POPULATION ONSET AND FINITE-BATCH RESOLUTION IN ONE-STEP FEATURE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A population gradient can point toward a task-relevant direction even when a finite empirical gradient cannot resolve it. We study this gap for one-step feature learning in two-layer networks with spiked Gaussian inputs. We identify two target-weighted covariance quantities with distinct roles. The population amplification A sets the learning-rate scale at which task-directed overlap becomes order one, whereas the ratio H/A^2 gives a sufficient independent-batch scale for recovering that movement empirically. These scales can vary independently, so anisotropy can amplify feature movement without making it easier to estimate from finite data. Under a narrower smooth activation class and explicit triangular conditions, we derive at the critical batch scale a joint random law separating coherent task alignment from total representation movement. For rank-one spikes, an exact alignment–anisotropy map characterizes where population amplification, finite-batch difficulty, and the critical law arise; the population and critical results also extend to fixed-rank spikes. Direct-onset, factorial, and geometry-wide experiments support the predicted normalizations and expose finite-dimensional deviations. The results concern one-step representation dynamics, not downstream risk, and the recovery scale is sufficient rather than necessary.

1 INTRODUCTION

A population gradient may point toward a task-relevant direction, yet a finite empirical gradient may fail to resolve it. A single gradient step is the simplest setting in which these questions can be separated: does the population update create task-directed movement, and can a fresh finite batch estimate that movement? Under isotropic data, signal and fluctuation scales are often tied to the same ambient geometry. Anisotropic covariance can amplify them differently.

Recent one-step theory characterizes the nearly rank-one structure and learning-rate regimes of the first update (Ba et al., 2022), batch thresholds indexed by information or leap exponents (Dandi et al., 2024; Arnaboldi et al., 2024), Hermite features created along a learning-rate staircase (Moniri et al., 2024), and the spectrum and downstream risk of learned spiked-feature models (Cui et al., 2024; Dandi et al., 2025). These results primarily develop the isotropic-input theory. Structured-data analyses show that covariance can alter gradient-based recovery (Mousavi-Hosseini et al., 2023) and post-step predictors (Demir & Doğan, 2025). What remains unresolved for the standard one-step model studied here is whether anisotropic geometry controls population movement and finite-batch resolvability through the same scale. We show that it does not.

Three directions make this interaction transparent: $\boldsymbol{\xi}$ is where the task lives; in $\boldsymbol{\Sigma} = \mathbf{I}_n + \theta\boldsymbol{\gamma}\boldsymbol{\gamma}^\top$, $\boldsymbol{\gamma}$ is where covariance is exceptional; and $\mathbf{v} = \boldsymbol{\Sigma}\boldsymbol{\xi}/\sqrt{\boldsymbol{\xi}^\top\boldsymbol{\Sigma}\boldsymbol{\xi}}$ is the covariance-weighted task direction seen by the gradient. Their alignment is $\alpha = \boldsymbol{\gamma}^\top\boldsymbol{\xi}$. This minimal spike adds one privileged data direction to the task direction, and the fixed-rank result shows the mechanism is not tied to one spike; we do not claim that low-rank spikes model all data.

For a two-layer network of hidden-layer width k, the answer separates two target-weighted covariance moments,

$$A = \frac{\boldsymbol{\xi}^\top\boldsymbol{\Sigma}^2\boldsymbol{\xi}}{\boldsymbol{\xi}^\top\boldsymbol{\Sigma}\boldsymbol{\xi}}, \quad H = \frac{\boldsymbol{\xi}^\top\boldsymbol{\Sigma}^3\boldsymbol{\xi}}{\boldsymbol{\xi}^\top\boldsymbol{\Sigma}\boldsymbol{\xi}}, \qquad \eta_{\text{onset}} \asymp \frac{\sqrt{k}}{A}, \quad m_{\text{res}} \asymp \frac{H}{A^2}.$$

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.

The factor A is population amplification and sets the learning-rate scale. The ratio H/A^2 is the signal-to-noise resolution coordinate: it is a sufficient independent-batch scale for recovering the population overlap, not a universal necessary sample complexity. A factorial experiment makes the distinction concrete. When A is matched but H/A^2 differs by a factor of about three, the population onset remains matched while the resolution batch changes. When H/A^2 is matched but A differs by a factor of about sixteen, the required raw learning rate changes by the same factor while normalized finite-batch behavior remains similar.

At $m_{\mathrm{FL}} \asymp H/A^2$, empirical fluctuations remain order one. Under a narrower smooth activation class and explicit triangular conditions, Theorem 4.2 gives a joint random limit for coherent alignment Q_a and total overlap energy Q_2. The same fluctuation drives both, but Q_2 also contains incoherent energy: coherent alignment is not the same as total movement. In the noise-dominated regime, order-one movement can persist while its coherent sign becomes unreliable.

Finally, the rank-one parameterization $\theta = c_\theta n^\beta$ and $\alpha = c_\alpha n^{-\delta}$ yields exact *alignment–anisotropy scaling regimes*. It identifies where anisotropy amplifies population movement, where finite-batch resolution becomes harder, and where the primitive and critical theorem conditions hold. In particular, $\beta = 1/2$ is a spectral crossover rather than a global theorem boundary. For fixed nonlinearities with nonzero first Gaussian channels, activation changes order-one coefficients but not the geometry exponents; the critical law uses its stated narrower class. All conclusions concern representations, not downstream prediction.

Our contributions are:

Covariance-controlled population onset. From the exact projected update, we show that for a width-k network the task-weighted amplification A sets the population overlap onset $\eta \asymp \sqrt{k}/A$, under explicit concentration and self-prediction conditions.

Finite-batch feature-gradient resolution. For an independent feature-learning batch, we show that H/A^2 is a sufficient scale for recovering the population overlap, separating gradient estimation from population movement.

Critical coherent-alignment law. At the finite-batch resolution scale, we derive under the narrower critical class a joint random limit that distinguishes coherent task alignment Q_a from total overlap energy Q_2.

Alignment–anisotropy regimes and validation. For polynomial spike/alignment scaling, we derive the exact theorem-admissibility regimes, extend the covariance-level results to fixed-rank spikes, and validate both scales through direct-onset, factorial, and geometry-wide experiments.

1.1 Related work

Frozen features and structured inputs. Random-feature and kernel limits characterize approximation, spectra, and high-dimensional risk (Rahimi & Recht, 2007; Louart et al., 2018; Mei & Montanari, 2022; Gerace et al., 2020; Hu & Lu, 2023; Montanari & Saeed, 2022; Ghorbani et al., 2021). Data or architectural symmetry can improve frozen models (Mei et al., 2021); under spiked covariance, task–spike alignment controls which nonlinear components are already present before adaptation (Demir & Doğan, 2024). The present work instead studies the label-dependent change produced by one update.

One-step feature learning. Gradient-based representation learning can separate neural networks from fixed kernels (Damian et al., 2022). One-step and early-time work has characterized aggressive learning-rate regimes, information-exponent batch thresholds, learned spectral spikes, and early alignment (Ba et al., 2022; Dandi et al., 2024; Cui et al., 2024; Moniri et al., 2024; Dandi et al., 2025; Boursier & Flammarion, 2025). Our object is different: the two target-weighted covariance moments of an ordinary unpreconditioned one-step update under anisotropic Gaussian design. Dandi et al. (2024) derive isotropic one-step batch thresholds controlled by the target's Hermite/leap complexity; we instead ask how anisotropic task–covariance geometry changes that signal-to-noise scale in the nonzero first-channel regime.

Structured gradients and stochastic limits. Covariance-normalized dynamics exploit task–spike alignment under spiked data (Mousavi-Hosseini et al., 2023); layer-wise preconditioning can remedy

SGD's statistical suboptimality under ill-conditioned inputs (Zhang et al., 2025). Vanilla SGD for anisotropic single-index models instead admits effective-dimension sample bounds (Braun et al., 2025). We isolate the first unpreconditioned two-layer update itself and its separate population and finite-batch scales A and H/A^2; structured Gaussian mixtures yield polynomial-equivalent post-step predictors (Demir & Doğan, 2025). Generic minibatch noise and wide-network fluctuation limits (Liu et al., 2022; Descours et al., 2024), critical high-dimensional SGD (Ben Arous et al., 2024), and signal-enhancing online procedures (Damian et al., 2023) provide complementary stochastic context; they do not identify the A versus H/A^2 separation studied here.

2 Problem Setup and Target-Weighted Geometry

We study the first-layer representation created by one gradient step in a two-layer network under structured Gaussian inputs. The normalization below is fixed throughout; in particular, the auxiliary network used to generate the feature gradient is distinct from any readout fitted in later numerical diagnostics. All objects may form triangular sequences indexed by n; in particular, $k = k_n$, $m_{\mathrm{FL}} = m_{\mathrm{FL},n}$, and the covariance parameters may depend on n. We suppress this dependence unless it matters.

2.1 Data and teacher

Let $\mathbf{x} \in \mathbb{R}^n$ satisfy

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \qquad \mathbf{\Sigma} \succ 0, \qquad \|\boldsymbol{\xi}\|_2 = 1, \tag{1}$$

and define the variance-one target coordinate and teacher

$$B := \boldsymbol{\xi}^\top \mathbf{\Sigma} \boldsymbol{\xi}, \qquad S := \frac{\boldsymbol{\xi}^\top \mathbf{x}}{\sqrt{B}}, \qquad y = \sigma^*(S). \tag{2}$$

Here $\boldsymbol{\xi}$ is the teacher or task direction. The teacher nonlinearity σ^* will satisfy the regularity and moment assumptions stated before Theorem 3.2. Our theorems cover the rank-one model $\mathbf{\Sigma} = \mathbf{I}_n + \theta \boldsymbol{\gamma}\boldsymbol{\gamma}^\top$ and its fixed-rank extension in Corollary 3.4. Although several identities below are valid for general covariance matrices, we do not claim a general deterministic-covariance theorem. Gaussianity makes the population contractions and finite-batch fluctuations analytically separable through Stein and Hermite identities; whether the same two-scale picture persists beyond Gaussian designs remains open.

2.2 Auxiliary network and one-step update

Put $s := \operatorname{tr} \mathbf{\Sigma}$. The initialized feature rows and auxiliary output coefficients are independent and satisfy

$$\mathbf{f}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_n/s), \qquad a_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad i \in [k]. \tag{3}$$

Let $\mathbf{a} = (a_1, \dots, a_k)^\top$ and $\mathbf{F} = (\mathbf{f}_1, \dots, \mathbf{f}_k)^\top$. Define

$$g_{\mathbf{F}}(\mathbf{x}) = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} a_i \sigma(\mathbf{f}_i^\top \mathbf{x}), \qquad \mathcal{L}_{\mathrm{pop}}(\mathbf{F}) = \frac{1}{2}\mathbb{E}\big[(y - g_{\mathbf{F}}(\mathbf{x}))^2\big]. \tag{4}$$

The population update is $\mathbf{F}_1^{\mathrm{pop}} = \mathbf{F}_0 - \eta \nabla_{\mathbf{F}} \mathcal{L}_{\mathrm{pop}}(\mathbf{F}_0)$. The coefficients a_i are used only to define this feature-learning gradient and the coherent observable below. They are not the independently fitted ridge readout used in the experiments. Our theoretical object is the first-layer representation after the update.

2.3 Independent empirical feature gradient

Let $\mathcal{D}_{\mathrm{FL}} = \{(\mathbf{x}_\ell, y_\ell)\}_{\ell=1}^{m_{\mathrm{FL}}}$ be independent of $(\mathbf{F}_0, \mathbf{a})$, with iid samples from (1)–(2). Define

$$\widehat{\mathcal{L}}_{\mathrm{FL}}(\mathbf{F}) = \frac{1}{2m_{\mathrm{FL}}} \sum_{\ell=1}^{m_{\mathrm{FL}}} \big(y_\ell - g_{\mathbf{F}}(\mathbf{x}_\ell)\big)^2, \qquad \mathbf{F}_1^{\mathrm{emp}} = \mathbf{F}_0 - \eta \nabla_{\mathbf{F}} \widehat{\mathcal{L}}_{\mathrm{FL}}(\mathbf{F}_0). \tag{5}$$

Independence of this feature-learning sample is part of every empirical-gradient theorem below. It isolates finite-sample error in estimating the feature gradient from effects caused by sample reuse in downstream fitting; we do not claim that independence is necessary in every formulation.

2.4 TARGET-WEIGHTED COVARIANCE GEOMETRY

The covariance between $\mathbf{x}$ and S is $\mathbf{v} := \mathbb{E}[\mathbf{x}S] = \boldsymbol{\Sigma}\boldsymbol{\xi}/\sqrt{B}$. It is the covariance-weighted target direction selected by the population gradient; it need not be parallel to the task direction $\boldsymbol{\xi}$. Three target-weighted moments will be used:

$$A := \|\mathbf{v}\|_2^2 = \frac{\boldsymbol{\xi}^\top \boldsymbol{\Sigma}^2 \boldsymbol{\xi}}{B}, \qquad H := \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v} = \frac{\boldsymbol{\xi}^\top \boldsymbol{\Sigma}^3 \boldsymbol{\xi}}{B}, \qquad C := \mathbf{v}^\top \boldsymbol{\Sigma}^2 \mathbf{v} = \frac{\boldsymbol{\xi}^\top \boldsymbol{\Sigma}^4 \boldsymbol{\xi}}{B}. \tag{6}$$

The quantity A measures amplification of the target-gradient direction and sets the population learning-rate scale. The moment H sets the projected empirical-gradient variance, so H/A^2 becomes the feature-gradient sample scale. By contrast, C, $\operatorname{tr} \boldsymbol{\Sigma}^2$, and $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$ control concentration and self-prediction admissibility; they are not additional headline scales. We also define $d_{\mathrm{eff}} := s^2 / \operatorname{tr} \boldsymbol{\Sigma}^2$.

2.5 REPRESENTATION OBSERVABLES

For each row let

$$q_i := \mathbf{f}_i^\top \mathbf{v}, \qquad Q_2(\mathbf{F}) := \frac{1}{k} \sum_{i=1}^{k} q_i^2, \qquad Q_a(\mathbf{F}) := \frac{1}{k} \sum_{i=1}^{k} a_i q_i. \tag{7}$$

Thus Q_2 is total overlap energy in the target-gradient direction, while Q_a records the component coherent with the auxiliary coefficients. Neither observable is a test risk, and no implication from overlap to downstream generalization is assumed.

3 POPULATION FEATURE-LEARNING ONSET

We first identify the population signal scale, then study its estimation from an independent feature-learning sample.

