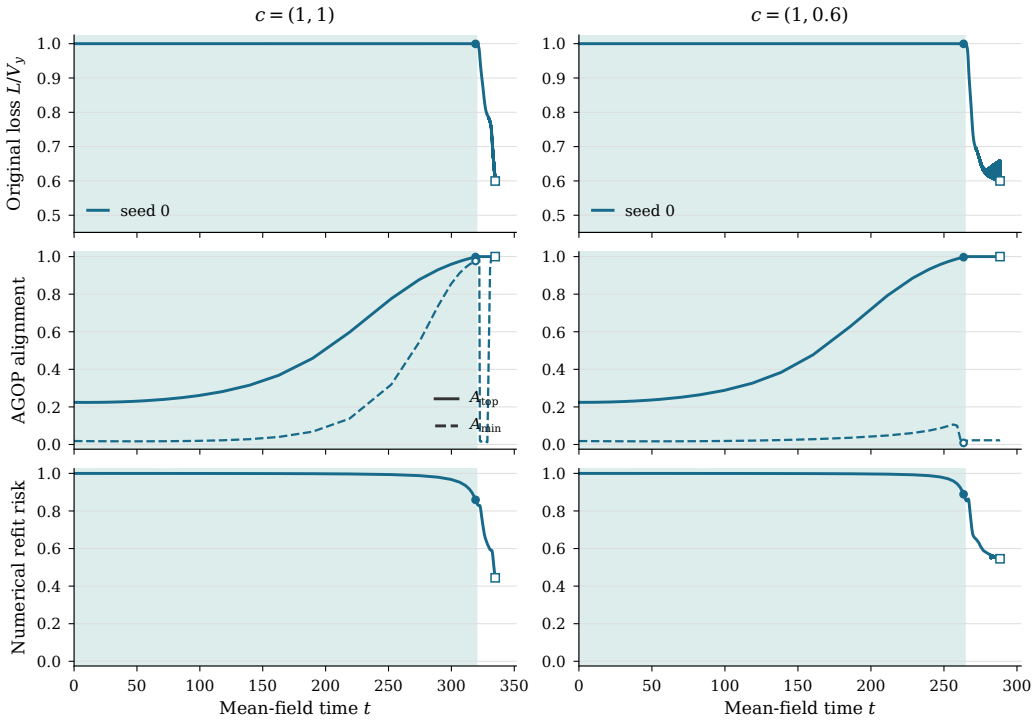


Figure 13: Fixed seed 0 for balanced and unbalanced rank-two cubic teachers. Solid curves are A_{top} and dashed curves are A_{min}. Strong leading-direction alignment can coexist with weak recovery of the least aligned teacher direction.

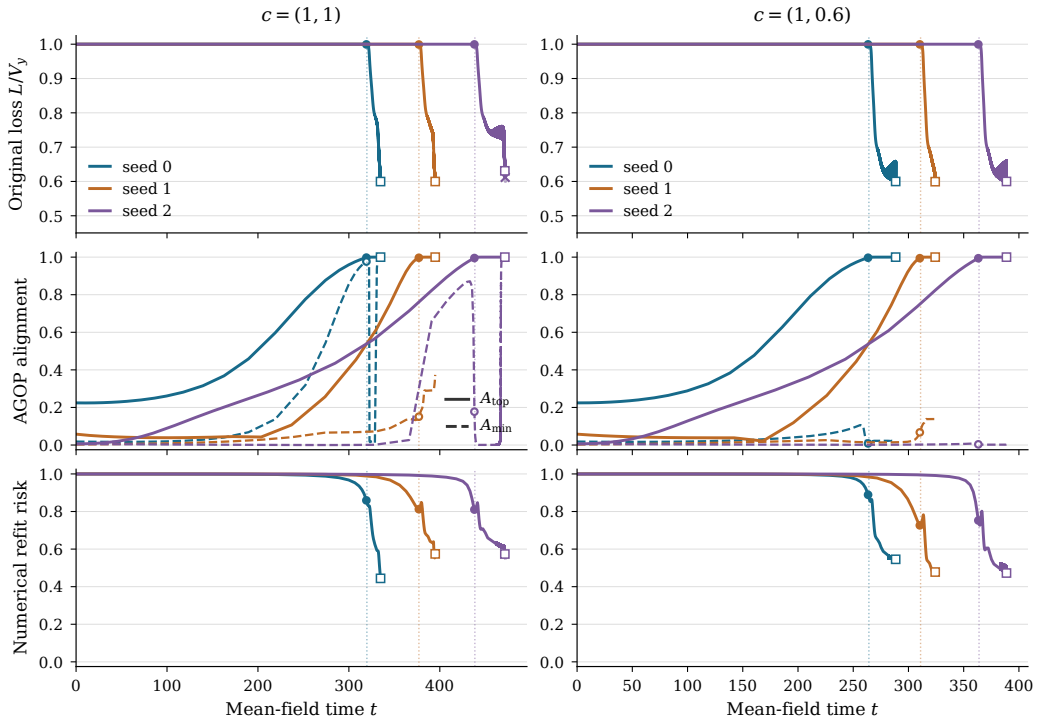


Figure 14: All three seeds for each rank-two coefficient vector. The weakest-direction statistic is retained for every seed, including cases with poor simultaneous recovery.

