SAME SPECTRAL BULK, DIFFERENT STABLE STEP SIZES IN WIDE NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

A common approach to understanding neural network training is to describe directional sensitivities through eigenvalue distributions and use this description to assess stable learning rates. This seems natural: the distribution records how many directions are relatively flat and how many are sensitive. We show that even this overall description can be insufficient to determine local training stability. We construct wide neural networks with the same limiting eigenvalue distribution and all fixed-order normalized spectral moments, yet the same learning rate is stable for one and unstable for the other. Stability can be constrained by a single exceptional direction whose contribution to the distribution vanishes as width grows. Our analysis shows that matching aggregate statistics does not match the critical directions: the network computation can selectively amplify a few directions that determine stability while leaving the limiting distribution unchanged. We characterize when these directions emerge and when they restrict the learning rate in Gaussian networks with two nonlinear layers, for a single input near zero training error with all weights trained. We further give explicit conditions under which a trainable output layer preserves the difference or averages it away. Experiments confirm that networks with nearly identical spectral bulk can admit different stable learning rates, and that a few forward-pass statistics capturing the exceptional directions improve prediction of this difference. These findings can help guide neural network training by identifying the information needed beyond aggregate spectral descriptions.

1 INTRODUCTION

Understanding how architecture and initialization affect training is a central goal of neural network theory. Wide-network limits give tractable accounts of signal and gradient propagation (Jacot et al., 2018; Yang, 2020a). Spectral descriptions summarize sensitivity across directions, suggesting a way to assess which learning rates permit stable training.

Experiments show why the choice of summary matters. On CIFAR-10, Pennington et al. (2017) find large differences in learning speed among deep networks initialized to preserve the same average gradient scale. Faster learning is associated with a more uniform response across input directions. In networks without batch normalization, Ghorbani et al. (2019) observe slow training together with gradients concentrated in a few directions along which the loss changes most sharply. These observations highlight an important distinction between typical behavior and the directions that can impede learning.

A natural response is to consider the entire eigenvalue distribution. For wide networks, its limit records the proportions of relatively flat and sensitive directions (Fan & Wang, 2020; Yang, 2020b). Yet a direction can have vanishing weight in this distribution and still determine stability. Near a zero-error solution of squared loss, a stable gradient step must damp small perturbations in every direction with positive curvature. One unusually sensitive direction can therefore restrict the learning rate. We ask:

> *Can networks with the same limiting spectral distribution have different stable learning rates? If so, what additional information is needed to predict the difference?*

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.

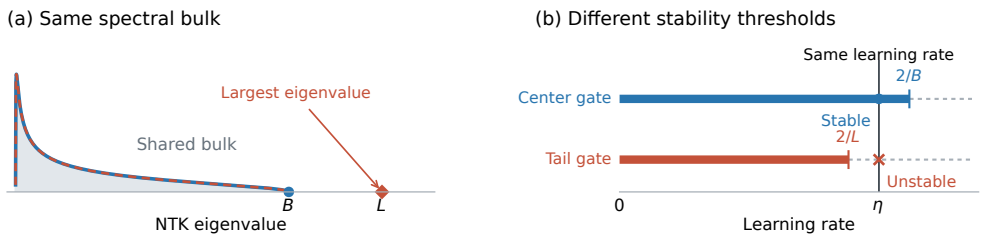


Figure 1: The same spectral bulk can conceal a different training threshold. Predictions for the neural tangent kernel (NTK) of two nonlinear Gaussian layers. (a) Center and tail gates share the limiting density but have different largest eigenvalues $B < L$. The zero atom is omitted; the outlier has vanishing spectral mass. (b) Solid segments show local stable ranges at zero training error. A common rate $\eta \in (2/L, 2/B)$ is stable for the center gate and unstable for the tail gate, with both weights trained. Parameters match Table 1.

In a simple Gaussian network example, the same learning rate damps perturbations in one model and amplifies a perturbation in the other, with each Jacobian held fixed. Interestingly, the difference persists as the networks become arbitrarily wide, despite agreement in their limiting eigenvalue distributions and all fixed-order spectral moments. Matching aggregate statistics does not match the critical directions. Selecting the same fraction of neurons can retain different amounts of input signal and selectively amplify a direction. The missing information comes from the dependence between the weights and the neuron responses they produce. Figure 1 illustrates the resulting separation in stability.

Our decomposition separates the directions involved in the network computation from the remaining Gaussian variation. For fixed depth and a fixed number of inputs, we determine the spectral boundary of the resulting reference model. Eigenvalues beyond this boundary are outliers. Classical perturbation theory describes their effect when the correction is given (Benaych-Georges & Nadakuditi, 2011). We derive this correction from the network computation. For two nonlinear Gaussian layers, we determine when an eigenvalue separates from the bulk, compute its limiting location, and quantify its eigenvector overlap with the retained preactivation vector. The same dependence thus explains the separation and makes it predictable.

The neural tangent kernel (NTK) describes how weight updates change the outputs locally. We calculate it to determine how these directions affect parameter training. In a single-input problem near zero training error, we obtain different local critical learning rates for networks with the same limiting NTK spectrum and all weights trained. A trainable Gaussian output layer can preserve this difference or average it away. We give an explicit condition relating the exceptional direction's strength to the output-to-hidden width ratio. Nonlinear training experiments test these boundaries. Forward-pass statistics improve their prediction at finite width. These results connect aggregate spectra and exceptional directions to local training stability.

2 Related Work

Wide-network limits make aspects of neural network training and signal propagation amenable to analysis. At fixed sample size, conjugate kernels and NTKs converge to deterministic kernels (Jacot et al., 2018; Yang, 2020a). Tensor Programs give laws for normalized Jacobian traces, including corrections from reused transposed weights (Yang, 2020b). Fan & Wang (2020) derive spectral laws for multilayer kernels and NTKs at proportional sample size and width. Deterministic equivalents give related sample-covariance descriptions (Hachem et al., 2007; Silverstein & Choi, 1995). We study the directional information needed to recover exceptional eigenvalues and local stability.

Outliers in nonlinear kernels have been studied by Wang et al. (2024); Cranston et al. (2026). In particular, Wang et al. (2024) characterize how input spikes propagate through nonlinear random layers and analyze a trained rank-one signal in a representation-learning model. Here conditioning on the network computation generates the perturbation. For two nonlinear layers, we evaluate its

directional resolvents, including a perturbation direction that depends on the Gaussian reference matrix, and compute the resulting parameter-NTK threshold.

Strong convergence controls operator norms and spectral support in addition to traces (Haagerup & Thorbjørnsen, 2005; Chen et al., 2026). We apply the operator-valued Gaussian theorem of Jekel et al. (2026) to the conditional reference model. Finite-rank perturbation theory then describes eigenvalues outside this support (Benaych-Georges & Nadakuditi, 2011; Belinschi & Shlyakhtenko, 2012). The concrete calculations use the spectral law of Gaussian products (Burda et al., 2011) and the generalized sample-covariance spike map (Bai & Yao, 2008). Conditioning supplies the strengths and directions required by these results.

Work on dynamical isometry relates input-Jacobian spectra to signal propagation and conditioning (Pennington et al., 2017; Collins & Hayase, 2021; Dadoun et al., 2025). The empirical NTK factorization of Hazelden et al. (2026) separates a parameter-space kernel from its map into output coordinates. We study the role of relative eigenspaces and compute the final NTK outlier for a specified trainable Gaussian output layer.

3 Theoretical analysis

Our explicit training comparisons use one input and squared loss $\frac{1}{2}\|f-y\|^2$, with the target equal to the initial network output. Let $\mathcal{J}$ be the Jacobian with respect to the trainable parameters and $\Theta = \mathcal{J}\mathcal{J}^*$. Linearized gradient descent contracts every positive-curvature direction exactly when $0 < \eta < 2/\lambda_1(\Theta)$; zero-curvature directions have multiplier one. We ask whether equal limiting spectral laws can give different critical steps. We first isolate the mechanism in one layer, then compute the parameter NTK for two nonlinear layers and a trainable output layer.

We use n for width. For self-adjoint $A \in \mathbb{C}^{d\times d}$, order its eigenvalues as $\lambda_1(A) \geq \cdots \geq \lambda_d(A)$, write $\operatorname{tr}_d A = d^{-1}\operatorname{Tr} A$, and let $\mu_A = d^{-1}\sum_j \delta_{\lambda_j(A)}$. Other notation is introduced as needed.

3.1 Equal spectral bulk does not imply equal stability

A derivative gate determines which neurons contribute to a perturbation of the output. We compare gates that retain the same fraction of neurons but select different preactivation values. This holds the limiting eigenvalue distribution fixed while allowing the network to respond differently along the input direction. A single Gaussian layer suffices to make the distinction explicit.

Let $W_n \in \mathbb{R}^{m_n\times n}$ have i.i.d. $\mathcal{N}(0,n^{-1})$ entries, let $m_n/n \to \alpha \in (0,\infty)$, and fix h_n with $\|h_n\|^2 = n$. Set $z_n = W_n h_n$. For a Borel set $A \subset \mathbb{R}$ satisfying $\mathbb{P}(Z\in A) > 0$ and $\mathbb{P}(Z \in \partial A) = 0$ for $Z \sim \mathcal{N}(0,1)$, define

$$\begin{aligned} D_{A,n} &= \operatorname{diag}\big(\mathbf{1}_{\{z_{n,i}\in A\}}\big), & K_{A,n} &= D_{A,n}W_nW_n^*D_{A,n}, \\ p_A &= \mathbb{P}(Z\in A), & q_A &= \mathbb{E}[Z^2\mathbf{1}_{\{Z\in A\}}], \\ y_A &= \alpha p_A, & \theta_A &= \alpha q_A. \end{aligned} \tag{1}$$

The gate activation probability p_A is the probability of a nonzero activation derivative; q_A measures the retained preactivation energy. To see how they enter the spectrum, rotate $h_n/\sqrt{n}$ to e_1 and write $W_n = n^{-1/2}[z_n\ X_n]$. The remaining columns $X_n \in \mathbb{R}^{m_n\times(n-1)}$ form a standard Gaussian matrix independent of z_n. Hence

$$\begin{aligned} B_{A,n} &:= n^{-1/2}[0\ D_{A,n}X_n], & s_{A,n} &:= n^{-1/2}D_{A,n}z_n, \\ D_{A,n}W_n &= B_{A,n} + s_{A,n}e_1^*, & K_{A,n} &= B_{A,n}B_{A,n}^* + s_{A,n}s_{A,n}^*. \end{aligned}$$

Conditional on the gate, the active block of $B_{A,n}B_{A,n}^*$ is Wishart. Its aspect ratio tends to $y_A = \alpha p_A$, while $\|s_{A,n}\|^2 \to \theta_A = \alpha q_A$. The spectrum is therefore a Marchenko–Pastur bulk with a rank-one correction whose strength is determined by the retained energy. We write $\mu_{\mathrm{MP},y}$ for the Marchenko–Pastur law with row-to-column aspect ratio y.

Theorem 3.1. *Under Equation (1), almost surely,*

$$\mu_{K_{A,n}} \Longrightarrow (1-p_A)\delta_0 + p_A\mu_{\mathrm{MP},y_A}, \tag{2}$$

and the largest eigenvalue satisfies

$$\lambda_{\max}(K_{A,n}) \longrightarrow \begin{cases} (1+\sqrt{y_A})^2, & \theta_A \le y_A + \sqrt{y_A}, \\ \dfrac{\theta_A(\theta_A+1-y_A)}{\theta_A - y_A}, & \theta_A > y_A + \sqrt{y_A}. \end{cases} \tag{3}$$

In the supercritical regime, if $v_{1,n}$ is a top eigenvector and $u_{A,n} := D_{A,n}z_n/\|D_{A,n}z_n\|$, then

$$|\langle v_{1,n}, u_{A,n}\rangle|^2 \longrightarrow 1 - \frac{y_A}{(\theta_A - y_A)^2}.$$

Gates with the same p_A therefore have the same limiting empirical spectral distribution and all fixed normalized moments. They can still have different largest eigenvalues, leading-eigenvector overlaps with $u_{A,n}$, and stable step-size ranges under gradient descent on the linearized model.

The proof is given in Appendix B. The theorem separates the number of selected neurons, which determines the bulk through p_A, from their energy, which determines the outlier through q_A. For frozen or linearized squared loss, the residual evolves as $r_{t+1} = (I - \eta K)r_t$. Every positive mode contracts exactly when $0 < \eta < 2/\lambda_{\max}(K)$. Thus, if two gates have top limits $L_1 > L_2$, any $\eta \in (2/L_1, 2/L_2)$ eventually contracts every positive mode of the second model and expands the top mode of the first, despite their equal bulk laws.

This difference also occurs in parameter training. For $\phi'_A = \mathbf{1}_A$, $D_A W$ is the input Jacobian of $\phi_A(Wh)$, whereas training W gives the NTK $\|h\|^2 D_A$. Instead, train U in $f_U(x) = \phi_A(W_n U x)$ with W_n frozen, $\|x\| = 1$, and deterministic $U_0 x = h_n$. Then

$$\mathrm{d}f_{U_0}[\Delta U] = D_{A,n} W_n \Delta U x, \qquad \Theta^{\mathrm{out}}_{U_0} = K_{A,n}.$$

For loss $\frac{1}{2}\|f_U(x) - y\|^2$ at $y = f_{U_0}(x)$, Theorem 3.1 therefore determines local stability. Training both weights in $\phi_A(GUx/\sqrt{n})$, with $G_0 = \sqrt{n}W_n$, gives $K_{A,n} + D_{A,n}$: every active eigenvalue shifts by one and the eigenvectors are preserved. Appendix C proves these identities.

The energy condition also explains when this dependence has no effect on the upper edge. For ReLU, $A = (0, \infty)$ and Gaussian symmetry gives $p_A = q_A = 1/2$. Hence $\theta_A = y_A < y_A + \sqrt{y_A}$, so the top eigenvalue remains at the bulk edge. An upper outlier requires $q_A > p_A + \sqrt{p_A/\alpha}$: the gate must retain enough energy, relative to the number of selected neurons, to create a separated mode.

3.2 Conditional Gaussian decomposition

Observing Wh reveals the weight matrix along one direction and leaves Gaussian randomness in its orthogonal complement. Joint conditioning extends this argument to the adaptive matrix–vector queries in a multilayer computation. The resulting reference weights are independent of the query history, with an error confined to the queried subspaces. Proposition D.1 bounds the resulting rank error in polynomial and rational expressions.

For fixed-depth Gaussian networks evaluated on a fixed number of inputs, we establish strong convergence of the reference weights together with the original derivative gates (Theorem D.2). The assumptions include proportional widths, nondegenerate limiting preactivation covariances, and ReLU or smooth Lipschitz activations with bounded derivatives. Interlacing then locates the original spectrum below the number of leading eigenvalues allowed by the rank bound (corollary D.3).

The leading eigenvalues depend on the projections of the correction directions onto the reference eigenspaces. For $K_n = H_n + U_n C_n U_n^*$ with a fixed-rank correction, these are encoded by $U_n^*(zI - H_n)^{-1}U_n$. Proposition D.4 converts its limit into a finite-dimensional outlier equation. The complete framework appears in Appendix D. We next evaluate these directional limits for two nonlinear Gaussian layers, obtaining predictions directly from the network parameters.

3.3 Outliers in two-layer Gaussian networks

We now compute the directional limits, outlier, and alignment for a ReLU first layer followed by the gated activation. Consider one deterministic input h_0 with $\|h_0\|^2 = n$. Let $G_1 \in \mathbb{R}^{m \times n}$

and $G_2 \in \mathbb{R}^{\ell \times m}$ have independent standard Gaussian entries, independently across widths, with $m/n \to \alpha > 0$ and $\ell/m \to \beta > 0$. Set

$$W_1 = \frac{G_1}{\sqrt{n}}, \quad W_2 = \sqrt{\frac{2}{m}} G_2, \quad z_1 = W_1 h_0, \quad h_1 = (z_1)_+, \quad z_2 = W_2 h_1, \quad f = \phi_A(z_2). \tag{4}$$

The second activation is $\phi_A(t) = \int_0^t \mathbf{1}_A(u)\,\mathrm{d}u$, where A is a finite union of intervals and $p = \mathbb{P}(Z \in A) \in (0,1)$ for $Z \sim \mathcal{N}(0,1)$. Define

$$D_1 = \mathrm{diag}(\mathbf{1}_{z_1>0}), \quad D_2 = \mathrm{diag}(\mathbf{1}_{z_2 \in A}), \quad q = \mathbb{E}[Z^2 \mathbf{1}_A(Z)],$$
$$c = \alpha/2, \quad d = \alpha\beta p, \quad \theta = 2\beta q.$$

The ratios c, d count active hidden and output neurons, respectively, relative to the input width. Here $J = D_2 W_2 D_1 W_1$ is the input Jacobian. Weight-update stability is governed by the parameter NTK Θ derived below, using all entries of G_1, G_2 as Euclidean parameters and no averaging of the loss over output coordinates.

As in the one-layer example, we rotate along the vector used to query a weight matrix. Here that vector is the hidden activation. On the active output rows, this gives $J = Z_n + a_n b_n^*$ with a Gaussian matrix product Z_n. The vector a_n is parallel to the gated second preactivation. The vector b_n contains the hidden activation norm and an independent Gaussian vector; Equation (73) gives their exact construction. Multiplying by the adjoint gives, with $H_n = Z_n Z_n^*$,

$$JJ^* = H_n + U_n C_n U_n^*, \quad U_n = [a_n\ Z_n b_n], \quad C_n = \begin{bmatrix} \|b_n\|^2 & 1 \\ 1 & 0 \end{bmatrix} \longrightarrow \begin{bmatrix} c+1 & 1 \\ 1 & 0 \end{bmatrix}.$$

Write $\mu_{c,d}$ for the limiting spectral law of H_n on these active rows. Lemma H.1 gives, almost surely for every polynomial P,

$$U_n^* P(H_n) U_n \longrightarrow \Omega(P) = \mathrm{diag}\left(\theta \int P(x)\,\mathrm{d}\mu_{c,d}(x), \quad d \int x P(x)\,\mathrm{d}\mu_{c,d}(x)\right). \tag{5}$$

Equation (5) shows why the directions matter. The first direction samples the reference eigenspaces according to their spectral mass. The second places more weight on larger eigenvalues, as expressed by the factor x in the integral. This bias follows from $\mathrm{Cov}(Z_n b_n \mid a_n, Z_n) = H_n/n$. Treating both directions as independent of the reference would lose this weighting and change the outlier equation.

To locate the outlier, let $m(z) = \int (z-x)^{-1}\,\mathrm{d}\mu_{c,d}(x)$ be the reference resolvent trace. Parameterize it by $\Lambda(w) = (1+w)(c+w)(d+w)/(cw)$. Let $w_* > 0$ minimize this function and set $B = \Lambda(w_*)$, the reference upper edge. For $z > B$, the compressed resolvent tends to $\mathrm{diag}(\theta m(z), w(z))$, where $z = \Lambda(w)$ on $0 < w < w_*$ and $m(z) = (d+w)/(dz)$. The determinant in Equation (52) reduces to

$$(q/p - 1)w(c + 1 + w) = c. \tag{6}$$

A root with $0 < w < w_*$ gives an eigenvalue above B. It reaches the edge when $w = w_*$. Thus the ratio q/p, the average squared preactivation among selected neurons, again controls the transition.

Theorem 3.2. *For Equation (4), the empirical law of $K = JJ^*$, normalized over all ℓ output coordinates, converges almost surely to $(1-p)\delta_0 + p\mu_{c,d}$. Every fixed normalized moment converges to the corresponding moment of this law. Define*

$$T_* := 1 + \frac{c}{w_*(c + 1 + w_*)}, \qquad T := q/p. \tag{7}$$

If $T \le T_*$, then $\lambda_1(K) \to B$. If $T > T_*$, set

$$w_o = \frac{\sqrt{(c+1)^2 + 4c/(T-1)} - (c+1)}{2}, \qquad L = \Lambda(w_o). \tag{8}$$

Then $0 < w_o < w_*$, $\lambda_1(K) \to L > B$, and $\lambda_2(K) \to B$. For the unit vector along $D_2 z_2$, $\widehat{a}_n = D_2 z_2 / \|D_2 z_2\|$ and a unit top eigenvector v_1, in the supercritical regime

$$|\langle v_1, \widehat{a}_n \rangle|^2 \longrightarrow \mathcal{O} = \frac{c + 1 + w_o}{c + 1 + 2w_o} \left[1 - \frac{w_o^2(1 + c + d + 2w_o)}{cd}\right]. \tag{9}$$

The overlap tends to zero in the subcritical and critical regimes. Gates with the same p have identical limiting bulk laws and fixed normalized moments, but can have different L and critical step sizes for the linearized dynamics $2/L$.

Adding the NTK contributions from G_1 and G_2 gives the corresponding parameter-training threshold.

Theorem 3.3. *With all entries of G_1, G_2 trained, the output NTK at initialization is*

$$\Theta = D_2 W_2 D_1 W_2^* D_2 + v_n D_2, \qquad v_n = 2\|h_1\|^2/m. \tag{10}$$

Set $y_A = 2\beta p$ and $\theta = 2\beta q$. Its limiting empirical law is $(1-p)\delta_0 + p(T_1)_{\#}\mu_{\mathrm{MP},y_A}$, where $T_1(x) = 1 + x$, and all fixed normalized moments converge. The largest eigenvalue satisfies

$$\lambda_1(\Theta) \longrightarrow 1 + \begin{cases} (1+\sqrt{y_A})^2, & \theta \le y_A + \sqrt{y_A}, \\ \dfrac{\theta(\theta+1-y_A)}{\theta - y_A}, & \theta > y_A + \sqrt{y_A}. \end{cases} \tag{11}$$

In the second case, the top eigenvector's squared overlap with $D_2 z_2/\|D_2 z_2\|$ tends to $1 - y_A/(\theta - y_A)^2$. At the interpolating target $y = f(G_{1,0}, G_{2,0})$, the local linear-stability threshold is $2/\lambda_1(\Theta)$. Thus equal-p gates can give different local stability thresholds when training all parameters in the same network with two nonlinear layers.

The G_2 contribution shifts the active eigenvalues by $v_n \to 1$, while the G_1 contribution is Wishart with a preactivation-dependent rank-one correction. The parameter NTK and input Jacobian therefore have different transitions. Table 1 reports this comparison, and Appendix H gives the proofs.

3.4 Outlier propagation through the output layer

A linear output layer mixes the exceptional direction with other hidden modes. We determine when it preserves the spectral separation while all weights are trained. Consider an independent standard Gaussian matrix $G_3 \in \mathbb{R}^{o\times\ell}$ and use the output

$$F_A(h_0) = R\psi_A(z_2), \qquad R = G_3/\sqrt{\ell} \in \mathbb{R}^{o\times\ell}, \tag{12}$$

All entries of G_1, G_2, G_3 are trained with Euclidean parameter metrics, a common step size, and loss $\frac{1}{2}\|F_A - y\|^2$. The input and local interpolation setting are as above. Write $\kappa_A = \mathbb{E}[\psi_A(Z)^2]$, with $\psi_A' = \mathbf{1}_A$. Training the output layer adds this second moment as a scalar shift of the limiting NTK. We match it to isolate the hidden outlier at a common output bulk law. For the center and tail primitives, which have zero Gaussian mean, the fixed shift

$$\psi_A = \phi_A + \sqrt{\kappa - e_A}, \qquad e_A = \mathbb{E}[\phi_A(Z)^2], \qquad \kappa = \max(e_c, e_t)$$

matches $\kappa_A = \kappa$ without changing the derivatives, hence preserving the gates and their p, q values.