3.1 EXACT PROJECTED UPDATE

Write $U_i = \mathbf{f}_i^\top \mathbf{x}$, $q_{0i} = \mathbf{f}_{0i}^\top \mathbf{v}$, and $r_i = \mathbf{f}_{0i}^\top \boldsymbol{\Sigma} \mathbf{v}$. Define

$$L_i := \mathbf{v}^\top \mathbb{E}[\sigma^*(S)\sigma'(U_i)\mathbf{x}], \qquad K_{ij} := \mathbf{v}^\top \mathbb{E}[\sigma(U_j)\sigma'(U_i)\mathbf{x}]. \tag{8}$$

Gaussian integration by parts gives the exact identities

$$\begin{aligned} L_i &= A\mathbb{E}[\sigma^{*\prime}(S)\sigma'(U_i)] + r_i \mathbb{E}[\sigma^*(S)\sigma''(U_i)], \\ K_{ij} &= r_j \mathbb{E}[\sigma'(U_j)\sigma'(U_i)] + r_i \mathbb{E}[\sigma(U_j)\sigma''(U_i)]. \end{aligned} \tag{9}$$

Proposition 3.1 (Exact projected update)**.** *For the full population loss, the projected descent direction is*

$$d_i = \frac{a_i}{\sqrt{k}} \left(L_i - \frac{1}{\sqrt{k}} \sum_{j=1}^{k} K_{ij} a_j \right), \qquad q_{1i} = q_{0i} + \eta d_i, \tag{10}$$

and, with $\Delta Q_2^{\mathrm{pop}} := Q_2(\mathbf{F}_1^{\mathrm{pop}}) - Q_2(\mathbf{F}_0)$*,*

$$\Delta Q_2^{\mathrm{pop}} = \frac{2\eta}{k} \sum_i q_{0i} d_i + \frac{\eta^2}{k} \sum_i d_i^2. \tag{11}$$

The self-prediction term is retained: conditionally on $\mathbf{F}_0$, $\mathbb{E}_a[d_i \mid \mathbf{F}_0] = -K_{ii}/k$. Its diagonal contraction is nonzero, so the self term is not exactly centered.

3.2 GENERIC FIRST-CHANNEL THEOREM

Teacher regularity (T). σ^* is locally absolutely continuous with weak derivative $(\sigma^*)'$; for some $\epsilon > 0$, $\mathbb{E}|\sigma^*(G)|^{2+\epsilon} < \infty$ and $\mathbb{E}|\sigma^{*\prime}(G)|^{1+\epsilon} < \infty$, with the corresponding Gaussian families uniformly integrable for variances near one.

Network regularity (P). $\sigma \in C^2$, and $\sigma, \sigma', \sigma''$ have polynomial growth. For $G \sim \mathcal{N}(0,1)$, set $m_* := \mathbb{E}[\sigma^{*\prime}(G)]$ and $m_\sigma := \mathbb{E}[\sigma'(G)]$.

Specialize first to

$$\mathbf{\Sigma} = \mathbf{I}_n + \theta \boldsymbol{\gamma}\boldsymbol{\gamma}^\top, \qquad \|\boldsymbol{\gamma}\|_2 = 1, \qquad \alpha := \boldsymbol{\gamma}^\top \boldsymbol{\xi}, \qquad k/n \to \phi \in (0, \infty). \tag{12}$$

Here $\boldsymbol{\gamma}$ is the covariance-spike direction, $\boldsymbol{\xi}$ is the task direction, and α is their alignment. Assume visibly the four primitive conditions

$$\frac{\operatorname{tr}(\mathbf{\Sigma}^2)\log k}{s^2} \to 0, \quad \frac{\|\mathbf{\Sigma}\|_{\mathrm{op}} \log k}{s} \to 0, \quad \frac{A \log k}{s} \to 0, \quad \frac{C \log k}{sA^2} \to 0. \tag{13}$$

The first two ratios give uniform concentration of initialized preactivation variances. The ratio $A \log k/s$ controls the maximal initial target overlap, whereas $C \log k/(sA^2)$ makes self-prediction negligible relative to the leading population signal.

Theorem 3.2 (Covariance-controlled population overlap onset)**.** *Assume: (i) the rank-one model (12) with $k/n \to \phi \in (0, \infty)$; (ii) T and P; (iii) the four primitive conditions (13); and (iv) the nondegenerate first-channel condition $m_* m_\sigma \neq 0$. Then the full population step satisfies*

$$\Delta Q_2^{\mathrm{pop}} = \frac{\eta^2 A^2}{k}\{m_*^2 m_\sigma^2 + o_p(1)\}. \tag{14}$$

Thus, for fixed $c > 0$ and $\eta = c\sqrt{k}/A$,

$$\Delta Q_2^{\mathrm{pop}} = c^2 m_*^2 m_\sigma^2 + o_p(1), \qquad \eta_{\mathrm{onset}} \asymp \frac{\sqrt{k}}{A}. \tag{15}$$

The theorem identifies a representation-level onset, not a risk transition. The constant c is any fixed positive number, so the adaptive step makes the leading overlap energy order one. If either first Gaussian channel vanishes, the displayed leading limit is zero and no higher-channel onset theorem is claimed. For matched tanh, $m_* = m_\sigma = \mu := \mathbb{E}\operatorname{sech}^2(G)$, recovering the coefficient μ^4. We deliberately focus on the nonzero first-channel regime to isolate the effect of task–covariance geometry from the separate information-exponent/Hermite-order difficulty studied in isotropic one-step theory (Dandi et al., 2024; Moniri et al., 2024). Appendices A–C give the proof, including the complete self contribution.

For a positive sequence X_n, write $\exp_n(X) = e$ when $X_n = n^{e+o(1)}$.

Corollary 3.3 (Alignment-dependent onset)**.** *Let $\theta = c_\theta n^\beta$ and $\alpha = c_\alpha n^{-\delta}$. In every regime satisfying (13),*

$$\exp_n A = \begin{cases} \beta, & 0 \le \delta \le \beta/2, \\ 2\beta - 2\delta, & \beta/2 < \delta < \beta, \\ 0, & \delta \ge \beta, \end{cases} \qquad \eta_{\mathrm{onset}} \asymp \begin{cases} n^{1/2-\beta}, & 0 \le \delta \le \beta/2, \\ n^{1/2-2\beta+2\delta}, & \beta/2 < \delta < \beta, \\ n^{1/2}, & \delta \ge \beta. \end{cases} \tag{16}$$

The scale H/A^2 is order one under strong alignment, grows for $\beta/2 < \delta < 3\beta/2$, and returns to order one for very weak alignment; the exact five-branch exponent and admissibility regions are in Appendix F. The value $\beta = 1/2$ is the spectral crossover $\theta^2 \asymp n$ in $\operatorname{tr}\mathbf{\Sigma}^2$, not the theorem's global upper boundary.

3.3 Fixed-rank extension

For fixed spike rank r_{sp}, let

$$\mathbf{\Sigma} = \mathbf{I}_n + \sum_{b=1}^{r_{\mathrm{sp}}} \theta_b \boldsymbol{\gamma}_b \boldsymbol{\gamma}_b^\top, \qquad \boldsymbol{\gamma}_b^\top \boldsymbol{\gamma}_{b'} = \mathbf{1}\{b = b'\}. \tag{17}$$

Corollary 3.4 (Fixed-rank population overlap onset)**.** *Under T/P, (17), $k/n \to \phi \in (0, \infty)$, the four conditions (13), and $m_* m_\sigma \neq 0$, the conclusion of Theorem 3.2 holds with coefficient $m_*^2 m_\sigma^2$ and $\eta_{\mathrm{onset}} \asymp \sqrt{k}/A$.*

The proof is covariance-level, but the result is only for fixed-rank spiked covariance, not arbitrary deterministic covariance; see Appendix E.

4 FINITE-SAMPLE FEATURE GRADIENTS

4.1 POPULATION RECOVERY

For an independent feature-learning sample, write $\widehat{d}_i = d_i + \varepsilon_i$. The exact shared-sample covariance and overlap expansion are in Appendix G. Define $\chi_{\mathrm{FL}} := H/(m_{\mathrm{FL}}A^2)$.

Theorem 4.1 (Empirical recovery of the population overlap)**.** *Assume: (i) T and P together with either the rank-one or fixed-rank population assumptions; (ii) $m_* m_\sigma \neq 0$; (iii) the feature-learning sample is independent of $(\mathbf{F}_0, \mathbf{a})$; and (iv)*

$$\chi_{\mathrm{FL}} = \frac{H}{m_{\mathrm{FL}}A^2} \to 0, \tag{18}$$

then at $\eta = c\sqrt{k}/A$,

$$\frac{\Delta Q_2^{\mathrm{emp}}}{\Delta Q_2^{\mathrm{pop}}} \to_p 1, \qquad \Delta Q_2^{\mathrm{emp}} = c^2 m_*^2 m_\sigma^2 + o_p(1). \tag{19}$$

Thus H/A^2 is a sufficient geometric recovery scale for fixed T/P activation pairs. The theorem does not assert universal necessity; the nondegenerate order-one interpretation below requires the narrower critical class.

4.2 CRITICAL TRIANGULAR REGIME

Critical teacher regularity (C–T). Strengthen T by requiring that, for some $\epsilon > 0$, $\mathbb{E}|\sigma^*(G)|^{8+\epsilon} < \infty$ and $\mathbb{E}|\sigma^{*\prime}(G)|^{2+\epsilon} < \infty$, uniformly near unit Gaussian variance.

Critical network regularity (C–P). Assume $\mathbb{E}\sigma(G) = 0$; σ is bounded C^3 with bounded derivatives through order three; put $m_1 := \mathbb{E}\sigma'(G) = m_\sigma \neq 0$ and $m_2 := \mathbb{E}\sigma'(G)^2$; and $\sigma'(G) - m_1$ has probabilists' Hermite rank exactly two. Put $\nu := \nu_{\sigma^*,\sigma} = \mathbb{E}[\sigma^*(G)^2] + \mathbb{E}[\sigma(G)^2]$. This rank-two condition makes the leading off-diagonal derivative covariance quadratic in the normalized Gram correlation, precisely the structure controlled in the growing-batch row law. Tanh, standard erf, and centered sigmoid satisfy C–P. ReLU is allowed as a teacher under C–T but not as the network activation in this theorem.

Theorem 4.2 (Critical coherent-alignment and overlap laws)**.** *Assume: (i) the rank-one population conditions; (ii) C–T and C–P; (iii) a feature-learning sample independent of $(\mathbf{F}_0, \mathbf{a})$; and (iv), with ρ denoting the signal-to-noise geometry ratio,*

$$\rho := \frac{A}{\sqrt{H}} \to 0, \quad m_{\mathrm{FL}} \to \infty, \quad m_{\mathrm{FL}}\rho^2 \to \chi^{-1} \in (0, \infty), \quad \frac{m_{\mathrm{FL}}}{k} \to 0, \quad \frac{m_{\mathrm{FL}}}{d_{\mathrm{eff}}} \to 0. \tag{20}$$

At $\eta = c\sqrt{k}/A$, the same $Z \sim \mathcal{N}(0,1)$ drives

$$Q_a(\mathbf{F}_1^{\mathrm{emp}}) - Q_a(\mathbf{F}_0) \Rightarrow c\, m_1 \left(m_* + \sqrt{\nu\chi}\, Z\right), \tag{21}$$

$$\Delta Q_2^{\mathrm{emp}} \Rightarrow c^2 \left[m_1^2 \left(m_* + \sqrt{\nu\chi}\, Z\right)^2 + \nu\chi(m_2 - m_1^2)\right]. \tag{22}$$

The same scalar Gaussian Z couples coherent alignment and total overlap; the second term in (22) is incoherent diagonal energy. This is a triangular-array limit for the one-step representation, not a generalization law. The restrictions $m_{\mathrm{FL}}/k \to 0$ and $m_{\mathrm{FL}}/d_{\mathrm{eff}} \to 0$ remain part of the theorem. The first is a growing-batch condition for row averaging and leave-one-row errors; the second controls sample-norm fluctuations in the variance-dependent derivative kernel. These are sufficient proof conditions, not asserted to be necessary. For matched tanh, $m_* = m_1 = \mu$, $m_2 = \mathbb{E}\,\mathrm{sech}^4(G)$, and $\nu = 2\mathbb{E}\tanh^2(G)$, exactly recovering the previous coefficients. The critical class is deliberately narrower than the population class.

Corollary 4.3 (Fixed-rank critical law)**.** *Under the fixed-rank population assumptions, C–T/C–P, and every condition in (20), the limits (21)–(22) hold with the fixed-rank moments A and H.*

This does not extend the theorem beyond fixed-rank spiked covariance.

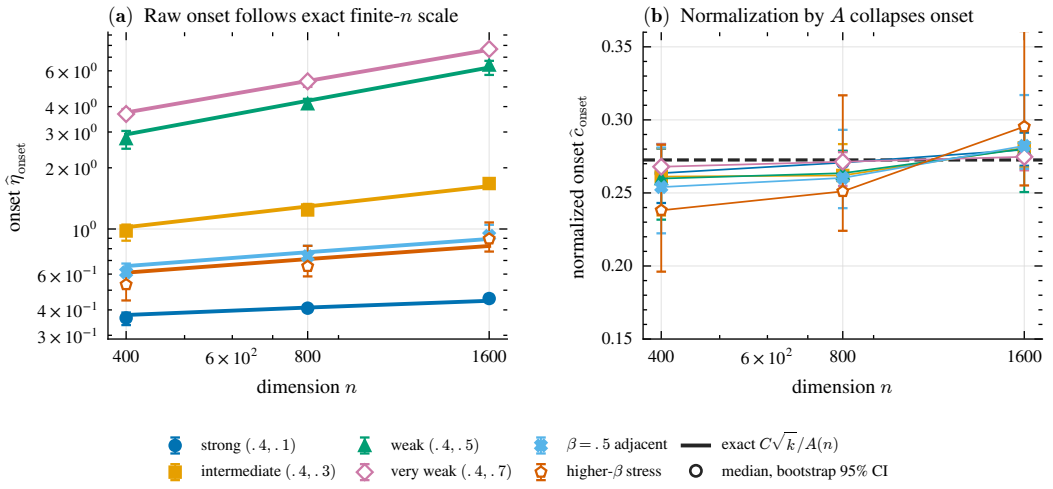


Figure 1: **Direct onset scaling.** (a) Median onset $\widehat{\eta}_{\text{onset}}$, defined by $Q_{\text{update}} = 0.01$, for six rank-one geometry sequences; bars are bootstrap 95% intervals from 60 trials per sequence and dimension. Curves show the exact finite-n prediction $C\sqrt{k}/A(n)$ with the single threshold constant $C = \sqrt{0.01/\mu^4} = 0.273$. (b) Normalization by $A/\sqrt{k}$; the dashed line is the matched-tanh prediction. Directions use $m_{\text{FL}} = 1000$ and are population-like proxies (realized $\chi_{\text{FL}} = 0.0010$–$0.0043$), not exact population gradients. Parenthesized pairs denote the spike and alignment exponents (β, δ) in $\theta = c_\theta n^{\beta}$ and $\alpha = c_\alpha n^{-\delta}$; the $\beta = .5$-adjacent and higher-β stress labels correspond to $(.5, .35)$ and $(.6, .45)$, respectively. The latter is a finite-size stress case.

4.3 Noise-dominated feature gradients

When $\chi_{\text{FL}} \to \infty$, the population normalization $\sqrt{k}/A$ is noise dominated. Under the growing-sample derivative-kernel controls, the alternative scale

$$\eta_{\text{noise}} = d\sqrt{\frac{km_{\text{FL}}}{H}} \tag{23}$$

preserves order-one representation energy while coherent alignment has random leading sign. The full proved coefficient remains in the matched- tanh specialization and is stated in Appendix J; no generic noise-dominated coefficient is claimed.

5 Numerical Validation

We test the population scale $\sqrt{k}/A$, the independent-batch coordinate $\chi_{\text{FL}} = H/(m_{\text{FL}}A^2)$, and the coherent critical law. The main experiments use matched tanh and independent trials within each cell; paired comparisons reuse initialization and sample randomness. No experiment is interpreted as a downstream generalization result.

