How Root Multiplicity Displaces the Finite-Width Edge of Chaos in MLPs and Message-Passing Networks

Anonymous authors
Paper under double-blind review

Abstract

How far a finite-width network's order/chaos crossing sits from its mean-field location is set by one integer, the multiplicity r of the mean-field root: an $O(1/F)$ correction to the Lyapunov exponent moves the crossing by $O(F^{-1/r})$. In message-passing networks at zero bias the root is double. We prove this through an exact identity, $G - 1 = \kappa_p(g-1)^2(1+o(1))$, whose constant is universal ($\kappa_2 = 1/3$) across smooth saturating odd activations, so the finite-width crossing lies at distance of order $F^{-1/2}$, not F^{-1}, from both the linearized boundary and mean field, which coincide. For MLPs and regular graphs with uniform input the network reduces exactly to a two-state Markov chain, and we prove the critical-window law $\log g = \log g_{\mathrm{lin}}(F) + \sqrt{3/F}\,(1+o(1))$. Between $g_{\mathrm{lin}}(F)$ and the crossing the exponent is at least $-(1+o(1))/(2F)$, sharply; at $\log g - \log g_{\mathrm{lin}}(F) = x/F$ the chain converges to an explicit diffusion, and $2F\chi^\star \to -\min(1, 2x)$ for every $x > 0$. Pre-registered measurements on full networks confirm the window law: the displacement decays as $F^{-1/2}$ at zero bias (fitted exponent -0.56) and as F^{-1} at nonzero bias (-1.09), and softsign is displaced more than $\tanh$ by the predicted factor 1.84. Bipartite graphs follow a predicted second branch without the $1/F$ cost, which a registered field-level test confirms. We prove, and a registered test at held-out widths confirms, a $5/(2F)$ shift for ReLU, whose root is simple.

1 Introduction

Information propagation theory classifies a randomly initialized network as ordered or chaotic by whether it contracts or amplifies input perturbations across depth (Poole et al., 2016; Schoenholz et al., 2017; Yang & Schoenholz, 2017). The boundary of *trainability* in hyperparameter space is fractal (Sohl-Dickstein, 2024), and at finite width so is the order/chaos boundary at initialization (D'Inverno et al., 2026); for message-passing networks, which mix nodes through a graph operator, it is open. Estimating a box dimension presupposes counting the right set, so the first question is where the boundary of a finite-width message-passing network actually lies.

The natural analytic locator is the linearization at the origin. The natural repair is mean field at the fixed point actually reached. At zero bias the second returns the first exactly (Proposition 3), and both miss the measured crossing by 45–91% at width $F = 6$. The gap shrinks with width but slowly, and a slowly shrinking gap needs an explanation: generic finite-width corrections to smooth quantities are $O(1/F)$. At a degenerate edge the correction becomes $O(F^{-1/2})$ in the boundary's *location*, and linearizing there inherits an error of that size.

Both halves of the explanation are proved. First, the mean-field edge is degenerate: an exact identity (Theorem 2) shows that the tangent gain touches 1 rather than crossing it, $G(g) = 1+\kappa_p(g-1)^2+\ldots$, with $\kappa_2 = 1/3$ for every smooth saturating odd activation and $\kappa_1 = \pi/32$ for softsign. Second, for MLPs and regular graphs with uniform input the state variance and the state–tangent alignment form an exact two-state Markov chain, and the stationary alignment that F channels cannot suppress costs the exponent exactly $1/F$. Setting the quadratic against this cost places the quenched crossing at $\log g = \log g_{\mathrm{lin}}(F) + \sqrt{3/F}\,(1+o(1))$ (Theorem 3, Figure 1). Prior finite-width theories re-tune the critical *kernel* perturbatively in $1/F$ (Roberts et al., 2022); our locus is the zero of the quenched top Lyapunov exponent, and an $O(1/F)$ correction to that exponent moves it by $F^{-1/2}$, non-analytically

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.