Let $y_A = 2\beta p \in (0,1)$, let L_A be the hidden NTK top limit in Equation (11), and write

$$\tau = (1-p)\delta_0 + p(T_1)_{\#}\mu_{\mathrm{MP},y_A}, \qquad b_\tau = 1 + (1+\sqrt{y_A})^2, \qquad T_1(x) = 1 + x.$$

Two scalar functions describe how the output layer transforms this spectrum. For $a > b_\tau$, define

$$Q(a) = \int \frac{t^2}{(a-t)^2}\,\mathrm{d}\tau(t), \qquad \Psi_\gamma(a) = a\left[\gamma + \int \frac{t}{a-t}\,\mathrm{d}\tau(t)\right].$$

Here γ is the limiting output-to-hidden width ratio. The function Q decreases from infinity to zero, so $Q(a_\gamma) = \gamma$ defines a unique $a_\gamma > b_\tau$. This value is the separation threshold for hidden eigenvalues. Above it, Ψ_γ maps a hidden eigenvalue to its output location, before the scalar shift from training G_3.

Theorem 3.4. *In Equation (12), the exact NTK is*

$$\Theta^{\mathrm{out}} = R\Theta R^* + \kappa_{A,n} I_o, \qquad \kappa_{A,n} = \|\psi_A(z_2)\|^2/\ell \longrightarrow \kappa_A. \tag{13}$$

Suppose $o/\ell \to \gamma \in (0,\infty)$. The bulk upper edge is $B_\gamma = \kappa_A + \Psi_\gamma(a_\gamma)$, and almost surely

$$\lambda_1(\Theta^{\mathrm{out}}) \longrightarrow \begin{cases} \kappa_A + \Psi_\gamma(L_A), & L_A > a_\gamma, \\ B_\gamma, & L_A \le a_\gamma. \end{cases} \tag{14}$$

Thus a separated hidden outlier survives exactly when

$$\gamma > \gamma_{\mathrm{crit}} := Q(L_A), \qquad L_A > b_\tau. \tag{15}$$

Let u_A be the hidden top eigenvector and v_A the output top eigenvector. In the surviving case,

$$\frac{|v_A^* R u_A|^2}{\|R u_A\|^2} \longrightarrow 1 - \frac{Q(L_A)}{\gamma}. \tag{16}$$

The overlap tends to zero in the suppressed case, including equality. Gates with the same p and κ_A have identical limiting output empirical laws and all fixed normalized moments, even when their top limits differ. At interpolation, their local linear-stability thresholds are $2/\lambda_1(\Theta^{\mathrm{out}})$.

The retained energy fixes L_A, and the output width ratio decides whether this mode determines the local critical step. The transformation Ψ_γ is the established generalized covariance spike map (Bai & Yao, 2008). Here its population law and spike strength are supplied by the hidden network. Appendix I gives the proof.

When the output dimension is small relative to the hidden width, Gaussian averaging removes the directional difference. If $o/\ell \to 0$, including fixed output dimension, proposition I.1 gives

$$\left\|\Theta^{\mathrm{out}} - (2p + \kappa_A) I_o\right\| \longrightarrow 0 \quad \text{almost surely.}$$

The matched networks then have the same limiting local linear-stability threshold: averaging over hidden directions removes the spectral difference. Appendix J.1 examines the role of relative eigenspaces in more general factored output kernels.

4 Experiments

We test whether retained preactivation energy creates an outlier at fixed spectral bulk and changes the local training threshold after the output map. Auxiliary spectral tests appear in Appendix M. Protocols and per-seed results appear in Appendices C.5, H.8, I.2 and L.

4.1 Two nonlinear layers

The model in Section 3.3 uses $m = 2n$ and $\ell = 8n$. We use fixed center and tail thresholds satisfying $\mathbb{P}(|Z| < a) = \mathbb{P}(|Z| > b) = p = 0.1$ for $Z \sim \mathcal{N}(0, 1)$. Their limiting active fractions agree, but finite-width counts may differ. Each width $n \in \{256, 512, 1024\}$ has eight independent Gaussian draws, with shared weights for the two gates within each draw. The spectral predictions use the model parameters.

Figure 2 shows the predicted transition: increasing retained energy separates the largest eigenvalue from the reference edge and increases its eigenvector alignment with the gated preactivation. The input Jacobian and parameter NTK undergo this transition at different gate energies, showing why their stability predictions must be distinguished. The directional-resolvent measurements in Table 7 support the calculation: the diagonal entries approach their distinct predicted limits, while the off-diagonal entry approaches zero.

Table 1: Predicted NTK top limits L_c, L_t, local critical steps η_c^*, η_t^*, and finite-width stability for center (c) and tail (t) gates. At the common step $\eta = (\eta_c^* + \eta_t^*)/2$, the last two columns count draws satisfying $\eta\lambda_1 < 2$ among 24 networks per gate. The network has $m/n = 2$, $\ell/m = 4$, and $p = 0.1$; the linear output uses $o/\ell = 0.5$ and matched activation second moments. All weights are trained for loss $\frac{1}{2}\|f - y\|^2$.

Network	L_c	L_t	η_c^*	η_t^*	NTK-stable draws Center	Tail
Two nonlinear layers	4.5889	5.8090	0.4358	0.3443	24/24	1/24
With linear output	2.9743	3.3151	0.6724	0.6033	24/24	7/24

Table 1 evaluates the finite-width NTK stability condition. We choose a common step between the two predicted critical values, which predicts stability for the center gate and instability for the tail gate. Every sampled center network and all but one tail network satisfy this prediction. Separate gradient-descent trajectories test the local dynamics by perturbing networks around their interpolating targets and training both Gaussian weights. The residuals follow the local multiplier of the

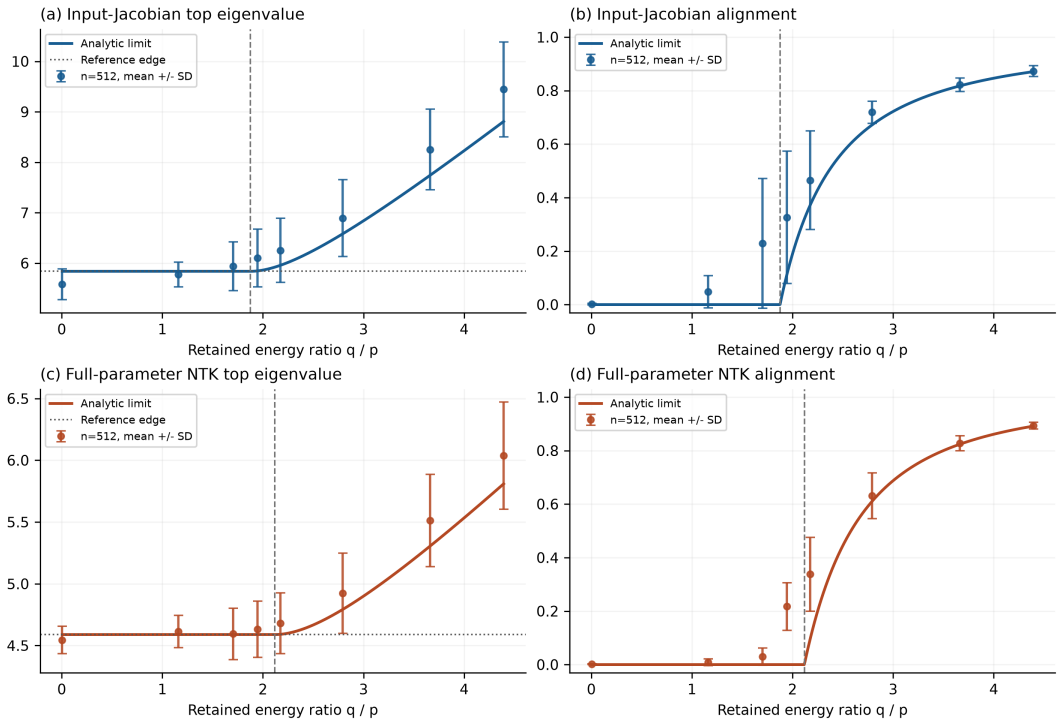


Figure 2: The input Jacobian and parameter NTK have different outlier thresholds in the same two-layer network at fixed gate probability $p = 0.1$. Panels (a,b) show JJ^* and panels (c,d) show Θ. Solid curves predict the largest eigenvalue and squared overlap with $D_2 z_2 / \|D_2 z_2\|$. Markers and error bars give means and sample standard deviations over eight networks at $n = 512$. Vertical dashed lines mark the Jacobian and parameter-NTK thresholds. Horizontal dotted lines mark the corresponding reference edges.

measured finite-width NTK, including in the stable tail case. Automatic differentiation verifies the NTK identity. Appendix H.8 gives the protocol and numerical checks.

The overlap measurements in Figure 2 test whether the leading eigenvector follows the preactivation direction identified by Gaussian conditioning. Agreement in both eigenvalue location and overlap supports the directional calculation. The different transitions for JJ^* and Θ show why input sensitivity and sensitivity to weight updates must be distinguished.

Appendix C reports the corresponding experiment for a linear first layer. Training the lower weight gives K, and training both weights gives $K + D$.

4.2 Gaussian linear output

We add the Gaussian layer in Equation (12), match the center and tail activation second moments, and train all three weights. Matching these moments and the gate activation probability gives a common predicted output bulk law. Figure 3(a,b) shows that the tail outlier separates only once the output-to-hidden width ratio exceeds the predicted threshold. Above this threshold, the top eigenvalue rises above the common edge and its eigenvector aligns with the projected hidden direction. Panel (d) gives the complementary case: scalar-output NTKs approach the common limit in proposition I.1.

Matching activation second moments isolates the effect of output width (Section 3.4). The scalar-output kernels approach a common limit despite their different hidden top eigenvalues.

Figure 3(c) compares predictions across finite networks. The population prediction uses limiting moments. The conditional predictor estimates the hidden spectral law and largest eigenvalue from observed gate counts, preactivation variance, and retained energy. It then applies the output-layer

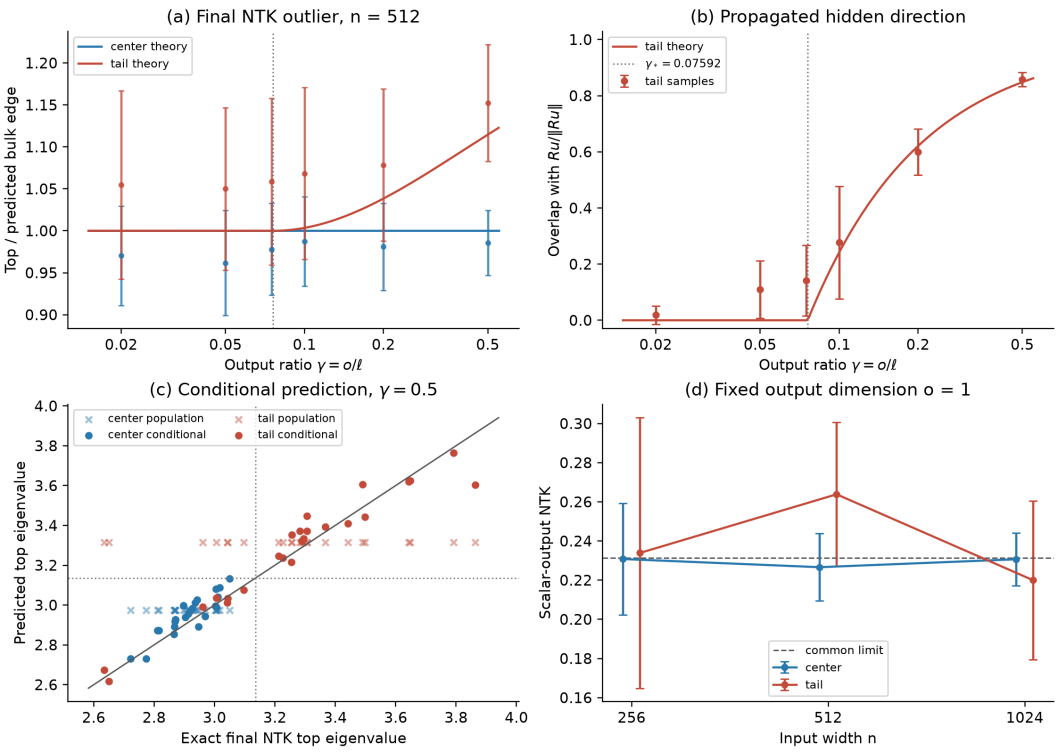


Figure 3: The output-to-hidden width ratio determines whether the hidden NTK outlier remains separated. (a) Largest output-NTK eigenvalue divided by the predicted bulk edge B_γ. (b) Squared overlap $|v_A^* R u_A|^2 / \|R u_A\|^2$ with the projected hidden top eigenvector. Error bars show sample standard deviations over eight paired draws at $n = 512$. (c) Population and conditional predictions against the measured largest eigenvalues of 24 paired networks at $\gamma = 0.5$. In (c), dotted lines mark the common-step stability threshold $2/\eta$ and the solid diagonal denotes exact agreement. (d) Scalar-output NTKs at three widths, compared with their common limit.

map in Theorem 3.4, using the observed activation second moment. The measured NTK spectrum is used only for evaluation.

These forward-pass statistics improve predictions without constructing the NTK. In Figure 3(c), this reduces the mean absolute relative eigenvalue error from 5.00% to 1.65%. More accurate eigenvalues also improve the stability decisions: agreement with the finite-width NTK rises from 41 of 48 cases to all 48, including the stable tail cases. Appendix I.2 specifies the estimator and evaluation protocol.

Updates of all three weights near interpolation follow the local NTK prediction (Table 8).

5 Conclusion

We have studied what a spectral description of a wide neural network must retain to predict local training stability. Gaussian networks can have the same limiting eigenvalue distribution and all fixed-order normalized moments, yet admit different stable learning rates. We identify the cause in the dependence between weights and derivative gates and show how conditioning on the computation recovers the relevant directions. For two nonlinear layers and a trainable Gaussian output, this yields explicit outlier and local parameter-training predictions, including a width threshold at which the output preserves the hidden spectral difference. The resulting lesson for spectral analysis is to account for both the distribution of modes and the perturbation directions identified by Gaussian conditioning. Extending the directional calculation to other activations and to interactions across a growing set of inputs would clarify how this mechanism operates in broader training settings.

AI USE STATEMENT

AI tools assisted literature searches, proof drafting and revision, experimental design, numerical implementation and code editing. The authors are responsible for the correctness, originality, and presentation of the submitted work.

REPRODUCIBILITY STATEMENT

Proofs appear in Appendices A to C and E to J. Experimental protocols, parameter settings, and numerical checks are given in the corresponding appendices. The accompanying code and per-seed records reproduce the parameter-training, two-layer, and linear-output experiments.

REFERENCES

S. T. Belinschi and D. Shlyakhtenko. Free probability of type B: analytic interpretation and applications. *American Journal of Mathematics*, 134(1):193–234, 2012.

F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low-rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

Z. Burda, A. Jarosz, G. Livan, M. A. Nowak, and A. Święch. Eigenvalues and singular values of products of rectangular Gaussian random matrices (the extended version). *Acta Physica Polonica B*, 42:939–985, 2011.

C.-F. Chen, J. Garza-Vargas, J. A. Tropp, and R. van Handel. A new approach to strong convergence. *Annals of Mathematics*, 203(2):555–602, 2026.

B. Collins and T. Hayase. Asymptotic freeness of layerwise Jacobians caused by invariance of multilayer perceptron: the Haar orthogonal case. *arXiv:2103.13466*, 2021.

C. Cranston, Z. Wang, T. Kemp, and W. M. Mahoney. Eigen-spike emergence and quadratic equivalents for conjugate kernels on nonlinearly separable data. In *Proceedings of the Conference on Learning Theory*, volume 336, pages 1476–1574, 2026.

B. Dadoun, S. Hayou, H. Salam, M. E. A. Seddik, and P. Youssef. On the stability of the Jacobian matrix in deep neural networks. *arXiv:2506.08764*, 2025.

Z. Fan and Z. Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

B. Ghorbani, S. Krishnan, and Y. Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2232–2241, 2019.

U. Haagerup and S. Thorbjørnsen. A new application of random matrices: $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ is not a group. *Annals of Mathematics*, 162(2):711–775, 2005.

W. Hachem, P. Loubaton, and J. Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

J. Hazelden, L. Driscoll, E. Shlizerman, and E. Shea-Brown. The global empirical NTK: self-referential bias and dimensionality of gradient descent learning. *arXiv:2605.08746*, 2026.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

D. Jekel, Y. Lee, B. Nelson, and J. Pi. Strong convergence to operator-valued semicirculars. *arXiv:2506.19940, version 3*, 2026.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

A. Nica and R. Speicher. *Lectures on the Combinatorics of Free Probability*. Cambridge University Press, 2006.

J. Pennington, S. S. Schoenholz, and S. Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *Advances in Neural Information Processing Systems*, 30, 2017.

J. W. Silverstein and S.-I. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

Z. Wang, D. Wu, and Z. Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. In *Proceedings of the Conference on Learning Theory*, 2024.

G. Yang. Tensor Programs II: neural tangent kernel for any architecture. *arXiv:2006.14548*, 2020a.

G. Yang. Tensor Programs III: neural matrix laws. *arXiv:2009.10685*, 2020b.

Z. Bai and J.-F. Yao. Limit theorems for sample eigenvalues in a generalized spiked population model. *arXiv:0806.1141*, 2008. Theorems 4.1–4.2 of the cited arXiv version.

ORGANIZATION OF THE APPENDIX

Appendix A collects the spectral tools used in the proofs. The one-layer result in Section 3.1 is proved in Appendix B, with its parameter-training realization in Appendix C. Appendix D states the general results summarized in Section 3.2; their proofs follow in Appendices E to G. Appendices H and I prove the two-layer and output-layer results in Sections 3.3 and 3.4 and give their experimental protocols. Appendix J extends the output analysis to factored kernels. Numerical evaluation and further experiments appear in Appendices K to M.

A SPECTRAL PRELIMINARIES

The proof separates two questions: where the reference spectrum ends, and whether the network creates eigenvalues beyond it. Joint Gaussian regression supplies the reference by isolating the directions observed during the computation. Its difference from the original weights has rank bounded by the query dimensions, and this bound propagates through fixed matrix expressions.

To determine the reference support, we prove trace and norm convergence of the derivative-gate algebra for fixed-depth MLPs. The associated covariance maps satisfy the hypotheses of Jekel et al. (2026, Theorem 4.12). Conditional application of that theorem gives strong convergence of the reference operators.

Interlacing then locates the original spectrum below the possible outliers. The remaining eigenvalues are roots of a determinant involving the reference resolvent along the perturbation directions. The one-layer and two-layer calculations evaluate these directional limits. Differentiating the network with respect to its parameters gives shifted Wishart matrices and their Gaussian output compressions, so the same spectral tools also determine the local training thresholds.

A.1 RANK AND INTERLACING LEMMAS

Lemma A.1. *For Hermitian $A, B \in \mathbb{C}^{N\times N}$,*

$$\sup_x |F_A(x) - F_B(x)| \le \frac{\operatorname{rank}(A-B)}{N}, \tag{17}$$

where F_A and F_B are the empirical spectral distribution functions.

Lemma A.2. *Let the eigenvalues be ordered decreasingly. If $A = A^*$, $B = B^*$, and $\operatorname{rank}(A-B) \le r$, then*

$$\lambda_{j+r}(B) \le \lambda_j(A) \le \lambda_{j-r}(B), \tag{18}$$

for every admissible j, with the usual endpoint conventions.

Combining lemmas A.1 and A.2 with positive limiting mass of the Gaussian reference operator near the upper support edge proves corollary D.3:

$$\limsup_{n\to\infty} \lambda_{r_n+1}(K_n) \le \limsup_{n\to\infty} \lambda_1(H_n) = b,$$
$$\liminf_{n\to\infty} \lambda_{r_n+1}(K_n) \ge \liminf_{n\to\infty} \lambda_{2r_n+1}(H_n) = b.$$

Strong convergence gives $\lambda_1(H_n) \to b$. The second equality follows because $2r_n = o(N_n)$ and every interval $(b-\varepsilon, b]$ has positive limiting mass under the Gaussian reference law.

A.2 MATRIX DETERMINANT IDENTITY

Lemma A.3. *Let $H \in \mathbb{C}^{N\times N}$, $U \in \mathbb{C}^{N\times r}$, and $C \in \mathbb{C}^{r\times r}$. For $\zeta \notin \operatorname{spec}(H)$,*

$$\det(\zeta I - H - UCU^*) = \det(\zeta I - H)\det\left(I - CU^*(\zeta I - H)^{-1}U\right). \tag{19}$$

A.3 INTERLACING FOR GRAM MATRICES

Lemma A.4. *Let $A, B \in \mathbb{C}^{N\times d}$, $K = AA^*$, and $H = BB^*$. If $\operatorname{rank}(A-B) \le r$, then the positive and negative inertia indices of $K - H$ are each at most r. Consequently,*

$$\lambda_{j+r}(H) \le \lambda_j(K) \le \lambda_{j-r}(H) \tag{20}$$

for admissible indices. Suppose along a sequence that $r_n = o(N_n)$, H_n converges strongly to a law with upper endpoint B, and every interval $(B-\varepsilon, B]$ has positive limiting mass. Then

$$\lambda_{r_n+1}(K_n) \longrightarrow B. \tag{21}$$

Proof. On $\mathcal{V} = \ker(A-B)^*$, one has $A^*x = B^*x$ and hence $x^*(K-H)x = 0$. Since $\mathcal{V}$ has codimension at most r, every subspace on which $K-H$ is positive or negative definite has dimension at most r. The min–max principle gives Equation (20). In particular,

$$\lambda_{2r_n+1}(H_n) \le \lambda_{r_n+1}(K_n) \le \lambda_1(H_n).$$

Both bounds tend to B by strong convergence and the endpoint-mass assumption. □

A.4 Rank-one perturbations and eigenvector overlap

For Hermitian A_n and a unit vector u_n, the directional resolvent has the spectral decomposition

$$g_n(z) := u_n^*(zI - A_n)^{-1}u_n = \sum_j \frac{|\langle u_n, v_{j,n}\rangle|^2}{z - \lambda_j(A_n)}. \tag{22}$$

Suppose a positively oriented contour Γ encloses only the simple top eigenvalue for all large n, and $\lambda_1(A_n) \to L$. If $g_n \to g$ uniformly on Γ and g is analytic inside except possibly at L, then

$$|\langle u_n, v_{1,n}\rangle|^2 = \frac{1}{2\pi \mathrm{i}} \oint_\Gamma g_n(z)\,\mathrm{d}z \longrightarrow \mathrm{Res}_{z=L}\, g(z), \tag{23}$$

At the upper edge, positivity in Equation (22) gives, for any unit top eigenvector and real $z > \lambda_1(A_n)$,

$$0 \le |\langle u_n, v_{1,n}\rangle|^2 \le (z - \lambda_1(A_n))g_n(z). \tag{24}$$

If $\lambda_1(A_n) \to B$ and $g_n(z) \to g(z)$ for fixed $z > B$, the overlap therefore tends to zero whenever $(z-B)g(z) \to 0$ as $z \downarrow B$. This argument allows a repeated top eigenvalue. The width limit is taken before approaching the edge.