5.1 Direct population-onset scaling

We define the empirical onset as the smallest bracketed learning rate at which the pure update energy $Q_{\text{update}} := k^{-1}\sum_i(\eta d_i)^2$ reaches 0.01. Figure 1 compares this onset across six rank-one geometry sequences and three dimensions. The raw onsets track the exact finite-n prediction $C\sqrt{k}/A(n)$ using one threshold-determined constant, and multiplication by $A/\sqrt{k}$ collapses the medians around that prediction. We emphasize the finite-n curves rather than fitted asymptotic exponents; the higher-β arm is deliberately retained as a finite-size stress case. The update directions use a large independent batch and are population-like empirical proxies, not exact population gradients.

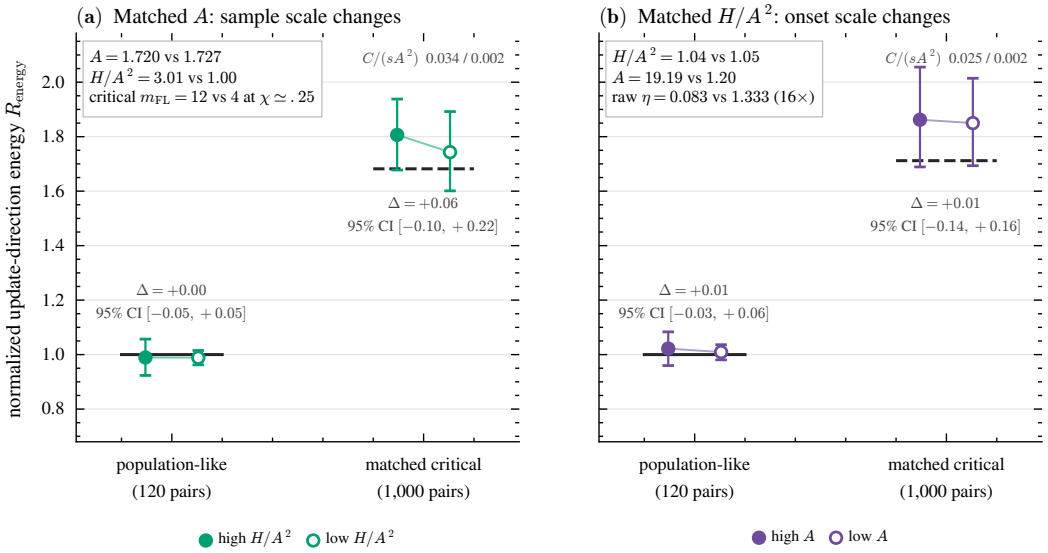


Figure 2: **Factorial separation of onset and resolution scales.** Left: nearly matched A but a factor-3.00 difference in H/A^2; matching $\chi_{\mathrm{FL}} \simeq 0.25$ requires $m_{\mathrm{FL}} = 12$ and 4. Right: matched H/A^2 but a factor-15.99 difference in A, hence in raw learning rate at fixed c. Points show R_{energy}; bars are paired-bootstrap 95% intervals from 120 population-like pairs and 1,000 critical pairs. Black segments mark the matched-tanh population value and critical-law mean. The comparison separates normalizations; it is not an equivalence test.

5.2 Separating population onset from finite-batch resolution

For matched tanh, define $R_{\mathrm{energy}} := Q_{\mathrm{update}}/(c^2\mu^4)$.

Figure 2 gives a factorial separation of the two scales. In the matched-A construction, the population learning-rate scale is nearly identical while H/A^2 differs by a factor of 3.00; reaching the same critical coordinate $\chi_{\mathrm{FL}} \simeq 0.25$ therefore requires batches of 12 and 4. In the matched-H/A^2 construction, the sample scales agree within 1.25%, while A—and hence the raw learning rate at fixed $c = \eta A/\sqrt{k}$—differs by a factor of 15.99. After the appropriate normalizations, neither population-like nor critical paired comparisons show a resolved large energy separation. This is evidence for distinct coordinates, not a proof that the learned representations are equivalent.

5.3 Critical-law coverage across geometry

Theorem 4.2 predicts a Gaussian coherent increment after normalization by finite-n geometry. Figure 3 tests six proved-region cells at $n = 800$ and 1600, with 120 independent trials in each cell. Across the 12 comparisons, one-sample KS distances are 0.053–0.126, and every bootstrap interval for the signed mean error contains zero. Variance-ratio intervals contain one in nine cells; three higher-β dimension-cells retain resolved variance inflation. The plot keeps these deviations visible. Its colored regions are theorem-admissibility regions, not empirically fitted boundaries.

5.4 Robustness and limitations

Appendix L reports representative CDF/QQ diagnostics, exact finite-dimensional geometry maps, activation-normalized tests, width ratios $k/n \in \{0.25, 0.5, 1, 2, 4\}$, and sensitivity to the primitive self ratio $C/(sA^2)$. The smooth activation pairs follow their activation-specific coefficients at the tested scales; tanh/ReLU is evaluated only for the population result. Width changes preserve the predicted order but not exact finite-width invariance, and discrepancies grow near order-one primitive ratios. Stress cells outside the theorem region are used only as limitations.

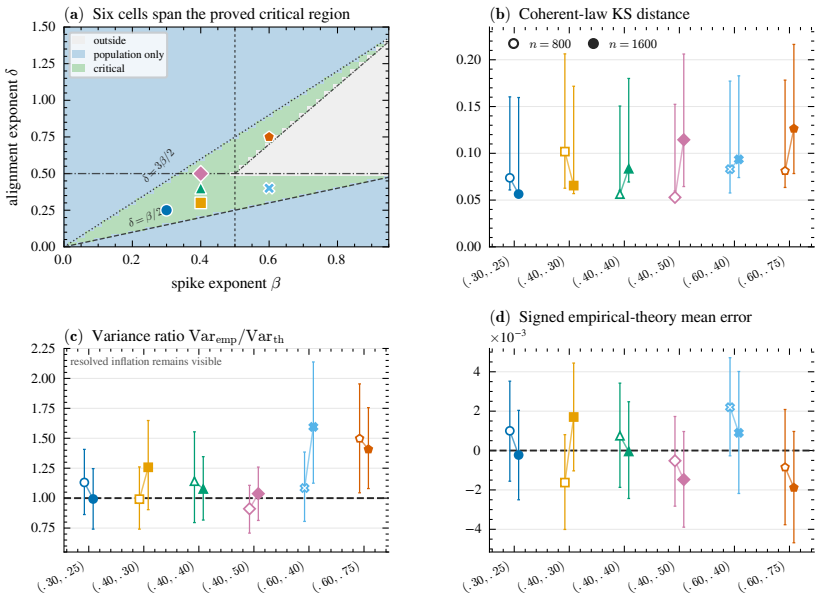


Figure 3: **Critical coherent-law coverage across geometry.** (a) Six tested (β, δ) cells over the exact proved rank-one regions (gray: outside; blue: population only; green: critical); boundaries are theoretical, not fitted. (b)–(d) KS distance, variance ratio, and mean error at $n = 800, 1600$, with 120 trials per cell and bootstrap 95% intervals. Three higher-β cells retain visible finite-size variance inflation.

6 Discussion and Limitations

The analysis separates population movement from finite-batch resolution and coherent alignment from total energy. The amplification A controls when the population representation moves, whereas H/A^2 is a sufficient batch scale for resolving that movement. At critical sampling, Q_a remains random while Q_2 also contains incoherent diagonal energy, so total movement is not a reliable proxy for directional learning. The factorial experiment varies the two geometric coordinates separately, and the critical-plane test keeps finite-size variance deviations visible.

These observables describe representation dynamics rather than test risk; fresh-readout diagnostics show no reproducible gain, and connecting alignment to prediction requires additional analysis. Likewise, the (β, δ) exponents organize asymptotic regimes, but finite-dimensional accuracy also depends on the primitive ratios in (13). Thus favorable exponents need not imply uniformly small error, and the region maps denote theorem conditions rather than empirical boundaries.

The theory assumes Gaussian inputs, fixed-rank spikes, and a nonzero first channel. The critical law further requires the narrower C–P activation class, an independent batch, and the stated effective-dimension and other triangular restrictions; it does not cover ReLU networks. We prove neither a downstream generalization result nor a multi-step theorem.

7 Conclusion

We show that anisotropic data geometry separates two difficulties that coincide in simpler isotropic settings: creating a task-directed population update and resolving that update from a finite batch. Different covariance statistics control these difficulties, so stronger population movement need not imply easier statistical estimation. At the resolution boundary, finite-batch fluctuations govern coherent alignment while also creating incoherent representation energy, making total movement an unreliable proxy for directional learning. Understanding how this separation evolves over multiple updates, and when coherent representation learning improves prediction, are natural next steps.

AI USE STATEMENT

Generative AI tools were used to assist with text polishing, equation and notation checks, double-checking mathematical derivations, and limited coding support. The authors reviewed and verified all resulting text, mathematics, and code and take full responsibility for the content of the submission.

REFERENCES

Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. Online learning and information exponents: The importance of batch size & Time/Complexity tradeoffs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1730–1762, 2024.

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 37932–37946, 2022.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. *Communications on Pure and Applied Mathematics*, 77(3):2030–2080, 2024. doi: 10.1002/cpa.22169.

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://www.jmlr.org/papers/v26/24-1523.html.

Guillaume Braun, Minh Ha Quang, and Masaaki Imaizumi. Learning a single index model from anisotropic data with vanilla stochastic gradient descent. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1216–1224, 2025. URL https://proceedings.mlr.press/v258/braun25a.html.

Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue Lu, Lenka Zdeborová, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9662–9695, 2024.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. Smoothing the landscape boosts the signal for SGD: Optimal sample complexity for learning single index models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 752–784, 2023.

Alexandru Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5413–5452, 2022.

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.

Yatin Dandi, Luca Pesce, Hugo Cui, Florent Krzakala, Yue Lu, and Bruno Loureiro. A random matrix theory perspective on the spectrum of learned features and asymptotic generalization capabilities. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2224–2232, 2025.

Samet Demir and Zafer Doğan. Input-Label correlation governs a linear-to-nonlinear transition in random features under spiked covariance. *arXiv preprint arXiv:2409.20250*, 2024.

Samet Demir and Zafer Doğan. Asymptotic analysis of two-layer neural networks after one gradient step under Gaussian mixtures data with structure. In *The Thirteenth International Conference on Learning Representations*, 2025.

Arnaud Descours, Arnaud Guillin, Manon Michel, and Boris Nectoux. Law of large numbers and central limit theorem for wide two-layer neural networks: The mini-batch and noisy case. *Journal of Machine Learning Research*, 25(208):1–76, 2024.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3452–3462, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021. doi: 10.1214/20-AOS1990.

Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023. doi: 10.1109/TIT.2022.3217698.

Ziyin Liu, Kangqiao Liu, Takashi Mori, and Masahito Ueda. Strength of minibatch noise in SGD. In *International Conference on Learning Representations*, 2022.

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018. doi: 10.1214/17-AAP1328.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022. doi: 10.1002/cpa.22008.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3351–3418, 2021.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36106–36159, 2024.

Andrea Montanari and Basil N. Saeed. Universality of empirical risk minimization. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4310–4312, 2022.

Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A. Erdogdu. Gradient-based feature learning under structured data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71449–71485, 2023.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Thomas Tck Zhang, Behrad Moniri, Ansh Nagwekar, Faraz Rahman, Anton Xue, Hamed Hassani, and Nikolai Matni. On the concurrence of layer-wise preconditioning methods and provable feature learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 75793–75833, 2025. URL https://proceedings.mlr.press/v267/zhang25bh.html.

A GAUSSIAN IDENTITIES AND EXACT DECOMPOSITION

We collect the identities used throughout the proofs. Let

$$B = \boldsymbol{\xi}^\top \boldsymbol{\Sigma} \boldsymbol{\xi}, \quad S = \frac{\boldsymbol{\xi}^\top \mathbf{x}}{\sqrt{B}}, \quad \mathbf{v} = \frac{\boldsymbol{\Sigma}\boldsymbol{\xi}}{\sqrt{B}}, \quad V = \mathbf{v}^\top \mathbf{x}, \quad s = \operatorname{tr} \boldsymbol{\Sigma}, \tag{24}$$

and

$$A = \mathbf{v}^\top \mathbf{v} = \frac{\boldsymbol{\xi}^\top \boldsymbol{\Sigma}^2 \boldsymbol{\xi}}{B}, \quad H = \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v} = \frac{\boldsymbol{\xi}^\top \boldsymbol{\Sigma}^3 \boldsymbol{\xi}}{B}, \quad C = \mathbf{v}^\top \boldsymbol{\Sigma}^2 \mathbf{v} = \frac{\boldsymbol{\xi}^\top \boldsymbol{\Sigma}^4 \boldsymbol{\xi}}{B}. \tag{25}$$

For an initialized row, put $U_i = \mathbf{f}_i^\top \mathbf{x}$, $\tau_i^2 = \mathbf{f}_i^\top \boldsymbol{\Sigma} \mathbf{f}_i$, $q_i = \mathbf{f}_i^\top \mathbf{v}$, and $r_i = \mathbf{f}_i^\top \boldsymbol{\Sigma} \mathbf{v}$. Conditional on $\mathbf{f}_i$, (S, V, U_i) is centered Gaussian and

$$\begin{aligned} \operatorname{Var}(S) = 1, \quad & \operatorname{Var}(V) = H, \quad \operatorname{Var}(U_i) = \tau_i^2, \\ \operatorname{Cov}(S, V) = A, \quad & \operatorname{Cov}(S, U_i) = q_i, \quad \operatorname{Cov}(V, U_i) = r_i. \end{aligned} \tag{26}$$

We repeatedly use multivariate Gaussian integration by parts in Sobolev form: if $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and h has an integrable weak derivative, then

$$\mathbb{E}[\mathbf{x} h(\mathbf{x})] = \boldsymbol{\Sigma} \mathbb{E}[\nabla h(\mathbf{x})]. \tag{27}$$

Taking $h(\mathbf{x}) = \sigma^*(S)\sigma'(U_i)$ gives the exact label projection

$$L_i := \mathbf{v}^\top \mathbb{E}[\sigma^*(S)\sigma'(U_i)\mathbf{x}] = A\, \mathbb{E}[\sigma^{*\prime}(S)\sigma'(U_i)] + r_i\, \mathbb{E}[\sigma^*(S)\sigma''(U_i)]. \tag{28}$$

The expectations on the right depend only on the bivariate Gaussian law of (S, U_i). Thus the first channel is multiplied exactly by A; the second channel is retained and controlled through r_i.

Similarly, differentiation of $h(\mathbf{x}) = \sigma(U_j)\sigma'(U_i)$ yields the exact self-prediction kernel

$$K_{ij} := \mathbf{v}^\top \mathbb{E}[\sigma(U_j)\sigma'(U_i)\mathbf{x}] = r_j \mathbb{E}[\sigma'(U_j)\sigma'(U_i)] + r_i \mathbb{E}[\sigma(U_j)\sigma''(U_i)]. \tag{29}$$

No independence between U_i and U_j is used here.

For the population loss, direct differentiation gives

$$-\nabla_{\mathbf{f}_i} \mathcal{L}_{\text{pop}}(\mathbf{F}_0) = \frac{a_i}{\sqrt{k}} \mathbb{E}[(y - g_{\mathbf{F}_0}(\mathbf{x}))\sigma'(U_i)\mathbf{x}]. \tag{30}$$

Projecting onto $\mathbf{v}$ and inserting $g_{\mathbf{F}_0}$ proves

$$d_i = \frac{a_i}{\sqrt{k}} \left(L_i - \frac{1}{\sqrt{k}} \sum_{j=1}^{k} K_{ij} a_j \right), \qquad q_{1i} = q_{0i} + \eta d_i. \tag{31}$$

Consequently,

$$\Delta Q_2^{\text{pop}} = \frac{2\eta}{k} \sum_i q_{0i} d_i + \frac{\eta^2}{k} \sum_i d_i^2. \tag{32}$$

This proves Proposition 3.1.