Table 3: Every declared SwiGLU trajectory at the loss-selected prefix with $\delta = 10^{-3}$. P is the last diagnostic inside the prefix, so t_P may precede the exact last admissible update (both clocks are retained in the audit). R is the attained prediction risk of the equilibrated 10^{-12} eigencutoff solve, not a certified unrestricted infimum. $--$ means the weakest-direction statistic duplicates A_{top} at rank one. The source smokes use $d = 4, m = 6$; the scientific runs use $d = 16, m = 64, \epsilon = .05$, except $m = 256, \epsilon = .1$. The three GD controls share seed 0; † marks the stated time/step analysis horizon reached before the loss stopping target, rather than convergence; ‡ denotes the retained numerical divergence of the $\Delta t = .3$ control. Loss values are divided by the teacher variance.

Case (seed)	t_P	$\ell_0 \to \ell_P$	$A^0_{\text{top}} \to A^P_{\text{top}}$	$A^0_{\text{min}} \to A^P_{\text{min}}$	$R_0 \to R_P$	ℓ_f
Smoke $r = 2$ (0)	445.6	1.000000 → 0.999418	0.193 → 0.995	0.121 → 0.256	1.000 → 0.961	0.8994
Smoke $r = 1$ (0)	235.4	1.000000 → 0.999461	0.113 → 0.987	–	1.000 → 0.960	0.8977
h_3 (0)	226.4	1.000000 → 0.999039	0.128 → 0.998	–	1.000 → 0.836	0.4999
h_3 (1)	266.2	1.000000 → 0.999403	0.022 → 0.994	–	1.000 → 0.629	0.4997
h_3 (2)	311.6	1.000000 → 0.999401	0.000 → 0.994	–	1.000 → 0.642	0.5000
$h_3, m = 256$ (0)	55.5	1.000000 → 0.999037	0.150 → 0.995	–	0.818 → 0.043	0.8994
$h_3 + .3h_2$ (0)	39.8	0.999999 → 0.999128	0.128 → 0.996	–	0.969 → 0.543	0.6972
$h_3 + .3h_2$ (1)	43.1	1.000000 → 0.999493	0.022 → 0.992	–	0.979 → 0.582	0.7000
$h_3 + .3h_2$ (2)	53.2	1.000000 → 0.999280	0.000 → 0.993	–	0.963 → 0.405	0.6972
ReLU (0)	20.3	0.999999 → 0.999130	0.128 → 0.990	–	0.698 → 0.063	0.6939
ReLU (1)	20.1	1.000001 → 0.999240	0.022 → 0.979	–	0.721 → 0.074	0.6932
ReLU (2)	21.3	0.999998 → 0.999129	0.000 → 0.997	–	0.483 → 0.055	0.6972
tanh (0)	21.4	1.000001 → 0.999313	0.128 → 0.980	–	0.590 → 0.072	0.6948
tanh (1)	22.7	1.000001 → 0.999306	0.022 → 0.983	–	0.745 → 0.069	0.6926
tanh (2)	15.7	0.999998 → 0.999397	0.000 → 0.956	–	0.558 → 0.098	0.6998
$r = 2, (1, 1)$ (0)	319.2	1.000000 → 0.999374	0.224 → 0.997	0.018 → 0.976	1.000 → 0.859	0.5997
$r = 2, (1, 1)$ (1)	377.0	1.000000 → 0.999060	0.058 → 0.997	0.003 → 0.151	1.000 → 0.812	0.5998
$r = 2, (1, 1)$ (2)	438.0	1.000000 → 0.999354	0.006 → 0.995	0.006 → 0.178	1.000 → 0.810	0.6119†
$r = 2, (1, .6)$ (0)	263.4	1.000000 → 0.999486	0.224 → 0.997	0.018 → 0.008	1.000 → 0.889	0.6000
$r = 2, (1, .6)$ (1)	310.3	1.000000 → 0.999465	0.058 → 0.994	0.003 → 0.067	1.000 → 0.726	0.5994
$r = 2, (1, .6)$ (2)	363.3	1.000000 → 0.999434	0.006 → 0.994	0.006 → 0.004	1.000 → 0.752	0.6000
GD $\Delta t = 0.3$ (0)	221.4	1.000000 → 0.999424	0.128 → 0.997	–	1.000 → 0.853	failed‡
GD $\Delta t = 0.1$ (0)	220.9	1.000000 → 0.999282	0.128 → 0.997	–	1.000 → 0.846	0.6504†
GD $\Delta t = 0.05$ (0)	220.3	1.000000 → 0.999507	0.128 → 0.996	–	1.000 → 0.859	1.0001†