Lemma A.5. *Let $S_n = S_n^*$ have eventually bounded norm and $\lambda_1(S_n) \to B$. Let $\|u_n\| = 1$, $\theta_n \to \theta > 0$, and $K_n = S_n + \theta_n u_n u_n^*$. Assume*

$$m_n(z) := u_n^*(zI - S_n)^{-1}u_n \longrightarrow m(z) = \int \frac{1}{z-x}\,\mathrm{d}\nu(x)$$

locally uniformly on $\mathbb{C} \setminus [-M, B]$, where ν is a probability measure supported in $[-M, B]$ and $\mathrm{spec}(S_n) \subset [-M, B + o(1)]$. If $\theta m(B^+) > 1$, the equation $\theta m(L) = 1$ has a unique solution $L > B$, and $\lambda_1(K_n) \to L$. For a unit top eigenvector $v_{1,n}$,

$$|\langle u_n, v_{1,n}\rangle|^2 \longrightarrow -\frac{1}{\theta^2 m'(L)} = \rho'(\theta), \qquad \theta m(\rho(\theta)) = 1. \tag{25}$$

The derivative holds with the reference transform m fixed. If $\theta m(B^+) \le 1$, then $\lambda_1(K_n) \to B$. In this regime the overlap tends to zero whenever

$$\lim_{z \downarrow B} (z-B)\frac{m(z)}{1-\theta m(z)} = 0. \tag{26}$$

In particular, this condition holds at and below threshold for a regular square-root edge:

$$m(z) = m(B^+) - a\sqrt{z-B} + O(z-B), \qquad 0 < m(B^+) < \infty, \quad a > 0.$$

Proof. The determinant lemma and Sherman–Morrison identity give

$$\frac{\det(zI - K_n)}{\det(zI - S_n)} = 1 - \theta_n m_n(z), \qquad g_n(z) := u_n^*(zI - K_n)^{-1}u_n = \frac{m_n(z)}{1 - \theta_n m_n(z)}. \tag{27}$$

On (B, ∞), m is positive and strictly decreasing to zero. A separated root therefore exists exactly under the stated strict inequality. Local uniform convergence gives an eigenvalue near this simple root, and rank-one interlacing permits at most one eigenvalue above $\lambda_1(S_n)$. In the absence of

a separated root, boundedness and the determinant identity exclude subsequential limits above B. The lower bound $\lambda_1(K_n) \geq \lambda_1(S_n)$ then gives convergence to B.

Away from the reference spectrum and the zeros of $1-\theta m$, the second identity gives locally uniform convergence to $g = m/(1-\theta m)$. At a separated root, Equation (23) yields

$$\mathrm{Res}_{z=L}\, g(z) = \frac{m(L)}{-\theta m'(L)} = -\frac{1}{\theta^2 m'(L)}.$$

Implicit differentiation of $\theta m(\rho(\theta)) = 1$, with m fixed, identifies this expression with $\rho'(\theta)$.

When $\lambda_1(K_n) \to B$, Equation (24) and Equation (26) give zero overlap. For a regular square-root edge, the denominator has a positive limit below threshold and is proportional to $\sqrt{z-B}$ at threshold, so the condition holds in both cases. □

We use the same concentration argument for the directional resolvents in the Gaussian applications. If a unit vector is uniform on the sphere and independent of a norm-bounded Hermitian matrix, sphere concentration gives

$$u_n^* P(S_n) u_n - \mathrm{tr}_{N_n} P(S_n) \longrightarrow 0$$

almost surely for every fixed polynomial P, provided $N_n \asymp n$. The tails at any fixed error are exponentially small in N_n. The conclusion also holds for a fixed direction independent of an orthogonally invariant matrix, by conditioning on its eigenvalues. Weak spectral convergence and the norm bound identify the limiting moments. Polynomial approximation and resolvent bounds then give the locally uniform convergence used above.

Finally, whenever $\mu_{A_n} \Rightarrow \mu$ and the Hermitian matrices A_n have eventually bounded norm, all fixed normalized moments converge. Indeed, their spectra lie in a common compact interval, on which each fixed power is a bounded continuous function. We use this observation after establishing the empirical law and norm bounds.

B Proof of Theorem 3.1

We prove Theorem 3.1 by conditioning on the preactivation vector, applying the rank-one perturbation result, and computing the resulting eigenvector overlap.

B.1 Rank-one decomposition

By right-orthogonal invariance, rotate $h_n/\sqrt{n}$ to e_1. Then

$$W_n = \frac{1}{\sqrt{n}} \begin{bmatrix} z_n & X_n \end{bmatrix},$$

where $z_n \sim \mathcal{N}(0, \mathrm{I}_{m_n})$ and $X_n \in \mathbb{R}^{m_n \times (n-1)}$ is an independent standard Gaussian matrix. Define

$$B_{A,n} := \frac{1}{\sqrt{n}} \begin{bmatrix} 0 & D_{A,n} X_n \end{bmatrix}, \qquad s_{A,n} := \frac{D_{A,n} z_n}{\sqrt{n}}, \qquad S_{A,n} := B_{A,n} B_{A,n}^*.$$

Set $u_{A,n} := s_{A,n}/\|s_{A,n}\|$ when $s_{A,n} \neq 0$, and set $u_{A,n} := 0$ otherwise. Let $v_{1,n}$ be a unit top eigenvector of $K_{A,n}$. The rank-one decomposition is

$$\begin{aligned} D_{A,n} W_n &= B_{A,n} + s_{A,n} e_1^*, \\ K_{A,n} &= (D_{A,n} W_n)(D_{A,n} W_n)^* = S_{A,n} + s_{A,n} s_{A,n}^*. \end{aligned}$$

The cross terms vanish because $B_{A,n} e_1 = 0$. Conditional on z_n, the mask and $s_{A,n}$ are fixed, whereas $S_{A,n}$ is generated by the independent matrix X_n and is a sample covariance matrix on the active rows.

B.2 Active Wishart block and isotropic resolvent

Let $I_n := \{i : z_{n,i} \in A\}$ and $k_n := |I_n|$. Define

$$\widehat{S}_n := \frac{1}{n} X_{I_n,n} X_{I_n,n}^*, \qquad t_n := \|s_{A,n}\|^2.$$

When $k_n > 0$, set $\widehat{u}_n := (z_{n,i})_{i\in I_n}/\|(z_{n,i})_{i\in I_n}\|$. For $k_n = 0$, take the empty zero vector. Because $p_A > 0$, almost surely $k_n > 0$ for all sufficiently large n. After permuting coordinates, the following identities hold:

$$S_{A,n} \simeq \widehat{S}_n \oplus 0_{m_n-k_n}, \qquad K_{A,n} \simeq (\widehat{S}_n + t_n \widehat{u}_n \widehat{u}_n^*) \oplus 0_{m_n-k_n}. \tag{28}$$

The Gaussian law of large numbers gives, almost surely,

$$\frac{k_n}{m_n} \to p_A, \qquad \frac{k_n}{n-1} \to y_A, \qquad t_n = \frac{1}{n}\sum_{i=1}^{m_n} z_{n,i}^2 \mathbf{1}_{\{z_{n,i}\in A\}} \to \theta_A. \tag{29}$$

Here $\widehat{S}_n = ((n-1)/n)(n-1)^{-1} X_{I_n,n} X_{I_n,n}^*$, so the change from $n-1$ to n does not affect the limit. Conditional on the full z-sequence, I_n, t_n, and $\widehat{u}_n$ are fixed, while $\widehat{S}_n$ is an independent orthogonally invariant Wishart matrix.

Set $y := y_A$ and $b_y := (1+\sqrt{y})^2$. The Marchenko–Pastur law (Marchenko & Pastur, 1967) and corollary F.4 with population δ_1 give $\lambda_1(\widehat{S}_n) \to b_y$. Conditional orthogonal invariance and the sphere-concentration argument in Appendix A yield

$$\widehat{u}_n^*(zI - \widehat{S}_n)^{-1}\widehat{u}_n \longrightarrow m_y(z) := \int \frac{1}{z-x}\,\mathrm{d}\mu_{\mathrm{MP},y}(x) \tag{30}$$

locally uniformly off the limiting support. The reference spectra converge to that support, so resolvent bounds give uniformity on each compact subset of its complement. Conditioning and Fubini give the joint almost-sure statement. The active-row proportion and rank-one decomposition imply

$$\mu_{K_{A,n}} \Longrightarrow (1-p_A)\delta_0 + p_A \mu_{\mathrm{MP},y_A}.$$

B.3 Outlier and eigenvector overlap

Applying Equation (68) with population δ_1 gives the MP inverse transform

$$z = \frac{1}{m_y(z)} + \frac{1}{1-y m_y(z)} = \rho_y(t) := t + 1 + \frac{y}{t-y}, \qquad t = 1/m_y(z). \tag{31}$$

The branch from infinity has $t > \theta_c := y + \sqrt{y}$. Since $\rho_y'(t) = 1 - y/(t-y)^2$, its endpoint is $b_y = \rho_y(\theta_c) = (1+\sqrt{y})^2$ and $m_y(b_y^+) = 1/\theta_c$. By lemma A.5, an upper outlier exists exactly when

$$\theta_A > y_A + \sqrt{y_A}. \tag{32}$$

Its location and overlap follow from the same inverse transform:

$$L = \rho_{y_A}(\theta_A) = \frac{\theta_A(\theta_A + 1 - y_A)}{\theta_A - y_A}, \tag{33}$$

$$|\langle v_{1,n}, u_{A,n}\rangle|^2 \longrightarrow \rho_{y_A}'(\theta_A) = 1 - \frac{y_A}{(\theta_A - y_A)^2}. \tag{34}$$

At and below threshold, the lemma gives $\lambda_1(K_{A,n}) \to b_y$. Strictly below threshold its zero-overlap criterion is immediate. At equality, $\rho_y(\theta_c) = b_y$ and $\rho_y'(\theta_c) = 0$, so

$$(z - b_y)\frac{m_y(z)}{1-\theta_c m_y(z)} = \frac{\rho_y(t) - \rho_y(\theta_c)}{t - \theta_c} \longrightarrow \rho_y'(\theta_c) = 0 \qquad (t \downarrow \theta_c).$$

Thus the overlap also vanishes at the critical point. The explicit density and transform formulas for numerical evaluation are given in Appendix K.

Finally, $\|K_{A,n}\| \le \|\widehat{S}_n\| + t_n$ is eventually bounded. The empirical law and compact-support observation in Appendix A give every fixed normalized moment. This completes the proof of Theorem 3.1.

C A network with a linear first layer

Consider a linear first layer followed by a nonlinear layer. Training the first weight gives the NTK in Theorem 3.1, and training both weights adds a scalar shift on the active outputs.

C.1 Model and parameter NTK

Fix $x \in \mathbb{R}^d$ with $\|x\| = 1$, a deterministic $h_n \in \mathbb{R}^n$ with $\|h_n\|^2 = n$, and a deterministic $U_0 \in \mathbb{R}^{n\times d}$ satisfying $U_0 x = h_n$, for example, $U_0 = h_n x^*$. Draw W_n as in Equation (1), independently of U_0. Consider one training example with vector output

$$f_U(x) = \phi_A(W_n U x) \in \mathbb{R}^{m_n}, \qquad \mathcal{L}(U) = \frac{1}{2}\|f_U(x) - y\|^2. \tag{35}$$

Only the entries of U are trained, using the Euclidean (Frobenius) parameter metric. The input, W_n, and activation are fixed. The loss is unaveraged. Averaging over output coordinates multiplies the admissible step sizes below by m_n.

For the center gate $A_c = (-a, a)$ use the clipped-linear activation $\phi_c(t) = \max(-a, \min(t, a))$. For the tail gate $A_t = (-\infty, -b) \cup (b, \infty)$ use the soft-thresholding activation $\phi_t(t) = (t-b)_+ - (-t-b)_+$. Their almost-everywhere derivatives are the required binary gates. Gaussian preactivations avoid their finitely many breakpoints almost surely at each finite width. These piecewise-linear activations satisfy the assumptions of Theorem 3.1.

Proposition C.1. *In Equation (35), the output NTK at U_0 is exactly $K_{A,n}$ from Equation (1). Its limiting empirical law, fixed normalized moments, largest eigenvalue, and supercritical directional overlap are therefore given by Theorem 3.1. For $y = f_{U_0}(x)$, the nonzero eigenvalues of $\nabla^2\mathcal{L}(U_0)$ equal those of $K_{A,n}$. The linearization of parameter gradient descent contracts every positive-curvature mode exactly when $0 < \eta < 2/\lambda_1(K_{A,n})$.*

Proof. Let $D = D_{A,n}$. The Jacobian, viewed as an operator from matrices with Frobenius inner product to output vectors, and its adjoint are

$$J_U[\Delta U] = DW_n \Delta U x, \qquad J_U^*[r] = W_n^* D r x^*.$$

Thus $J_U J_U^*[r] = \|x\|^2 DW_n W_n^* Dr = K_{A,n} r$. At the interpolating target the residual-dependent second-derivative term in the loss Hessian vanishes, so $\nabla^2\mathcal{L}(U_0) = J_U^* J_U$. The nonzero eigenvalues of $J_U^* J_U$ and $J_U J_U^*$ agree. The derivative of the gradient-descent map is $I - \eta J_U^* J_U$, which yields the stated criterion. Null parameter directions have multiplier one. □

For these piecewise-linear activations the statement has an exact local nonlinear interpretation. At each finite width, U_0 lies inside an open activation cell. Throughout that cell $f_U(x) - f_{U_0}(x) = J_U[U - U_0]$ and the loss is quadratic. Below $2/\lambda_1(K_{A,n})$, a sufficiently small ball around U_0 stays inside the cell under gradient descent, since $\|I - \eta J_U^* J_U\| \le 1$. Its positive modes contract. Above that threshold, arbitrarily small perturbations along the top right singular direction expand until they leave the cell. The activation-cell radius depends on the finite-width initialization.

C.2 Training both weight matrices

To train both weights, write

$$f_{U,G}(x) = \phi_A(n^{-1/2} GUx), \qquad G_0 = \sqrt{n} W_n, \tag{36}$$

and train all entries of U and G at the same step size with their Euclidean parameter metrics. Thus G_0 has i.i.d. standard Gaussian entries, while $U_0 x = h_n$ is deterministic. This scaling is part of the model and fixes the relative learning rates of the effective weights.

Corollary C.2. *At (U_0, G_0), the output NTK in Equation (36) is*

$$\Theta_{A,n}^{\text{both}} = K_{A,n} + D_{A,n}. \tag{37}$$

Its limiting law is $(1 - p_A)\delta_0 + p_A (T_1)_{\#}\mu_{\text{MP},y_A}$, where $T_1(t) = t + 1$. Gates with the same p_A have the same limiting fixed normalized moments. The largest eigenvalue is $1 + \lambda_1(K_{A,n})$, and its supercritical eigenvector overlap is unchanged. At the interpolating target, the limiting critical step size for linearized gradient descent at interpolation is $2/(1 + L_A)$, where L_A is the limit in Equation (3).

Proof. The U block contributes $K_{A,n}$ by proposition C.1. The G differential is $J_G[\Delta G] = n^{-1/2} D_{A,n} \Delta G h_n$, so $J_G J_G^* = n^{-1}\|h_n\|^2 D_{A,n}^2 = D_{A,n}$. Independent parameter coordinates contribute additively to the NTK. On the active rows $D_{A,n}$ is the identity; on the inactive rows both matrices vanish. The shift therefore preserves active eigenvectors and shifts every active eigenvalue by one. The empirical law and moment conclusions follow from Theorem 3.1 and the limiting active fraction. At zero residual the Hessian is the full parameter Jacobian Gram matrix, giving the stated threshold. □

With both weights trained, the network is bilinear within each activation cell. At interpolation, the initial NTK gives the derivative of the gradient-descent map. The experiments below test this local prediction using small parameter perturbations.

C.3 Predictions from network parameters

Choose $m_n = 8n$, $p = 0.1$, $a = \Phi^{-1}((1+p)/2)$, and $b = \Phi^{-1}(1-p/2)$, where Φ and φ are the standard Gaussian distribution and density. Then

$$q_c = p - 2a\varphi(a), \qquad q_t = p + 2b\varphi(b), \qquad y = 8p = 0.8.$$

Both NTKs have limiting law $(1-p)\delta_0 + p\mu_{\mathrm{MP},0.8}$. The center gate is subcritical, whereas the tail gate has $\theta_t = 8q_t = 3.514289$. The resulting predictions, without fitting finite-width eigenvalues, are

$$\begin{aligned} B = (1+\sqrt{0.8})^2 = 3.588854, \quad & L_t = 4.809025, \\ \eta_c^* = 2/B = 0.557281, \quad & \eta_t^* = 2/L_t = 0.415885, \\ |\langle v_1, u_t\rangle|^2 \longrightarrow 0.891413, \quad & \eta = \tfrac{1}{2}(\eta_c^* + \eta_t^*) = 0.486583. \end{aligned} \tag{38}$$

The bulk-only prediction admits this η for both networks, while the original tail NTK has an expanding top mode. The two activations use the same architecture, loss, and parameter metric.

For the both-weight parameterization, the corresponding bulk and tail edges are $B+1$ and L_t+1. We use the independently specified step

$$\eta_{\mathrm{both}} = \tfrac{1}{2}\left(\frac{2}{B+1} + \frac{2}{L_t+1}\right).$$

This again lies between the two true limiting stability thresholds. Both activation choices use the same loss and parameter metric. The targets equal the respective initial outputs, and the experiment measures stability under small parameter perturbations.

C.4 Reference spectrum and interlacing

Rotate so that $h_n = \sqrt{n}e_1$ and write $W_n = n^{-1/2}[z_n \; X_n]$. The reference Gram matrix is $H_n^0 = n^{-1} D_A X_n X_n^* D_A$, and $K_{A,n} = H_n^0 + s_A s_A^*$ with $s_A = n^{-1/2} D_A z_n$. Thus $\lambda_2(K_{A,n}) \to B$ by rank-one interlacing and edge mass. A full Gaussian reference, independent of the gate history, is obtained by replacing z_n with a fresh independent standard Gaussian column g_n:

$$H_n = D_A G_n G_n^* D_A, \qquad G_n = n^{-1/2}[g_n \; X_n].$$

The correction $K_{A,n} - H_n = s_A s_A^* - t_A t_A^*$ has rank at most two, where $t_A = n^{-1/2} D_A g_n$. Accordingly $\lambda_1(H_n) \to B$ and $\lambda_3(K_{A,n}) \to B$, while $\lambda_1(K_{A_t,n}) \to L_t > B$. Only the last quantity determines the actual tail network's local step-size boundary. Both reference constructions converge strongly by the conditional Wishart argument used in Appendix B.

For training both weights, the reference is $H_n + D_A$ and all these active upper-edge quantities are shifted by one. The rank bounds and outlier overlap are unchanged.

C.5 Experimental protocol

The supplementary experiment uses $n \in \{256, 512, 1024\}$, $m = 8n$, $d = 3$, $x = (1,2,2)^*/3$, and $U_0 = \sqrt{n}e_1 x^*$. The eight seeds are 101, 211, 307, 401, 503, 601, 701, and 809. Each width–seed pair initializes NumPy's generator with `SeedSequence([seed,n])`. Center and tail gates

share the same Gaussian weight. The two training regimes reuse these paired draws. Thresholds and step sizes are fixed from the limiting formulas, with no fit to observed eigenvalues. All computations use double precision.

Table 2: Parameter NTK with only the lower weight trained: mean $\pm$ sample standard deviation over eight seeds. K is the original output NTK and H its full Gaussian reference. The rank-two correction allows the trimmed quantity $\lambda_3(K)$. The overlap is with the retained preactivation direction. For both-weight training, add one to the three displayed eigenvalue columns. The overlap is unchanged.

n	Gate	$\lambda_1(K)$	$\lambda_1(H)$	$\lambda_3(K)$	Overlap
256	Center	3.566 ± 0.101	3.573 ± 0.102	3.272 ± 0.059	0.003 ± 0.002
256	Tail	4.780 ± 0.220	3.537 ± 0.129	3.375 ± 0.135	0.902 ± 0.015
512	Center	3.607 ± 0.101	3.609 ± 0.101	3.472 ± 0.097	0.001 ± 0.002
512	Tail	4.798 ± 0.175	3.549 ± 0.070	3.451 ± 0.078	0.898 ± 0.009
1024	Center	3.538 ± 0.028	3.540 ± 0.029	3.452 ± 0.021	0.001 ± 0.002
1024	Tail	4.884 ± 0.127	3.579 ± 0.053	3.531 ± 0.052	0.890 ± 0.013

The reference replaces the Gaussian column along the input direction and shares the remaining columns with the original matrix. Eigenvalues are computed by symmetric dense eigensolvers on the active block. Inactive eigenvalues are zero. For each seed, we also evaluate $\lambda_2(K)$, the reference upper edge, retained preactivation energy, critical step, and finite-width stability.

Table 3: Steps chosen from the asymptotic network parameters and finite-width decisions. Each count pools 24 width–seed draws within its regime. The NTK criterion contracts all positive modes for the center gate and expands the top tail mode. The bulk-only threshold predicts stability for both gates.

Trained weights	Bulk threshold	Tail threshold	Step	Stable center	Stable tail
Lower only	0.5573	0.4159	0.4866	24/24	0/24
Both	0.4358	0.3443	0.3901	24/24	0/24

We verify the parameter Jacobians by automatic differentiation. At $n = 12$, $m = 96$, $d = 3$, seed 20260908, PyTorch automatic differentiation constructs the full parameter Jacobian directly for center, tail, and ReLU activations. We check $J_U J_U^* = K$, the direct-W identity $J_W J_W^* = nD$, and the both-weight identity $JJ^* = K + D$ independently of the large-matrix eigensolver. Across these checks the maximum relative Frobenius error is below 4×10^{-16}.