The coherent observable satisfies an equally useful exact identity:

$$Q_a(\mathbf{F}_1^{\text{pop}}) - Q_a(\mathbf{F}_0) = \frac{\eta}{k} \sum_i a_i d_i. \tag{33}$$

It isolates movement paired with the auxiliary coefficients, whereas Q_2 records total projected energy. Finally, the self term is not centered: conditional on $\mathbf{F}_0$, Gaussian averaging over $\mathbf{a}$ gives

$$\mathbb{E}_a[d_i \mid \mathbf{F}_0] = -\frac{K_{ii}}{k}, \qquad \mathbb{E}_a \left[\frac{1}{k} \sum_i q_{0i} d_i \middle| \mathbf{F}_0 \right] = -\frac{1}{k^2} \sum_i q_{0i} K_{ii}. \tag{34}$$

The bounds below control this diagonal correction together with the centered self contribution.

B Proof of the Generic Rank-One Population Theorem

Recall $\mathbf{\Sigma} = \mathbf{I}_n + \theta\boldsymbol{\gamma}\boldsymbol{\gamma}^\top$, $k/n \to \phi \in (0, \infty)$, and the four primitive conditions

$$\frac{\mathrm{tr}(\mathbf{\Sigma}^2)\log k}{s^2} \to 0, \quad \frac{\|\mathbf{\Sigma}\|_{\mathrm{op}}\log k}{s} \to 0, \quad \frac{A\log k}{s} \to 0, \quad \frac{C\log k}{sA^2} \to 0. \tag{35}$$

They are precisely the assumptions in (13).

Initialization concentration. Write $\mathbf{f}_i = s^{-1/2}\mathbf{z}_i$, $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. Gaussian quadratic-form concentration and a union bound imply

$$\max_{i\le k} |\tau_i^2 - 1| = o_p(1) \tag{36}$$

from the first two conditions in (35). Moreover, (q_i, r_i) is exactly centered Gaussian with

$$\mathrm{Var}(q_i) = A/s, \qquad \mathrm{Cov}(q_i, r_i) = H/s, \qquad \mathrm{Var}(r_i) = C/s. \tag{37}$$

Gaussian maxima and the last two primitive conditions give

$$\max_i |q_i| = O_p\Big(\sqrt{A\log k/s}\Big) = o_p(1), \qquad \max_i \frac{|r_i|}{A} = O_p\Big(\sqrt{C\log k/(sA^2)}\Big) = o_p(1). \tag{38}$$

In particular, $Q_2(\mathbf{F}_0) = O_p(A/s) = o_p(1)$.

Label channel. Put $b_i = L_i/A$. By (28), (36)–(38), continuity of Gaussian expectations and the uniform-integrability clauses in T/P,

$$\max_i |b_i - m_* m_\sigma| = o_p(1). \tag{39}$$

Quadratic contribution. With $\mathbf{K} = (K_{ij})$, (31) becomes

$$d_i = \frac{Aa_i}{\sqrt{k}}\left(b_i - \frac{(\mathbf{Ka})_i}{\sqrt{k}A}\right). \tag{40}$$

Appendix C proves

$$E_{\mathrm{self}} := \frac{1}{k}\sum_i a_i^2\left\{\frac{(\mathbf{Ka})_i}{\sqrt{k}A}\right\}^2 = o_p(1). \tag{41}$$

For $\|\mathbf{z}\|_{a,k}^2 = k^{-1}\sum_i a_i^2 z_i^2$, the Gaussian law of large numbers and (39) give $\|\mathbf{b}\|_{a,k} \to_p |m_* m_\sigma|$. Both directions of the triangle inequality, together with (41), therefore imply

$$\frac{1}{k}\sum_i a_i^2\left(b_i - \frac{(\mathbf{Ka})_i}{\sqrt{k}A}\right)^2 \to_p m_*^2 m_\sigma^2. \tag{42}$$

Consequently

$$\frac{1}{k}\sum_i d_i^2 = \frac{A^2}{k}\{m_*^2 m_\sigma^2 + o_p(1)\}. \tag{43}$$

Cross term. For the label part, conditioning on $\mathbf{F}_0$ gives zero mean over $\mathbf{a}$ and

$$\mathrm{Var}_a\left(\frac{1}{k\sqrt{k}}\sum_i q_i a_i L_i \,\middle|\, \mathbf{F}_0\right) = \frac{1}{k^3}\sum_i q_i^2 L_i^2 = O_p\left(\frac{A^3}{k^2 s}\right). \tag{44}$$

At $\eta = c\sqrt{k}/A$, its contribution is $o_p(1)$. Appendix C controls the entire self cross term, including the conditional mean in (34), and gives

$$\eta\left|\frac{1}{k}\sum_i q_i d_i^{\mathrm{self}}\right| \le c\sqrt{Q_2(\mathbf{F}_0)}E_{\mathrm{self}}^{1/2} = o_p(1). \tag{45}$$

Substitution of (43) and (45) into (32) proves

$$\Delta Q_2^{\mathrm{pop}} = \frac{\eta^2 A^2}{k}\{m_*^2 m_\sigma^2 + o_p(1)\}. \tag{46}$$

Taking $\eta = c\sqrt{k}/A$ yields $\Delta Q_2^{\mathrm{pop}} = c^2 m_*^2 m_\sigma^2 + o_p(1)$. When $m_* m_\sigma \neq 0$, the leading energy vanishes when $\eta A/\sqrt{k} \to 0$ and is order one when this normalized step is order one; hence $\eta_{\mathrm{onset}} \asymp \sqrt{k}/A$. This completes the proof of Theorem 3.2.

C Self-Prediction Kernel Bounds

We prove the self estimates used in Appendix B. The exact Stein representation (29), Network P, and uniform row variance concentration give, for a finite activation-dependent constant C_σ independent of n and k,

$$|K_{ij}| \le C_\sigma(|r_i| + |r_j|), \qquad \|\mathbf{K}\|_{\mathrm{F}}^2 \le 2C_\sigma^2 k\|\mathbf{r}\|_2^2, \quad \mathbf{r} := (r_1, \dots, r_k)^\top. \tag{47}$$

Since $r_i \sim \mathcal{N}(0, C/s)$ independently across rows, $\|\mathbf{r}\|_2^2 = O_p(kC/s)$. Thus

$$\|\mathbf{K}\|_{\mathrm{F}}^2 = O_p(k^2 C/s). \tag{48}$$

Conditional on $\mathbf{F}_0$, Wick's formula gives the exact fourth-order average

$$\mathbb{E}_a\left[\frac{1}{k^2A^2}\sum_i a_i^2(\mathbf{Ka})_i^2 \middle| \mathbf{F}_0\right] = \frac{\|\mathbf{K}\|_{\mathrm{F}}^2 + 2\sum_i K_{ii}^2}{k^2A^2} = O_p\left(\frac{C}{sA^2}\right). \tag{49}$$

The fourth primitive condition, followed by conditional Markov, proves

$$E_{\mathrm{self}} = \frac{1}{k}\sum_i a_i^2 \left\{\frac{(\mathbf{Ka})_i}{\sqrt{k}A}\right\}^2 = o_p(1). \tag{50}$$

This is the primitive replacement for the earlier abstract self-energy condition. Reverse triangle in the weighted norm then prevents cancellation between the label vector and the self vector and yields (42); no separate anti-cancellation assumption is needed.

For completeness, the self cross term is

$$T_{\mathrm{self}} = -\frac{1}{k^2}\mathbf{a}^\top \mathbf{D}_q \mathbf{Ka}, \qquad \mathbf{D}_q = \mathrm{Diag}(q_1, \dots, q_k). \tag{51}$$

It is *not* exactly centered. Its conditional mean is

$$\mathbb{E}_a[T_{\mathrm{self}} \mid \mathbf{F}_0] = -\frac{1}{k^2}\sum_i q_i K_{ii}. \tag{52}$$

The centered Gaussian quadratic form may be bounded using

$$\|\operatorname{sym}(\mathbf{D}_q\mathbf{K})\|_{\mathrm{F}} \le \|\mathbf{q}\|_\infty \|\mathbf{K}\|_{\mathrm{F}} = O_p\left(\frac{k\sqrt{AC\log k}}{s}\right), \tag{53}$$

and $\mathrm{Var}_a(T_{\mathrm{self}} \mid \mathbf{F}_0) \le 2k^{-4}\|\operatorname{sym}(\mathbf{D}_q\mathbf{K})\|_{\mathrm{F}}^2$, where $\mathbf{q} = (q_1, \dots, q_k)^\top$. An estimate controlling the centered and diagonal pieces simultaneously is more direct. Cauchy–Schwarz, $Q_2(\mathbf{F}_0) = O_p(A/s)$, and (50) give, at $\eta = c\sqrt{k}/A$,

$$\eta|T_{\mathrm{self}}| \le c\sqrt{Q_2(\mathbf{F}_0)}E_{\mathrm{self}}^{1/2} = o_p(1). \tag{54}$$

Thus the primitive ratios replace all self-cross, self-energy, and non-cancellation assumptions in the theorem's region.

The argument deliberately requires $C/(sA^2) \to 0$ (with logarithmic slack for uniform row control). When $C/(sA^2) = \Theta(1)$, (49) only gives order-one self energy, so the coefficient may change. That boundary is not covered by the theorem.

D Activation Classes and the ReLU Boundary

Table 1 summarizes the network examples used in the generalized results. Here $m_1 = \mathbb{E}\sigma'(G)$, $m_2 = \mathbb{E}\sigma'(G)^2$, and the rank is that of $\sigma'(G) - m_1$.

Tanh, erf, and centered sigmoid are centered, bounded C^3 activations with bounded derivatives and derivative Hermite rank exactly two. A ReLU teacher belongs to T and C–T through its weak derivative $\mathbf{1}_{\{t>0\}}$; it has $m_* = 1/2$ and $\mathbb{E}[\sigma^*(G)^2] = 1/2$.

Network activation	m_1	m_2	Derivative rank and coverage
$\tanh t$	$\mathbb{E}\,\mathrm{sech}^2 G$	$\mathbb{E}\,\mathrm{sech}^4 G$	rank 2; population and critical
$\mathrm{erf}(t)$	$2/\sqrt{3\pi}$	$4/(\pi\sqrt{5})$	rank 2; population and critical
$\frac{1}{2}\tanh(t/2)$	$\frac{1}{4}\mathbb{E}\,\mathrm{sech}^2(G/2)$	$\frac{1}{16}\mathbb{E}\,\mathrm{sech}^4(G/2)$	rank 2; population and critical
t_+	$1/2$	$1/2$	rank 1; separate population result only

Table 1: Examples of network activations. The erf convention is $\mathrm{erf}(t) = 2\pi^{-1/2}\int_0^t e^{-u^2}\,du$; the third row is the centered logistic sigmoid $\{1+e^{-t}\}^{-1} - \frac{1}{2} = \frac{1}{2}\tanh(t/2)$.

Proposition D.1 (ReLU-network population onset)**.** *Let the network activation be $\sigma(t) = t_+$ and let the teacher satisfy T. Under the population primitive conditions, if $m_* \neq 0$, then*

$$\Delta Q_2^{\mathrm{pop}} = \frac{\eta^2 A^2}{k}\left\{\frac{m_*^2}{4} + o_p(1)\right\}. \tag{55}$$

In particular, at $\eta = c\sqrt{k}/A$ the limit is $c^2 m_*^2/4$.

Proof. The ReLU weak derivative is $\mathbf{1}_{\{t>0\}}$ and its distributional second derivative is δ_0. Gaussian integration by parts in BV form gives

$$\begin{aligned} L_i &= A\mathbb{E}[\sigma^{*\prime}(S)\mathbf{1}_{\{U_i>0\}}] + r_i\mathbb{E}[\sigma^*(S)\delta_0(U_i)], \\ K_{ij} &= r_j\mathbb{E}[\mathbf{1}_{\{U_j>0\}}\mathbf{1}_{\{U_i>0\}}] + r_i\mathbb{E}[(U_j)_+\delta_0(U_i)]. \end{aligned} \tag{56}$$

Here $\mathbb{E}[T\delta_0(U_i)] = p_{U_i}(0)\mathbb{E}[T \mid U_i = 0]$. For $i = j$, the second kernel term is zero and the first is $r_i/2$, so the nonzero diagonal is retained. Gaussian conditioning on the row-concentration event gives $|K_{ij}| \le K(|r_i| + |r_j|)$ and $\max_i |L_i/A - m_*/2| = o_p(1)$. The generic self-kernel and cross-term bounds then prove the result. □

Critical boundary. For a ReLU network, $\sigma'(G) - 1/2 = \mathbf{1}_{\{G>0\}} - 1/2$ has Hermite rank one. Uncentered ReLU also contributes a random intercept to $g_{\mathbf{F}_0}$. Hence the smooth rank-two critical proof does not apply: *no ReLU-network critical theorem is claimed.*

E Fixed-Rank Extension

Let

$$\mathbf{\Sigma} = \mathbf{I}_n + \sum_{b=1}^{r_{\mathrm{sp}}} \theta_b \boldsymbol{\gamma}_b \boldsymbol{\gamma}_b^\top, \qquad \boldsymbol{\gamma}_b^\top \boldsymbol{\gamma}_{b'} = \mathbf{1}\{b = b'\}, \qquad \alpha_b = \boldsymbol{\gamma}_b^\top \boldsymbol{\xi}, \qquad r_{\mathrm{sp}} \text{ fixed.} \tag{57}$$

Writing $\lambda_b = 1 + \theta_b$, spectral calculus gives, for $p \ge 1$,

$$M_p := \boldsymbol{\xi}^\top \mathbf{\Sigma}^p \boldsymbol{\xi} = 1 + \sum_{b=1}^{r_{\mathrm{sp}}} (\lambda_b^p - 1)\alpha_b^2. \tag{58}$$

Consequently,

$$\begin{aligned} B &= 1 + \sum_b \theta_b \alpha_b^2, \\ A &= \frac{1 + \sum_b (2\theta_b + \theta_b^2)\alpha_b^2}{B}, \\ H &= \frac{1 + \sum_b (3\theta_b + 3\theta_b^2 + \theta_b^3)\alpha_b^2}{B}, \\ C &= \frac{1 + \sum_b (4\theta_b + 6\theta_b^2 + 4\theta_b^3 + \theta_b^4)\alpha_b^2}{B}. \end{aligned} \tag{59}$$

Also

$$s = n + \sum_b \theta_b, \quad \operatorname{tr} \mathbf{\Sigma}^2 = n + \sum_b (2\theta_b + \theta_b^2), \quad \|\mathbf{\Sigma}\|_{\mathrm{op}} = \max\{1, \max_b (1 + \theta_b)\}. \tag{60}$$

The population proof is covariance-level. For $\mathbf{f}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n/s)$, the exact laws

$$\mathrm{Var}(q_i) = A/s, \quad \mathrm{Cov}(q_i, r_i) = H/s, \quad \mathrm{Var}(r_i) = C/s, \quad \mathbb{E}\tau_i^2 = 1, \quad \mathrm{Var}(\tau_i^2) = 2\operatorname{tr}(\mathbf{\Sigma}^2)/s^2 \tag{61}$$

are unchanged. The four conditions

$$\frac{\operatorname{tr}(\mathbf{\Sigma}^2)\log k}{s^2} \to 0, \quad \frac{\|\mathbf{\Sigma}\|_{\mathrm{op}} \log k}{s} \to 0, \quad \frac{A \log k}{s} \to 0, \quad \frac{C \log k}{sA^2} \to 0 \tag{62}$$

therefore reproduce (36)–(38). The exact generic label identity gives $L_i/A \to m_* m_\sigma$ uniformly, while (47)–(54) depend only on C, s, A and hence also apply. The quadratic coefficient converges to $m_*^2 m_\sigma^2$, and both cross terms vanish. Thus

$$\Delta Q_2^{\mathrm{pop}} = \frac{\eta^2 A^2}{k}\{m_*^2 m_\sigma^2 + o_p(1)\}, \qquad \eta_{\mathrm{onset}} \asymp \frac{\sqrt{k}}{A}, \tag{63}$$

which proves Corollary 3.4.