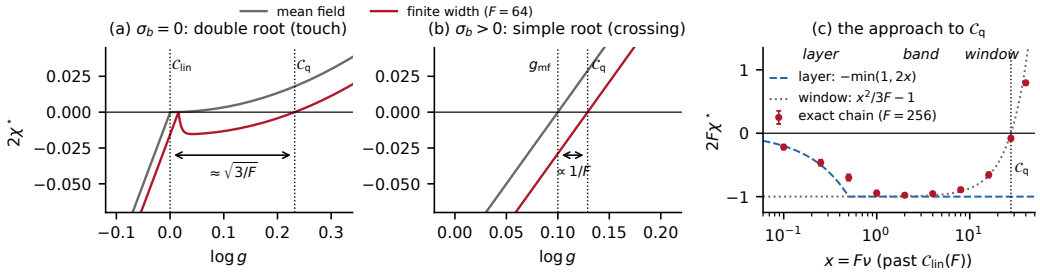


Figure 1: Why root multiplicity sets the displacement. (a, b) To scale at $F = 64$. At zero bias the mean-field exponent touches zero (a double root), and lowering it by $1/F$ moves the crossing by $\sqrt{3/F}$; at nonzero bias it crosses with nonzero slope, and the same lowering moves it by $O(1/F)$. (c) The approach past $\mathcal{C}_{\text{lin}}(F)$ on the exact chain ($F = 256$, 32 seeds, ± 2 SE): the boundary layer $-\min(1, 2x)$ (Theorem 5, Proposition 8; the approach is slow, $O(1/\log F)$, at the kink $x = \frac{1}{2}$), the band at -1 (Corollary 1), and the window $x^2/3F - 1$ (Theorem 3).

in $1/F$. This reconciles our measurements with reports that finite-width MLPs follow the mean-field threshold perturbatively (Terin, 2025). The principle is that root multiplicity sets the exponent, and it makes three predictions we fixed before measuring. The argument has three steps: root multiplicity sets the exponent (§3), the exact chain sets the constant (§4), and registered measurements test both (§5). Nonzero bias makes the root simple, so the decay steepens to F^{-1}. Softsign's smaller κ makes its displacement 1.84 times that of $\tanh$. On a bipartite regular graph a second tangent block, with weakened alignment, removes the $1/F$ cost, so the displacement collapses to $o(F^{-1/2})$. The first two hold on full networks. For the third, the measured fields behave as predicted, but an audit of our own estimator left its registered crossing tests inconclusive, and a shallower test was falsified by its own rule (§5).

Graphs are the testbed, not the source of the effect, and enter in two ways. Non-bipartite regular graphs with uniform input are exactly the MLP, and bipartite ones form a distinct branch. Irregular graphs differ at small width, where every one we tested crosses and the MLP does not; at large width we find no difference that survives an audit of our estimator, except on hub-dominated graphs, displaced more at the widths tested (Appendix J).

Notation. Five loci (Appendix A): $\mathcal{C}_{\text{lin}}$ and $\mathcal{C}_{\text{lin}}(F)$, the linearization at the origin at infinite and at finite width; $\mathcal{C}_{\text{ind}}$, mean field at the fixed point reached; $\mathcal{C}_{\text{ann}}$, the annealed boundary; $\mathcal{C}_{\text{q}}$, the zero of the quenched top Lyapunov exponent $\chi^\star$; and $\widehat{\mathcal{C}}_{L,S}$, the first sign change of $\chi^\star$ estimated at depth L from S realizations, which is what we measure. $\mathcal{C}_{\text{q}}$ is computed exactly only for MLPs and regular graphs with uniform input; **elsewhere we do not claim to have located it**. Distance past $\mathcal{C}_{\text{lin}}(F)$ is $\nu = \log g - \log g_{\text{lin}}(F)$, read on two scales: the window $s = \nu\sqrt{F}$, where the crossing sits ($s \to \sqrt{3}$), and the boundary layer $x = F\nu$ next to $\mathcal{C}_{\text{lin}}(F)$ (Figure 1(c)).

Contributions.

- *Proved:*
  - the exact mean-field identity, whose null is a double root with coefficient κ_p (Theorem 2), and the root shift: an $O(1/F)$ cost at a root of multiplicity r displaces the crossing by $(a/\kappa F)^{1/r}$, so double roots give $F^{-1/2}$ and simple roots F^{-1} (§3);
  - the critical-window law for MLPs and regular graphs with uniform input, both branches, constant included (Theorem 3), and the $O(1/F)$ cost at nonzero bias (Proposition 14);
  - $|2F\chi^\star| \le 1 + o(1)$ between $\mathcal{C}_{\text{lin}}(F)$ and $\mathcal{C}_{\text{q}}$, sharp (Corollary 1); at $\nu = x/F$ the chain's limit is an explicit diffusion with $2F\chi^\star \to -\min(1, 2x)$ for every $x > 0$ (Proposition 7, Theorem 5, Proposition 8); and the $\tanh$ $F^{-1/2}$ term of Λ (Proposition 9);
  - the exact finite-width linear boundary for every symmetric operator (Theorem 1), the quenched bound for MLPs and regular graphs (Theorem 4), and direction bounds for

general graphs at every depth and bias, including an almost-sure tangent-weighted one and a sharp one-step trade-off (Propositions 2, 4, 12, 13).