We test the local stability predictions by gradient descent. At $n = 512$, all eight seeds and both gates are tested in both training regimes, giving 32 trajectories. Each uses the interpolating target at its initialization and starts with a perturbation along the top right singular vector of the appropriate parameter Jacobian. If ℓ denotes its top NTK eigenvalue, $T = 12$, and δ is the minimum preactivation distance to an activation breakpoint, the joint parameter perturbation has norm

$$\epsilon_0 = \frac{0.02\,\delta}{\|W\|_F \max\{1, |1 - \eta\ell|\}^T}.$$

This finite-width, spectrum-dependent amplitude is used only to probe local stability. The learning rate is selected from the limiting formulas. For both-weight training, a unit joint parameter direction is

$$\Delta U = \frac{W^* D v_1 x^*}{\sqrt{\lambda_1(K) + 1}}, \qquad \Delta G = \frac{D v_1 h_n^*}{\sqrt{n(\lambda_1(K) + 1)}}.$$

We differentiate the nonlinear network loss and update its parameters at every step. It recomputes all preactivations and records any gate changes. None occur in these trajectories. With the upper weight frozen, the cellwise residual recursion is exact. With both weights trained, the initial NTK provides a local approximation to the bilinear parameter dynamics. The maximum relative discrepancy between the measured residual-norm ratio and $|1 - \eta\ell|^t$ over all recorded steps is 0.000033.

D GAUSSIAN CONDITIONING AND LOW-RANK PERTURBATIONS

This appendix gives the precise statements behind Section 3.2. The Gaussian decomposition identifies a finite-rank correction; strong convergence determines the reference spectrum; a compressed resolvent locates the remaining outliers. The one-layer calculation works because observing Wh reveals the weight matrix along one direction and leaves Gaussian randomness in the orthogonal complement. A multilayer computation reveals several such directions, with later vectors depending on earlier outputs. We retain all of these observations by treating the forward and backward products as adaptive queries. Conditional on their history, the observed products impose linear constraints on each weight matrix. This makes it possible to isolate the dependence without discarding the activations that generated it.

Let $W \in \mathbb{R}^{m\times n}$ have i.i.d. centered Gaussian entries of a common variance, and fix $T < \infty$. At step t, choose right and left query vectors a_t and b_t measurably from the preceding queries and responses. Let $\mathscr{T}$ be the sigma-field generated by $\{a_t, Wa_t, b_t, W^*b_t : t \le T\}$. Let $V \in \mathbb{R}^{n\times r}$ and $U \in \mathbb{R}^{m\times s}$ be orthonormal bases of the spans of the right and left queries, respectively. Finally, set $Y = WV$ and $R = W^*U$, and write $\Pi_{\mathcal{U}}$ for orthogonal projection onto a subspace $\mathcal{U}$.

Proposition D.1. *Conditioned on $\mathscr{T}$,*

$$\begin{aligned} W \mid \mathscr{T} &\overset{d}{=} W^{\parallel} + \Pi_{U^\perp}\widetilde{W}\Pi_{V^\perp}, \\ W^{\parallel} &= YV^* + UR^*(I - VV^*), \qquad \operatorname{rank}(W^{\parallel}) \le r + s, \end{aligned} \tag{39}$$

where $\widetilde{W}$ is a Gaussian matrix independent of $\mathscr{T}$.

To compare full matrix expressions, couple W to a Gaussian reference matrix G^{ref} that is independent of $\mathscr{T}$, so that

$$W = G^{\mathrm{ref}} + E, \qquad \operatorname{rank}(E) \le r + s. \tag{40}$$

Let F be a fixed polynomial or rational matrix expression in the weight blocks with query-history-measurable coefficients, and assume that every inverse is well defined. Replacing each weight block by its Gaussian reference gives

$$\operatorname{rank}\big(F(W;\mathscr{T}) - F(G^{\mathrm{ref}};\mathscr{T})\big) \le \sum_e c_e(F)(r_e + s_e), \tag{41}$$

where $c_e(F)$ is the number of occurrences of weight block e.

For a finite family of independent matrices queried in a predictable sequence, the coupling can be chosen jointly. The reference matrices are mutually independent and independent as a family of the query history. The blockwise rank bound is proved in lemma E.1.

The coupling preserves the part of W that the computation has not observed: W and G^{ref} share their $U^\perp \times V^\perp$ block. Their difference is confined to the queried subspaces, which gives the rank bounds in Equations (40) and (41). Joint conditioning makes the same construction valid when queries move between layers.

For later use, write

$$H_n := F(G_n^{\mathrm{ref}};\mathscr{T}_n), \qquad K_n := F(W_n;\mathscr{T}_n).$$

Let N_n be the common dimension of H_n and K_n. The matrix H_n is the Gaussian reference expression, and K_n is the original expression. When $\operatorname{rank}(K_n - H_n) = o(N_n)$, the two matrices have the same limiting empirical spectral distribution, although their largest eigenvalues may differ. For rational expressions, the inverse factors in the original, reference, and limiting systems must satisfy a common uniform bound. If $A_{j,n}$, $A_{j,n}^{\mathrm{ref}}$, and a_j denote the corresponding matrices, we assume

$$\max\big\{\|A_{j,n}^{-1}\|, \|(A_{j,n}^{\mathrm{ref}})^{-1}\|, \|a_j^{-1}\|\big\} \le \delta_j^{-1} \qquad \text{for all sufficiently large } n. \tag{42}$$

For self-adjoint or normal denominators, this condition is equivalent to a uniform gap between zero and the spectrum. For nonnormal denominators, a spectral gap alone does not control the inverse norm.

To recognize a new outlier, we need to know where the reference spectrum ends. A limiting distribution alone would leave the same uncertainty about exceptional eigenvalues that motivated our

original question. We therefore establish strong convergence, which controls operator norms as well as normalized traces. The following network assumptions allow us to determine the reference boundary in this stronger sense.

Consider a fixed-depth, bias-free Gaussian feedforward network. The entries within each weight matrix are i.i.d., and the weights are independent across layers and widths. Assume $n_\ell/n \to c_\ell \in (0,\infty)$, and fix the set $[k] = \{1,\dots,k\}$ of input points. Its activations and derivative gates are

$$\begin{aligned}(W_{\ell,n})_{ij} &\sim \mathcal{N}\left(0, \frac{\sigma_\ell^2}{n_{\ell-1}}\right), \qquad z_\ell^{(a)} = W_{\ell,n} h_{\ell-1}^{(a)}, \\ h_\ell^{(a)} &= \phi_\ell(z_\ell^{(a)}), \qquad\qquad D_\ell^{(a)} = \operatorname{diag}\left(\phi_\ell'(z_\ell^{(a)})\right).\end{aligned} \tag{43}$$

Assume the deterministic inputs satisfy

$$\frac{1}{n_0}\langle h_0^{(a)}, h_0^{(b)}\rangle \longrightarrow Q_0^{ab}. \tag{44}$$

Each activation ϕ_ℓ is either ReLU, with its almost-everywhere derivative gate, or a Lipschitz C^1 function with bounded continuous derivative. Starting from Q_0, define

$$\begin{aligned}Q_1^{ab} &= \sigma_1^2 Q_0^{ab}, \\ Q_{\ell+1}^{ab} &= \sigma_{\ell+1}^2 \mathbb{E}_{g\sim\mathcal{N}(0,Q_\ell)}[\phi_\ell(g_a)\phi_\ell(g_b)], \qquad 1 \le \ell < L.\end{aligned} \tag{45}$$

Assume in addition that the limiting preactivation covariance matrices are uniformly nondegenerate:

$$Q_\ell \succeq \kappa I_k, \qquad \ell = 1,\dots,L, \tag{46}$$

for a constant $\kappa > 0$. Process the forward queries in increasing layer order and the subsequent backward queries in decreasing layer order. Each query is then measurable with respect to the preceding observations. By lemma E.1, the matrices $(G_{\ell,n}^{\mathrm{ref}})_{\ell=1}^L$ can be chosen mutually independent and independent as a family of the complete query history. Define

$$J_n^{\mathrm{ref},(a)} = D_L^{(a)} G_{L,n}^{\mathrm{ref}} \cdots D_1^{(a)} G_{1,n}^{\mathrm{ref}}. \tag{47}$$

Write $J_n^{\mathrm{ref}} = (J_n^{\mathrm{ref},(1)},\dots,J_n^{\mathrm{ref},(k)})$. Hermitian dilation represents each rectangular block and its adjoint in one self-adjoint system of dimension $N_n := \sum_{\ell=0}^L n_\ell$, with block projections marking source and target spaces.

Theorem D.2. *Fix L and k. Assume proportional widths, the input-Gram limit in Equation (44), the activation conditions above, the covariance recursion in Equation (45), and the nondegeneracy condition in Equation (46). Then the Gaussian reference weights, the derivative gates from the original network, and the block projections converge jointly and strongly almost surely to a tracial operator system $(\mathcal{A},\tau)$. Let j^{ref} be the limiting Gaussian reference Jacobian tuple. For every fixed d and $p \in M_d(\mathbb{C})\langle J, J^*\rangle$,*

$$\begin{aligned}(\mathrm{tr}_d \otimes \mathrm{tr}_{N_n})p(J_n^{\mathrm{ref}}, J_n^{\mathrm{ref}*}) &\longrightarrow (\mathrm{tr}_d \otimes \tau)p(j^{\mathrm{ref}}, j^{\mathrm{ref}*}), \\ \|p(J_n^{\mathrm{ref}}, J_n^{\mathrm{ref}*})\| &\longrightarrow \|p(j^{\mathrm{ref}}, j^{\mathrm{ref}*})\|.\end{aligned} \tag{48}$$

For self-adjoint p, the spectra converge in Hausdorff distance. The same conclusion holds for each fixed rational expression whenever every inverted matrix satisfies the uniform inverse bound in Equation (42). The statement includes ReLU derivative gates under Equation (46).

Strong convergence makes the reference edge a spectral boundary: the reference has no eigenvalues a fixed distance beyond the limiting support. The proof establishes joint convergence of the gates and then applies conditional operator-valued strong convergence. Appendix F.4 verifies the assumptions of Jekel et al. (2026, Theorem 4.12).

We can now transfer this boundary to the original matrix. The finite-rank correction may move some leading eigenvalues, but interlacing controls the spectrum below them.

Corollary D.3. *Let $K_n = K_n^* \in \mathbb{C}^{N_n\times N_n}$ arise from a fixed polynomial or rational expression, and let $H_n = H_n^*$ be its Gaussian reference. Suppose H_n converges strongly to $h = h^*$ with limiting spectral law μ_h, and suppose Equation (41) gives $\mathrm{rank}(K_n - H_n) \le r_n = o(N_n)$. Then K_n and H_n have the same limiting empirical spectral distribution. Let $b = \max \mathrm{spec}(h)$. If $\mu_h((b-\varepsilon, b]) > 0$ for every $\varepsilon > 0$, then*

$$\lambda_{r_n+1}(K_n) \longrightarrow b. \tag{49}$$

It remains to locate the exceptional modes themselves. The rank bound tells us how many eigenvalues can move, but their locations depend on how the correction acts on the reference eigenspaces. A resolvent records this dependence by weighting each eigenspace according to its distance from a candidate eigenvalue ζ. For a fixed-rank correction, the matrix determinant identity gives

$$\begin{aligned} K_n &= H_n + U_n C_n U_n^*, \qquad M_n(\zeta) := U_n^*(\zeta I - H_n)^{-1} U_n, \\ \frac{\det(\zeta I - K_n)}{\det(\zeta I - H_n)} &= \det(I - C_n M_n(\zeta)), \qquad \zeta \notin \operatorname{spec}(H_n). \end{aligned} \tag{50}$$

The small matrix $M_n(\zeta)$ retains the projections of the correction directions onto the reference eigenspaces. These are the additional data needed beyond the eigenvalue distribution. The following proposition makes this statement precise: convergence of the directional moments gives a limiting equation for the separated eigenvalues.

Proposition D.4. *Assume $H_n = H_n^*$ converges strongly to $h = h^*$, $U_n \in \mathbb{C}^{N_n \times r}$ has fixed r and uniformly bounded norm, and $C_n = C_n^* \to C = C^*$. Set $K_n = H_n + U_n C_n U_n^*$. If, for every polynomial p,*

$$U_n^* p(H_n) U_n \longrightarrow \Omega(p), \tag{51}$$

then Ω extends continuously from polynomials to $C^*(h)$, and $M_n(\zeta)$ converges locally uniformly off $\operatorname{spec}(h)$ to $\Omega((\zeta - h)^{-1})$. Define the limiting determinant function by

$$D(\zeta) := \det\big(I - C\,\Omega((\zeta - h)^{-1})\big). \tag{52}$$

Each isolated simple real zero of D outside $\operatorname{spec}(h)$ produces one nearby separated eigenvalue of K_n. Conversely, every convergent eigenvalue sequence that stays a positive distance from $\operatorname{spec}(h)$ can accumulate only at a zero of D.

The remaining task is to calculate Ω from the network, so that the outlier equation becomes a prediction in terms of its parameters. Section 3.3 carries this out for two nonlinear Gaussian layers. The calculation also determines the leading eigenvector alignment, which tests whether the exceptional mode is the one identified by the conditioning argument. Appendix G treats higher-rank corrections.

E Gaussian Conditioning and Rank Bounds

E.1 Joint regression under adaptive queries

The joint construction below includes the single-matrix decomposition in proposition D.1 as the case $L = 1$.

Lemma E.1. *For $\ell \in [L]$, let*

$$W_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}, \qquad (W_\ell)_{ij} \sim \mathcal{N}\left(0, \frac{\sigma_\ell^2}{n_{\ell-1}}\right),$$

with entries independent within each block, and assume that the matrices $(W_\ell)_{\ell=1}^L$ are mutually independent. Let $\mathscr{F}_0$ be independent of these matrices. At each time $t \le T$, an $\mathscr{F}_{t-1}$-measurable rule chooses a layer ℓ_t, a query type, and a corresponding $\mathscr{F}_{t-1}$-measurable vector. Define the completed sigma-fields by

$$\mathscr{F}_t = \begin{cases} \sigma(\mathscr{F}_{t-1}, W_{\ell_t} a_t), & \text{for a right query } M \mapsto M a_t, \\ \sigma(\mathscr{F}_{t-1}, W_{\ell_t}^* b_t), & \text{for a left query } M \mapsto M^* b_t. \end{cases}$$

No other weight-dependent information is added. Write $\mathscr{T} := \mathscr{F}_T$ for the complete query history.

For each layer, let V_ℓ and U_ℓ be orthonormal bases measurable with respect to the query history for the spans of its right and left queries. Set

$$Y_\ell := W_\ell V_\ell, \qquad R_\ell := W_\ell^* U_\ell, \qquad r_\ell := \dim V_\ell, \qquad s_\ell := \dim U_\ell,$$

and define

$$W_\ell^{\parallel} := Y_\ell V_\ell^* + U_\ell R_\ell^* (I - V_\ell V_\ell^*).$$

Then, jointly over all layers,

$$(W_\ell)_{\ell=1}^{L} \big| \mathscr{T} \overset{d}{=} \left(W_\ell^{\|} + \Pi_{U_\ell^\perp} \widetilde{W}_\ell \Pi_{V_\ell^\perp} \right)_{\ell=1}^{L}, \tag{53}$$

where the matrices $(\widetilde{W}_\ell)_{\ell=1}^{L}$ are mutually independent, have the respective laws of $(W_\ell)_{\ell=1}^{L}$, and are independent of $\mathscr{T}$.

On an enlarged probability space, there are also matrices $(G_\ell^{\mathrm{ref}})_{\ell=1}^{L}$ with the respective laws of the weights. These matrices are mutually independent and jointly independent of $\mathscr{T}$, and

$$W_\ell = G_\ell^{\mathrm{ref}} + E_\ell, \qquad \mathrm{rank}(E_\ell) \le r_\ell + s_\ell \quad (\ell \in [L]). \tag{54}$$

The Gaussian reference family may remain dependent on the original weight family through the coupling.

Proof. Equip $\mathsf{H}_\ell = \mathbb{R}^{n_\ell \times n_{\ell-1}}$ with the Frobenius inner product multiplied by $n_{\ell-1}/\sigma_\ell^2$. Then $\mathsf{W} = (W_\ell)_{\ell=1}^{L}$ is a standard Gaussian vector in $\mathsf{H} = \bigoplus_\ell \mathsf{H}_\ell$.

Induct on the query number. Conditional on $\mathscr{F}_{t-1}$, the next query is a deterministic linear map supported on one summand. Gaussian regression updates the conditional mean and projects the residual covariance onto the kernel of this new constraint. The conditional law therefore remains Gaussian on the affine space consistent with the observed responses. At the final step its unobserved subspace is

$$\mathsf{N}_{\mathscr{T}} = \bigoplus_{\ell=1}^{L} \Pi_{U_\ell^\perp} \mathsf{H}_\ell \Pi_{V_\ell^\perp}.$$

Let $\mathcal{P}_{\mathscr{T}}$ be its orthogonal projection. The minimum-norm point in the affine constraint space is $\mu_{\mathscr{T}} = (W_\ell^{\|})_\ell$. Indeed, $U_\ell^* Y_\ell = R_\ell^* V_\ell$ shows that $W_\ell^{\|} V_\ell = Y_\ell$ and $(W_\ell^{\|})^* U_\ell = R_\ell$, and $\mathcal{P}_{\mathscr{T}} \mu_{\mathscr{T}} = 0$. Thus

$$\mathsf{W} \mid \mathscr{T} \sim N(\mu_{\mathscr{T}}, \mathcal{P}_{\mathscr{T}}).$$

Its block-diagonal residual covariance proves Equation (53).

Take a standard Gaussian copy $\widetilde{\mathsf{W}}$ independent of the weights and history, and define

$$\mathsf{G}^{\mathrm{ref}} = \mathcal{P}_{\mathscr{T}} \mathsf{W} + (I - \mathcal{P}_{\mathscr{T}}) \widetilde{\mathsf{W}}. \tag{55}$$

Conditional on $\mathscr{T}$, the two terms are independent centered Gaussians on complementary subspaces. For $v \in \mathsf{H}$ their characteristic function is

$$\mathbb{E}[e^{\mathrm{i}\langle v, \mathsf{G}^{\mathrm{ref}} \rangle} \mid \mathscr{T}] = e^{-\|\mathcal{P}_{\mathscr{T}} v\|^2/2} e^{-\|(I-\mathcal{P}_{\mathscr{T}})v\|^2/2} = e^{-\|v\|^2/2}.$$

The conditional law is the original product Gaussian law, independent of the history. This proves both mutual independence of the reference blocks and independence of the entire family from $\mathscr{T}$.

Finally, $\mathsf{W} - \mathsf{G}^{\mathrm{ref}} = (I - \mathcal{P}_{\mathscr{T}})(\mathsf{W} - \widetilde{\mathsf{W}})$. On block ℓ,

$$(I - \mathcal{P}_{\mathscr{T},\ell}) M = M V_\ell V_\ell^* + U_\ell U_\ell^* M (I - V_\ell V_\ell^*),$$

whose two terms have ranks at most r_ℓ and s_ℓ. This gives Equation (54). The reference and original weights share their unqueried blocks through Equation (55). □

E.2 Rank bounds for matrix expressions

The proof follows the recursive construction of F. A weight block uses Equation (54). Left and right multiplication preserve or decrease rank. For a product,

$$\begin{aligned} AB - A^{\mathrm{ref}} B^{\mathrm{ref}} &= (A - A^{\mathrm{ref}}) B + A^{\mathrm{ref}} (B - B^{\mathrm{ref}}), \\ \mathrm{rank}(AB - A^{\mathrm{ref}} B^{\mathrm{ref}}) &\le \mathrm{rank}(A - A^{\mathrm{ref}}) + \mathrm{rank}(B - B^{\mathrm{ref}}). \end{aligned} \tag{56}$$

For an inverse, whenever both matrices are invertible,

$$\begin{aligned} A^{-1} - (A^{\mathrm{ref}})^{-1} &= A^{-1} (A^{\mathrm{ref}} - A) (A^{\mathrm{ref}})^{-1}, \\ \mathrm{rank}\big(A^{-1} - (A^{\mathrm{ref}})^{-1}\big) &\le \mathrm{rank}(A - A^{\mathrm{ref}}). \end{aligned} \tag{57}$$

Finite-n invertibility suffices for this rank identity. Passing strong convergence through $x \mapsto x^{-1}$ additionally uses a uniform inverse bound. Addition uses rank subadditivity, and taking adjoints preserves rank. This proves proposition D.1.

F Proof of Theorem D.2

We prove Theorem D.2 by controlling the coefficients left after conditioning on the network computation. These coefficients are the derivative gates and the projections onto layer spaces. We establish their trace and norm limits, then verify the covariance conditions of the Gaussian operator-valued theorem of Jekel et al. (2026). This determines the reference spectrum. The transfer to the original matrix uses the interlacing argument in Appendix A.

F.1 Block matrix representation

Embed the layer spaces into

$$\mathcal{H}_n = \bigoplus_{\ell=0}^{L} \mathbb{R}^{n_\ell},$$

set $N_n := \dim \mathcal{H}_n$, and let $e_{\ell,n}$ be the projection onto the ℓth block. For the almost-sure triangular-array realization, take the original weight families independently over width indices and let each width-n query history use only its width-n weights. Apply lemma E.1 at every width and take the product of the enlarged probability spaces. The resulting Gaussian reference array $(G^{\mathrm{ref}}_{\ell,n})_{\ell,n}$ is independent of the full sequence of query histories and has the product law of the original weights over both indices. For $G_{\ell,n} := G^{\mathrm{ref}}_{\ell,n}$, use the Hermitian dilation

$$X_{\ell,n} = e_{\ell,n} G_{\ell,n} e_{\ell-1,n} + e_{\ell-1,n} G^*_{\ell,n} e_{\ell,n}. \tag{58}$$

Conditional on the complete query history, $(X_{\ell,n})_{\ell=1}^{L}$ is a jointly centered Gaussian family with the product law of the independent layer dilations. Define the layer-diagonal algebra

$$\mathcal{D}^{\mathrm{lay}}_n := \left\{ b \in M_{N_n} : b = \sum_{r=0}^{L} e_{r,n} b e_{r,n} \right\} = \bigoplus_{r=0}^{L} e_{r,n} M_{N_n} e_{r,n}.$$

Conditioned on the realized query history, the coefficient algebra is

$$\mathcal{B}_n = C^*\Big(\{e_{\ell,n}\}_{\ell=0}^{L}, \{D^{(a)}_{\ell,n}\}_{\ell,a} \Big) \subseteq \mathcal{D}^{\mathrm{lay}}_n.$$

The fixed inputs enter through the covariance recursion and the realized gates. Fixed matrix-valued coefficients are handled by $M_d(\mathbb{C}) \otimes \mathcal{B}_n$ and remain within the layer-diagonal base algebra. The joint Gaussian reference family is independent of this algebra.

For the original network, define the empirical preactivation covariance

$$Q^{ab}_{\ell,n} := \frac{\sigma^2_\ell}{n_{\ell-1}} \langle h^{(a)}_{\ell-1}, h^{(b)}_{\ell-1} \rangle.$$

Conditional on the previous layer, the row tuples

$$\mathbf{z}_{\ell,i,n} := \left(z^{(1)}_{\ell,i}, \dots, z^{(k)}_{\ell,i} \right)$$

are i.i.d. $\mathcal{N}(0, Q_{\ell,n})$.