For a compact description of polynomial regimes, let $\theta_b = c_b n^{\beta_b}$, $\alpha_b = d_b n^{-\delta_b}$ and define

$$D = \max\{0, \max_b (\beta_b - 2\delta_b)\}, \qquad P_p = \max\{0, \max_b (p\beta_b - 2\delta_b)\}. \tag{64}$$

When $\max_b \beta_b < 1$, the target moments have exponents $\exp_n A = P_2 - D$, $\exp_n H = P_3 - D$, and $\exp_n C = P_4 - D$. The self-smallness exponent is

$$E_R = P_4 + D - 1 - 2P_2, \tag{65}$$

so $E_R < 0$ (with logarithmic slack) gives the final primitive condition. This classification shows the distinction precisely: A determines the leading onset scale inside the admissible region, whereas C, trace moments, and the operator norm determine whether that region applies. No result for arbitrary deterministic covariance is asserted.

F Exact Rank-One Spike–Alignment Regimes

Let

$$\mathbf{\Sigma} = \mathbf{I}_n + \theta \boldsymbol{\gamma}\boldsymbol{\gamma}^\top, \qquad \theta = c_\theta n^\beta, \qquad \alpha = \boldsymbol{\gamma}^\top \boldsymbol{\xi} = c_\alpha n^{-\delta}, \tag{66}$$

where $c_\theta > 0$, $c_\alpha \neq 0$, and $\beta, \delta \geq 0$. The exact target-weighted moments are

$$\begin{aligned} B &= 1 + \theta\alpha^2, \\ A &= \frac{1 + (2\theta + \theta^2)\alpha^2}{1 + \theta\alpha^2}, \\ H &= \frac{1 + (3\theta + 3\theta^2 + \theta^3)\alpha^2}{1 + \theta\alpha^2}, \\ C &= \frac{1 + (4\theta + 6\theta^2 + 4\theta^3 + \theta^4)\alpha^2}{1 + \theta\alpha^2}. \end{aligned} \tag{67}$$

Write $\exp_n(X) = e$ when $X = n^{e + o(1)}$, and define

$$D = (\beta - 2\delta)_+, \qquad P_p = (p\beta - 2\delta)_+. \tag{68}$$

Then

$$\exp_n(B) = D, \quad \exp_n(A) = P_2 - D, \quad \exp_n(H) = P_3 - D, \quad \exp_n(C) = P_4 - D. \tag{69}$$

F.1 Moment and sample exponents

Alignment range	$\exp_n B$	$\exp_n A$	$\exp_n H$	$\exp_n C$
$0 \le \delta < \beta/2$	$\beta - 2\delta$	β	2β	3β
$\delta = \beta/2$	0	β	2β	3β
$\beta/2 < \delta < \beta$	0	$2\beta - 2\delta$	$3\beta - 2\delta$	$4\beta - 2\delta$
$\delta = \beta$	0	0	β	2β
$\beta < \delta < 3\beta/2$	0	0	$3\beta - 2\delta$	$4\beta - 2\delta$
$3\beta/2 \le \delta < 2\beta$	0	0	0	$4\beta - 2\delta$
$\delta \ge 2\beta$	0	0	0	0

All leading coefficients are nonnegative, so no exponent-level cancellation occurs. Corollary 3.3 follows from $\eta_{\text{onset}} \asymp n^{1/2-\exp_n(A)}$.

For the feature-gradient sample scale,

$$\frac{H}{A^2} = n^{\rho_m + o(1)}, \qquad \rho_m = P_3 + D - 2P_2, \tag{70}$$

where

$$\rho_m = \begin{cases} 0, & 0 \le \delta \le \beta/2, \\ 2\delta - \beta, & \beta/2 < \delta < \beta, \\ \beta, & \delta = \beta, \\ 3\beta - 2\delta, & \beta < \delta < 3\beta/2, \\ 0, & \delta \ge 3\beta/2. \end{cases} \tag{71}$$

Thus the scale grows exactly for $\beta > 0$ and $\beta/2 < \delta < 3\beta/2$; exponent zero means an order-one scale, not necessarily convergence to one.

F.2 Population and critical admissibility

For $\beta < 1$, the trace-square, operator, and target-overlap primitive conditions hold automatically. The self-prediction ratio has exponent

$$e_{\text{self}} = \begin{cases} \beta - 1, & 0 \le \delta < \beta/2, \\ 2\delta - 1, & \beta/2 \le \delta < \beta, \\ 4\beta - 2\delta - 1, & \beta \le \delta < 2\beta, \\ -1, & \delta \ge 2\beta. \end{cases} \tag{72}$$

The logarithmic primitive requires the strict inequality $e_{\text{self}} < 0$. Hence the complete population-theorem region is

$$\boxed{\begin{array}{ll} 0 \le \beta < 1/2: & \delta \ge 0, \\ \beta = 1/2: & \delta \in [0, 1/2) \cup (1/2, \infty), \\ 1/2 < \beta < 1: & \delta \in [0, 1/2) \cup (2\beta - 1/2, \infty), \\ \beta \ge 1: & \text{no points under the current primitives.} \end{array}} \tag{73}$$

The effective dimension has exponent

$$\rho_d = \begin{cases} 1, & 0 \le \beta \le 1/2, \\ 2 - 2\beta, & 1/2 < \beta < 1, \\ 0, & \beta \ge 1. \end{cases} \tag{74}$$

At critical sampling, the additional requirements are $\rho_m > 0$, $\rho_m < 1$, and $\rho_m < \rho_d$. Intersecting them with (73) gives

$$\boxed{\begin{array}{ll} 0 < \beta < 1/2: & \beta/2 < \delta < 3\beta/2, \\ \beta = 1/2: & \delta \in (1/4, 1/2) \cup (1/2, 3/4), \\ 1/2 < \beta < 1: & \delta \in (\beta/2, 1/2) \cup (2\beta - 1/2, 3\beta/2), \\ \text{otherwise}: & \varnothing. \end{array}} \tag{75}$$

The sample-ratio conditions remain assumptions of the general critical theorem even though they introduce no further interior cut in this polynomial rank-one family.

F.3 Equality ledger and interpretation

- At $\beta = 1/2$, population theory holds except at $\delta = 1/2$. This is the crossover $\theta^2 \asymp n$ in $\operatorname{tr} \mathbf{\Sigma}^2$, not the global upper boundary.
- At $\delta = \beta/2$ and $\delta = 3\beta/2$, $\rho_m = 0$, so the growing-batch critical theorem does not apply.
- At $\delta = \beta$, population and critical theory both apply only when $0 < \beta < 1/2$; the common point $\beta = \delta = 1/2$ fails self control.
- For $\beta \geq 1/2$, the line $\delta = 1/2$ is excluded because the self exponent is zero and the logarithm diverges.
- For $\beta > 1/2$, the line $\delta = 2\beta - 1/2$ is likewise excluded.
- At $\beta = 1$, the global trace/operator exponents are zero, so their logarithmic conditions exclude every alignment.

For independent generic unit directions, $|\boldsymbol{\gamma}^\top \boldsymbol{\xi}| = \Theta_p(n^{-1/2})$, corresponding at exponent level to $\delta = 1/2$. When $\beta < 1/2$, this orientation receives no polynomial onset benefit. At and above the crossover it lies on the current self-prediction boundary and is not covered. This is only an exponent benchmark; it is not a claim about every random teacher/spike construction or its order-one prefactor.

G Exact Empirical-Gradient Decomposition

Condition on $(\mathbf{F}_0, \mathbf{a})$ and set

$$\mathcal{R}(\mathbf{x}) = y - g_{\mathbf{F}_0}(\mathbf{x}), \qquad Z_i(\mathbf{x}) = \mathcal{R}(\mathbf{x})\sigma'(U_i)V, \qquad \mu_i = \mathbb{E}[Z_i(\mathbf{x}) \mid \mathbf{F}_0, \mathbf{a}]. \tag{76}$$

For an independent sample $\{\mathbf{x}_\ell\}_{\ell=1}^{m_{\mathrm{FL}}}$, the projected empirical direction is exactly

$$\widehat{d}_i = \frac{a_i}{\sqrt{k}} \frac{1}{m_{\mathrm{FL}}} \sum_{\ell=1}^{m_{\mathrm{FL}}} Z_i(\mathbf{x}_\ell) = d_i + \varepsilon_i, \qquad d_i = \frac{a_i}{\sqrt{k}} \mu_i, \tag{77}$$

where

$$\varepsilon_i = \frac{a_i}{\sqrt{k}} \left\{ \frac{1}{m_{\mathrm{FL}}} \sum_{\ell=1}^{m_{\mathrm{FL}}} Z_i(\mathbf{x}_\ell) - \mu_i \right\}. \tag{78}$$

This uses the full residual, including self prediction.

Conditional iid sampling gives

$$\mathbb{E}[\varepsilon_i \mid \mathbf{F}_0, \mathbf{a}] = 0, \qquad \operatorname{Cov}(\varepsilon_i, \varepsilon_j \mid \mathbf{F}_0, \mathbf{a}) = \frac{a_i a_j}{k m_{\mathrm{FL}}} \Omega_{ij}, \tag{79}$$

with

$$\Omega_{ij} = \mathbb{E}[\mathcal{R}(\mathbf{x})^2 \sigma'(U_i)\sigma'(U_j)V^2 \mid \mathbf{F}_0, \mathbf{a}] - \mu_i \mu_j. \tag{80}$$

The rows are therefore not conditionally independent: they share the same feature-learning sample.

Under T/P, row concentration, Gaussian moment bounds, and exact averaging over the auxiliary coefficients give

$$\frac{1}{k} \sum_i \varepsilon_i^2 = O_p\left(\frac{H}{k m_{\mathrm{FL}}} \right). \tag{81}$$

Indeed, after division by H, the conditional expected noise energy is bounded by the row average of $a_i^2 \mathbb{E}[\mathcal{R}^2 \sigma'(U_i)^2 V^2 / H \mid \mathbf{F}_0, \mathbf{a}]$. The label part is controlled by Teacher T and polynomial Gaussian moments. For the self part, Gaussian Wick contraction gives exactly

$$\mathbb{E}_a[a_i^2 g_{\mathbf{F}_0}(\mathbf{x})^2 \mid \mathbf{F}_0, \mathbf{x}] = \frac{1}{k} \left\{ \sum_j \sigma(U_j)^2 + 2\sigma(U_i)^2 \right\}, \tag{82}$$

which is uniformly bounded in row average under Network P. No rowwise independence is used. The population direction instead satisfies

$$\frac{1}{k} \sum_i d_i^2 = \frac{A^2}{k} \{ m_*^2 m_\sigma^2 + o_p(1) \}. \tag{83}$$

Their ratio is, up to bounded activation/residual constants,

$$\chi_{\mathrm{FL}} = \frac{H}{m_{\mathrm{FL}} A^2}, \qquad \mathrm{SNR}_{\mathrm{grad}} \asymp \frac{m_{\mathrm{FL}} A^2}{H}. \tag{84}$$

Thus H/A^2 is the sufficient generic recovery scale. A two-sided critical interpretation uses the narrower C–T/C–P assumptions. This is a feature-gradient estimation scale, not a generalization sample complexity.

Finally, $q_1 = q + \eta(d + \varepsilon)$ gives the exact expansion

$$\Delta Q_2^{\mathrm{emp}} = \Delta Q_2^{\mathrm{pop}} + \frac{2\eta}{k} \sum_i q_i \varepsilon_i + \frac{2\eta^2}{k} \sum_i d_i \varepsilon_i + \frac{\eta^2}{k} \sum_i \varepsilon_i^2. \tag{85}$$

The last term is nonnegative and has order $\eta^2 H/(k m_{\mathrm{FL}})$; relative to population quadratic energy its order is χ_{FL}. Equation (85) is the starting point for the recovery proof.

H Proof of Population Recovery

Assume T/P, the generic rank-one population theorem or its fixed-rank extension, $m_* m_\sigma \neq 0$, and let $\chi_{\mathrm{FL}} = H/(m_{\mathrm{FL}} A^2) \to 0$. Define the normalized population and noise energies

$$P_k := \frac{k}{A^2} \frac{1}{k} \sum_i d_i^2, \qquad N_k := \frac{k}{A^2} \frac{1}{k} \sum_i \varepsilon_i^2. \tag{86}$$

The population theorem gives $P_k \to_p m_*^2 m_\sigma^2$, while (81) gives $N_k = O_p(\chi_{\mathrm{FL}}) = o_p(1)$.

At $\eta = c\sqrt{k}/A$, the pure empirical-noise term in (85) equals $c^2 N_k = o_p(1)$. The signal–noise term obeys

$$\left| \frac{2\eta^2}{k} \sum_i d_i \varepsilon_i \right| \le 2c^2 \sqrt{P_k N_k} = o_p(1). \tag{87}$$

Likewise, using $Q_2(\mathbf{F}_0) = O_p(A/s) = o_p(1)$,

$$\left| \frac{2\eta}{k} \sum_i q_i \varepsilon_i \right| \le 2c \sqrt{Q_2(\mathbf{F}_0)} \sqrt{N_k} = o_p(1). \tag{88}$$

All three corrections in (85) vanish, hence

$$\Delta Q_2^{\mathrm{emp}} = \Delta Q_2^{\mathrm{pop}} + o_p(1) = c^2 m_*^2 m_\sigma^2 + o_p(1). \tag{89}$$

Because the denominator converges to the positive constant $c^2 m_*^2 m_\sigma^2$, Slutsky's theorem also yields $\Delta Q_2^{\mathrm{emp}} / \Delta Q_2^{\mathrm{pop}} \to_p 1$. The proof uses only the covariance-level population and noise bounds, so it applies verbatim to the fixed-rank setting under the corresponding primitive assumptions. This proves Theorem 4.1.