- *Computed:* $\mathcal{C}_{\mathrm{q}}$ for MLPs to $F = 4096$ from the exact reduction (Table 1).
- *Measured, with predictions fixed before running:*
  - $F^{-1/2}$-consistent decay at zero bias (fitted exponent -0.56 $[-0.66, -0.46]$);
  - the F^{-1} law at nonzero bias (-1.09 $[-1.36, -0.92]$);
  - the softsign prefactor (1.7–2.2 against 1.84 predicted);
  - the bipartite branch, by a registered field test that matches the exact chain (crossing-location tests are inconclusive after an audit of our estimator, and the shallower $F = 1024$ one was falsified);
  - nearby inputs merging below $\mathcal{C}_{\mathrm{q}}$ and separating above it, and small-width graph effects.
- *Conjectured:* the quenched bound on irregular graphs, certified on $199/200$ cells.
- *Negative:* the crossing is not a better initialization, and at depth $L \geq 4F$ a worse one (post hoc).

2 Setup, a trap, and the exact linear boundary

Fix a graph with symmetric propagation operator A, spectral radius $\mu_{\max}(A) = \max_k |\lambda_k(A)|$, and consider

$$H^{\ell+1} = \phi\big(\sigma_{\mathrm{s}} H^{\ell} W_s^{\ell} + \sigma_{\mathrm{n}} A H^{\ell} W_n^{\ell} + \sigma_{\mathrm{b}} \mathbf{1} b^{\ell\top}\big)\,, \tag{1}$$

with $W_\bullet^{\ell} \in \mathbb{R}^{F\times F}$ i.i.d. $\mathcal{N}(0, 1/F)$ entries, $b^{\ell} \sim \mathcal{N}(0, I_F)$, $H^0 = \mathbf{1}$, and $\sigma_{\mathrm{b}} = 0$ except in §5 and §K. An MLP is the case $n = 1$. The diagnostic is the top tangent Lyapunov exponent,

$$v^{\ell+1} = \phi'(Z^{\ell}) \odot \big(\sigma_{\mathrm{s}} v^{\ell} W_s^{\ell} + \sigma_{\mathrm{n}} A v^{\ell} W_n^{\ell}\big)\,, \qquad \chi^{\star} = \lim_{L\to\infty} \tfrac{1}{L} \sum_{\ell} \log \|v^{\ell}\|, \tag{2}$$

estimated by finite-L normalization over the second half of the chain and averaged over S weight realizations. Activations are odd with $|\phi'| \leq \phi'(0) = 1$ (Appendix E explains why oddness is necessary). **The trap:** at $\sigma_{\mathrm{n}} = 0$ the graph term vanishes and a sweep measures a width-F MLP; the tell is $\max \chi^{\star}$ agreeing to four decimal places across operators. We use the *graph-active ray* $\sigma_{\mathrm{n}} = 1/(2\mu_{\max}(A))$, on which $g = \sigma_{\mathrm{s}}^2 + \mu_{\max}^2 \sigma_{\mathrm{n}}^2 = \sigma_{\mathrm{s}}^2 + 1/4$ and $\mathcal{C}_{\mathrm{lin}}$ sits at $\sigma_{\mathrm{s}} = 0.8660$ for every graph, while 16 (graph, operator) cells give 15 distinct values of $\max \chi^{\star}$.

Proposition 1 ($\mathcal{C}_{\mathrm{lin}}$ in closed form)**.** *Let ϕ be odd and C^1 with $\phi'(0) = 1$, and let A be symmetric. At $\sigma_{\mathrm{b}} = 0$ the origin is a fixed point of equation 1, and the annealed stability boundary of equation 2 about it is $\mathcal{C}_{\mathrm{lin}} : \sigma_{\mathrm{s}}^2 + \mu_{\max}(A)^2 \sigma_{\mathrm{n}}^2 = 1$.*

Theorem 1 (Exact finite-width linear boundary)**.** *For $\phi = \mathrm{id}$ and any symmetric A, the quenched top exponent of equation 2 is, almost surely and for every initial tangent with a component on a top eigenvector, $\chi^{\star}_{\mathrm{lin}}(F) = \frac{1}{2}\log g + \gamma_F$ with $\gamma_F = \frac{1}{2}\big(\psi(F/2) - \log(F/2)\big) = -\frac{1}{2F} + O(F^{-2})$. Its zero set is $\mathcal{C}_{\mathrm{lin}}(F) : \sigma_{\mathrm{s}}^2 + \mu_{\max}^2 \sigma_{\mathrm{n}}^2 = g_{\mathrm{lin}}(F) := e^{-2\gamma_F} = 1 + F^{-1} + O(F^{-2})$.*

The proof block-diagonalizes the tangent in the eigenbasis of A; each block is a product of i.i.d. Gaussian matrices, whose exponent is Newman (1986)'s (Appendix C.2). The boundary is graph-independent at every width, and it replaces the empirical interpolation used by the paper we build on; an independent random-matrix computation agrees to Monte-Carlo error (Appendix G). All displacements below are quoted over $\mathcal{C}_{\mathrm{lin}}(F)$, which on the ray is $\sqrt{g_{\mathrm{lin}}(F) - 1/4}$: 0.9707 at $F = 6$, 0.8683 at $F = 256$.