Lemma F.1. *Under the assumptions of Theorem D.2, there are deterministic matrices $Q_1, \dots, Q_L$ such that*

$$Q_{\ell,n} \longrightarrow Q_\ell \qquad \text{a.s.}, \tag{59}$$

for every fixed layer. They satisfy

$$Q^{ab}_{\ell+1} = \sigma^2_{\ell+1} \mathbb{E}\Big[\phi_\ell(Z^{(a)}_\ell) \phi_\ell(Z^{(b)}_\ell) \Big], \qquad Z_\ell \sim \mathcal{N}(0, Q_\ell). \tag{60}$$

Proof. The base case is Equation (44). Conditional on the previous layer, the preactivation tuples are i.i.d. Gaussian with uniformly bounded covariance. Since ϕ_ℓ is Lipschitz, each activation coordinate is uniformly sub-Gaussian, and each product $\phi_\ell(z^{(a)}_{\ell,i}) \phi_\ell(z^{(b)}_{\ell,i})$ is uniformly sub-exponential. Conditional Bernstein bounds, followed by Borel–Cantelli over the finitely many input pairs, give convergence of every empirical covariance entry to its conditional expectation. Continuity of Gaussian expectations in the covariance matrix closes the induction. □

F.2 Strong convergence of the gate algebra

For a polynomial r in the k gate coordinates at layer ℓ, set

$$f_{\ell,r}(x_1,\dots,x_k) := r(\phi'_\ell(x_1),\dots,\phi'_\ell(x_k))\,.$$

Under the smooth-gate assumptions this function is bounded and continuous.

Lemma F.2. *For every fixed ℓ and every such r,*

$$\frac{1}{n_\ell}\sum_{i=1}^{n_\ell} f_{\ell,r}(\mathbf{z}_{\ell,i,n}) \longrightarrow \mathbb{E} f_{\ell,r}(Z_\ell) \qquad \text{a.s.}, \tag{61}$$

and

$$\|\mathrm{diag}(f_{\ell,r}(\mathbf{z}_{\ell,1,n}),\dots,f_{\ell,r}(\mathbf{z}_{\ell,n_\ell,n}))\| \longrightarrow \sup_{x\in\mathbb{R}^k}|f_{\ell,r}(x)|. \tag{62}$$

Because $Q_\ell \succeq \kappa I_k$, the right-hand side equals the essential supremum under $\mathcal{N}(0,Q_\ell)$. Consequently, on an event of probability one for the query history, the block projections and gate diagonals converge strongly as a deterministic coefficient family.

Proof. For the trace, conditional Hoeffding bounds are summable in n because $f_{\ell,r}$ is bounded. The conditional expectation converges by Equation (59) and weak continuity of Gaussian laws. A countable dense polynomial family and uniform boundedness give these trace limits simultaneously for all fixed polynomials.

Let ν_ℓ be the limiting gate distribution and

$$\mathcal{K}_\ell = \overline{\{(\phi'_\ell(x_1),\dots,\phi'_\ell(x_k)) : x\in\mathbb{R}^k\}}.$$

Boundedness makes $\mathcal{K}_\ell$ compact. Continuity and the full support of $\mathcal{N}(0,Q_\ell)$ give $\operatorname{supp}\nu_\ell = \mathcal{K}_\ell$. Every finite-width gate tuple lies in this set. For $R_n = r(D^{(1)}_{\ell,n},\dots,D^{(k)}_{\ell,n})$ and $M_r = \max_{\mathcal{K}_\ell}|r|$, the deterministic upper bound and the trace lower bound give, for every fixed integer $j\ge 1$,

$$\|R_n\| \le M_r, \qquad \liminf_n \|R_n\| \ge \left[\int_{\mathcal{K}_\ell} |r(x)|^{2j}\,\mathrm{d}\nu_\ell(x)\right]^{1/(2j)}.$$

Letting $j\to\infty$ after the trace limit makes the right-hand side tend to $\|r\|_{L^\infty(\nu_\ell)} = M_r$, proving Equation (62). For a fixed matrix-valued polynomial, use the same support bound and $(\mathrm{tr}_d\otimes\mathrm{tr}_{\ell,n})[(R_n^*R_n)^j]$. Faithfulness of the limiting matrix trace gives the same norm conclusion. □

For ReLU derivative gates, $\phi'(x) = \mathbf{1}_{\{x>0\}}$. The condition $Q_\ell \succeq \kappa I_k$ gives positive probability to every orthant. Here $\mathcal{K}_\ell = \{0,1\}^k$ and $\operatorname{supp}\nu_\ell = \mathcal{K}_\ell$. Orthant boundaries have Gaussian measure zero, so the trace limits hold as above, and the same moment lower bound proves norm convergence.

The mutually orthogonal layer projections decompose every coefficient polynomial as a direct sum of layer polynomials. This gives strong convergence of the full coefficient tuple. Its normalized trace and norm are, respectively,

$$\sum_{\ell=0}^{L}\frac{n_\ell}{N_n}\,\mathrm{tr}_{\ell,n}(\,\cdot\,), \qquad \max_{0\le\ell\le L}\|\cdot\|_\ell.$$

The width-ratio limits and lemma F.2 give convergence of both quantities. The matrix-valued argument in the lemma covers every fixed $M_d(\mathbb{C})$ amplification.

F.3 Rectangular Gaussian matrices as GOE corners

Let S_n be a standard GOE matrix on $\mathcal{H}_n$: its upper-triangular entries are independent, with variance $1/N_n$ off the diagonal and $2/N_n$ on the diagonal. The reference weights can be represented exactly as

$$G^{\mathrm{ref}}_{\ell,n} = \sigma_\ell\sqrt{\frac{N_n}{n_{\ell-1}}}\,e_{\ell,n}S_n e_{\ell-1,n}. \tag{63}$$

The variance of each corner entry is $\sigma_\ell^2/n_{\ell-1}$. Different layers of the feedforward chain use different unordered block pairs, hence disjoint independent GOE entries. To preserve the coupling in lemma E.1, fill these corners with the already constructed reference weights, scaled as in Equation (63), and complete the remaining entries with independent Gaussians. The resulting S_n is independent of the full query history, while the original weight decomposition remains unchanged. For a general family with repeated unordered block pairs, use a separate independent GOE matrix for each weight.

Lemma F.3. *Let $\mathcal{H}_n = \bigoplus_{j=0}^{L} \mathbb{R}^{n_j}$ with $n_j/n \to c_j > 0$. Let b_n be a deterministic finite block-diagonal tuple containing the block projections and converging strongly to a tuple in a tracial algebra with faithful trace. Adjoin a fixed finite family of independent rectangular Gaussian matrices between distinct blocks, each with i.i.d. entries of variance $\sigma_e^2/n_{\mathrm{source}(e)}$, where $\sigma_e > 0$ is fixed. Their Hermitian dilations and b_n converge jointly and strongly almost surely. The statement includes every fixed matrix amplification.*

Proof. We first adjoin independent standard GOE matrices $S_{1,n}, \dots, S_{q,n}$ to the coefficient tuple. For one such matrix, Gaussian contraction gives

$$\eta_n(A) := \mathbb{E}[S_n A S_n] = \mathrm{tr}_{N_n}(A) I + \frac{A^{\mathsf{T}}}{N_n}, \qquad \eta_n^0(A) := \mathrm{tr}_{N_n}(A) I. \tag{64}$$

Mixed covariance maps vanish. The maps η_n are completely positive, trace-symmetric, and uniformly bounded. Transposition preserves operator norm, so $\|\eta_n - \eta_n^0\|_{\mathrm{op}\to\mathrm{op}} = N_n^{-1}$. For a fixed covariance polynomial F, induction over its additions, products, adjoints and covariance-map applications gives

$$\|F(\eta_n, b_n) - F(\eta_n^0, b_n)\| = O_F(N_n^{-1}), \tag{65}$$

with the diagonal covariance maps understood for the independent family. All intermediate expressions are uniformly bounded. Each replacement contributes $O(N_n^{-1})$, and F has fixed depth. The expression $F(\eta_n^0, b_n)$ uses only ordinary polynomials and normalized traces. Strong convergence of b_n therefore gives its trace and norm limits, with covariance $\eta_{ij}(a) = \delta_{ij}\tau(a)1$. This proves strong covariance-law convergence without a rate assumption on b_n.

For the Choi convention used in Jekel et al. (2026), let F_{N_n} be the tensor flip. Then

$$T_n = \sum_{i,j} \eta_n(E_{ij}) \otimes E_{ij} = \frac{I + \mathsf{F}_{N_n}}{N_n}, \qquad \|T_n\| = \frac{2}{N_n}. \tag{66}$$

For every fixed independent GOE subfamily, the Choi matrix is block diagonal with these blocks. Hence $(\log N_n)^3 \|T_{F,n}\| \to 0$. The hypotheses of Jekel et al. (2026, Theorem 4.12) now hold, giving joint strong convergence of b_n and the GOE family to b and semicircular elements free from b and from one another.

Each rectangular matrix is a corner of one of these GOE matrices. Corner extraction and Hermitian dilation are fixed polynomial operations, and the factors $\sigma_e\sqrt{N_n/n_{\mathrm{source}(e)}}$ converge to positive constants. This proves the asserted joint strong convergence. Fixed amplifications follow from the matrix-valued form of strong convergence. □

The corner representation also identifies the operator-valued limit. Write $c_\Sigma = \sum_j c_j$, $g_\ell = \sigma_\ell\sqrt{c_\Sigma/c_{\ell-1}}\, e_\ell s e_{\ell-1}$ and $x_\ell = g_\ell + g_\ell^*$, using the single semicircular element for the chain. On the limiting block-diagonal coefficient algebra $\mathcal{B}$, one has $E_{\mathcal{B}}(sas) = \tau(a)1$. Thus, with $\tau_j(b) = \tau(e_j b e_j)/\tau(e_j)$,

$$E_{\mathcal{B}}(x_\ell b x_j) = \delta_{\ell j}\left[\sigma_\ell^2 \tau_{\ell-1}(b) e_\ell + \sigma_\ell^2 \frac{c_\ell}{c_{\ell-1}} \tau_\ell(b) e_{\ell-1}\right]. \tag{67}$$

The mixed terms vanish because the block pairs are distinct. These $\mathcal{B}$-linear combinations of s form a $\mathcal{B}$-valued semicircular family with the displayed covariance. This is the same covariance governing the rectangular reference model and its scalar equations below.

Corollary F.4. *Let $T_n \in \mathbb{R}^{N_n \times N_n}$ be positive semidefinite and converge strongly almost surely to a compactly supported law ν. Let $R_n = G_n/\sqrt{N_n} \in \mathbb{R}^{M_n \times N_n}$, where the Gaussian family*

$(G_n)_n$ has independent standard entries and is independent of the full sequence $(T_n)_n$. Suppose $N_n \asymp n$ and $M_n/N_n \to \gamma > 0$. Then $H_n = R_n T_n R_n^*$ converges strongly almost surely. Its limiting resolvent trace, normalized by M_n, satisfies

$$z = \frac{1}{m(z)} + \int \frac{t}{1 - \gamma t m(z)} \, \mathrm{d}\nu(t), \qquad m(z) \sim z^{-1}. \tag{68}$$

The branch has $\Im m(z) < 0$ for $\Im z > 0$.

Proof. Condition on a sequence of T_n with strong convergence. Functional calculus gives strong convergence of $T_n^{1/2}$, so lemma F.3 applies to the two block projections, $T_n^{1/2}$ and R_n, using N_n as the width scale. The product H_n therefore converges strongly. Conditioning and Fubini give the joint conclusion.

For the Hermitian dilation of $R_n T_n^{1/2}$, the two block resolvents satisfy the operator-valued semicircular Dyson equation

$$r(\zeta) = \left[\zeta - \int t \, s(t, \zeta) \, \mathrm{d}\nu(t) \right]^{-1}, \qquad s(t, \zeta) = \frac{1}{\zeta - \gamma t r(\zeta)}.$$

The Schur complement gives $r(\zeta) = \zeta m(\zeta^2)$. Substitution yields Equation (68), with the branch selected at infinity. □

Write $b_\nu = \max \operatorname{supp} \nu$ and, for $a > b_\nu$,

$$\Psi_{\gamma,\nu}(a) = a \left[\gamma + \int \frac{t}{a - t} \, \mathrm{d}\nu(t) \right], \qquad Q_\nu(a) = \int \frac{t^2}{(a - t)^2} \, \mathrm{d}\nu(t).$$

With $a = 1/(\gamma m(z))$, Equation (68) gives $z = \Psi_{\gamma,\nu}(a)$ and $\Psi'_{\gamma,\nu}(a) = \gamma - Q_\nu(a)$. If $Q_\nu(b_\nu^+) > \gamma$, the branch from infinity first turns at the unique $a_* > b_\nu$ with $Q_\nu(a_*) = \gamma$. This gives the upper edge $\Psi_{\gamma,\nu}(a_*)$ (Silverstein & Choi, 1995). Since $\Psi''_{\gamma,\nu}(a_*) > 0$, it is a regular square-root edge. The populations δ_1, $\mu_{\mathrm{MP},c}$ and the shifted MP mixture in Appendix I all have $Q_\nu(b_\nu^+) = \infty$.

F.4 Application of the operator-valued strong-convergence theorem

Let $\mathscr{F}_{\mathrm{tr}}$ be generated by the full sequence of query-history coefficients. On the probability-one event in lemma F.2, conditioning on $\mathscr{F}_{\mathrm{tr}}$ makes the gates and block projections a deterministic strongly convergent tuple. By lemma E.1, the reference weights retain their independent Gaussian product law. The GOE-corner proof of lemma F.3 applies with the covariance-law and Choi conditions verified above. Thus, conditionally almost surely, for every fixed matrix size d and polynomial p,

$$\begin{aligned} (\mathrm{tr}_d \otimes \mathrm{tr}_{N_n}) p(X_{1,n}, \ldots, X_{L,n}; \mathcal{B}_n) &\longrightarrow (\mathrm{tr}_d \otimes \tau) p(x_1, \ldots, x_L; \mathcal{B}), \\ \| p(X_{1,n}, \ldots, X_{L,n}; \mathcal{B}_n) \| &\longrightarrow \| p(x_1, \ldots, x_L; \mathcal{B}) \|, \end{aligned} \tag{69}$$

Let E be the common convergence event for all positive integers d and a countable dense family of matrix-valued polynomials. Regular conditional probabilities and Fubini give

$$\mathbb{P}(E \mid \mathscr{F}_{\mathrm{tr}}) = 1 \quad \text{almost surely}, \qquad \mathbb{P}(E) = \mathbb{E}[\mathbb{P}(E \mid \mathscr{F}_{\mathrm{tr}})] = 1.$$

Uniform norm bounds extend convergence to every fixed polynomial. Applying this to the Gaussian reference Jacobians gives Equation (48) and completes the proof of Theorem D.2.

F.5 Rational matrix expressions

For the strongly convergent Gaussian reference system, an invertible limiting denominator already ensures eventual uniform invertibility. Indeed, suppose a square reference denominator A_n converges strongly to an invertible a. Choose $M > \|a\|$. Then $q = \|I - a^* a / M^2\| < 1$, and strong convergence gives $\|I - A_n^* A_n / M^2\| \le q_0 < 1$ eventually. Thus

$$A_n^{-1} = \frac{1}{M^2} \sum_{j=0}^{\infty} \left(I - \frac{A_n^* A_n}{M^2} \right)^j A_n^*.$$

If $C = \sup_{n \geq n_0} \|A_n\|$, truncation after J has error at most $Cq_0^{J+1}/[M^2(1-q_0)]$, uniformly for $n \geq n_0$. First apply polynomial strong convergence at fixed J, then let $J \to \infty$. Induction over sums, products, adjoints and square inverses proves joint strong convergence for each fixed rational expression whose limiting denominators are invertible.

For the original system, finite-width invertibility suffices for rank propagation in proposition D.1. Uniform inverse bounds are still needed when controlling its norms and fixed normalized moments, as required in Equation (42). A rank-one change from I_n to $\mathrm{diag}(n^{-1}, 1, \ldots, 1)$ preserves the reference limit but makes the original inverse norm equal to n.

G Finite-rank perturbations and extensions

G.1 Finite-rank determinant reduction

Proof of proposition D.4. Let $C_U = \sup_n \|U_n\|$. Strong convergence and the directional moment limits imply

$$\|\Omega(p)\| \leq C_U^2 \lim_n \|p(H_n)\| = C_U^2 \|p(h)\|.$$

Thus $p(h) = 0$ implies $\Omega(p) = 0$, and Ω extends uniquely and continuously to $C^*(h)$.

Fix a compact set $\mathcal{K} \subset \mathbb{C} \setminus \mathrm{spec}(h)$. Strong convergence permits a compact neighborhood $\mathcal{S}$ of $\mathrm{spec}(h)$ with $\mathrm{dist}(\mathcal{K}, \mathcal{S}) > 0$ and $\mathrm{spec}(H_n) \subset \mathcal{S}$ for all sufficiently large n.

For each $\zeta \in \mathcal{K}$, the function $f_\zeta(x) := (\zeta - x)^{-1}$ is continuous on $\mathcal{S}$. Polynomial approximation on the real compact set $\mathcal{S}$ gives, for every $\varepsilon > 0$, a polynomial $p_{\zeta,\varepsilon}$ such that

$$\sup_{x \in \mathcal{S}} \left| (\zeta - x)^{-1} - p_{\zeta,\varepsilon}(x) \right| \leq \varepsilon.$$

The uniform bound on U_n and the directional moment hypothesis in Equation (51) then give

$$\begin{aligned} M_n(\zeta) &= U_n^* p_{\zeta,\varepsilon}(H_n) U_n + O(\varepsilon) \\ &\longrightarrow \Omega(p_{\zeta,\varepsilon}) + O(\varepsilon) \\ &= \Omega((\zeta - h)^{-1}) + O(\varepsilon). \end{aligned}$$

Resolvent Lipschitz bounds and a finite covering of $\mathcal{K}$ make the convergence locally uniform in ζ.

Set

$$D_n(\zeta) := \det(I - C_n M_n(\zeta)), \qquad D(\zeta) := \det\big(I - C\,\Omega((\zeta - h)^{-1})\big)\,.$$

The determinant identity in Equation (50) and the locally uniform convergence above imply $D_n \to D$ away from $\mathrm{spec}(h)$. If ζ_0 is an isolated simple zero of D, choose a small disk around ζ_0 disjoint from $\mathrm{spec}(h)$ and containing no other zero. Rouché's theorem gives exactly one zero of D_n in that disk for all large n, and the determinant identity identifies it with a separated eigenvalue of K_n. Conversely, every convergent sequence of eigenvalues that remains a positive distance from $\mathrm{spec}(h)$ has a limit satisfying $D(\zeta) = 0$. This proves the stated separated-eigenvalue correspondence. □

The first extension assumes locally uniform control of a compressed resolvent. The second assumes jointly Gaussian forward and backward feature vectors.

G.2 Growing but sublinear perturbation rank

Let $U_n \in \mathbb{C}^{N_n \times r_n}$ and

$$K_n = H_n + U_n C_n U_n^*, \qquad r_n \longrightarrow \infty, \qquad \frac{r_n}{N_n} \longrightarrow 0,$$

and define, on any analytic domain avoiding the relevant spectra,

$$M_n(\zeta) := U_n^*(\zeta I - H_n)^{-1} U_n, \qquad \mathcal{L}_n(\zeta) := \frac{1}{r_n} \log\det(I - C_n M_n(\zeta))\,.$$

On a branch where $I - C_n M_n(\zeta)$ is invertible, differentiation of the finite-rank determinant identity gives the exact relation

$$\begin{aligned}\mathcal{L}_n'(\zeta) &= \frac{1}{r_n}\operatorname{Tr}\big[(I - C_n M_n(\zeta))^{-1}\big(-C_n M_n'(\zeta)\big)\big] \\ &= \frac{1}{r_n}\left[\operatorname{Tr}(\zeta I - K_n)^{-1} - \operatorname{Tr}(\zeta I - H_n)^{-1}\right].\end{aligned} \tag{70}$$

Suppose on an open set $\mathcal{O}$ away from the relevant spectra that $I - C_n M_n(\zeta)$ is locally uniformly invertible and that $\mathcal{L}_n$ converges locally uniformly to an analytic function $\mathcal{L}$. Analytic convergence then permits differentiation, and Equation (70) shows that the normalized resolvent-trace difference converges locally uniformly to $\mathcal{L}'(\zeta)$.

G.3 Gaussian parameter-feature covariance

For jointly centered real Gaussian forward and backward states a and d, define $\Sigma_A := \mathbb{E}[aa^*]$, $\Sigma_D := \mathbb{E}[dd^*]$, $C := \mathbb{E}[ad^*]$, and the centered parameter feature $\xi := a \otimes d - \mathbb{E}[a \otimes d]$. In this real setting, the adjoint is the transpose. Identifying a parameter-pair vector with a matrix argument X, Wick's formula gives the covariance map

$$\eta_{\text{pair}}(X) = \Sigma_A X \Sigma_D + C X^* C. \tag{71}$$

This exact identity gives the covariance of a dense parameter block whenever its forward and backward feature vectors are jointly Gaussian.

H Proofs for two-layer Gaussian networks

This appendix proves Theorems 3.2 and 3.3. We first calculate the directional moments for the input Jacobian, then derive the parameter NTK and its stability threshold.

The two-layer calculation has one additional dependence to resolve: a perturbation direction contains the Gaussian reference itself. We first derive the conditional matrix representation and its product law. Computing the spectral weights of both perturbation directions then gives the determinant and overlap formulas for the input Jacobian. The same representation reduces the parameter NTK to a shifted Wishart matrix with a rank-one correction.

H.1 Network model

We use the model in Equation (4), with $m/n \to \alpha > 0$ and $\ell/m \to \beta > 0$. The standard real Gaussian matrices $G_1 \in \mathbb{R}^{m\times n}$ and $G_2 \in \mathbb{R}^{\ell\times m}$ are independent across layers and widths.

Here h_0 is deterministic with $\|h_0\|^2 = n$, A is a finite union of intervals with $p := \mathbb{P}(Z \in A) \in (0,1)$ for $Z \sim \mathcal{N}(0,1)$, and $\phi_A(t) = \int_0^t \mathbf{1}_A(u)\,\mathrm{d}u$. Endpoints have Gaussian probability zero. Thus both activations are nonlinear and Lipschitz. The deterministic scaling $\sqrt{2/m}$ makes the second preactivation variance converge to one. Write

$$\begin{gathered}D_1 = \operatorname{diag}(\mathbf{1}_{z_1>0}), \quad D_2 = \operatorname{diag}(\mathbf{1}_{z_2\in A}), \quad q = \mathbb{E}[Z^2\mathbf{1}_A(Z)], \\ c = \alpha/2, \quad d = \alpha\beta p, \quad \theta = 2\beta q.\end{gathered}$$

The quantities p, q refer to the limiting standard Gaussian preactivation. The input Jacobian is $J = D_2 W_2 D_1 W_1$, and $K = JJ^*$. We study one fixed input and ℓ output coordinates using the conditional Gaussian representation below.