I Proof of the Critical Finite-Sample Law

We prove Theorem 4.2 under exactly its stated triangular conditions. Write $m = m_{\mathrm{FL}}$ and assume

$$\rho = \frac{A}{\sqrt{H}} \to 0, \quad m \to \infty, \quad m\rho^2 \to \chi^{-1} \in (0, \infty), \quad \frac{m}{k} \to 0, \quad \frac{m}{d_{\mathrm{eff}}} \to 0, \tag{90}$$

in addition to the primitive population assumptions. Put

$$h_i(\mathbf{x}) = \sigma'(\mathbf{f}_i^\top \mathbf{x}), \quad W(\mathbf{x}) = \{\sigma^*(S) - g_{\mathbf{F}_0}(\mathbf{x})\} V, \quad B_i = \frac{1}{m} \sum_{\ell=1}^{m} W(\mathbf{x}_\ell) h_i(\mathbf{x}_\ell). \tag{91}$$

Then $\widehat{d}_i = a_i B_i / \sqrt{k}$ exactly. Write

$$\kappa_* = \mathbb{E}\sigma^*(G)^2, \qquad \kappa_\sigma = \mathbb{E}\sigma(G)^2, \qquad \nu = \kappa_* + \kappa_\sigma. \tag{92}$$

I.1 LEAVE-ONE-ROW-OUT REDUCTION

Define

$$g_0^{(-i)}(\mathbf{x}) = \frac{1}{\sqrt{k}} \sum_{j \neq i} a_j \sigma(\mathbf{f}_j^\top \mathbf{x}), \qquad W^{(-i)}(\mathbf{x}) = \{\sigma^*(S) - g_0^{(-i)}(\mathbf{x})\}V. \tag{93}$$

The identities

$$g_0 - g_0^{(-i)} = \frac{a_i}{\sqrt{k}} \sigma(U_i), \qquad W - W^{(-i)} = -\frac{a_i}{\sqrt{k}} \sigma(U_i) V \tag{94}$$

are exact. Boundedness under C–P and Gaussian moments of V imply

$$\frac{1}{k} \sum_i a_i^2 |B_i - B_i^{(-i)}|^2 = O_p(H/k). \tag{95}$$

After division by A^2 this is $O_p(H/(kA^2)) = o_p(1)$, and the associated bilinear errors vanish by Cauchy–Schwarz.

Conditional on $(\mathbf{F}_0, \mathbf{x})$, $g_0^{(-i)}(\mathbf{x})$ is exactly Gaussian over $\mathbf{a}_{-i}$, with variance

$$K_k^{(-i)}(\mathbf{x}, \mathbf{x}) = \frac{1}{k} \sum_{j \neq i} \sigma^2(\mathbf{f}_j^\top \mathbf{x}). \tag{96}$$

Let $r_{\mathbf{x}} = \|\mathbf{x}\|_2^2/s$. Quadratic-form concentration gives $r_{\mathbf{x}} - 1 = O_p(d_{\text{eff}}^{-1/2})$, and conditional Hoeffding concentration gives, in batch-averaged mean square,

$$K_k^{(-i)}(\mathbf{x}, \mathbf{x}) - \kappa_\sigma = O_p(k^{-1/2} + d_{\text{eff}}^{-1/2}). \tag{97}$$

Thus the auxiliary predictor has asymptotic variance κ_σ.

Conditional on $\mathbf{f}_i$, the covariance matrix of $(S, V/\sqrt{H}, U_i)$ is

$$\begin{pmatrix} 1 & \rho & q_i \\ \rho & 1 & r_i/\sqrt{H} \\ q_i & r_i/\sqrt{H} & \tau_i^2 \end{pmatrix}. \tag{98}$$

The primitive conditions yield uniformly

$$\tau_i^2 = 1 + o_p(1), \quad q_i = o_p(1), \quad \frac{|r_i|}{\sqrt{H}} = O_p\Big(\sqrt{C \log k/(sH)}\Big) = o_p(1), \tag{99}$$

where $C \leq \|\mathbf{\Sigma}\|_{\text{op}} H$. The remaining covariance between U_i and $g_0^{(-i)}$ is

$$\frac{1}{\sqrt{k}} \sum_{j \neq i} a_j (\mathbf{f}_i^\top \mathbf{\Sigma} \mathbf{f}_j) \mathbb{E}\sigma'\big(N(0, \tau_j^2)\big), \tag{100}$$

whose conditional second moment is $O_p(d_{\text{eff}}^{-1})$.

Lemma I.1 (Fixed-row mixed-moment decoupling)**.** *For fixed i, let $G_i = g_0^{(-i)}(\mathbf{x})$ and $\widetilde{W}_i = \{\sigma^*(S) - G_i\}V$. Under the primitive conditions and the critical weak-alignment regime,*

$$\begin{aligned} &\text{Cov}(U_i, G_i \mid \mathbf{F}_0, \mathbf{a}) = O_p(d_{\text{eff}}^{-1/2}), \\ &\mathbb{E}[\widetilde{W}_i h_i] - m_1 \mathbb{E}\widetilde{W}_i = o_p(A), \qquad \mathbb{E}[\widetilde{W}_i^2 h_i^2] - m_2 \mathbb{E}\widetilde{W}_i^2 = o_p(H). \end{aligned} \tag{101}$$

The same orders hold after the row averages used below.

Indeed, the first bound follows from the displayed conditional second moment. For the two mixed moments, expand the relevant Gaussian functions in Hermite series. Every nonconstant contraction contains at least one of q_i, $r_i/\sqrt{H}$, or $\text{Cov}(U_i, G_i)$; Cauchy–Schwarz, C–T/C–P moments, and (99) give respectively $o_p(A)$ and $o_p(H)$. This is an annealed decoupling statement appropriate to the unconditional theorem.

I.2 Variance-dependent derivative kernel

For independent inputs $\mathbf{x}, \mathbf{x}'$, define

$$r_{\mathbf{x}} := \frac{\|\mathbf{x}\|_2^2}{s}, \qquad R(\mathbf{x}, \mathbf{x}') := \frac{\mathbf{x}^\top \mathbf{x}'}{s}, \qquad \varrho(\mathbf{x}, \mathbf{x}') := \frac{R(\mathbf{x}, \mathbf{x}')}{\sqrt{r_{\mathbf{x}} r_{\mathbf{x}'}}}, \tag{102}$$

and, for $G \sim \mathcal{N}(0,1)$,

$$b_p(r) := \mathbb{E}\big[\{\sigma'(\sqrt{r}G) - m_1\} H_p(G)\big]\,. \tag{103}$$

Lemma I.2 (Variance-dependent derivative kernel)**.** *Under C–P, conditional on $\mathbf{x}, \mathbf{x}'$, Mehler's identity gives*

$$C_{\mathbf{F}}(\mathbf{x}, \mathbf{x}') = b_0(r_{\mathbf{x}}) b_0(r_{\mathbf{x}'}) + \sum_{p \geq 1} \frac{b_p(r_{\mathbf{x}}) b_p(r_{\mathbf{x}'})}{p!} \varrho(\mathbf{x}, \mathbf{x}')^p. \tag{104}$$

Moreover, $b_0(1) = b_1(1) = 0$ and $b_2(1) =: c_2 \neq 0$. Uniformly for $r, r' \in [1/2, 3/2]$,

$$|b_0(r)| + |b_1(r)| \leq C_\sigma |r - 1|, \qquad |b_2(r) - c_2| \leq C_\sigma |r - 1|, \tag{105}$$

and consequently, with $D_{\mathbf{x}} := r_{\mathbf{x}} - 1$,

$$|C_{\mathbf{F}}(\mathbf{x}, \mathbf{x}')| \leq C_\sigma \big(|D_{\mathbf{x}} D_{\mathbf{x}'}| + R(\mathbf{x}, \mathbf{x}')^2\big). \tag{106}$$

Proof. The standardized pair $(\mathbf{f}^\top \mathbf{x}/\sqrt{r_{\mathbf{x}}}, \mathbf{f}^\top \mathbf{x}'/\sqrt{r_{\mathbf{x}'}})$ is jointly standard Gaussian with correlation $\varrho(\mathbf{x}, \mathbf{x}')$, proving (104). Hermite rank two at unit variance gives the three identities at $r = 1$. Differentiation under the Gaussian integral gives $b_p'(r) = (2\sqrt{r})^{-1} \mathbb{E}[\sigma''(\sqrt{r}G) G H_p(G)]$ for $p = 0, 1, 2$; boundedness in C–P proves (105). Parseval bounds the remaining coefficients uniformly. More explicitly,

$$\sum_{p \geq 3} \frac{|b_p(r) b_p(r')|}{p!} |\varrho|^p \leq |\varrho|^3 \left(\sum_{p \geq 3} \frac{b_p(r)^2}{p!} \right)^{1/2} \left(\sum_{p \geq 3} \frac{b_p(r')^2}{p!} \right)^{1/2} \leq C_\sigma |\varrho|^3.$$

For $r, r' \in [1/2, 3/2]$, $|\varrho| \leq 2|R|$ and $|\varrho|^3 \leq 2|\varrho|^2 \leq 8R^2$. Separating the first three terms and using (105) therefore proves (106). □

This lemma is essential: Hermite rank two holds at unit variance, but the conditional row variances fluctuate. The generally false envelope $|C_{\mathbf{F}}| \lesssim R^2$ is not used.

I.3 Growing-batch row law

Set $\zeta_i(\mathbf{x}) = h_i(\mathbf{x}) - m_1$ and

$$C_{\mathbf{F}}(\mathbf{x}_\ell, \mathbf{x}_r) = \mathbb{E}_{\mathbf{f}}[\zeta(\mathbf{f}^\top \mathbf{x}_\ell) \zeta(\mathbf{f}^\top \mathbf{x}_r) \mid \mathbf{x}_\ell, \mathbf{x}_r]. \tag{107}$$

The derivative field has Hermite rank two only at unit variance. For $r_\ell = \|\mathbf{x}_\ell\|_2^2/s$, $D_\ell = r_\ell - 1$, and $R_{\ell r} = \mathbf{x}_\ell^\top \mathbf{x}_r / s$, Lemma I.2 gives, on $\max_{\ell \leq m} |D_\ell| \leq 1/2$,

$$|C_{\mathbf{F}}(\mathbf{x}_\ell, \mathbf{x}_r)| \leq C_\sigma \{|D_\ell D_r| + R_{\ell r}^2\}, \qquad \ell \neq r. \tag{108}$$

The event has probability tending to one because $\mathbb{P}(\max_\ell |D_\ell| > 1/2) \leq 8m/d_{\text{eff}}$. Notice that (108) contains the variance-mismatch constant and does not assert the false bound $|C_{\mathbf{F}}| \leq C_\sigma R^2$.

There is also a mean-shift term because $\mathbb{E}_{\mathbf{f}} \zeta(\mathbf{f}^\top \mathbf{x}_\ell) = b_0(r_\ell)$ need not vanish. Indeed, conditional on the batch,

$$\mathbb{E}_{\mathbf{f}}[h(\mathbf{f}^\top \mathbf{x}_\ell) h(\mathbf{f}^\top \mathbf{x}_r)] = m_1^2 + m_1 \{b_0(r_\ell) + b_0(r_r)\} + C_{\mathbf{F}}(\mathbf{x}_\ell, \mathbf{x}_r). \tag{109}$$

Writing $W_\ell := W(\mathbf{x}_\ell)$ and $S_0 = m^{-1} \sum_\ell W_\ell b_0(r_\ell)$, the linear contribution is $2m_1 \bar{W} S_0 / A^2$. Gaussian quadratic-form moments, the C–T/C–P moment bounds, and Lemma I.2 give

$$T_2^{\text{pert}} = \frac{1}{A^2 m^2} \sum_{\ell \neq r} W_\ell W_r \frac{1}{2} \left\{ \frac{b_2(r_\ell) b_2(r_r)}{r_\ell r_r} - c_2^2 \right\} R_{\ell r}^2,$$

$$T_{\geq 3} = \frac{1}{A^2 m^2} \sum_{\ell \neq r} W_\ell W_r \sum_{p \geq 3} \frac{b_p(r_\ell) b_p(r_r)}{p!} \left\{ \frac{R_{\ell r}}{\sqrt{r_\ell r_r}} \right\}^p, \tag{110}$$

and

$$\frac{S_0}{A} = O_p(\sqrt{m/d_{\text{eff}}}), \qquad \frac{1}{A^2m^2}\sum_{\ell\neq r} W_\ell W_r b_0(r_\ell)b_0(r_r) = O_p(m/d_{\text{eff}}),$$
$$T_2^{\text{pert}} = O_p(m/d_{\text{eff}}^{3/2}), \qquad T_{\geq 3} = O_p(m/d_{\text{eff}}^{3/2}), \tag{111}$$

and the diagonal variance replacement costs $O_p(d_{\text{eff}}^{-1/2})$. Here are the rate calculations suppressed by this summary. On the good event, (105) makes the perturbed quadratic coefficient $O(|D_\ell| + |D_r|)$, while the Hermite-tail estimate in Lemma I.2 is $O(|R_{\ell r}|^3)$. Gaussian quadratic-form moments give $\mathbb{E}|D|^4 = O(d_{\text{eff}}^{-2})$ and $\mathbb{E}|R_{\ell r}|^6 = O(d_{\text{eff}}^{-3})$; C–T/C–P and the fixed-row lemma give $\mathbb{E}W^4 = O(H^2)$. Hölder applied termwise, followed by averaging the $m(m-1)$ off-diagonal pairs, thus bounds both normalized sums by

$$O_p\left(\frac{H/A^2}{d_{\text{eff}}^{3/2}}\right) = O_p\left(\frac{m}{d_{\text{eff}}^{3/2}}\right),$$

where the last equality uses $H/A^2 \asymp m$ at critical sampling. Truncation outside the good event is absorbed by the same fourth/eighth-moment bounds. The constant unit-variance rank-two term is $O_p(m/d_{\text{eff}}) = o_p(1)$ by the same absolute moment bounds. Appendix K gives an optional sharper signed calculation for that term, but it is not needed here.

Define $\mathcal{R}_{k,m}$ as the maximum of the absolute errors in the quadratic and linear row identities displayed below. Conditional row averaging after the leave-two-rows-out replacement contributes $O_p(\sqrt{m/k})$; (95) contributes $O_p(\sqrt{H/(kA^2)})$ by Cauchy–Schwarz. Combining those terms with (111) gives the completed bound

$$\mathcal{R}_{k,m} = O_p\Big(\sqrt{m/d_{\text{eff}}} + \sqrt{m/k} + \sqrt{H/(kA^2)} + \varepsilon_{\text{prim}}\Big) = o_p(1), \tag{112}$$

where $\mathcal{R}_{k,m}$ is the normalized error in the row-averaged quadratic identity and $\varepsilon_{\text{prim}} \to 0$ collects (97)–(99): row-variance, q_i, $r_i/\sqrt{H}$, and leave-one-out predictor-covariance errors. Each component vanishes under the four population primitive ratios; no additional critical-scale assumption is hidden in $\varepsilon_{\text{prim}}$. At critical sampling $H/A^2 \asymp m$, so every displayed term vanishes under the stated assumptions. Truncating the batch fourth moment of $W/\sqrt{H}$, applying the conditional row law of large numbers, and removing the truncation with the C–T/C–P eighth-moment bound justifies the conditional averaging step. The effective-dimension condition is used by the variance mean shift and the absolute remainder bounds; the signed rank-two argument does not remove it for the full kernel.

The resulting joint row law is

$$\frac{1}{k}\sum_i a_i^2(B_i/A)^2 - m_1^2(\bar{W}/A)^2 - \frac{m_2 - m_1^2}{A^2m^2}\sum_{\ell=1}^m W(\mathbf{x}_\ell)^2 \to_p 0, \tag{113}$$
$$\frac{1}{k}\sum_i a_i^2 B_i/A - m_1\bar{W}/A \to_p 0, \tag{114}$$

where $\bar{W} = m^{-1}\sum_\ell W(\mathbf{x}_\ell)$.

I.4 Scalar triangular limit

Lemma I.3 (Scalar triangular moments)**.** *Conditional on a typical initialization, the $W(\mathbf{x}_\ell)$ are iid and, for some $\epsilon > 0$,*

$$\frac{\mathbb{E}W}{A} \to m_*, \qquad \frac{\text{Var}(W)}{H} \to \nu, \qquad \sup_n \mathbb{E}|W/\sqrt{H}|^{2+\epsilon} < \infty \tag{115}$$

in annealed probability.