Proposition 2 (Annealed domination)**.** *Let ϕ be odd and differentiable with $|\phi'| \leq \phi'(0) = 1$ and A symmetric. Then $\mathbb{E}\|v^{\ell+1}\|^2 \leq g\,\mathbb{E}\|v^{\ell}\|^2$ for the unnormalized recursion, so $\mathcal{C}_{\mathrm{ann}}$ lies on or outside $\mathcal{C}_{\mathrm{lin}}$.*

3 Root multiplicity: the mean-field edge is a double root

Linearizing at the fixed point the recursion *actually reaches* gives

$$q^{\star} = g\,\mathbb{E}\Big[\phi\big(\sqrt{q^{\star}}\,\xi\big)^2\Big] + \sigma_{\mathrm{b}}^2, \qquad \mathcal{C}_{\mathrm{ind}} : G(g) := g\,m(q^{\star}) = 1, \quad m(q) = \mathbb{E}[\phi'(\sqrt{q}\,\xi)^2]. \tag{3}$$

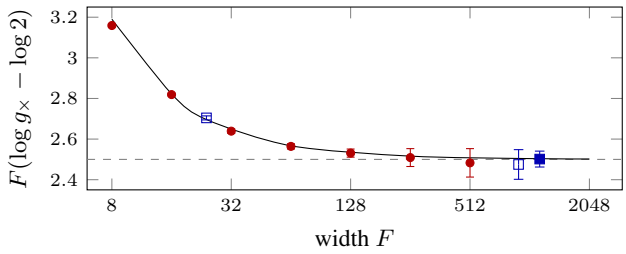


Figure 2: ReLU, a simple root: the crossing shift times F on a ReLU MLP. Dashed: $5/2$, the limit proved in Proposition 10; solid: the exact χ^2 value $P(F)$. Red circles: exploratory runs; blue squares: registered held-out widths ($F = 24$; $F = 1024$, open: underpowered first run, filled: powered follow-up, offset in F). Bars: 95% intervals. The shift is $O(1/F)$, not $F^{-1/2}$.

Theorem 2 (Exact mean-field identity)**.** *Let ϕ be odd with $|\phi'| \le 1$, $\psi(x) = \phi(x)/x$ and $X \sim \mathcal{N}(0, q^\star)$. Then*

$$G(g) - 1 = g\,\mathbb{E}\big[X^2\psi'(X)^2\big] - \sigma_{\mathrm{b}}^2/q^\star. \tag{4}$$

If $\psi(x) = 1 - c|x|^p + O(|x|^{p+\delta})$ near 0 with $c > 0$, and the corresponding expansion of ψ', then at $\sigma_{\mathrm{b}} = 0$, $G(g) - 1 = \kappa_p(g-1)^2(1 + o(1))$ as $g \downarrow 1$, where $\kappa_p = p^2 M_{2p}/(4M_{p+2}^2)$ and $M_s = \mathbb{E}|\xi|^s$.

Proposition 3 (Mean field predicts no displacement)**.** *Let ϕ be odd and nonlinear with $|\phi'| \le \phi'(0) = 1$, A symmetric and $\sigma_{\mathrm{b}} = 0$. Then $G(1) = 1$ and $G(g) > 1$ for every $g > 1$ at which the nontrivial fixed point exists, so $\mathcal{C}_{\mathrm{ind}} = \mathcal{C}_{\mathrm{lin}}$.*

The proof of Theorem 2 is Stein's lemma applied to $x\psi(x)^2$ (Appendix C.5). At $\sigma_{\mathrm{b}} = 0$ the right side of equation 4 is non-negative and vanishes only for linear ϕ, which is Proposition 3. On the 17 graphs of §5, $\mathcal{C}_{\mathrm{ind}}$ and $\mathcal{C}_{\mathrm{lin}}$ have identical median relative error, 36.3%. The expansion says how G leaves 1: quadratically, for every such activation, with $\kappa_2 = 1/3$ for $\tanh$, erf, $\arctan$ and sublinear ($p = 2$) and $\kappa_1 = \pi/32$ for softsign ($p = 1$). The null is therefore a double root on the branch where the fixed point is nontrivial ($G = g$ below $g = 1$). At $\sigma_{\mathrm{b}} > 0$ the term $-\sigma_{\mathrm{b}}^2/q^\star$ makes the root simple.