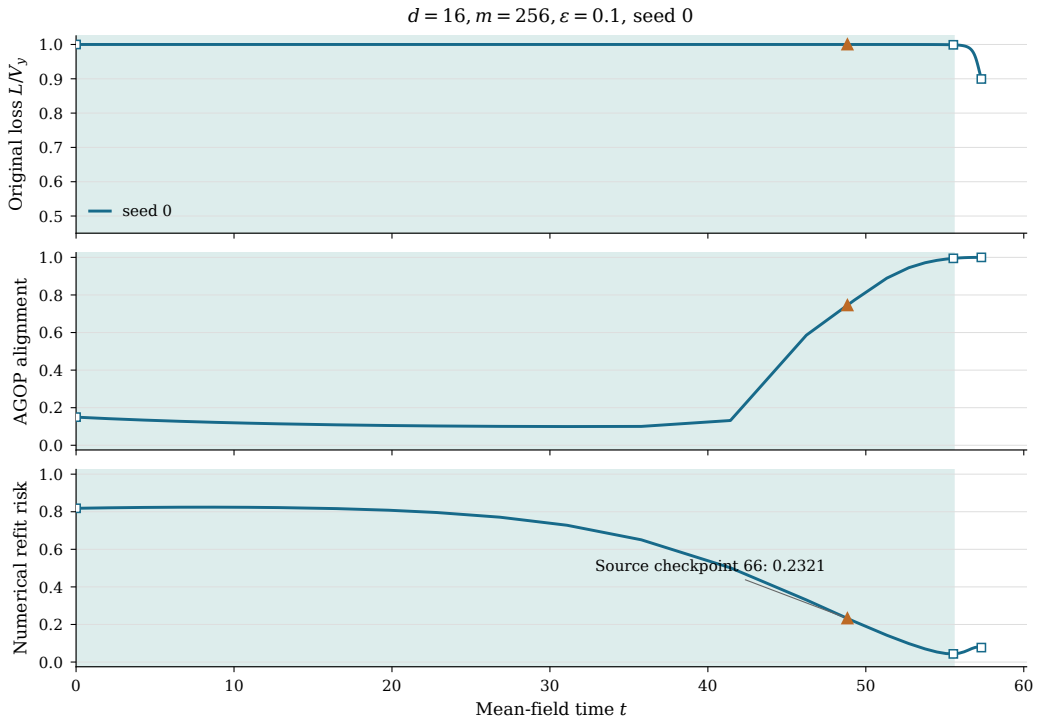


Figure 15: Width-256 rank-one source reproduction ($d = 16, \epsilon = .1$, seed 0). The refit panel reports attained prediction risk from the unregularized numerical head solve. The triangle is the separately validated source checkpoint 66, whose corrected risk is .232108; it does not change the primary prefix. The width exceeds $(d+1)(d+2)/2 = 153$. Quadrature and spectral-cutoff checks accompany these finite-initialization values; the display does not assert an exact zero-risk fit.