For the mean, Gaussian integration by parts gives $\mathbb{E}[\sigma^*(S)V] = Am_*$, while Lemma I.1 makes the leave-one-row predictor contribution $o_p(A)$. For the variance, $\rho = A/\sqrt{H} \to 0$ decouples $V/\sqrt{H}$ from the teacher coordinate and centered auxiliary predictor; their squared terms converge to κ_* and

κ_σ, and the mixed term vanishes. C–T's $8+\epsilon$ teacher moment, bounded C–P, Gaussian moments of $V/\sqrt{H}$, and conditional Gaussianity of the leave-one-row predictor yield the uniform $(2+\epsilon)$ bound. Consequently Lindeberg's condition and uniform integrability hold, so the triangular Lindeberg CLT and conditional LLN give jointly

$$\frac{\bar{W}}{A} \Rightarrow m_* + \sqrt{\nu\chi}\, Z, \qquad \frac{1}{mH}\sum_{\ell=1}^{m} W(\mathbf{x}_\ell)^2 \to_p \nu. \tag{116}$$

I.5 Final overlap laws

Combining (113) and (116) yields

$$\frac{1}{k}\sum_i a_i^2 (B_i/A)^2 \Rightarrow m_1^2 (m_* + \sqrt{\nu\chi}\, Z)^2 + \nu\chi(m_2 - m_1^2). \tag{117}$$

At $\eta = c\sqrt{k}/A$, the quadratic update equals c^2 times the left side of (117). The initial-overlap cross term is $o_p(1)$ by Cauchy–Schwarz, since $Q_2(\mathbf{F}_0) = o_p(1)$ and the quadratic factor is tight. This proves the stated limit for $\Delta Q_2^{\mathrm{emp}}$. Finally, (33) and (114) give

$$Q_a(\mathbf{F}_1^{\mathrm{emp}}) - Q_a(\mathbf{F}_0) \Rightarrow c\, m_1 (m_* + \sqrt{\nu\chi}\, Z). \tag{118}$$

The fixed-rank corollary follows because each step above uses the same covariance-level primitive bounds and exact Gaussian covariances; all conditions in (90), especially $m/d_{\mathrm{eff}} \to 0$, remain in force.

J Noise-Dominated Regime

This detailed proposition remains in the matched-tanh specialization. Put $\mu = \mathbb{E}\,\mathrm{sech}^2(G)$, $\mu_2 = \mathbb{E}\,\mathrm{sech}^4(G)$, and $\kappa = \mathbb{E}\tanh^2(G)$. Assume the rank-one primitive conditions, $m = m_{\mathrm{FL}} \to \infty$, and the same growing-sample derivative-kernel conditions used in Appendix I. Suppose $\chi_{\mathrm{FL}} = H/(mA^2) \to \infty$. At the population step $\eta = c\sqrt{k}/A$, the pure empirical-noise contribution has scale

$$\frac{\eta^2 H}{km} = c^2 \chi_{\mathrm{FL}} \to \infty, \tag{119}$$

so the population normalization is not statistically usable.

Instead take

$$\eta_{\mathrm{noise}} = d\sqrt{\frac{km}{H}}, \qquad d > 0. \tag{120}$$

The population signal under this scaling is smaller than the empirical-noise scale by $\chi_{\mathrm{FL}}^{-1/2}$ and vanishes. The row law of Appendix I, applied after normalization by $\sqrt{H/m}$, remains valid after the variance-dependent kernel repair. Indeed, the variance mean shift is $O_p(\sqrt{m/d_{\mathrm{eff}}})$, the complete off-diagonal absolute remainder is $O_p(m/d_{\mathrm{eff}})$, conditional row averaging is $O_p(\sqrt{m/k})$, and the leave-one-row replacement is $O_p(\sqrt{m/k})$ at this normalization. Thus the same assumptions $m/k \to 0$ and $m/d_{\mathrm{eff}} \to 0$ close the growing-batch row law. Because $\chi_{\mathrm{FL}} \to \infty$ and $m \to \infty$ imply $A/\sqrt{H} \to 0$, the centered sample residual obeys the required scalar CLT. Therefore

$$\Delta Q_2^{\mathrm{emp}} \Rightarrow 2\kappa d^2\{\mu^2 Z^2 + \mu_2 - \mu^2\}, \tag{121}$$

and the row-linear counterpart gives

$$Q_a(\mathbf{F}_1^{\mathrm{emp}}) - Q_a(\mathbf{F}_0) \Rightarrow d\mu\sqrt{2\kappa}\, Z. \tag{122}$$

The initial-overlap cross term again vanishes because $Q_2(\mathbf{F}_0) = o_p(1)$ and the rescaled quadratic energy is tight. The coherent limit is centered and has random sign: order-one energy here is not evidence of reliable target learning or improved prediction. The result does not cover bounded m and does not relax the derivative-kernel or effective-dimension conditions used for the growing-sample row law. No generic activation coefficient is claimed for this regime.

K Optional Signed Rank-Two Sharpening

This optional section records completed refinements of the leading derivative kernel. They explain why $m_{\mathrm{FL}}/d_{\mathrm{eff}} \to 0$ may be conservative, but they do *not* remove it from Theorem 4.2. The explicit signed calculations below are for the matched-tanh specialization; no stronger generic critical theorem is inferred from them.

Let $\mathbf{x}, \mathbf{x}'$ be independent $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, put

$$R(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^\top \mathbf{x}'}{s}, \qquad \zeta(u) = \operatorname{sech}^2(u) - \mu, \qquad W(\mathbf{x}) = \{\tanh S - g_{\mathbf{F}_0}(\mathbf{x})\}V. \tag{123}$$

For a unit-variance Gaussian argument, the centered even field has expansion

$$\zeta(u) = \sum_{p \geq 2,\ p \text{ even}} \frac{c_p}{p!} H_p(u), \qquad c_p = \mathbb{E}[\zeta(G) H_p(G)]. \tag{124}$$

Twice integrating by parts and using $(\operatorname{sech}^2 u)'' = 4\operatorname{sech}^2 u - 6\operatorname{sech}^4 u$ gives the exact rank-two coefficient

$$c_2 = \mathbb{E}[\operatorname{sech}^2(G)(G^2 - 1)] = 4\mu - 6\mu_2. \tag{125}$$

Hence for jointly standard Gaussian arguments with correlation R,

$$\mathbb{E}[\zeta(G_1)\zeta(G_2)] = \frac{c_2^2}{2} R^2 + \sum_{p \geq 4,\ p \text{ even}} \frac{c_p^2}{p!} R^p. \tag{126}$$

K.1 Signed rank-two matrix identity

Condition on $(\mathbf{F}_0, \mathbf{a})$ and define

$$\mathbf{M} = \mathbb{E}[W(\mathbf{x})\mathbf{x}\mathbf{x}^\top \mid \mathbf{F}_0, \mathbf{a}]. \tag{127}$$

Coordinate expansion and independence of $\mathbf{x}, \mathbf{x}'$ give the exact identity

$$\mathbb{E}[W(\mathbf{x})W(\mathbf{x}')R(\mathbf{x}, \mathbf{x}')^2 \mid \mathbf{F}_0, \mathbf{a}] = \frac{\|\mathbf{M}\|_{\mathrm{F}}^2}{s^2}. \tag{128}$$

Thus the leading signed statistic is a squared matrix norm rather than a sum of absolute Gram entries; no factor m appears in its expectation.

To identify its structure, let $\mathbf{u} = \boldsymbol{\xi}/\sqrt{B}$, $\mathbf{v} = \mathbf{\Sigma}\mathbf{u}$, and $\mathbf{z} = \mathbf{\Sigma}\mathbf{v}$. Second-order Stein applied to the label part $W_y = \tanh(S)V$ gives

$$\mathbf{M}_y = A\mu\mathbf{\Sigma} + A\nu_3 \mathbf{v}\mathbf{v}^\top + \mu(\mathbf{v}\mathbf{z}^\top + \mathbf{z}\mathbf{v}^\top), \qquad \nu_3 = \mathbb{E}\tanh'''(G). \tag{129}$$

Since $\mathbb{E}W_y = A\mu$, centering removes the bulk $A\mu\mathbf{\Sigma}$ exactly. The remainder lies in $\operatorname{span}\{\mathbf{v}, \mathbf{z}\}$ and

$$\|\mathbf{v}\mathbf{v}^\top\|_{\mathrm{F}}^2 = A^2, \qquad \|\mathbf{v}\mathbf{z}^\top + \mathbf{z}\mathbf{v}^\top\|_{\mathrm{F}}^2 = 2AC + 2H^2 \leq 4AC. \tag{130}$$

After normalization by $A^2 s^2$, these terms vanish under the population primitives. The uncentered bulk contributes $(\mathbb{E}W/A)^2 \operatorname{tr}(\mathbf{\Sigma}^2)/s^2 = O_p(d_{\mathrm{eff}}^{-1})$.

K.2 Self-prediction matrix and residual moment

For a fixed row $\mathbf{f}$, write $\mathbf{p} = \mathbf{\Sigma}\mathbf{f}$, $r = \mathbf{f}^\top\mathbf{\Sigma}\mathbf{v}$, and $\tau^2 = \mathbf{f}^\top\mathbf{\Sigma}\mathbf{f}$. If $\mu_\tau = \mathbb{E}\operatorname{sech}^2(N(0, \tau^2))$ and $\nu_{3,\tau} = \mathbb{E}\tanh'''(N(0, \tau^2))$, Stein gives

$$\mathbf{M}_{\mathbf{f}} := \mathbb{E}[\tanh(\mathbf{f}^\top\mathbf{x})V\mathbf{x}\mathbf{x}^\top \mid \mathbf{f}] = r\mu_\tau\mathbf{\Sigma} + r\nu_{3,\tau}\mathbf{p}\mathbf{p}^\top + \mu_\tau(\mathbf{p}\mathbf{z}^\top + \mathbf{z}\mathbf{p}^\top). \tag{131}$$

The auxiliary predictor matrix is $\mathbf{M}_g = k^{-1/2}\sum_j a_j \mathbf{M}_{\mathbf{f}_j}$.

Lemma K.1 (Self-prediction matrix)**.** *There is a numerical $C_0 < \infty$ such that*

$$\mathbb{E}_{\mathbf{F}_0, \mathbf{a}}\|\mathbf{M}_g\|_{\mathrm{F}}^2 \leq C_0 \left\{ \frac{C\operatorname{tr}(\mathbf{\Sigma}^2)}{s} + \frac{C\{\operatorname{tr}(\mathbf{\Sigma}^2)\}^2}{s^3} \right\}. \tag{132}$$

Consequently $\|\mathbf{M}_g\|_{\mathrm{F}}^2/(A^2 s^2) \to_p 0$ under the primitive conditions. The same holds for its conditionally centered version.

Proof. Conditional on the rows, Gaussian averaging over $\mathbf{a}$ gives $\mathbb{E}_a\|\mathbf{M}_g\|_{\mathrm{F}}^2 = k^{-1}\sum_j \|\mathbf{M}_{\mathbf{f}_j}\|_{\mathrm{F}}^2$ with no cross-row terms. The three terms in (131) have squared norms bounded by constants times $r^2 t_2$, $r^2\|\mathbf{p}\|_2^4$, and $C\|\mathbf{p}\|_2^2$, where $t_2 = \mathrm{tr}(\mathbf{\Sigma}^2)$. Wick's formula gives $\mathbb{E}r^2 = C/s$, $\mathbb{E}\|\mathbf{p}\|_2^2 = t_2/s$, and $\mathbb{E}[r^2\|\mathbf{p}\|_2^4] \le C_0(C/s)(t_2/s)^2$, proving (132). Markov and the primitive ratios give the probability statement. Subtracting the conditional mean adds only $\bar{g}\mathbf{\Sigma}$, with $\mathbb{E}\bar{g}^2 \le C/s$, which obeys the same bound. □

Lemma K.2 (Conditional full-residual fourth moment)**.** *For the complete residual weight W,*

$$\mathbb{E}_{\mathbf{x}}[W(\mathbf{x})^4 \mid \mathbf{F}_0, \mathbf{a}] = O_p(H^2). \tag{133}$$

Proof. Conditional on $(\mathbf{F}_0, \mathbf{x})$, $g_{\mathbf{F}_0}(\mathbf{x})$ is centered Gaussian over $\mathbf{a}$ with variance $k^{-1}\sum_j \tanh^2(\mathbf{f}_j^\top \mathbf{x}) \le 1$. Hence, for $G_8 = \mathbb{E}_{\mathbf{x}}[|g_{\mathbf{F}_0}(\mathbf{x})|^8 \mid \mathbf{F}_0, \mathbf{a}]$, $\mathbb{P}(G_8 > L) \le 105/L$. Using $|u-v|^4 \le 8(|u|^4 + |v|^4)$, Cauchy–Schwarz, $\mathbb{E}V^4 = 3H^2$, and $\mathbb{E}V^8 = 105H^4$ gives

$$\mathbb{E}_{\mathbf{x}}[W^4] \le 24H^2 + 8\sqrt{105}\, H^2 G_8^{1/2}, \tag{134}$$

which proves the claim while retaining label, self-prediction, and all mixed terms. □

K.3 Hoeffding projections of the rank-two term

Define the conditional symmetric kernel

$$h(\mathbf{x}, \mathbf{x}') = \frac{W(\mathbf{x})W(\mathbf{x}')}{A^2} R(\mathbf{x}, \mathbf{x}')^2. \tag{135}$$

Its first Hoeffding projection is

$$h_1(\mathbf{x}) = \frac{W(\mathbf{x})}{A^2 s^2}\mathbf{x}^\top \mathbf{M}\mathbf{x} - \frac{\|\mathbf{M}\|_{\mathrm{F}}^2}{A^2 s^2}. \tag{136}$$

For $\mathbf{B}_M = \mathbf{\Sigma}^{1/2}\mathbf{M}\mathbf{\Sigma}^{1/2}$, Gaussian quadratic-form moments give

$$\{\mathbb{E}(\mathbf{x}^\top \mathbf{M}\mathbf{x})^4\}^{1/2} \le C_0\{(\mathrm{tr}\,\mathbf{B}_M)^2 + \|\mathbf{B}_M\|_{\mathrm{F}}^2\} \le C_0 t_2 \|\mathbf{M}\|_{\mathrm{F}}^2. \tag{137}$$

Together with the fourth-moment lemma and the critical relation $H/(mA^2) = O(1)$, this yields

$$\frac{1}{m}\mathbb{E}[h_1(\mathbf{x})^2 \mid \mathbf{F}_0, \mathbf{a}] = o_p(1). \tag{138}$$

For the degenerate projection, conditional Cauchy–Schwarz gives

$$\frac{1}{m^2}\mathbb{E}[h(\mathbf{x}, \mathbf{x}')^2 \mid \mathbf{F}_0, \mathbf{a}] \le \frac{\mathbb{E}W^4}{m^2 A^4}\{\mathbb{E}R(\mathbf{x}, \mathbf{x}')^8\}^{1/2}. \tag{139}$$

Conditioning on $\mathbf{x}$ and using Gaussian quadratic-form moments yields $\mathbb{E}R^8 \le C_0(t_2/s^2)^4$. Therefore

$$\frac{1}{m^2}\mathbb{E}[h(\mathbf{x}, \mathbf{x}')^2 \mid \mathbf{F}_0, \mathbf{a}] = O_p\left[\left(\frac{H}{mA^2}\right)^2 \left(\frac{t_2}{s^2}\right)^2\right] = o_p(1). \tag{140}$$

Equations (138)–(140) close both projections of the exact constant-coefficient rank-two kernel without an m/d_{eff} condition.