Root multiplicity sets the exponent. *Root shift:* if $X_F(g) = c\,\nu^r - a/F + o(F^{-1})$ uniformly for $\nu = g - g_0(F) = O(F^{-1/r})$, its zeros with $\nu > 0$ there sit at $\nu = (a/cF)^{1/r}(1 + o(1))$, as substituting $\nu = sF^{-1/r}$ shows. So an $O(1/F)$ cost moves a double root by $F^{-1/2}$ and a simple one by F^{-1}. At $\sigma_{\mathrm{b}} = 0$, $r = 2$ for every activation of the form in Theorem 2, so no such odd activation gives $r = 3$; at $\sigma_{\mathrm{b}} > 0$, $r = 1$. ReLU, outside this class, has a simple root ($G = g/2$) and a proved $5/(2F)$ shift (Proposition 10, Figure 2). There the $O(1/F)$ cost is proved for the MLP, with an explicit constant (Proposition 14), and the crossing it predicts lies inside the measured interval at four of five widths (Appendix K). What remains is the size of the $O(1/F)$ cost at zero bias.

4 The exact chain sets the constant

What moves the crossing. At finite width, the state variance q and the cosine ρ between state and tangent of an MLP, or of a regular graph with uniform input, form an exact two-state Markov chain. Twice the exponent equals the stationary mean of a one-step functional $\Lambda(q) - \rho^2 qK(q) + \dots$, where Λ is the mean-field log-gain and $K(0) = -(\phi'^2)''(0) = 12|a|$ with $\phi(x) = x + ax^3 + \dots$ (Appendix C.9). Two effects enter at order $1/F$: the exact linear penalty of Theorem 1 shifts the stationary state, so the mean-field term is evaluated at $\nu = \log g - \log g_{\mathrm{lin}}(F)$ rather than at $\log g$; and finitely many channels cannot suppress the random alignment between state and tangent: it contracts at a rate proportional to q and is re-excited by sampling noise of variance $1/F$, so its stationary cost is $1/F + o(1/F)$ (Figure 3). The double root turns that $O(1/F)$ cost into an $O(F^{-1/2})$ displacement (Figure 1); Theorem 3 supplies $r = 2$, $c = 1/3$, $a = 1$ in the root shift, hence $\sqrt{3/F}$.

Theorem 3 (Critical-window law)**.** *Let ϕ be odd, bounded and C^5 with bounded derivatives, $|\phi'| \le \phi'(0) = 1$, $\phi' > 0$ with ϕ'^2 non-increasing in $|x|$, $\phi'''(0) = 6a < 0$, $|\log \phi'(x)| \le C(1 + x^2)$ and $\phi(y)/y \ge c_\psi/(1 + |y|)$. Take an MLP or a connected d-regular graph with node-uniform input,*

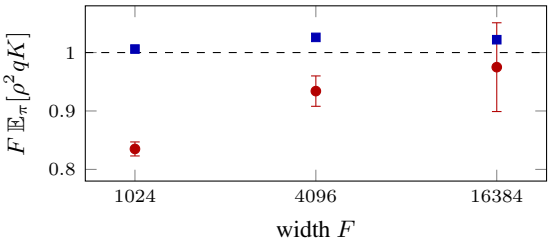


Figure 3: The mechanism in numbers: the stationary alignment cost $F\,\mathbb{E}_\pi[\rho^2 qK(q)]$ at the predicted crossing, on the exact chain (tanh, 8 seeds; red circles, ± 2 SE; dashed: the limit 1). It tends to 1, the $1/F$ cost of Proposition 5 that the double root turns into the $\sqrt{3/F}$ shift. With K frozen at $K(0)$ (blue squares) the cost is already 1 within 3%; the gap between the two series is the $O(\mathbb{E}_\pi[q^2\rho^2])$ correction.

$\sigma_b = 0$, $\sigma_s, \sigma_n > 0$, *and let* $\nu = \log g - \log g_{\mathrm{lin}}(F)$. *Uniformly for* $\nu\sqrt{F} \in [s_0, s_1]$, $0 < s_0 < s_1$: *(i) if the graph is bipartite,* $2F\chi^\star = F\nu^2/3 + o(1)$*; (ii) otherwise, and in particular for the MLP,* $2F\chi^\star = F\nu^2/3 - 1 + o(1)$. *Hence in case (ii) every sign change of* $\chi^\star$ *in the window lies at* $\log g_\times = \log g_{\mathrm{lin}}(F) + \sqrt{3/F}\,(1 + o(1))$*; in case (i) there is none, and the crossing is displaced by* $o(F^{-1/2})$.