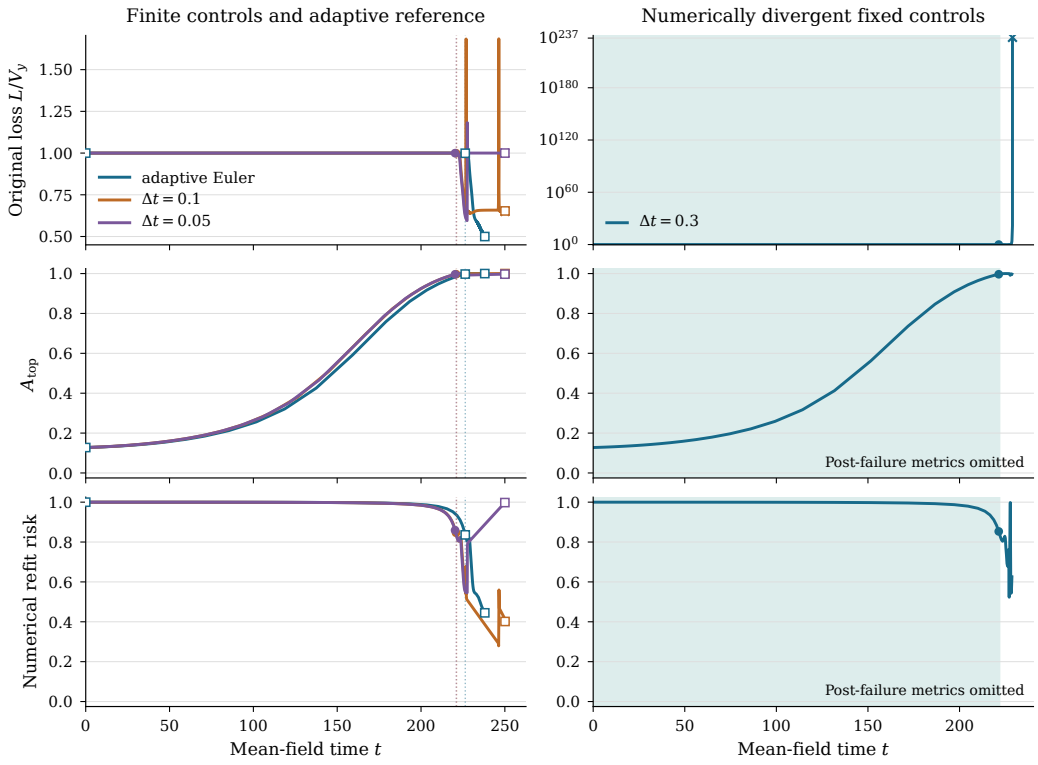


Figure 16: Fixed mean-field steps $\Delta t = .3, .1, .05$ versus the adaptive flow approximation, all with the same rank-one cubic seed 0 initialization. Parameter-space learning rates are $m\Delta t$. The x-axis is accumulated mean-field time, not optimizer step count. The right column retains numerically divergent controls on their full loss scale, while omitting alignment/refit after the source safeguard activates; the left column displays finite controls and the adaptive reference. The nominal time horizon is 250, checked on the accumulated update clock with at most one-step overshoot (actual endpoint times are retained); any trajectory reaching that horizon is censored, and no long-time convergence is claimed. These controls do not convert measured release into a fixed-step release theorem.

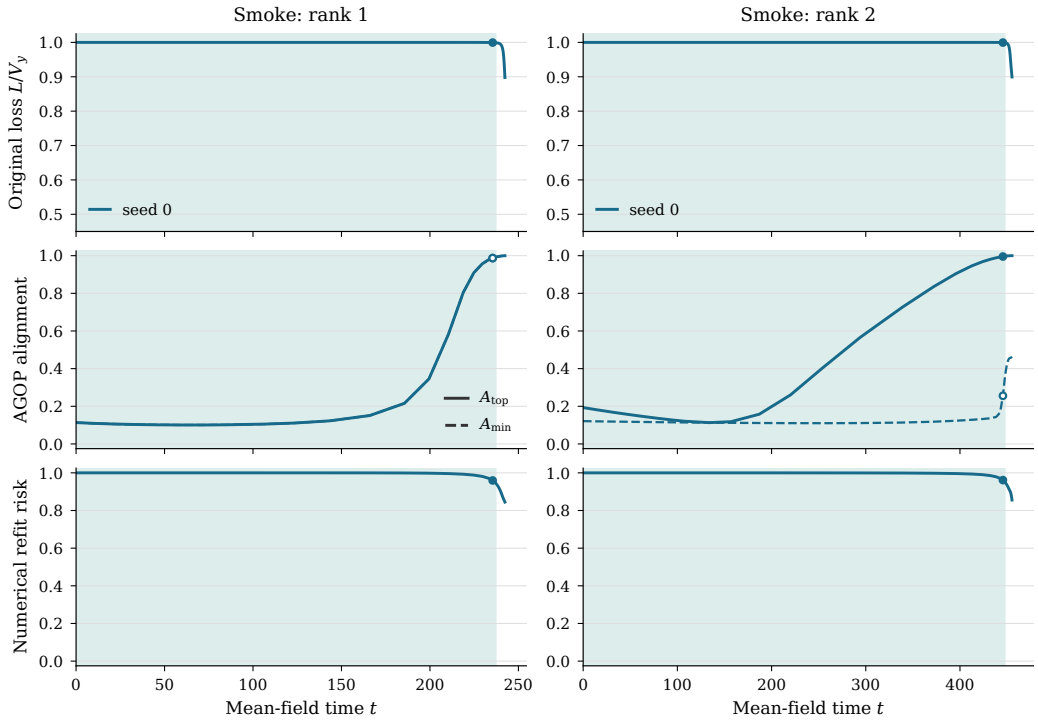


Figure 17: The two predeclared source smoke checks ($d = 4, m = 6$, seed 0), retained for completeness. The rank-one smoke uses $\epsilon = .1$ and the rank-two smoke $\epsilon = .05$.