H.2 Conditional Gaussian decomposition

Rotate $h_0/\sqrt{n}$ to e_1, let $I_1 = \{i : z_{1,i} > 0\}$, $k = |I_1|$, and $r = \|h_1\|$. Almost surely, $k/n \to c$ and $r^2/n \to c$. Conditional on z_1, rotate the k active hidden coordinates so that $h_1|_{I_1} = re_1$. The rotated first active weight block has the exact distribution

$$W_1|_{I_1} \simeq \frac{1}{\sqrt{n}}\begin{bmatrix} r & \xi^* \\ 0 & X\end{bmatrix}, \qquad \xi \in \mathbb{R}^{n-1}, \quad X \in \mathbb{R}^{(k-1)\times(n-1)}, \tag{72}$$

where ξ, X have independent standard Gaussian entries. The identity $h_1 = D_1 z_1$ places the first active preactivation in the same direction as the hidden activation.

In the same hidden basis the second active-column weight block is $\sqrt{2/m}[g\ Y_{\text{all}}]$, with independent standard Gaussian entries, independent of ξ, X, z_1. Set $v_n = 2r^2/m \to 1$. The second preactivation is $\sqrt{v_n}g$, so the active output set is $I_2 = \{i : \sqrt{v_n}g_i \in A\}$, of size s, and $s/n \to d$ almost surely. Conditioning on z_1, g fixes I_1, I_2 and leaves $\xi, X, Y := Y_{\text{all}}|_{I_2}$ mutually independent Gaussian arrays. On the active outputs, the original Jacobian is therefore

$$J = Z_n + a_n b_n^*, \qquad Z_n = \left[0\ \sqrt{\frac{2}{mn}} YX\right], \quad a_n = \sqrt{\frac{2}{m}} g|_{I_2}, \quad b_n = \frac{1}{\sqrt{n}}\begin{bmatrix} r \\ \xi \end{bmatrix}. \tag{73}$$

Inactive output rows are zero. In particular $\|a_n\|^2 \to \theta$ and $\|b_n\|^2 \to c+1$ almost surely. The direction of a_n is exactly that of $D_2 z_2$ on its active rows.

These representations are used at sufficiently large widths, when $k, s \geq 2$ almost surely. Arbitrary conventions at the finitely many exceptional initial widths do not affect any limit.

Let $H_n = Z_n Z_n^*$, $t_n = Z_n b_n$, and $U_n = [a_n\ t_n]$. Then

$$K = H_n + U_n C_n U_n^*, \qquad C_n = \begin{bmatrix} \|b_n\|^2 & 1 \\ 1 & 0 \end{bmatrix} \longrightarrow C = \begin{bmatrix} c+1 & 1 \\ 1 & 0 \end{bmatrix}. \tag{74}$$

The direction t_n depends on H_n through Z_n. This dependence leads to the factor x in its directional spectral measure in Equation (77).

H.3 Limiting spectrum of the Gaussian matrix product

Denote by $\mu_{c,d}$ the active-row limiting law of H_n, and set $m(z) = \int (z-x)^{-1}\, \mathrm{d}\mu_{c,d}(x)$. Define

$$\Lambda(w) = \frac{(1+w)(c+w)(d+w)}{cw}, \qquad \Lambda'(w) = \frac{2w+1+c+d-cd/w^2}{c}. \tag{75}$$

There is a unique positive w_* solving $\Lambda'(w_*) = 0$, equivalently $2w_*^3 + (1+c+d)w_*^2 - cd = 0$. The upper edge is $B = \Lambda(w_*)$. For $z > B$, the physical branch is the unique $w = w(z) \in (0, w_*)$ with $z = \Lambda(w)$, and

$$m(z) = \frac{d + w(z)}{dz}, \qquad w(z) = d \int \frac{x}{z-x}\, \mathrm{d}\mu_{c,d}(x). \tag{76}$$

To derive these formulas, write the exact reference Gram matrix as

$$H_n = \frac{2(k-1)(n-1)}{mn} \left(\frac{Y}{\sqrt{k-1}}\right)\left(\frac{XX^*}{n-1}\right)\left(\frac{Y^*}{\sqrt{k-1}}\right).$$

Conditional on the histories, X, Y are independent Gaussian matrices. The middle factor converges strongly to $\mu_{\text{MP},c}$, by corollary F.4 with population δ_1. Apply that corollary again with population $\mu_{\text{MP},c}$ and aspect ratio $\gamma = d/c$. The leading scalar tends to one, so H_n converges strongly. Conditioning and Fubini give the joint almost-sure conclusion.

Set $a = 1/(\gamma m(z))$ and $u = \int x/(a-x)\, \mathrm{d}\mu_{\text{MP},c}(x)$. The Gaussian covariance equation and the MP quadratic equation give, respectively,

$$z = a(\gamma + u), \qquad a = \frac{(1+u)(1+cu)}{u}.$$

Substituting $w = cu$ yields Equations (75) and (76); in particular $\int x\, \mathrm{d}\mu_{c,d}(x) = 1$. Since $Q_{\mu_{\text{MP},c}}(b^+_{\text{MP},c}) = \infty$, the branch from $w = 0$ first turns at w_*. Thus $\lambda_1(H_n) \to B = \Lambda(w_*)$, and the limiting law has mass arbitrarily close to B.

H.4 Directional moments

Lemma H.1. *For the network-generated U_n in Equation (74), almost surely for every polynomial P,*

$$U_n^* P(H_n) U_n \longrightarrow \Omega(P) := \begin{bmatrix} \theta \int P(x)\, \mathrm{d}\mu_{c,d}(x) & 0 \\ 0 & d \int xP(x)\, \mathrm{d}\mu_{c,d}(x) \end{bmatrix}. \tag{77}$$

Consequently, locally uniformly outside $\operatorname{supp} \mu_{c,d}$,

$$U_n^*(zI - H_n)^{-1}U_n \longrightarrow \begin{bmatrix} \theta m(z) & 0 \\ 0 & w(z) \end{bmatrix}. \tag{78}$$

In Equation (78), off the upper real branch $w(z)$ *denotes its analytic continuation* $d(zm(z) - 1)$.

Proof. Condition on z_1, g so that a_n and the dimensions are fixed. The finite-union boundary condition, $v_n \to 1$, and Gaussian concentration give $s/n \to d$ and $\|a_n\|^2 \to \theta$ almost surely: the conditional expectations are $\mathbb{P}(\sqrt{v_n}Z \in A)$ and $\mathbb{E}[Z^2 \mathbf{1}_A(\sqrt{v_n}Z)]$, which tend to p, q by dominated convergence. The corresponding empirical deviations have summable exponential tails.

Left orthogonal invariance of Y makes H_n orthogonally invariant, independently of a_n. Haar quadratic-form concentration and the strong product law give

$$a_n^* P(H_n) a_n - \|a_n\|^2 \operatorname{tr}_s P(H_n) \longrightarrow 0.$$

For the second direction, use $Z_n e_1 = 0$. Write $\widetilde{Z}_n$ for Z_n without its zero first column. Then $t_n = \widetilde{Z}_n \xi/\sqrt{n}$, with ξ independent of $\widetilde{Z}_n, a_n$, so $\operatorname{Cov}(t_n \mid a_n, Z_n) = H_n/n$. Conditional Gaussian quadratic-form concentration gives

$$t_n^* P(H_n) t_n - \frac{1}{n} \operatorname{Tr}[H_n P(H_n)] \longrightarrow 0.$$

The second trace converges to $d \int x P(x) \, \mathrm{d}\mu_{c,d}(x)$. Conditional on $a_n, \widetilde{Z}_n$, the cross term $a_n^* P(H_n) \widetilde{Z}_n \xi/\sqrt{n}$ is centered Gaussian (or its real and imaginary parts are), with variance bounded by $n^{-1}\|a_n\|^2 \|P(H_n)\|^2 \|H_n\| = O(n^{-1})$ on the eventual norm-bounded event. Its tails are summable. The same norm bounds give summable Haar and quadratic-form tails for each fixed polynomial. Apply Borel–Cantelli on a countable polynomial basis and then linearity, followed by Fubini over the histories. This proves Equation (77). Also $\|t_n\| \le \|Z_n\| \|b_n\|$ is eventually bounded. Polynomial approximation using strong convergence, as in proposition D.4, now gives the locally uniform resolvent limit and its derivatives. □

H.5 Outlier location and eigenvector overlap

Proof of Theorem 3.2. The rank-two identity and the strong product law give the empirical limit. Since $\|W_1\|, \|W_2\|$ are eventually bounded and $\|D_i\| \le 1$, so is $\|K\|$. All fixed normalized moments therefore converge by the compact-support observation in Appendix A.

Since $\operatorname{rank}(J - Z_n) \le 1$, lemma A.4 gives

$$\lambda_3(H_n) \le \lambda_2(K) \le \lambda_1(H_n).$$

Strong convergence and endpoint mass imply $\lambda_2(K) \to B$. Thus there is at most one upper outlier and $\liminf \lambda_1(K) \ge B$.

Set $M_n(z) = U_n^*(zI - H_n)^{-1}U_n$ and $D_n(z) = \det(I - C_n M_n(z))$. Woodbury gives $U_n^*(zI - K)^{-1}U_n = M_n(I - C_n M_n)^{-1}$. Expanding its $(1,1)$ entry yields the exact identities

$$\begin{aligned} D_n &= (1 - M_{12,n})(1 - M_{21,n}) - M_{11,n}(\|b_n\|^2 + M_{22,n}), \\ g_n(z) &:= \widehat{a}_n^*(zI - K)^{-1}\widehat{a}_n = \frac{M_{11,n}(z)}{\|a_n\|^2 D_n(z)}. \end{aligned} \tag{79}$$

By lemma H.1 and $\|a_n\|^2 \to \theta$, this gives

$$D_n(z) \longrightarrow \mathcal{D}(z) := 1 - \theta m(z)(c + 1 + w(z)), \qquad g_n(z) \longrightarrow \frac{m(z)}{\mathcal{D}(z)}. \tag{80}$$

The second convergence is locally uniform off the reference spectrum and the zeros of $\mathcal{D}$.

Using $\theta c/d = T = q/p$, define $P_T(w) = c - (T - 1)w(c + 1 + w)$. Then

$$\mathcal{D}(\Lambda(w)) = \frac{P_T(w)}{(1 + w)(c + w)}, \qquad m(\Lambda(w)) = \frac{cw}{d(1 + w)(c + w)}. \tag{81}$$

On $0 < w < w_*$, the function $\theta m(\Lambda(w))(c+1+w) = T[1 - c/((1+w)(c+w))]$ is strictly increasing. Hence $P_T(w) = 0$ has a root on this branch exactly when $T > T_*$. It is simple and gives Equation (8). The determinant identity and proposition D.4 locate the unique upper outlier. If there is no root above B, the same identity and boundedness exclude any top limit above B.

At the separated root L, Equation (23) gives

$$|\langle \widehat{a}_n, v_1 \rangle|^2 \longrightarrow \frac{m(L)}{\mathcal{D}'(L)} = \frac{cw_o\Lambda'(w_o)}{dP_T'(w_o)} = -\frac{w_o^2(c+1+w_o)\Lambda'(w_o)}{d(c+1+2w_o)}.$$

Substitution of Λ' yields Equation (9). Its two factors give $0 < \mathcal{O} < 1$ because $0 < w_o < w_*$ and $w_*^2(1+c+d+2w_*) = cd$.

For $T < T_*$, $\mathcal{D}(B^+) > 0$, so $(z-B)m(z)/\mathcal{D}(z) \to 0$. At $T = T_*$, $P_{T_*}(w_*) = 0$, $P'_{T_*}(w_*) = -(T_*-1)(c+1+2w_*) < 0$, and $\Lambda'(w_*) = 0$. Since $g(\Lambda(w)) = cw/[dP_T(w)]$, the ordinary difference quotients give

$$(z-B)g(z) = \frac{cw}{d}\frac{[\Lambda(w)-\Lambda(w_*)]/(w-w_*)}{[P_{T_*}(w)-P_{T_*}(w_*)]/(w-w_*)} \longrightarrow \frac{cw_*}{d}\frac{\Lambda'(w_*)}{P'_{T_*}(w_*)} = 0.$$

Here $w \uparrow w_*$ along the upper branch, after the width limit at fixed $z > B$. Equation (24) proves zero overlap in both cases, including the critical point. □

H.6 Parameter NTK

We now train all entries of G_1, G_2 in Equation (4), using their Euclidean parameter metrics and a common step size. The loss is $\mathcal{L} = \frac{1}{2}\|f - y\|^2$, without averaging over output coordinates.

Proof of Theorem 3.3. The parameter differentials are

$$\mathrm{d}f[\Delta G_1, \Delta G_2] = D_2 W_2 D_1 \frac{\Delta G_1 h_0}{\sqrt{n}} + \sqrt{\frac{2}{m}} D_2 \Delta G_2 h_1.$$

The two disjoint parameter blocks contribute additively to the NTK. Using $\|h_0\|^2 = n$ and $D_i^2 = D_i$ gives Equation (10). The same hidden rotation as above reduces its active block exactly to

$$\Theta|_{I_2} = \frac{2}{m} YY^* + a_n a_n^* + v_n I_s. \tag{82}$$

Conditional on the histories, Y is independent standard Gaussian, $(k-1)2/m \to 1$, $s/(k-1) \to 2\beta p$, $v_n \to 1$, and $\|a_n\|^2 \to 2\beta q$. Apply lemma A.5 to the conditional Wishart block and then add the scalar shift. The transform calculation in Appendix B gives the edge and overlap formulas. The directional limit is

$$a_n^* P\big(\tfrac{2}{m} YY^* + v_n I_s\big)\, a_n \longrightarrow \theta \int P(1+x)\, \mathrm{d}\mu_{\mathrm{MP},y_A}(x),$$

which determines the outlier equation. At zero residual the Hessian is the parameter Jacobian Gram matrix, with the same nonzero eigenvalues as Θ. The derivative of the gradient-descent map therefore contracts all positive modes exactly for $0 < \eta < 2/\lambda_1(\Theta)$. □

Within a fixed activation cell, the network is bilinear in G_1, G_2. The NTK therefore describes the derivative of the gradient-descent map at interpolation. Positive-curvature modes contract below the critical step, and null modes have multiplier one.

H.7 A family of derivative gates

Take $\alpha = 2$, $\beta = 4$, and $p = 0.1$. Center and tail gates use the activations in Appendix C. Table 4 gives the Jacobian and NTK predictions, and Table 1 reports the corresponding parameter-training thresholds.

To test the phase transition at fixed p, use

$$A_\rho = (-a_\rho, a_\rho) \cup (-\infty, -b_\rho) \cup (b_\rho, \infty), \quad a_\rho = \Phi^{-1}\left(\frac{1+p(1-\rho)}{2}\right), \quad b_\rho = \Phi^{-1}(1 - p\rho/2),$$

Table 4: Spectral predictions for the two-layer model with $\alpha = 2$, $\beta = 4$, and $p = 0.1$. Overlap is with the normalized gated second preactivation. The transition parameters refer to the gate family A_ρ defined below.

Operator	Bulk edge	Tail outlier	Overlap	Critical q/p	Critical ρ
JJ^*	5.8386	8.8078	0.8708	1.8762	0.2854
Θ	4.5889	5.8090	0.8914	2.1180	0.3375

for $\rho \in [0, 1]$, with empty tails at $\rho = 0$. Its activation is the sum of a clipped-linear and a soft-threshold function, and

$$q(\rho) = p - 2a_\rho\varphi(a_\rho) + 2b_\rho\varphi(b_\rho),$$

where the last term is zero at $\rho = 0$. Increasing ρ transfers gate activation probability from the center to the tails. The bulk remains fixed, while the Jacobian and parameter NTK cross their respective outlier thresholds. ReLU has $p = q = 1/2$ and is subcritical for both operators.

H.8 Experimental protocol

The width experiment uses $n \in \{256, 512, 1024\}$, $m = 2n$, and $\ell = 8n$. The seeds are 101, 211, 307, 401, 503, 601, 701, and 809, with NumPy seed sequences $(20260909, \text{seed}, n)$. Center and tail gates share each Gaussian network, giving 24 independent draws and 48 gate evaluations. Uncertainties are sample standard deviations across seeds. The predictions use population parameters (α, β, p, q). In the tables, $H = Z_n Z_n^*$ is the Jacobian reference and $H_\Theta = (2/m)YY^* + v_n I$ is the active parameter-NTK reference.

Table 5: Input-Jacobian Gram: mean $\pm$ sample standard deviation over eight seeds. The overlap is with the retained second-preactivation direction. Stable counts draws satisfying the exact finite-width criterion for contraction in positive-curvature directions at the midpoint of the limiting critical steps.

n	Gate	Largest eigenvalue	$\lambda_1(H)$	Overlap	Stable
256	Center	5.692 ± 0.387	5.693 ± 0.388	0.003 ± 0.003	8/8
256	Tail	8.586 ± 1.017	5.454 ± 0.707	0.876 ± 0.026	1/8
512	Center	5.631 ± 0.272	5.631 ± 0.273	0.001 ± 0.002	8/8
512	Tail	8.773 ± 0.951	5.594 ± 0.430	0.873 ± 0.021	0/8
1024	Center	5.697 ± 0.135	5.697 ± 0.135	0.001 ± 0.001	8/8
1024	Tail	8.527 ± 0.729	5.683 ± 0.310	0.867 ± 0.012	0/8

Table 6: Full-parameter output NTK: mean $\pm$ sample standard deviation over eight seeds. The overlap is with the retained second-preactivation direction. Stable counts draws satisfying the exact finite-width criterion for contraction in positive-curvature directions at the midpoint of the limiting critical steps.

n	Gate	Largest eigenvalue	$\lambda_1(H_\Theta)$	Overlap	Stable
256	Center	4.493 ± 0.163	4.493 ± 0.163	0.010 ± 0.013	8/8
256	Tail	5.613 ± 0.492	4.371 ± 0.365	0.889 ± 0.028	1/8
512	Center	4.534 ± 0.118	4.534 ± 0.118	0.002 ± 0.003	8/8
512	Tail	5.752 ± 0.408	4.462 ± 0.311	0.898 ± 0.013	0/8
1024	Center	4.518 ± 0.049	4.518 ± 0.049	0.002 ± 0.002	8/8
1024	Tail	5.714 ± 0.254	4.545 ± 0.197	0.887 ± 0.004	0/8

For each operator, the step size is the midpoint of its center and tail limiting critical values. The stability counts in the tables use the finite-width criterion $\eta\lambda_1 < 2$.

To evaluate the directional limits, we construct the hidden rotation, Z_n, a_n, b_n, and $t_n = Z_n b_n$. We record all entries of $U_n^*(zI - H_n)^{-1}U_n$ at $z/B \in \{1.2, 1.5, 2\}$, and the directional moments for

Table 8: Checks of the two nonlinear-layer parameter NTKs and local gradient descent. Jacobian errors are relative Frobenius errors against automatic differentiation. Trajectory error compares the observed residual-norm ratio with $|1 - \eta\lambda_1|^t$. Each model has 16 trajectories at $n = 256$. Stable counts use all 24 width–seed draws per gate at the midpoint of the limiting critical steps.

Network	NTK error	Trajectory error	Stable center	Stable tail
Two nonlinear layers	1.0×10^{-16}	4.1×10^{-7}	24/24	1/24
With linear output	2.7×10^{-16}	1.0×10^{-6}	24/24	7/24

$P(x) = 1, x, x^2$. The moments of $\mu_{c,d}$ used in the predictions are

$$\mu_0 = \mu_1 = 1, \qquad \mu_2 = A_0, \qquad \mu_3 = A_0^2 + \frac{d(1 + c + d)}{c}, \qquad A_0 = 1 + d + d/c.$$

The diagonal entries are compared with Equation (77), and the predicted off-diagonal entry is zero.

Table 7: Tail-gate compressed resolvent at $z = 1.5B$: seed means. The final column records the mean off-diagonal entry divided by the geometric mean of its two positive diagonal entries within each draw.

n	M_{00}	$\theta m(z)$	M_{11}	$w(z)$	Normalized M_{01}
256	0.44497	0.47107	0.14307	0.13916	0.00215
512	0.46785	0.47107	0.14120	0.13916	0.00227
1024	0.45992	0.47107	0.13328	0.13916	-0.00799

The transition experiment uses eight additional networks at $n = 512$, with NumPy seed sequences $(20260910, \text{seed}, 512)$. Each is evaluated at $\rho \in \{0, 0.15, 0.25, 0.30, 0.35, 0.50, 0.75, 1\}$, giving 64 paired gate evaluations. The curves in Figure 2 use $p = 0.1$ and the Gaussian integral $q(\rho)$.

For automatic-differentiation checks, PyTorch constructs the input Jacobian and the full (G_1, G_2) parameter Jacobian at $n = 8, m = 16, \ell = 64$, seed 20260909, for center, tail, and ReLU activations. We also run 16 training trajectories of 12 steps at $n = 256$, covering all seeds and both gates. The input is $\sqrt{n}e_1$, so only the first column of G_1 has a nonzero gradient. Updating that column and all of G_2 implements full-parameter gradient descent.

Each trajectory uses the unperturbed output as its interpolating target and starts with a small perturbation along the top parameter mode. For a unit top NTK eigenvector v_1 and $L_n = \lambda_1(\Theta)$, the unit joint parameter direction is

$$\Delta G_1 = \frac{D_1 W_2^* D_2 v_1 h_0^*}{\sqrt{nL_n}}, \qquad \Delta G_2 = \sqrt{\frac{2}{mL_n}} D_2 v_1 h_1^*.$$

Let δ be the minimum distance of either layer's preactivations to its activation breakpoints. The perturbation amplitude is

$$\epsilon_0 = \frac{0.02\delta}{(1 + \|W_2\|_F + \sqrt{2/m}\|h_1\|) \max\{1, |1 - \eta L_n|\}^{12}}.$$

This amplitude keeps the trajectories near interpolation. The learning rate is fixed from the population parameters. Both activations and parameter gradients are recomputed at each step. The gates remain unchanged throughout these trajectories, and the residuals follow their finite-width NTK predictions. Table 8 summarizes the errors and stability counts.

All calculations use double precision on the CPU. The tables report averages over all sampled seeds, including finite-width cases in which the tail gate remains stable.

I Proofs for the Gaussian output layer

We prove Theorem 3.4 and establish the common limit when the output dimension is sublinear in the hidden width. The experimental protocol follows the proofs.

Table 9: Activation parameters for the linear-output experiment at $p = 0.1$. Here $e_A = \mathbb{E}[\phi_A(Z)^2]$ and $\psi_A = \phi_A + \sqrt{\kappa - e_A}$, with $\kappa = \max_A e_A$.

Gate	Threshold	q_A	e_A	Shift
Center	0.125661	0.000525	0.014737	0.128574
Tail	1.644854	0.439286	0.031268	0.000000

Conditioning on the hidden network reduces the output NTK to a Gaussian covariance matrix with a rank-one perturbation. Recall Equation (12): $h_1 = \mathrm{ReLU}(G_1 h_0/\sqrt{n})$, $z_2 = \sqrt{2/m} G_2 h_1$, and $F_A = G_3 \psi_A(z_2)/\sqrt{\ell}$.