Proof sketch (Appendix C.9). (1) Given the past, channel pairs (pre-activation, tangent increment) are i.i.d. bivariate Gaussian, so (q, ρ) is a Markov chain with a unique stationary law π and $2\chi^\star = \mathbb{E}_\pi \log(gB/F)$ (Lemma 1). (2) Expanding in q, $2F\chi^\star = F\,\mathbb{E}_\pi\Lambda(q) - F\,\mathbb{E}_\pi[\rho^2 qK(q)] + o(1)$, and stationarity of $\log q$ with the double root of Theorem 2 gives $F\,\mathbb{E}_\pi\Lambda = F\nu^2/3 + o(1)$. (3) Stationarity of $\log(1-\rho^2)$, made rigorous by truncation because the contraction degenerates at $|\rho| = 1$, balances a population contraction $(K/S)q\rho^2$ (Lemma 3) against a sampling cost of exactly $1/F$ (Lemmas 4–7, via the Bartlett identity), so the alignment term tends to 1 (Proposition 5). On a bipartite graph a second tangent block cancels it. (4) The root shift turns $F\nu^2/3 - 1$ into $\sqrt{3/F}$. On the chain, $(\log g_\times - \log g_{\mathrm{lin}})\sqrt{F/3} = 1.15$, 1.04, 1.02 and 1.005 ± 0.004 at $F = 64$, 256, 1024 and 4096, and softsign, outside the smoothness class, follows the same law with $\kappa_1 = \pi/32$ (Figure 4(b)). The chain's stationary crossing is the quenched boundary $\mathcal{C}_q$. For MLPs and regular graphs we therefore compute $\mathcal{C}_q$ to Monte Carlo precision (Table 1), and the bipartite case is a large-width graph effect that the MLP cannot show (bipartite $\mathcal{C}_q$ displacement 1.78%, 0.87% and 0.43% at $F = 256$, 512 and 1024, against 7.65%, 5.32% and 3.71%).

The full profile. Past $\mathcal{C}_{\mathrm{lin}}(F)$ the exponent has one profile on three scales (Figure 1(c)): $2F\chi^\star \to -\min(1, 2x)$ at $\nu = x/F$, where the chain converges to an explicit diffusion (Theorem 5, Proposition 8); $2F\chi^\star \geq -1 - o(1)$ up to the crossing, sharply (Corollary 1); and $2F\chi^\star = s^2/3 - 1 + o(1)$ at $\nu = s/\sqrt{F}$.

5 Registered measurements of the $F^{-1/2}$ law in MLPs and graphs

Protocol. The width corpus has 24 graphs up to $n = 97$, ten from TU benchmarks, swept at $F \in \{4, \ldots, 256\}$ on the graph-active ray at $L = 160$ with 8 seed groups of $S = 8$. Its grid step (0.027 in σ_s) is as large as the whole displacement at $F = 256$, so a second study interpolates crossings on a grid of step $\approx 1\%$: an MLP of the same gain at $F \in \{16, \ldots, 2048\}$, $L \in \{160, 640, 2560\}$, and 8 of the 24 graphs at $F \in \{16, 256, 1024\}$, each with a linear control.

The measured displacement is consistent with the predicted $F^{-1/2}$ scaling, and is not a finite-depth transient. On the 24 graphs the tanh displacement is +37.3%, +22.0%, +11.2% and +8.2% at $F = 16$, 48, 128 and 256, with fitted exponent -0.56, 95% CI $[-0.66, -0.46]$ by bootstrap over graphs; the other four activations give -0.54 to -0.66 (all excluding -1; arctan and sublinear also exclude $-1/2$, on the steep side). The fitted exponent is a finite-range measurement, not the asymptote of Theorem 3: the exact chain's own displacement has slope -0.61 over $F \in [16, 2048]$ and reaches $-1/2$ only asymptotically; the data do not distinguish $-1/2$ from -0.56 at these widths.