Table 4: Every declared SwiGLU trajectory at the loss-selected prefix with $\delta = 10^{-4}$. P is the last diagnostic inside the prefix, so t_P may precede the exact last admissible update (both clocks are retained in the audit). R is the attained prediction risk of the equilibrated 10^{-12} eigencutoff solve, not a certified unrestricted infimum. $--$ means the weakest-direction statistic duplicates A_{top} at rank one. The source smokes use $d = 4, m = 6$; the scientific runs use $d = 16, m = 64, \epsilon = .05$, except $m = 256, \epsilon = .1$. The three GD controls share seed 0; † marks the stated time/step analysis horizon reached before the loss stopping target, rather than convergence; ‡ denotes the retained numerical divergence of the $\Delta t = .3$ control. Loss values are divided by the teacher variance.

Case (seed)	t_P	$\ell_0 \to \ell_P$	$A^0_{\text{top}} \to A^P_{\text{top}}$	$A^0_{\text{min}} \to A^P_{\text{min}}$	$R_0 \to R_P$	ℓ_f
Smoke $r = 2$ (0)	432.5	$1.000000 \to 0.999930$	$0.193 \to 0.981$	$0.121 \to 0.137$	$1.000 \to 0.985$	0.8994
Smoke $r = 1$ (0)	224.9	$1.000000 \to 0.999938$	$0.113 \to 0.910$	–	$1.000 \to 0.988$	0.8977
h_3 (0)	221.5	$1.000000 \to 0.999945$	$0.128 \to 0.986$	–	$1.000 \to 0.933$	0.4999
h_3 (1)	262.4	$1.000000 \to 0.999927$	$0.022 \to 0.974$	–	$1.000 \to 0.706$	0.4997
h_3 (2)	307.6	$1.000000 \to 0.999930$	$0.000 \to 0.977$	–	$1.000 \to 0.859$	0.5000
$h_3, m = 256$ (0)	52.7	$1.000000 \to 0.999945$	$0.150 \to 0.945$	–	$0.818 \to 0.098$	0.8994
$h_3 + .3h_2$ (0)	34.8	$0.999999 \to 0.999949$	$0.128 \to 0.977$	–	$0.969 \to 0.723$	0.6972
$h_3 + .3h_2$ (1)	39.3	$1.000000 \to 0.999941$	$0.022 \to 0.975$	–	$0.979 \to 0.759$	0.7000
$h_3 + .3h_2$ (2)	49.7	$1.000000 \to 0.999914$	$0.000 \to 0.980$	–	$0.963 \to 0.561$	0.6972
ReLU (0)	13.7	$0.999999 \to 0.999941$	$0.128 \to 0.936$	–	$0.698 \to 0.135$	0.6939
ReLU (1)	16.1	$1.000001 \to 0.999909$	$0.022 \to 0.944$	–	$0.721 \to 0.112$	0.6932
ReLU (2)	13.7	$0.999998 \to 0.999939$	$0.000 \to 0.968$	–	$0.483 \to 0.097$	0.6972
tanh (0)	15.7	$1.000001 \to 0.999917$	$0.128 \to 0.947$	–	$0.590 \to 0.121$	0.6948
tanh (1)	16.8	$1.000001 \to 0.999918$	$0.022 \to 0.956$	–	$0.745 \to 0.112$	0.6926
tanh (2)	12.2	$0.999998 \to 0.999927$	$0.000 \to 0.864$	–	$0.558 \to 0.140$	0.6998
$r = 2, (1, 1)$ (0)	314.3	$1.000000 \to 0.999926$	$0.224 \to 0.990$	$0.018 \to 0.958$	$1.000 \to 0.916$	0.5997
$r = 2, (1, 1)$ (1)	370.9	$1.000000 \to 0.999944$	$0.058 \to 0.976$	$0.003 \to 0.135$	$1.000 \to 0.850$	0.5998
$r = 2, (1, 1)$ (2)	433.2	$1.000000 \to 0.999925$	$0.006 \to 0.982$	$0.006 \to 0.871$	$1.000 \to 0.908$	0.6119†
$r = 2, (1, .6)$ (0)	259.1	$1.000000 \to 0.999940$	$0.224 \to 0.988$	$0.018 \to 0.100$	$1.000 \to 0.945$	0.6000
$r = 2, (1, .6)$ (1)	305.9	$1.000000 \to 0.999935$	$0.058 \to 0.975$	$0.003 \to 0.029$	$1.000 \to 0.782$	0.5994
$r = 2, (1, .6)$ (2)	358.7	$1.000000 \to 0.999934$	$0.006 \to 0.979$	$0.006 \to 0.008$	$1.000 \to 0.904$	0.6000
GD $\Delta t = 0.3$ (0)	216.6	$1.000000 \to 0.999949$	$0.128 \to 0.985$	–	$1.000 \to 0.936$	failed‡
GD $\Delta t = 0.1$ (0)	217.2	$1.000000 \to 0.999923$	$0.128 \to 0.988$	–	$1.000 \to 0.923$	0.6504†
GD $\Delta t = 0.05$ (0)	216.4	$1.000000 \to 0.999940$	$0.128 \to 0.987$	–	$1.000 \to 0.931$	1.0001†