Here G_3 is independent standard Gaussian and $\psi_A' = \mathbf{1}_A$. We train all entries of G_1, G_2, G_3 with Euclidean metrics, a common step size, and loss $\frac{1}{2}\|F_A - y\|^2$, at one deterministic input with $\|h_0\|^2 = n$.

I.1 Activation second moments

Training the output weight contributes the second moment $\kappa_A = \mathbb{E}[\psi_A(Z)^2]$. We match it across the gates to preserve the output bulk law. The center and tail primitives $\phi_c(z) = \max(-a, \min(z, a))$ and $\phi_t(z) = \mathrm{sign}(z)(|z| - b)_+$ have zero means and energies

$$e_c = p - 2a\varphi(a) + a^2(1-p), \qquad e_t = p(1+b^2) - 2b\varphi(b).$$

Here $a = \Phi^{-1}((1+p)/2)$ and $b = \Phi^{-1}(1 - p/2)$. Choose the fixed activations

$$\psi_A(z) = \phi_A(z) + \sqrt{\kappa - e_A}, \qquad \kappa = \max(e_c, e_t). \tag{83}$$

The shift preserves p, q and gives $\mathbb{E}[\psi_A(Z)^2] = \kappa$ for both activations. Table 9 lists the experimental parameters.

Let Θ denote the hidden-vector NTK in Equation (10), $y_A = 2\beta p$, and

$$\tau = (1-p)\delta_0 + p(T_1)_{\#}\mu_{\mathrm{MP},y_A}, \qquad b_\tau = 1 + (1 + \sqrt{y_A})^2.$$

Assume $0 < y_A < 1$, so the nonzero bulk is one interval separated from zero. Let L_A be the top limit in Equation (11). For $a > b_\tau$ define

$$I(a) = \int \frac{t}{a-t}\,\mathrm{d}\tau(t), \qquad Q(a) = \int \frac{t^2}{(a-t)^2}\,\mathrm{d}\tau(t), \qquad \Psi_\gamma(a) = a[\gamma + I(a)]. \tag{84}$$

These functions can be evaluated from the Marchenko–Pastur transform. With $m_y(u) = \int (u - x)^{-1}\,\mathrm{d}\mu_{\mathrm{MP},y}(x)$,

$$I(a) = p[a m_{y_A}(a-1) - 1], \qquad Q(a) = p[-a^2 m'_{y_A}(a-1) - 2a m_{y_A}(a-1) + 1]. \tag{85}$$

For example $m_y(u) = 2/[u + y - 1 + \sqrt{(u-1-y)^2 - 4y}]$ above the MP edge. The function Q decreases continuously from infinity to zero, so there is a unique $a_\gamma > b_\tau$ with $Q(a_\gamma) = \gamma$.

Proof of Theorem 3.4. The chain rule gives $R\,\mathrm{d}\psi_A$ for the first two parameter blocks and $\Delta G_3 \psi_A/\sqrt{\ell}$ for the third. Their Gram contributions add, proving Equation (13). Conditional on h_1, the entries of z_2 are independent $N(0, v_n)$ with $v_n \to 1$. Gaussian concentration and the at-most-linear growth of ψ_A give $\kappa_{A,n} \to \kappa_A$ almost surely.

Let (L_n, u_n) be the top eigenpair of Θ_n, and set

$$T_n = \Theta_n - L_n u_n u_n^*, \qquad H_n = R T_n R^*, \qquad r_n = R u_n. \tag{86}$$

Condition on the hidden weights and use an eigenbasis of Θ_n whose first vector is u_n. The first column of the rotated Gaussian output weight gives r_n and the remaining columns give H_n. These are conditionally independent, $r_n/\|r_n\|$ is uniform on the output sphere, and $\|r_n\|^2 \to \gamma$.

The matrix T_n converges strongly to the law τ. Indeed, the active block of Θ_n is a shifted Wishart matrix plus a positive rank-one term. For $y_A < 1$, its Wishart support is one positive interval. Interlacing and convergence of both Wishart edges place every non-top active eigenvalue in a vanishing

neighborhood of that interval. Replacing the top eigenvalue by zero introduces a point already in $\operatorname{supp}\tau$, since $p < 1$. The empirical law is unchanged in the limit, and its support has positive mass in every relative neighborhood. Thus the spectra of T_n converge to $\operatorname{supp}\tau$, which together with weak convergence gives strong convergence of this single matrix. Uniform polynomial approximation also gives strong convergence of $T_n^{1/2}$.

Apply corollary F.4 conditionally with population τ. It gives strong convergence of H_n and the scalar equation Equation (68). Here m_γ is the limiting resolvent trace normalized by o, including any zero eigenvalues. Set $a = 1/(\gamma m_\gamma(z))$. The branch from infinity satisfies

$$z = \Psi_\gamma(a), \qquad m_\gamma(\Psi_\gamma(a)) = \frac{1}{\gamma a}, \qquad \Psi'_\gamma(a) = \gamma - Q(a).$$

Since $Q(b_\tau^+) = \infty$, the edge calculation following corollary F.4 gives a regular square-root upper edge $B_0 = \Psi_\gamma(a_\gamma)$, with $\lambda_1(H_n) \to B_0$ and $m_\gamma(B_0^+) = 1/(\gamma a_\gamma)$.

The output NTK is now the rank-one perturbation

$$\Theta_n^{\text{out}} = H_n + L_n r_n r_n^* + \kappa_{A,n} I_o.$$

Sphere concentration gives the directional transform m_γ along $r_n/\|r_n\|$, while the perturbation strength tends to γL_A. By lemma A.5, the outlier survives exactly when

$$(\gamma L_A) m_\gamma(B_0^+) > 1 \quad \Longleftrightarrow \quad L_A > a_\gamma.$$

Its unshifted limit is $\Psi_\gamma(L_A)$. At fixed τ, γ, the overlap is the derivative with respect to the rank-one strength:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\Psi_\gamma(\theta/\gamma)\Big|_{\theta=\gamma L_A} = \frac{\Psi'_\gamma(L_A)}{\gamma} = 1 - \frac{Q(L_A)}{\gamma}.$$

The same lemma gives convergence to B_0 and zero overlap at and below threshold. Adding $\kappa_{A,n} \to \kappa_A$ proves the eigenvalue and overlap claims. Conditioning and Fubini give the joint almost-sure conclusions.

The reference law depends only on τ and γ. Rank-one perturbation and the scalar shift therefore give the claimed equality of limiting output laws. Bounded norms yield all fixed normalized moments. At interpolation, the loss Hessian is the parameter-Jacobian Gram matrix, so its local linear-stability threshold is $2/\lambda_1(\Theta_n^{\text{out}})$. □

This recovers the generalized spiked-covariance map of Bai & Yao (2008), with law τ, spike L_A, and shift κ_A supplied by the network.

Proposition I.1. *If $o/\ell \to 0$, including a fixed scalar or fixed-dimensional output, then*

$$\left\|\Theta^{\text{out}} - (2p + \kappa_A) I_o\right\| \longrightarrow 0 \quad \textit{almost surely}.$$

Consequently the center and tail networks with matched activation second moments have the same limiting local linear-stability threshold $2/(2p + \kappa)$ in this single-input regime.

Proof. Conditionally on the hidden weights, $R\Theta R^*$ is a Gaussian sample covariance with bounded population norm. A Gaussian quadratic-form bound and an ε-net of the o-dimensional unit sphere give $\|R\Theta R^* - (\operatorname{Tr}\Theta/\ell) I_o\| \to 0$: the tail exponent is of order ℓ for any fixed error, whereas the net logarithm is $O(o) = o(\ell)$. The probabilities are summable because ℓ grows proportionally to n. Also $\operatorname{Tr}\Theta/\ell \to \int t \,\mathrm{d}\tau(t) = 2p$ and $\kappa_{A,n} \to \kappa_A$. In particular a bounded hidden finite-rank mode has vanishing image norm in this regime. □

I.2 Experimental protocol

Predictions use Equations (84) and (85). Table 1 and Figure 3 report the thresholds and dependence on output dimension. We use $m = 2n$, $\ell = 8n$, $p = 0.1$, and the activations in Equation (83). Each width $n \in \{256, 512, 1024\}$ uses the eight seeds listed in Appendix H.8, with NumPy seed sequences $(20260911, \text{seed}, n)$. Each of the 24 independent network draws supplies both gates and nested row subsets of one Gaussian output weight. We test $o = 1$ and $\gamma \in \{0.02, 0.5\}$ at every

width, together with $\gamma \in \{0.05, 0.075, 0.1, 0.2\}$ at $n = 512$, rounding $o = \gamma\ell$. This gives 208 paired evaluations. Error bars and table uncertainties are sample standard deviations across eight seeds.

Table 10: Final NTK largest eigenvalues with all three weights trained, mean ± sample standard deviation across eight seeds.

Output regime	Gate	$n = 256$	$n = 512$	$n = 1024$
$o = 1$	center	0.2308 ± 0.0285	0.2266 ± 0.0172	0.2307 ± 0.0135
$o = 1$	tail	0.2339 ± 0.0691	0.2638 ± 0.0368	0.2200 ± 0.0404
$\gamma = 0.02$	center	0.4540 ± 0.0267	0.4724 ± 0.0288	0.4682 ± 0.0095
$\gamma = 0.02$	tail	0.4410 ± 0.0773	0.5135 ± 0.0547	0.4600 ± 0.0569
$\gamma = 0.5$	center	2.9031 ± 0.0788	2.9318 ± 0.1151	2.9036 ± 0.0477
$\gamma = 0.5$	tail	3.1914 ± 0.3955	3.4272 ± 0.2069	3.2074 ± 0.2585

The forward pass gives $k = |I_1|$, $s = |I_2|$, $v_n = 2\|h_1\|^2/m$, and $\theta_n = 2\|z_{2,I_2}\|^2/(mv_n)$. Before constructing the NTK, set

$$\widehat{p} = s/\ell, \quad \widehat{y} = s/(k-1), \quad \sigma_n = 2(k-1)/m, \quad \widehat{\tau} = (1-\widehat{p})\delta_0 + \widehat{p}(x \mapsto v_n + \sigma_n x)_{\#}\mu_{\mathrm{MP},\widehat{y}}.$$

The hidden top estimate is $v_n + \sigma_n(1+\sqrt{\widehat{y}})^2$ when $\theta_n \le \sigma_n(\widehat{y} + \sqrt{\widehat{y}})$. Above this threshold it is

$$\widehat{L} = v_n + \frac{\theta_n[\theta_n + \sigma_n(1-\widehat{y})]}{\theta_n - \sigma_n\widehat{y}}.$$

Substituting $\widehat{\tau}, \widehat{L}, \kappa_{A,n}$ and o/ℓ into the output map gives a deterministic-equivalent estimate using only forward-pass statistics. Figure 3(c) and Section 4.2 report its accuracy at $o/\ell = 0.5$. The 48 evaluations comprise center and tail gates on 24 independently sampled networks. Eigenvalue error is the mean absolute relative error. Stability agreement compares the predicted and exact finite-width NTK decisions at the common step $\eta = 0.637862$.

At $n = 256$, $\gamma = 0.5$, we train all three weights for 16 trajectories of eight steps under the unaveraged loss, using the initialized output as target. The perturbation follows the unit direction of the parameter Jacobian transpose applied to the top output eigenvector. Its amplitude is set by the initial gate margin to stay near interpolation. For input $\sqrt{n}e_1$, only the first column of G_1 has nonzero gradients. We update this column and all entries of G_2, G_3. Gates remain unchanged, and residuals follow $|1 - \eta\lambda_1|^t$ (Table 8).

Automatic differentiation of all three parameter blocks at $n = 8, m = 16, \ell = 64, o = 12$ verifies Equation (13). Direct quadrature of the Marchenko–Pastur density checks the output-map integrals.

J Factorization of the output NTK

J.1 Factored output kernels

In the models studied by Hazelden et al. (2026), the empirical neural tangent kernel on the global state factorizes as $\Theta_n^S = \mathcal{P}_{S,n}\mathcal{C}_n\mathcal{P}_{S,n}^*$. Here the weight-site variables collect local variations associated with the trainable weights. The matrix $\mathcal{C}_n$ is their kernel, and $\mathcal{P}_{S,n}$ maps these variations to the stacked network states. If $\mathcal{R}_n$ extracts the outputs and $\mathcal{P}_n := \mathcal{R}_n\mathcal{P}_{S,n}$, then

$$\Theta_n^{\mathrm{out}} = \mathcal{R}_n\Theta_n^S\mathcal{R}_n^* = \mathcal{P}_n\mathcal{C}_n\mathcal{P}_n^*. \tag{87}$$

For a positive-semidefinite matrix A, let π_A denote the orthogonal projection onto its range. We write $\mathrm{spec}^+(A)$ for its positive spectrum, $\lambda_{\min}^+(A)$ for its smallest positive eigenvalue, and $\sigma_{\min}^+(B)$ for the smallest positive singular value of B.

Proposition J.1. *Let $\mathcal{C} \in \mathbb{C}^{d\times d}$ be positive semidefinite, let $\mathcal{P} \in \mathbb{C}^{m\times d}$, let $\pi_{\mathcal{C}}$ be the support projection of $\mathcal{C}$, and set $\Theta = \mathcal{P}\mathcal{C}\mathcal{P}^*$. Then*

$$\begin{aligned} &\mathrm{rank}(\Theta) = \mathrm{rank}(\mathcal{P}\pi_{\mathcal{C}}), \quad \mathrm{spec}^+(\Theta) = \mathrm{spec}^+\left(\mathcal{C}^{1/2}\mathcal{P}^*\mathcal{P}\mathcal{C}^{1/2}\big|_{\pi_{\mathcal{C}}}\right), \\ &\lambda_{\max}(\Theta) = \|\mathcal{P}\mathcal{C}^{1/2}\|^2. \end{aligned} \tag{88}$$

If $\Theta \neq 0$, then

$$\lambda^{+}_{\min}(\Theta) = \sigma^{+}_{\min}(\mathcal{P}\mathcal{C}^{1/2})^2. \tag{89}$$

These identities express both output-NTK edges through the singular values of $\mathcal{P}\mathcal{C}^{1/2}$, which combine the scales, supports, and relative alignment of the two factors.

To pass these identities to the large-width limit, assume that the propagator is a fixed rational expression satisfying Equation (42). Assume also that the coefficients in $\mathcal{C}_n$ and $\mathcal{P}_n$, which depend on the recorded queries, converge jointly in the system of Theorem D.2. Holding the core and query history fixed, replace the weight blocks in $\mathcal{P}_n$ by their Gaussian references and set $\Theta^{\mathrm{ref}}_n := \mathcal{P}^{\mathrm{ref}}_n \mathcal{C}_n \mathcal{P}^{\mathrm{ref}*}_n$. The strong law and the rank bound in Equation (41) give

$$\operatorname{rank}(\mathcal{P}_n - \mathcal{P}^{\mathrm{ref}}_n) \le r_n \quad \Longrightarrow \quad \operatorname{rank}(\Theta^{\mathrm{out}}_n - \Theta^{\mathrm{ref}}_n) \le 2r_n.$$

Let N^{out}_n be the dimension of Θ^{out}_n. If $r_n = o(N^{\mathrm{out}}_n)$ and the limiting law of the strongly convergent Θ^{ref}_n places positive mass arbitrarily close to its upper endpoint b, the rank bound gives the same limiting empirical spectral distribution. Applying lemma A.4 to $\mathcal{P}_n \mathcal{C}_n^{1/2}$ and $\mathcal{P}^{\mathrm{ref}}_n \mathcal{C}_n^{1/2}$ gives

$$\lambda_{r_n+1}(\Theta^{\mathrm{out}}_n) \longrightarrow b.$$

The finite-dimensional identities in Equations (87) and (88) hold independently of the asymptotic assumptions. Joint convergence and the inverse bound allow these identities to pass to the large-width Gaussian reference model. In both settings, relative eigenvector alignment is part of the information required to determine the smallest positive eigenvalue and the largest eigenvalue.

Separated eigenvalues require the directional resolvent in proposition D.4. The models in Sections 3.3 and 3.4 evaluate these limits and determine the largest eigenvalue of their parameter NTKs.

The free-projection model gives an analytically tractable example of this alignment dependence. Let e and f be free projections with $\tau(e) = \rho$ and $\tau(f) = \gamma$, where $\rho, \gamma \in (0, 1)$. For scales $a, b > 0$, set $\mathcal{C} = af$ and $\mathcal{P} = \sqrt{b}\, e$. Write $\mu_{\Theta|e}$ for the spectral law in $e\mathcal{A}e$ under $\tau_e(\,\cdot\,) = \tau(\,\cdot\,)/\tau(e)$, and let $\lambda^{+}_{-}(\Theta) := \inf(\operatorname{spec}(\Theta) \cap (0, \infty))$. Then

$$\begin{aligned} \mu_{\Theta|e}(\{0\}) &= \frac{(\rho - \gamma)_+}{\rho}, \\ \lambda^{+}_{-}(\Theta) &= ab\left(\sqrt{\rho(1-\gamma)} - \sqrt{\gamma(1-\rho)}\right)^2. \end{aligned} \tag{90}$$

As ρ approaches γ, the lower positive edge approaches zero quadratically. In frozen or linearized full-batch least-squares dynamics, the corresponding small positive eigenvalues produce slow residual modes, and $\lambda_{\max}(\Theta^{\mathrm{out}}_n)$ determines the stable constant-step range.

J.2 Output compression and factorization identities

The global-state factorization of Hazelden et al. (2026) gives the following output compression:

$$\Theta^{S} = \mathcal{P}_S \mathcal{C} \mathcal{P}_S^*, \qquad \Theta^{\mathrm{out}} = \mathcal{R}\Theta^{S}\mathcal{R}^* = (\mathcal{R}\mathcal{P}_S)\mathcal{C}(\mathcal{R}\mathcal{P}_S)^*.$$

Set $\mathcal{P} := \mathcal{R}\mathcal{P}_S$.

Set $A := \mathcal{P}\mathcal{C}^{1/2}$ and let $\pi_{\mathcal{C}}$ be the support projection of $\mathcal{C}$. Then

$$\Theta = AA^*, \qquad A^*A = \mathcal{C}^{1/2}\mathcal{P}^*\mathcal{P}\mathcal{C}^{1/2}.$$

The nonzero spectra of AA^* and A^*A coincide, and their extreme nonzero eigenvalues are the squared positive singular values of A. Also $\mathcal{C}^{1/2}$ maps $\pi_{\mathcal{C}}\mathcal{H}$ bijectively onto itself and vanishes on its orthogonal complement, so

$$\operatorname{rank}(\Theta) = \operatorname{rank}(A) = \operatorname{rank}(\mathcal{P}\mathcal{C}^{1/2}) = \operatorname{rank}(\mathcal{P}\pi_{\mathcal{C}}).$$

These identities prove proposition J.1. In particular, they retain the relative eigenspaces of the core and propagator, which determines the extreme nonzero eigenvalues.

J.3 Rank bound for the output NTK

Let $\Delta\mathcal{P}_n := \mathcal{P}_n - \mathcal{P}_n^{\mathrm{ref}}$. With the realized core frozen,

$$\begin{aligned}\Theta_n^{\mathrm{out}} - \Theta_n^{\mathrm{ref}} &= \mathcal{P}_n\mathcal{C}_n\mathcal{P}_n^* - \mathcal{P}_n^{\mathrm{ref}}\mathcal{C}_n\mathcal{P}_n^{\mathrm{ref}*} \\ &= \Delta\mathcal{P}_n\mathcal{C}_n\mathcal{P}_n^* + \mathcal{P}_n^{\mathrm{ref}}\mathcal{C}_n\Delta\mathcal{P}_n^*, \\ \operatorname{rank}(\Theta_n^{\mathrm{out}} - \Theta_n^{\mathrm{ref}}) &\le 2\operatorname{rank}(\Delta\mathcal{P}_n).\end{aligned}$$

If the state propagator is the inverse of a lifted state Jacobian and the output map $\mathcal{R}_n$ is fixed, the inverse identity in Equation (57) transfers the query-history rank bound to $\Delta\mathcal{P}_n = \mathcal{R}_n(\mathcal{P}_{S,n} - \mathcal{P}_{S,n}^{\mathrm{ref}})$.

For the upper edge, the common Gram factorization gives a sharper index. Set $A_n = \mathcal{P}_n\mathcal{C}_n^{1/2}$ and $B_n = \mathcal{P}_n^{\mathrm{ref}}\mathcal{C}_n^{1/2}$. Then $\operatorname{rank}(A_n - B_n) \le r_n$ whenever $\operatorname{rank}(\Delta\mathcal{P}_n) \le r_n$. Lemma A.4 yields $\lambda_{r_n+1}(\Theta_n^{\mathrm{out}}) \to b$ under the reference convergence and endpoint-mass assumptions in the main text. The more conservative indices used in the experiments remain valid.

J.4 Alignment counterexample

Fix $0 < \varepsilon < 1$. The identities

$$\begin{gathered}\mathcal{P} = \operatorname{diag}(1,\varepsilon), \qquad \mathcal{C}_1 = \operatorname{diag}(1,\varepsilon), \qquad \mathcal{C}_2 = \operatorname{diag}(\varepsilon,1), \\ \mathcal{P}\mathcal{C}_1\mathcal{P}^* = \operatorname{diag}(1,\varepsilon^3), \qquad \mathcal{P}\mathcal{C}_2\mathcal{P}^* = \operatorname{diag}(\varepsilon,\varepsilon^2)\end{gathered}$$

show that the two isospectral cores produce smallest positive eigenvalues ε^3 and ε^2 under the same propagator. Their ratio is ε^{-1} even though both marginal spectra are unchanged. This proves that marginal core and propagator laws do not determine the smallest positive output-NTK eigenvalue.

J.5 Random-subspace calculation

For free projections e, f with traces ρ, γ, the continuous part of efe is the free-Jacobi law with support

$$x_\pm = \left(\sqrt{\rho(1-\gamma)} \pm \sqrt{\gamma(1-\rho)}\right)^2. \tag{91}$$

Writing $s(x)$ for the support projection of a positive operator x, and $\mu_{x|e}$ for the spectral law of $x \in e\mathcal{A}e$ under $\tau_e(\,\cdot\,) = \tau(\,\cdot\,)/\tau(e)$, the free-projection intersection formula gives (Nica & Speicher, 2006)

$$\tau(s(efe)) = \min\{\rho,\gamma\}, \qquad \mu_{efe|e}(\{0\}) = \frac{(\rho-\gamma)_+}{\rho}.$$

The lower edge approaches zero quadratically as ρ approaches γ, because

$$\begin{aligned}x_- &= \left(\sqrt{\rho(1-\gamma)} - \sqrt{\gamma(1-\rho)}\right)^2 \\ &= \frac{(\rho-\gamma)^2}{\left[\sqrt{\rho(1-\gamma)} + \sqrt{\gamma(1-\rho)}\right]^2}.\end{aligned} \tag{92}$$

When $\gamma > \rho$, the core spans more directions than the output subspace can carry. The map $e|_{f\mathcal{H}}$ then has a kernel, while the lower positive edge remains nonzero away from $\rho = \gamma$. The positive-singular-value identity in Equation (88) applies in this range.