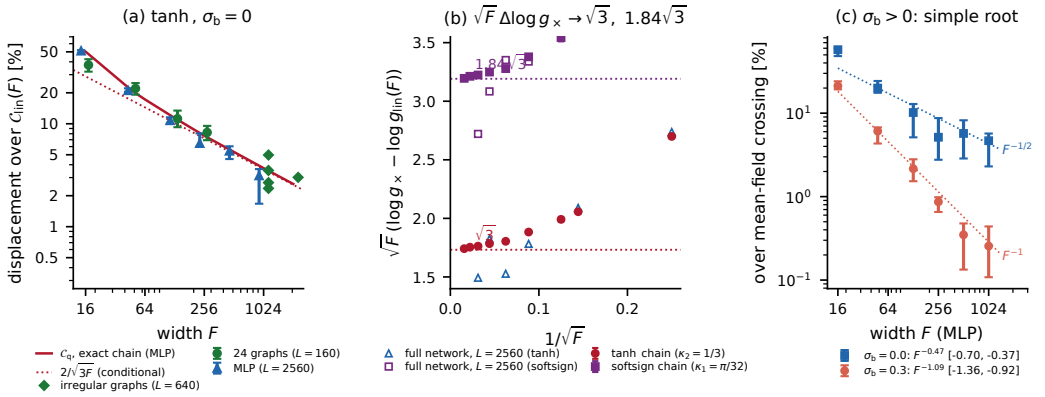


Figure 4: **(a)** Displacement over the exact linear boundary $\mathcal{C}_{\mathrm{lin}}(F)$ at $\sigma_{\mathrm{b}} = 0$, $\tanh$, log–log. Circles: the 24-graph corpus ($L = 160$); triangles: an MLP at $L = 2560$; both pooled over seed groups (Table 1). Diamonds: four irregular graphs re-measured at $L = 640$ (pooled). Solid: exact-chain $\mathcal{C}_{\mathrm{q}}$; dotted: the $2/\sqrt{3F}$ asymptote. **(b)** Critical-window convergence: $y = \sqrt{F}\,(\log g_{\times} - \log g_{\mathrm{lin}}(F))$ against $F^{-1/2}$. Filled: the exact chain (± 1 SE) for $\tanh$ and softsign, descending onto the predicted asymptotes $\sqrt{3}$ and $1.84\sqrt{3}$ — the gap between them is the predicted prefactor $\sqrt{\kappa_2/\kappa_1}$; in units of its asymptote the $\tanh$ chain reads 1.005 ± 0.004 at $F = 4096$. Open: full networks at $L = 2560$ (pooled over seed groups), scattered about the chain within their intervals. **(c)** Pre-registered MLP bias test, with displacement over the mean-field crossing (pooled over seed groups). Bars are 95% bootstrap intervals over graphs or seed groups.

Table 1: Displacement (%) over $\mathcal{C}_{\mathrm{lin}}(F)$, $\tanh$, $\sigma_{\mathrm{b}} = 0$. $\mathcal{C}_{\mathrm{q}}$: the MLP's quenched crossing from the exact two-state chain (standard errors ≤ 0.3); it tends to $2/\sqrt{3F}$ (Theorem 3). Measured columns are pooled: the crossing of the mean of all fields in a cell (8 seed groups of $S = 8$ for the MLP, $S = 4$ for the 8 graphs, $S = 8$ on the corpus grid of step 0.027), from a pre-registered re-run (Appendix J). Brackets: 95% bootstrap intervals over seed groups (MLP) or graphs. Averaging each group's first sign change instead gave, at $F = 16, 48, 128, 256, 512, 1024$, MLP $L = 2560$: 47.9, 20.6, 10.4, 6.9, 4.5, 3.4; corpus: 31.1, 15.5, 9.5, 6.8. †Above the MLP, mostly on bipartite trees; at $L = 640$ the four re-measured irregular graphs read 2.4–5.0% (§5).

F	$\mathcal{C}_{\mathrm{q}}$ (exact)	MLP, $L = 160$	MLP, $L = 2560$	24 graphs, $L = 160$	8 graphs, $L = 160$
16	50.2	50.8 [41.3, 53.2]	51.1 [49.0, 51.9]	37.3 [32.1, 42.4]	28.2 [23.8, 32.1]
48	20.7	17.8 [16.0, 23.8]	21.1 [20.4, 22.1]	22.0 [19.3, 24.9]	—
128	11.4	10.1 [4.2, 11.7]	10.8 [10.2, 11.6]	11.2 [9.3, 13.3]	—
256	7.7	5.4 [2.7, 9.1]	6.5 [6.0, 8.1]	8.2 [6.9, 9.5]	6.6 [5.5, 7.8]
512	5.3	7.1 [2.3, 7.8]	5.4 [4.5, 6.0]	—	—
1024	3.7	4.2 [2.0, 5.2]	3.1 [1.7, 3.6]	—	6.7 [4.9, 8.8]†
2048	2.6	2.8 [2.6, 4.5]	—	—	—