Table 5: Final observed outcomes for all 24 conditions. t_f, ℓ_f use the complete update history inside the stated analysis budget; D is the latest saved diagnostic within that budget. For balanced rank-two seed 2, $t_f = 471.9$ is update 3000, whereas the last diagnostic is update 2928 at $t_D = 471.5$; the raw 60-update tail remains archived and dotted in the plot. Failed-control post-safeguard metrics are omitted. Refit risks are numerical attained risks, subject to the quadrature limitations in the text; in particular the fixed .1 endpoint risk changes by .01133 on doubling quadrature.

Case (seed)	t_f	ℓ_f	t_D	A^D_{top}	$A^D_{\min}$	R_D	Status
Smoke $r = 2$ (0)	455.2	0.8994	455.2	1.000	0.463	0.856	loss stop
Smoke $r = 1$ (0)	242.4	0.8977	242.4	1.000	–	0.845	loss stop
h_3 (0)	238.0	0.4999	238.0	1.000	–	0.445	loss stop
h_3 (1)	277.7	0.4997	277.7	1.000	–	0.291	loss stop
h_3 (2)	323.8	0.5000	323.8	1.000	–	0.332	loss stop
$h_3, m = 256$ (0)	57.3	0.8994	57.3	1.000	–	0.077	loss stop
$h_3 + .3h_2$ (0)	42.4	0.6972	42.4	1.000	–	0.533	loss stop
$h_3 + .3h_2$ (1)	46.2	0.7000	46.2	1.000	–	0.521	loss stop
$h_3 + .3h_2$ (2)	56.0	0.6972	56.0	1.000	–	0.404	loss stop
ReLU (0)	22.2	0.6939	22.2	0.999	–	0.035	loss stop
ReLU (1)	22.0	0.6932	22.0	0.999	–	0.043	loss stop
ReLU (2)	24.1	0.6972	24.1	0.999	–	0.033	loss stop
tanh (0)	24.1	0.6948	24.1	0.997	–	0.048	loss stop
tanh (1)	25.3	0.6926	25.3	0.999	–	0.036	loss stop
tanh (2)	17.6	0.6998	17.6	0.998	–	0.083	loss stop
$r = 2, (1, 1)$ (0)	334.8	0.5997	334.8	1.000	1.000	0.444	loss stop
$r = 2, (1, 1)$ (1)	394.9	0.5998	394.9	1.000	0.373	0.574	loss stop
$r = 2, (1, 1)$ (2)	471.9	0.6119	471.5	1.000	0.999	0.574	censored
$r = 2, (1, .6)$ (0)	288.4	0.6000	288.4	1.000	0.022	0.546	loss stop
$r = 2, (1, .6)$ (1)	324.2	0.5994	324.2	1.000	0.138	0.478	loss stop
$r = 2, (1, .6)$ (2)	388.7	0.6000	388.7	1.000	0.002	0.473	loss stop
GD $\Delta t = 0.3$ (0)	228.9	6.11×10^{237}	–	–	–	–	failure
GD $\Delta t = 0.1$ (0)	250.1	0.6504	250.1	1.000	–	0.413	censored
GD $\Delta t = 0.05$ (0)	250.0	1.0001	250.0	0.997	–	0.998	censored

T.8 TWENTY-SEED RANK-ONE EXTENSION

We use d for input dimension, m for student width, κ for the output scale, ε for initialization scale, n for the optimizer update index, and t for accumulated mean-field time. The diagnostics are normalized trained loss $\ell_n = L(\theta_n)/\operatorname{Var}(y)$, leading alignment $A_{\text{top}} = \|U^\top e_1(M)\|^2$, and normalized numerical refit MSE R as defined in Section T.6. Here M is the population AGOP, U is the rank-one teacher basis, and $e_1(M)$ is its computed unit leading eigenvector. In the tables, $\Delta A = \Delta A_{\text{top}} = A_{\text{top}}(P) - A_{\text{top}}(0)$ and $\Delta R = R(0) - R(P)$, where P is the last saved diagnostic inside the primary loss prefix. The teacher polynomials are $h_2(s) = (s^2 - 1)/\sqrt{2}$ and $h_3(s) = (s^3 - 3s)/\sqrt{6}$.