L Additional Numerical Validation

This appendix complements the three main experiments with distributional, finite-dimensional, activation, width, and stability checks. Its purpose is not to introduce additional headline claims, but to show where the asymptotic normalizations are accurate, where finite-size deviations remain, and how the theorem assumptions appear in finite dimensions. Figure 4 shows representative critical-law shapes; Figure 5 displays exact finite-n geometry and theorem margins; Figures 6 and 7 test activation and width robustness; the final diagnostics record self-prediction sensitivity, stability, and reproducibility limitations.

L.1 CRITICAL DISTRIBUTION DIAGNOSTICS

Figure 3 summarizes geometry-wide KS, variance, and mean diagnostics, whereas Figure 4 shows the full empirical distribution for two representative cells. We choose one clean interior cell and one higher-β finite-size stress cell. The reader should compare both the empirical CDF and the QQ plot with the critical-law reference: the first cell shows close agreement in shape, while the harder cell makes the residual variance inflation visible. Neither reference distribution is fitted to the observations.

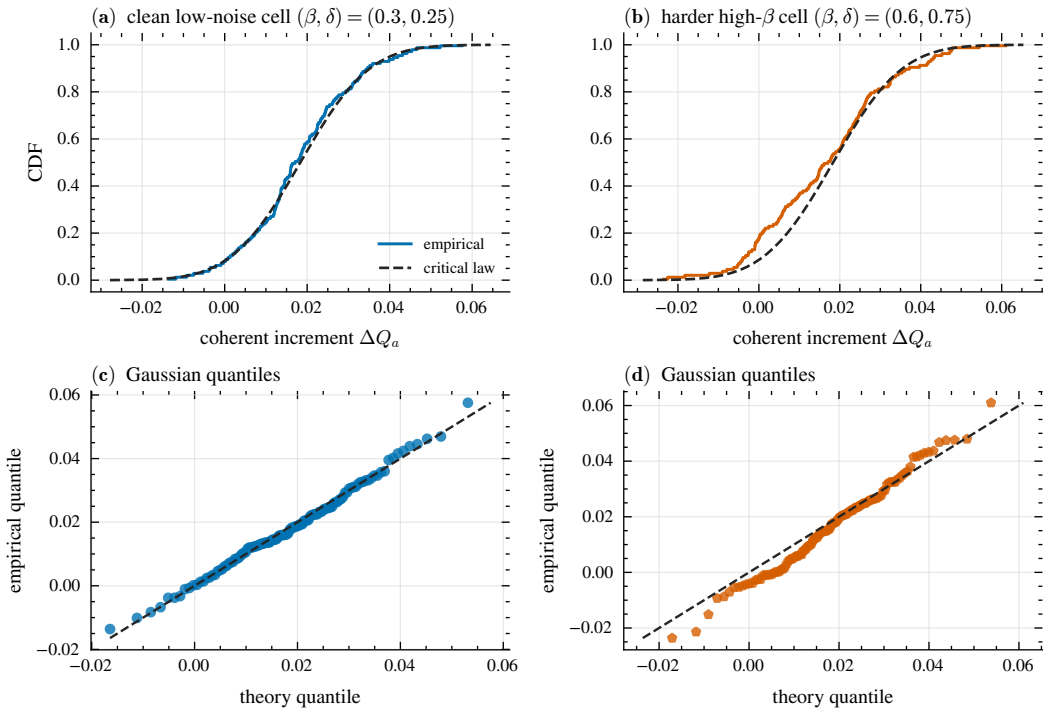


Figure 4: **Representative critical-law distribution diagnostics.** Empirical CDFs (top) and Gaussian QQ plots (bottom) for ΔQ_a in a clean $(\beta, \delta) = (0.3, 0.25)$ cell and a harder $(0.6, 0.75)$ cell, both at $n = 1600$ and $\chi_{\mathrm{FL}} \simeq 0.25$. Each panel uses 120 trials. Dashed curves and quantiles use finite-n parameters in the critical-law formula, not fitted Gaussians. The first cell is a clean theorem-interior check; the second is a higher-β finite-size stress check whose variance inflation is also summarized in Figure 3.

L.2 FINITE-DIMENSIONAL GEOMETRY

The four deterministic maps in Figure 5 explain why asymptotic exponent labels do not fully determine finite-n accuracy. Panel (a) shows the population amplification A: larger A lowers the raw onset $\sqrt{k}/A$. Panel (b) shows the finite-batch resolution difficulty H/A^2: larger values require larger m_{FL} to keep $\chi_{\mathrm{FL}} = H/(m_{\mathrm{FL}}A^2)$ small. Panel (c) shows the primitive self ratio $C/(sA^2)$, a self-prediction and theorem-margin diagnostic rather than a third headline learning scale. Panel (d) shows the effective dimension d_{eff} controlling sample-norm fluctuations in the critical theorem. Together they show that two geometries with similar asymptotic exponent labels can have substantially different finite-n theorem margins.

The overlaid curves are analytic, not fitted. The lines $\delta = \beta/2$ and $\delta = 3\beta/2$ delimit the nontrivial critical resolution window; $\delta = 1/2$ and $\delta = 2\beta - 1/2$ enter the population self-prediction admissibility condition above the spectral crossover; and $\beta = 1/2$ is the crossover landmark. Their intersections determine the population and critical admissibility regions stated in Appendix F; none is an empirical phase boundary.

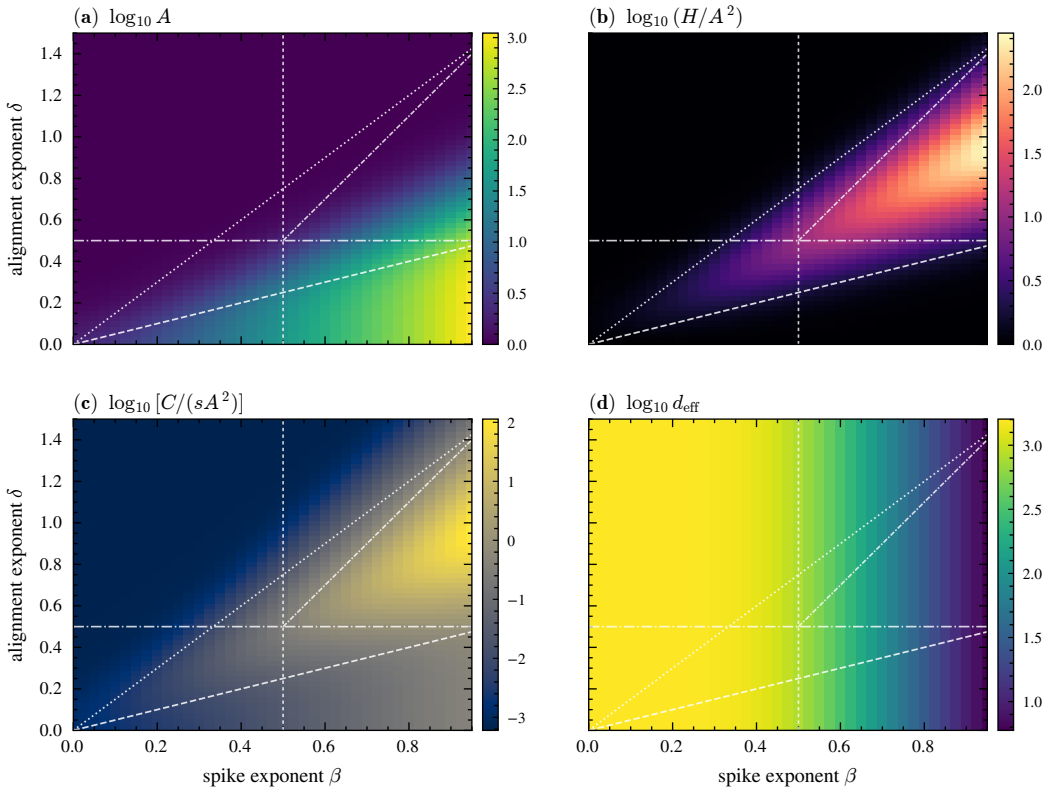


Figure 5: **Exact finite-dimensional geometry at $n = 1600$.** With $c_\theta = c_\alpha = 1$ and stored width ratio $k/n = 0.5$, the panels map (a) population amplification $\log_{10} A$, which lowers the raw onset $\sqrt{k}/A$; (b) finite-batch difficulty $\log_{10}(H/A^2)$, which enters χ_{FL}; (c) the self-prediction margin $\log_{10}[C/(sA^2)]$; and (d) $\log_{10} d_{\text{eff}}$, which controls sample-norm fluctuations. Panels (c)–(d) diagnose theorem validity, not additional headline signal scales. Overlaid analytic lines mark the population/critical admissibility boundaries and the $\beta = 1/2$ spectral crossover described in the text. These are deterministic maps with no stochastic error bars and no empirically fitted phase boundaries.

L.3 Activation robustness

Figure 6 tests whether activation-specific coefficients remove the leading non-geometric variation when the relevant first channels are nonzero. Throughout the figure, pairs are ordered as teacher/network activations. The population panel divides each update energy by its activation-specific population coefficient. The critical panel instead compares Q_a and Q_2 with their activation-specific critical-law means and therefore includes only network activations satisfying C–P. A ReLU teacher with a tanh network is allowed in that panel; a tanh teacher with a ReLU network is population-only, and no ReLU-network critical claim is made. The population and critical panels use different dimensions and campaigns, so they are separate normalization checks rather than a convergence comparison.

L.4 Width robustness

The critical theorem assumes $k/n \to \phi \in (0, \infty)$. Figure 7 therefore checks the critical normalization across the substantial finite range $k/n \in \{0.25, 0.5, 1, 2, 4\}$ while holding the geometry, target χ_{FL}, and all non-width parameters fixed. The 40 trials reuse paired randomness across widths. Both observables remain on the predicted order, but the broad Q_2 intervals support robustness of the scale only, not exact width invariance.

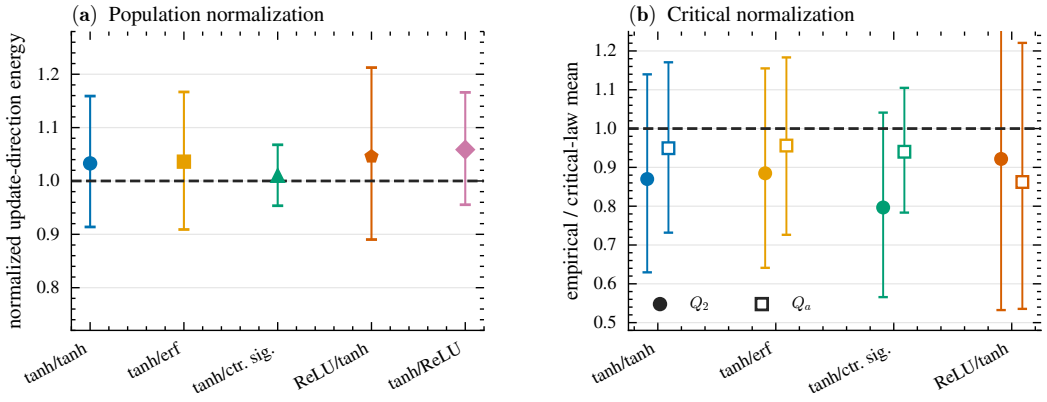


Figure 6: **Activation-normalized checks.** Population-like update energy divided by its activation-specific coefficient at $n = 1600$ (30 trials per pair), and critical Q_2, Q_a increments divided by their activation-specific means at $n = 800$ (40 paired trials). Bars are bootstrap 95% intervals. Pairs are teacher/network activations. The critical panel contains only tanh, erf, or centered-sigmoid networks satisfying C–P; ReLU/tanh is included because ReLU is the teacher, whereas tanh/ReLU is population-only. The panels use different dimensions and campaigns and are not a convergence comparison.

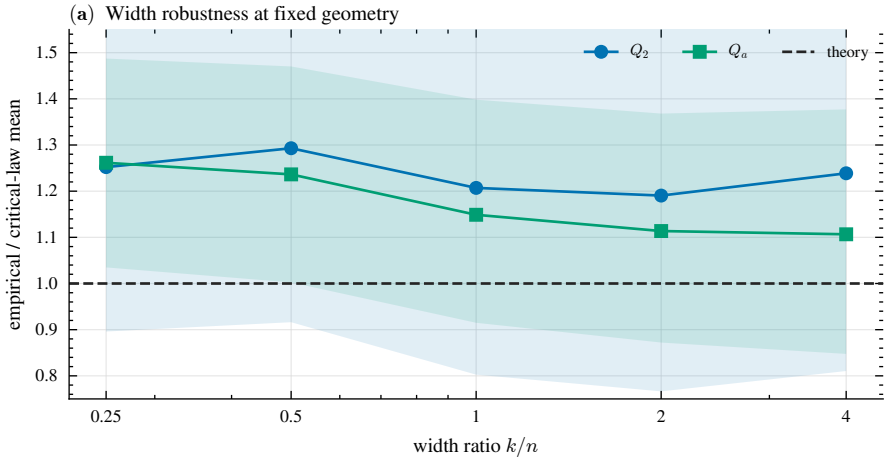


Figure 7: **Width robustness at fixed geometry.** At $(\beta, \delta) = (0.4, 0.4)$, $n = 800$, and realized $\chi_{\mathrm{FL}} = 0.257$, the width ratio ranges over $k/n \in \{0.25, 0.5, 1, 2, 4\}$. The plotted Q_a and Q_2 increments are divided by their respective critical-law means. Bands are paired-bootstrap 95% intervals from 40 trials whose randomness is shared across widths; the dashed line marks exact agreement at one. The broad Q_2 uncertainty supports robustness of order and scaling, not exact width invariance.

L.5 Additional theorem-margin and stability diagnostics

The stored prefactor-sensitivity campaign gives the most direct finite-dimensional limitation. Across two exponent cells at $n = 800$, large normalized population discrepancies appear as the exact self ratio $C/(sA^2)$ becomes order one, although the 20-trial cell-mean error is not monotone over all nine prefactors. This is a diagnostic of theorem margins, not a fitted admissibility boundary; the raw records are retained.

At the noise-rescaled step $\eta_{\mathrm{noise}} = d\sqrt{km_{\mathrm{FL}}/H}$, selected cells show order-one overlap energy while the coherent increment has standard deviation comparable to or larger than its mean. Because the

available batches contain only two to eight samples, we use this as qualitative evidence for unreliable coherent sign rather than precise coefficient matching.

Targeted full-matrix runs at $c = 0.05$ remain numerically controlled. Mean row updates range from 0.013 to 0.058 in the intermediate rank-one model and from 0.024 to 0.079 in a rank-four model. Mean post-update preactivation variances remain in $[1.002, 1.031]$ and $[1.001, 1.020]$, respectively. Fresh independent ridge readouts and nonlinear Hermite-channel diagnostics show no reproducible gain, consistent with the paper's representation-level scope rather than a downstream risk claim.

L.6 Reproducibility

The validated campaigns retain primary raw records for every reported point. Each record stores the seed, physical covariance parameters, $A, H, C, d_{\mathrm{eff}}$, raw η, normalized c, realized χ_{FL}, $m_{\mathrm{FL}}/d_{\mathrm{eff}}$, width, dimension, and trial identifier. Within each cell, trials are independent; factorial and robustness comparisons explicitly reuse randomness across paired arms.