K Numerical evaluation of limiting spectral laws

This appendix gives the deterministic calculations used to draw the theoretical curves in Figures 4 to 6.

K.1 Marchenko–Pastur law and a rank-one outlier

For an active-row aspect ratio $y \in (0,1)$, the positive Marchenko–Pastur density is

$$\rho_{\mathrm{MP},y}(x) = \frac{\sqrt{(b_y - x)(x - a_y)}}{2\pi y x}\,\mathbf{1}_{[a_y,b_y]}(x), \tag{93}$$

where

$$\begin{aligned} a_y &= (1-\sqrt{y})^2, \\ b_y &= (1+\sqrt{y})^2. \end{aligned} \tag{94}$$

For any $y > 0$, the resolvent on the upper real branch can be evaluated as

$$m_y(z) = \frac{2}{z + y - 1 + \sqrt{(z-a_y)(z-b_y)}}, \qquad z > b_y,$$

with the positive square root. For the rank-one perturbation in Theorem 3.1, a supercritical strength $\theta > y + \sqrt{y}$ produces the outlier

$$\rho_{\mathrm{out}}(\theta; y) = \frac{\theta(\theta + 1 - y)}{\theta - y}. \tag{95}$$

The finite-width histograms in Figure 5 include only positive eigenvalues. The point mass at zero contributed by inactive neurons is described analytically by Equation (2) and omitted from the conditional density plot.

K.2 Gaussian reference for a two-layer tanh network

For Figure 6, the activation is $\phi(x) = \tanh x$. With $Z \sim \mathcal{N}(0,1)$, define

$$\begin{aligned} q_1 &= \mathbb{E}\tanh^2(Z) = 0.39429449\ldots, \\ A_1 &= \operatorname{sech}^4(Z), \\ A_2 &= \operatorname{sech}^4(\sqrt{q_1}\,Z). \end{aligned} \tag{96}$$

The Gaussian comparison matrix conditioned on the realized gates is

$$J_n^{\mathrm{ref}} = D_{2,n} G_{2,n} D_{1,n} G_{1,n}, \qquad K_n^{\mathrm{ref}} = J_n^{\mathrm{ref}} J_n^{\mathrm{ref}*}, \tag{97}$$

with square Gaussian blocks normalized by $n^{-1/2}$. Its limit is the multiplicative covariance law generated by A_1, A_2 and two square Marchenko–Pastur factors. We evaluate the two-step separable-covariance deterministic equivalent of Hachem et al. (2007) through the following coupled equations. For $\zeta \in \mathbb{C}^+$, set $w = -1/\delta_2(\zeta)$ and solve

$$\begin{aligned} \delta_1(w) &= \mathbb{E}\left[\frac{A_1}{-w(1+\widetilde{\delta}_1(w)A_1)}\right], & \widetilde{\delta}_1(w) &= \frac{1}{-w(1+\delta_1(w))}, \\ m_1(w) &= \mathbb{E}\left[\frac{1}{-w(1+\widetilde{\delta}_1(w)A_1)}\right], & \delta_2(\zeta) &= \mathbb{E}\left[\frac{A_2}{-\zeta(1+\widetilde{\delta}_2(\zeta)A_2)}\right], \\ \widetilde{\delta}_2(\zeta) &= -\frac{1}{\zeta}\left[\frac{1}{\delta_2(\zeta)} - \frac{m_1(-1/\delta_2(\zeta))}{\delta_2(\zeta)^2}\right], & m_2(\zeta) &= \mathbb{E}\left[\frac{1}{-\zeta(1+\widetilde{\delta}_2(\zeta)A_2)}\right]. \end{aligned} \tag{98}$$

Here m_2 uses the $(K-\zeta I)^{-1}$ convention and has positive imaginary part. The resolvent in Equation (102) uses $(\zeta I - H)^{-1}$ and has negative imaginary part. The solution in this subsection is selected by $\Im m_2(\zeta) > 0$ and continuity from large imaginary part. The smoothed density is $\pi^{-1}\Im m_2(x + \mathrm{i}\eta)$. The upper edge is the real solution at which the Jacobian of the fixed-point residual in $(\delta_1, \widetilde{\delta}_1, \delta_2, \widetilde{\delta}_2)$ becomes singular. Gauss–Hermite quadrature gives

$$b_+^{(2)} = 2.6102625058\ldots. \tag{99}$$

K.3 Free-Jacobi density

For free projections with traces ρ and γ, define

$$x_- = \left(\sqrt{\rho(1-\gamma)} - \sqrt{\gamma(1-\rho)}\right)^2, \qquad x_+ = \left(\sqrt{\rho(1-\gamma)} + \sqrt{\gamma(1-\rho)}\right)^2. \tag{100}$$

The continuous component of the free-Jacobi law has density proportional to

$$\rho_{\mathrm{J}}(x) = \frac{\sqrt{(x_+ - x)(x - x_-)}}{2\pi x(1-x)}\, \mathbf{1}_{[x_-, x_+]}(x). \tag{101}$$

For the plots, Equation (101) is normalized by its numerical integral over $[x_-, x_+]$, because the histograms condition on the nontrivial continuous eigenvalues and remove the zero and unit atoms.

K.4 Dyson equation and finite-rank perturbations

For a Gaussian reference operator $H = A + S$ whose random part has an operator-valued semicircular limit over the base algebra, with covariance map η, use the same resolvent convention as in Equation (50):

$$R_H(\zeta) = \left(\zeta \mathrm{I} - A - \eta[R_H(\zeta)]\right)^{-1}, \qquad \zeta = x + \mathrm{i}\varepsilon, \tag{102}$$

where $R_H(\zeta)$ is the limiting conditional expectation of $(\zeta \mathrm{I} - H)^{-1}$. A numerical density is obtained from

$$\rho_\varepsilon(x) = -\frac{1}{\pi} \Im \tau R_H(x + \mathrm{i}\varepsilon). \tag{103}$$

We select the admissible branch by continuation from large to small imaginary part, enforcing $\Im R_H(\zeta) \preceq 0$ and the normalization $\zeta R_H(\zeta) \to \mathrm{I}$ as $|\zeta| \to \infty$.

For a finite-rank correction, one first evaluates the compressed resolvent of the Gaussian reference operator

$$M(\zeta) = U^*(\zeta \mathrm{I} - H)^{-1} U, \tag{104}$$

and then locates the separated roots of

$$\det(\mathrm{I} - CM(\zeta)) = 0. \tag{105}$$

Reported roots lie outside the support of the Gaussian reference law and correspond to simple crossings, with nonzero slope, of an eigenvalue of $CM(\zeta)$ through one.

L Additional experimental details

L.1 Statistical protocol and experimental units

All computations use double precision unless stated otherwise. The fixed seeds are 101, 211, 307, 401, 503, 601, 701, and 809. Table 11 summarizes the experimental units and uncertainty calculations. Bootstrap resampling is stratified within the design variables listed in the table.

Table 11: Experimental units, replication, and uncertainty summaries.

Study	Experimental unit and design	Replication	Summary and uncertainty
Network vs. reference	One tuple indexed by architecture, depth, width, input count, activation, and Gram family	8 seeds, 2,304 units	Width median, 2,000 stratified bootstrap draws for panel (a) of Figure 6
Reference vs. Dyson limit	One independently seeded reference spectrum at a fixed width	8 per width	Median edge error, 4,000 bootstrap draws for panel (d) of Figure 6
Output-kernel alignment	One base network with interventions nested within the realization	8 MLPs, 8 RNNs	Paired comparisons and pooled random-control distributions
Random subspaces	One independently drawn subspace pair at a fixed rank gap	20 per gap, 140 total	Medians across draws and exact nullity comparisons

L.2 Gate comparison at fixed gate activation probability

The comparison changes alignment with the preactivation direction while keeping the limiting active fraction fixed. The experiment uses $n = 900$, $m = 3600$, and a common Gaussian draw for the preactivation direction and orthogonal innovation. The center and tail thresholds satisfy

$$\mathbb{P}(|Z| \le a) = 0.1, \qquad \mathbb{P}(|Z| \ge b) = 0.1.$$

The shuffled condition permutes the tail mask before pairing it with the preactivation values, preserving its active-row count. The theoretical quantities computed from the model parameters are

$$y = 0.4, \quad b_y = (1 + \sqrt{0.4})^2 = 2.6649, \quad \rho_{\text{out}} = 3.0519.$$

The observed bulk upper edge, largest eigenvalue, and directional energy $\|s_{A,n}\|^2$ used in Figure 5 appear in Table 12.

Table 12: Finite-width gate-spectrum comparison.

Gate	Active rows	$\|s_{A,n}\|^2$	Bulk edge	Largest eigenvalue
Center	373	0.0021	2.6528	2.6528
Tail	346	1.7033	2.6101	2.9272
Shuffled tail	346	0.4125	2.5496	2.5504

L.3 Network and Gaussian reference operators

For input points $a = 1, \dots, k$, let $J_n^{(a)}$ be the network Jacobian and form $J_n = \bigoplus_{a=1}^{k} J_n^{(a)}$ and $K_n = J_n J_n^* \in \mathbb{R}^{N_n \times N_n}$, where $N_n = kn$. Its Gaussian reference is $H_n = J_n^{\text{ref}} J_n^{\text{ref}*}$ with $J_n^{\text{ref}} = \bigoplus_a J_n^{\text{ref},(a)}$. It keeps the realized activations, gates, and query spaces and replaces each reused weight by the coupled Gaussian reference matrix in proposition D.1. The rank bound gives

$$\operatorname{rank}(J_n - J_n^{\text{ref}}) \le 2Lk^2, \qquad \operatorname{rank}(K_n - H_n) \le R_K := 4Lk^2.$$

We test MLPs and residual networks at depths $\{2, 4, 6\}$, widths $\{256, 512, 1024, 2048\}$, $k \in \{2, 4, 8\}$ input points, tanh and ReLU activations, and two deterministic input-Gram families held fixed across widths.

Table 13: Bulk-discrepancy medians and width log–log slopes. Intervals use stratified bootstrap resampling over experimental units.

Metric	Width 256	Width 2,048	Slope	95% CI
Kolmogorov–Smirnov distance	6.8×10^{-3}	9.8×10^{-4}	-0.942	$[-1.000, -0.923]$
Wasserstein distance	5.5×10^{-3}	5.8×10^{-4}	-1.084	$[-1.261, -0.866]$

Table 14: Gaussian-reference comparisons with their limiting laws. The two-layer tanh Dyson law has predicted upper edge 2.6103. Intervals are 95% bootstrap intervals computed from the independently seeded edge errors.

Model	Width	KS distance	Upper-edge error (95% interval)
MP baseline	2,048	0.00861	0.678%
Two-layer tanh	256		11.57% [4.62%, 13.08%]
	512		7.66% [2.36%, 10.95%]
	1,024		3.20% [2.37%, 4.46%]
	2,048		2.44% [1.63%, 2.83%]

Writing the eigenvalues in decreasing order, all eigenvalues including zeros enter the two bulk distances,

$$d_{\mathrm{KS}} = \sup_x |F_{K_n}(x) - F_{H_n}(x)|,$$
$$W_1 = \frac{1}{N_n} \sum_{j=1}^{N_n} |\lambda_j(K_n) - \lambda_j(H_n)|,$$
$$e_{\mathrm{edge}} = \frac{|\lambda_{R_K+1}(K_n) - \lambda_1(H_n)|}{\lambda_1(H_n)}.$$

The last quantity measures the upper-edge error after excluding the number of eigenvalues allowed by the rank bound in corollary D.3. Figure 6 reports the width-wise medians, and Table 13 summarizes the fitted bulk-discrepancy slopes. Numerical ranks use absolute and relative tolerances 10^{-10} and 10^{-8}. Widths at most 512 use an exact singular-value decomposition. Larger widths use a Gaussian range sketch with ten oversampling directions. Exact singular-value decompositions on eight stratified width-2,048 cases returned the same numerical ranks as the sketch.

L.4 Convergence to the limiting reference law

As an unconditioned baseline, we compare a square Gaussian reference operator with the Marchenko–Pastur law of aspect ratio one. The baseline agrees closely with the limiting prediction. Table 14 reports its numerical summary together with the neural-network comparison.

Panels (c)–(d) of Figure 6 use a two-layer square tanh network. For each seed, the realized network generates $z_{1,n}, z_{2,n}$ and the frozen gates

$$D_{1,n} = \mathrm{diag}(\mathrm{sech}^2 z_{1,n}), \qquad D_{2,n} = \mathrm{diag}(\mathrm{sech}^2 z_{2,n}).$$

Independent Gaussian reference matrices then generate the operator in Equation (97). Eight independent seeds are used at each width. The coupled Dyson system in Equation (98) is evaluated by 100-node Gauss–Hermite quadrature, and its upper edge is found from the stability-Jacobian singularity.

The comparison links the realized nonlinear computation to its Gaussian reference and then to the coupled Dyson law. The edge error decreases across all tested widths, as shown in Figure 6d and Table 14.

L.5 Factorization validation

We use depth-two width-32 multilayer perceptrons and shared-weight recurrent neural networks, each evaluated on 32 inputs in dimension 32. The recurrent block is reused for sequences of length three. The site map B and propagator P are extracted from the computation graph, while the complete parameter Jacobian J_{AD} is constructed independently by automatic differentiation. We evaluate the relative Frobenius errors of $J_{\mathrm{AD}} = PB$ and $J_{\mathrm{AD}}J_{\mathrm{AD}}^{*} = PBB^{*}P^{*}$, together with their actions on 16 fixed Gaussian probes. Every network and probe check agrees at machine precision. Table 15 reports the maximum discrepancies.

L.6 Controlled alignment interventions

These experiments isolate the dependence of the lower positive edge on the relative eigenspaces of the core and $P^{*}P$. The extremal isospectral comparison pairs the two fixed spectra in the commuting orders that maximize and minimize their smallest positive product. All random interventions remain nested within the corresponding base realization.

The full-dimensional permutation control changes the eigenvalue pairing without changing either spectrum. For the reduced-Haar comparison, each factor spectrum is replaced by its 128 empirical quantiles and one reduced factor is conjugated by a Haar orthogonal matrix. For each base realization, the architecture control evaluates eight permutations restricted to weight-site blocks of the same layer and type and reports

$$R_{\mathrm{block}} = \frac{\max_s \lambda_{\min}^{+}(\Theta_s)}{\min_s \lambda_{\min}^{+}(\Theta_s)}.$$

Panels (a)–(b) of Figure 4 show the paired interventions and pooled nested-control distributions. Their numerical summaries appear in Table 15.

L.7 Finite-dimensional random-subspace comparison

The ambient dimension is 512. For target gaps $\delta \in \{0.02, 0.03, 0.045, 0.067, 0.10, 0.15, 0.20\}$, we set $r_e = \mathrm{round}(512(0.35 + \delta/2))$, $r_f = \mathrm{round}(512(0.35 - \delta/2))$, and $(\rho, \gamma) = (r_e/512, r_f/512)$. The resulting rank pairs are

$$(r_e, r_f) \in \{(184, 174), (187, 172), (191, 168), (196, 162), \\ (205, 154), (218, 141), (230, 128)\}.$$

For each pair we draw 20 independent Haar subspace pairs by Gaussian QR and analyze efe on the range of e. The predicted positive edge is given by Equation (91). Panels (c)–(d) of Figure 4 show that the finite-dimensional medians follow the predicted quadratic closing and that every observed nullity matches the rank identity.

Table 15: Independent factorization check and paired alignment interventions.

Check or intervention	Sampling	Reported statistic
Independent $J = PB$ construction	16 networks	Maximum relative error: 6.47×10^{-17} for J and 1.52×10^{-15} for JJ^{*}
Extremal isospectral alignment	16 paired bases	Median lower-edge ratio $101.764\times$
Full-dimensional permutations	2,048 nested draws	Median symmetric fold change $4.49\times$
Reduced 128-dimensional Haar rotations	1,024 nested draws	Median symmetric fold change $2.76\times$
Layer- and site-matched shuffles	16 bases, 8 shuffles each	Median per-base range ratio $1.405\times$ (5/16 above $2\times$)

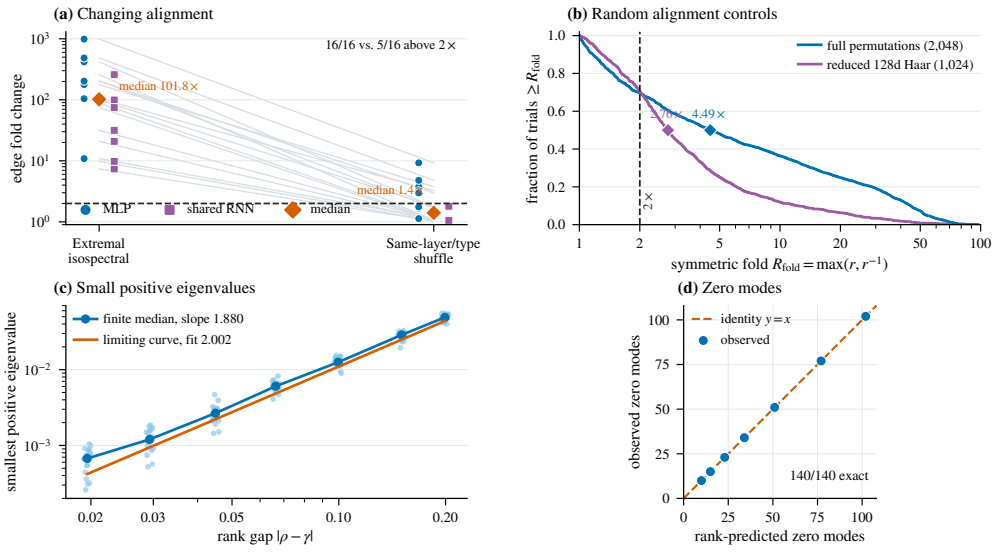


Figure 4: Operator-level alignment tests at fixed factor spectra. (a) Extremal pairings and layer-matched shuffles. (b) Random permutation and Haar controls. (c) Quadratic closing of the smallest positive eigenvalue in the random-subspace model. (d) Agreement between observed and rank-predicted nullities.

M Additional spectral experiments

M.1 Gaussian reference approximation

The preceding comparisons test the predictions in trainable networks. We next isolate the dependence that produces the outlier. Figure 5 compares the one-layer center and tail gates with a shuffled mask. All have the same gate activation probability and predicted Marchenko–Pastur bulk. Tail gating produces a preactivation-aligned outlier, while shuffling the mask preserves the active-unit count and removes both the alignment and the outlier. The spectral difference therefore depends on which preactivations are selected, as the retained-energy calculation predicts.

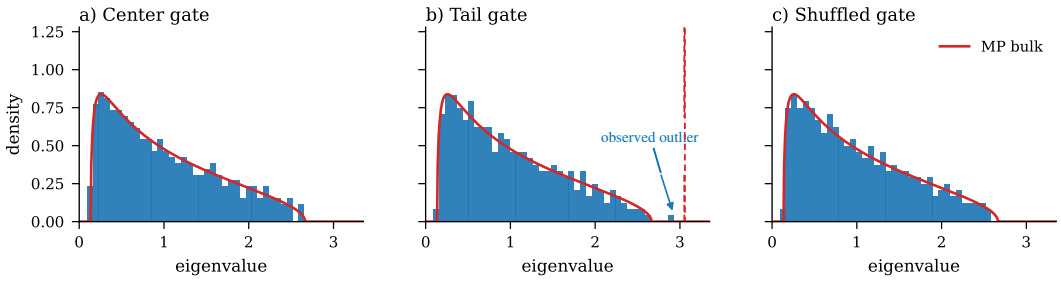


Figure 5: One-layer spectra for center, tail, and shuffled gates. The gates have the same gate activation probability. Tail gating produces a preactivation-aligned outlier at the predicted location, marked by the dashed line. Shuffling preserves the active-unit count and removes the alignment.

The multilayer reference construction determines the spectrum below the possible exceptional modes. We test this statement in Figure 6 by removing the number of leading eigenvalues allowed by the rank bound and comparing the remaining upper edge with the reference edge. Both the bulk discrepancy and the interlacing error

$$\frac{|\lambda_{R_K+1}(K_n) - \lambda_1(H_n)|}{\lambda_1(H_n)}$$

decrease with width. Here R_K is the algebraic rank bound, so the second metric tests corollary D.3. The correction ranks satisfy their bounds in all configurations. The multilayer-perceptron experiments follow the assumptions of Theorem D.2, and residual variants exhibit the same width trend. In a two-layer tanh model, the reference density and upper edge approach the coupled Dyson prediction. Tables 13 and 14 report the width dependence and bootstrap intervals.

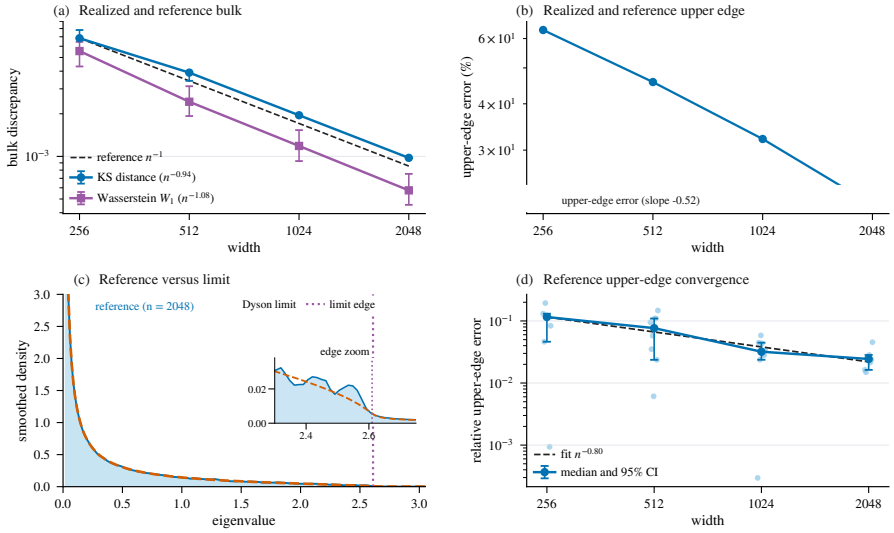


Figure 6: Gaussian reference approximation. (a) Bulk discrepancies and (b) the interlacing error $|\lambda_{R_K+1}(K_n) - \lambda_1(H_n)|/\lambda_1(H_n)$ decrease with width. (c,d) In the two-layer tanh model, the reference spectrum approaches the Dyson law and its upper edge. Error bars in (a,d) are bootstrap intervals.

M.2 Relative eigenspaces in factored kernels

We verify $J = PB$ by automatic differentiation in the tested multilayer perceptrons and recurrent networks (Table 15). Holding the spectra of the core and P^*P fixed, we vary their relative eigenvectors at the matrix level. The output edges change as predicted by proposition J.1. A random-subspace model also exhibits the predicted quadratic closing of the lower edge. Figure 4 and Appendices L.6 and L.7 give these operator comparisons and rank checks.