The four rank-one teacher configurations are extended from seeds 0–2 to seeds 0–19, giving eighty outcomes. The original twelve runs are retained exactly, and sixty-eight new runs use the same initial distribution, population force, adaptive step rule, quadrature and stopping criteria as Section T.6. The other twelve smoke, rank-two, wide-student and fixed-step control conditions are unchanged. Thus the combined study has ninety-two distinct conditions. The new rank-one initializations are a declared consecutive extension, with no selection based on alignment, loss or plot appearance. Fifty-nine trajectories were completed locally and twenty-one on a CPU server, with NumPy 1.24.3 and SciPy 1.10.1 in both environments and unchanged numerical kernels. The final termination counts are 80 loss-threshold stops.

For each run, $d = 16$, $m = 64$, $\kappa = 1$ and $\varepsilon = .05$. The relative-change step cap is .01 and the maximum mean-field step is 50. The cubic stops at normalized loss .5, the other links at .7, subject to ceilings of 20,000 updates and mean-field time 10^6. These are adaptive Euler flow approximations, with the output intercept profiled exactly. They do not instantiate the fixed-step theorem at a quantified practical initialization scale.

Uniform displays and distinct meanings. The colors, row order, typography and seed-summary convention match Figure 1. Loss and refit divide by teacher variance, the refit coefficients are unrestricted, and rank one makes the leading and weakest-direction alignments identical. The teacher-plane snapping and two-axis angle diagnostics in the ReLU figure do not apply to this rank-one comparison.

Each loss curve includes every recorded update. AGOP and corrected refit diagnostics use their original saved states; linear interpolation onto integer update indices is used only for drawing seed means and pointwise empirical 10th–90th percentiles. Such interpolation adds no observations. Means and bands stop at the earliest terminal update among all twenty seeds for that teacher. Individual recorded tails remain visible beyond that common support. There is no endpoint holding, extrapolation, survivor-only averaging or time alignment of individual trajectories. The display coordinate is $\log(1 + n/100)$ with actual update ticks; adaptive update index is a numerical-work clock, not elapsed flow time.

Numerical checks and endpoint accounting. The corrected prediction risk uses the same equilibrated Gram solve at relative cutoff 10^{-12}, with sensitivity checks at 10^{-8}, 10^{-10} and 10^{-14}. Every new seed also doubles all quadrature orders at initialization, its primary-prefix diagnostic and its final state. These are checks of stored states, not retraining or certified integration-error bounds. Across the 207 resolved comparisons, the largest absolute changes in normalized loss, leading alignment and normalized refit risk are 1.09×10^{-6}, 3.85×10^{-8} and 2.88×10^{-3}, respectively; 0 comparisons remain unresolved. Nonfinite or negative prediction risks, nonpositive Gram spectra, and numerically unresolved AGOP leading directions are flagged. The AGOP flag uses relative leading gap $> 10^{-8}$ and smallest-eigenvalue ratio $\geq -10^{-8}$; it is a numerical screening rule, not an error certificate. An unresolved seed suppresses the corresponding ensemble statistic instead of being dropped. The recorded unresolved-diagnostic count is 0.

For each seed, the primary plateau is the full-update initial prefix $|\ell_n - \ell_0| \leq 10^{-3}$. Endpoint gains use the last saved diagnostic inside that prefix; they are never evaluated at an interpolated threshold crossing. Table 6 summarizes these gains. The joint count requires alignment gain at least .5 and normalized refit-MSE gain at least .399 at one common saved in-prefix checkpoint. It counts verified successes among all twenty declared seeds, with unresolved cases kept distinct from resolved failures. Table 7 reports every seed's endpoint gains. Later loss changes are measured

outcomes, not conclusions of a trained-loss release theorem. The $.5/.399$ thresholds are illustrative summaries of this unrestricted numerical readout; they do not test the bounded-refit ReLU theorem. Unresolved endpoint alignment counts by teacher are 0, 0, 0, 0; unresolved joint-event counts are 0, 0, 0, 0, in the table's order.

Table 6: Twenty-seed rank-one outcomes. Alignment and refit gains are median [minimum, maximum] at the final saved in-prefix diagnostic, using all twenty values unless a parenthesized resolved count n is shown. $A_{\text{top}} > .95$ is an endpoint count; joint successes may occur at any common saved in-prefix checkpoint under the stated thresholds.

Teacher	$A_{\text{top}} > .95$	ΔA_{top}	ΔR	Joint
h_3	20/20	0.962 [0.745, 0.996]	0.222 [0.146, 0.491]	1/20
$h_3 + .3h_2$	20/20	0.956 [0.742, 0.996]	0.475 [0.320, 0.672]	16/20
ReLU	20/20	0.927 [0.719, 0.997]	0.519 [0.414, 0.683]	20/20
tanh	20/20	0.942 [0.709, 0.983]	0.476 [0.299, 0.676]	19/20

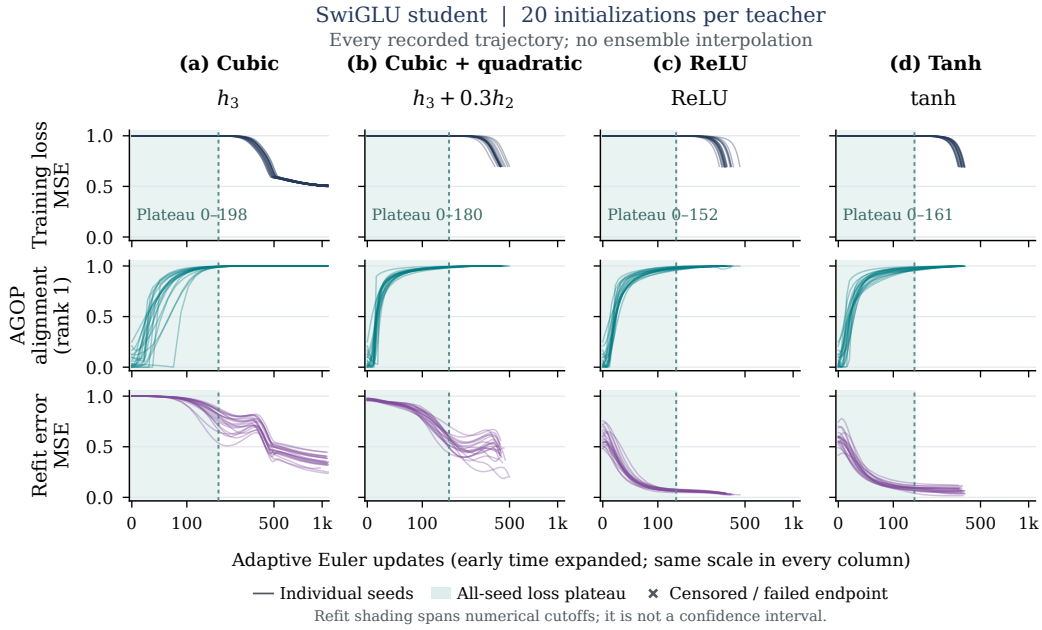


Figure 18: All twenty recorded trajectories per rank-one teacher. Individual curves retain their own saved grids and terminal times. The three rows and numerical meanings are those of Figure 2; all outcomes remain visible.

Table 7: Every seed's endpoint changes in leading AGOP alignment and normalized numerical refit MSE. Positive ΔR means lower error. Values are rounded numerical diagnostics, not error certificates; $--$ indicates an unresolved value.

	h_3		$h_3 + .3h_2$		ReLU		tanh	
Seed	ΔA	ΔR	ΔA	ΔR	ΔA	ΔR	ΔA	ΔR
0	0.870	0.164	0.869	0.426	0.862	0.635	0.852	0.518
1	0.972	0.371	0.969	0.397	0.957	0.647	0.961	0.676
2	0.994	0.358	0.993	0.558	0.997	0.428	0.956	0.460
3	0.981	0.235	0.979	0.483	0.969	0.551	0.955	0.459
4	0.745	0.239	0.742	0.360	0.719	0.486	0.709	0.438
5	0.899	0.146	0.894	0.459	0.882	0.619	0.869	0.539
6	0.786	0.208	0.783	0.506	0.766	0.514	0.770	0.457
7	0.988	0.150	0.979	0.575	0.973	0.447	0.978	0.428
8	0.996	0.349	0.996	0.557	0.984	0.442	0.983	0.510
9	0.963	0.312	0.956	0.466	0.927	0.534	0.926	0.470
10	0.979	0.181	0.977	0.501	0.973	0.588	0.969	0.564
11	0.978	0.288	0.980	0.505	0.975	0.637	0.972	0.623
12	0.995	0.203	0.992	0.499	0.977	0.525	0.972	0.498
13	0.993	0.186	0.992	0.320	0.983	0.572	0.967	0.504
14	0.873	0.491	0.873	0.468	0.871	0.683	0.836	0.633
15	0.962	0.156	0.956	0.449	0.926	0.470	0.951	0.483
16	0.924	0.264	0.922	0.438	0.907	0.464	0.892	0.457
17	0.941	0.367	0.935	0.539	0.913	0.472	0.933	0.420
18	0.826	0.175	0.830	0.672	0.823	0.414	0.793	0.299
19	0.896	0.174	0.895	0.325	0.882	0.466	0.878	0.451