Among single-parameter laws, cF^{-1}, $cF^{-1/3}$ and $c/\log F$ are rejected (weighted squared error 119, 39 and 72 against 4.2 for $cF^{-1/2}$), and a plateau term is not supported (fitted offset -1.9 points). At $F = 256$ every one of the 24 graphs still crosses with its own seed-group CI above $\mathcal{C}_{\mathrm{lin}}(F)$: $+8.2\%$ $[+6.9, +9.5]$ on average, $+9.9\%$ $[+7.7, +11.8]$ on the ten TU graphs. Table 1 sets the measurements against the exact $\mathcal{C}_{\mathrm{q}}$ of the MLP. Increasing depth from $L = 160$ to 2560 moves no estimate toward zero, so the effect is not a finite-depth transient. Measured correctly, the network sits on $\mathcal{C}_{\mathrm{q}}$: at $L = 2560$ the MLP's interval contains the exact $\mathcal{C}_{\mathrm{q}}$ at five of six widths, and from $F = 48$ to 256 the corpus means lie within 1.3 points of it. Our earlier estimates read 0–16% (MLP, $L = 2560$) and 12–38% (corpus) below $\mathcal{C}_{\mathrm{q}}$. They averaged each seed group's *first* sign change, which noise fires early where the field is flat, and a pre-registered re-run of every cell, pooling each cell's fields before locating the crossing, removes the gap (Appendix J). Linear controls recover $\mathcal{C}_{\mathrm{lin}}(F)$ to within 0.15% at every width.

Table 2: Predictions registered before measuring, with the rule as written and the outcome. Six hold, eight are inconclusive, and one each is neither confirmed nor refuted, falsified at a depth too shallow for the width, refuted for lack of power, or void. †Changed by a pre-registered audit of our estimator, which pools each graph's fields instead of averaging per-group first sign changes (§5). §Superseded by the field test. The bias, softsign and rate rows are pooled by a second audit (verdicts unchanged). ‡Detail, and all stored and per-group values, in Appendix J.

prediction	rule	measured	outcome
simple root at $\sigma_{\mathrm{b}} > 0$ gives F^{-1}	CI near -1; containing $-1/2$ falsifies	-1.09 $[-1.36, -0.92]$	confirmed
$\kappa_1 = \pi/32$ makes softsign $1.84\times$ $\tanh$	CI contains 1.84, excludes 1.3	2.24, 1.71 at $F = 256, 512$	confirmed
bipartite crossing: (i) classes separate; (ii) equals the chain's 0.92%; (iii) survives irregularity	(i) CI $< 2.66\%$; (ii) CI $\ni 0.92$; (iii) both pairs separate	(i) 1.5, 2.1%; (ii) $[0.40, 3.51]$; (iii) 2.87, 2.54 vs 5.52, 6.11	inconclusive; (ii) too wide†§
… (i) at $F = 1024$, $L = 640$	falsified if a mean lies inside its partner's CI	all three do	**falsified**
bipartite branch removes the $1/F$ cost (field)	network = chain, gap > 0.5	0.78–0.88 (chain 0.79–0.88)	confirmed
bipartite branch is a distinct *rate*	bipartite slope excludes $-1/2$, below partner's	-0.51 $[-0.91, -0.09]$; resolved, -1.95 $[-2.39, +0.78]$ vs -0.63 $[-0.75, -0.50]$	**refuted** twice*
training cost at $\mathcal{C}_{\mathrm{lin}}$ grows with L/F	linear control detects its gap at $L = 16F$ in ≥ 2 of 3 widths	1 of 3	**void**
near inputs merge below $\mathcal{C}_{\mathrm{q}}$, separate above ($L = 16F$)	switch CI excludes $\mathcal{C}_{\mathrm{lin}}(F)$, ≥ 0.7 of $\mathcal{C}_{\mathrm{q}}$	1.23, 1.03 vs 1.26, 1.09	confirmed
$\mathcal{C}_{\mathrm{q}}$'s gradient spread slows SGD, not per-layer SGD	SGD gap > 0 in ≥ 2 of 3 widths, per-layer gap $\ni 0$	both > 0; most arms at chance	inconclusive
trainability collapses at $\mathcal{C}_{\mathrm{q}}$ ($L = 4F$)	slope of $\log \Delta_c$ on $\log F$ $\ni -1/2$, $\not\ni 0$	-0.88 $[-1.05, -0.79]$	inconclusive
$F^{-1/2}$ law on 61 graphs to $n = 500$	every family's slope in $[-0.8, -0.4]$; spread falls	3 of 8 families; spread rises	inconclusive‡
layer law on irregular graphs	± 0.1 of the exact chain	$6/6$ graphs	confirmed
reservoirs track the log-gain	$\geq 5/6$ datasets, both tests	$5/6$, $4/6$	inconclusive
hub/ER gap closes by $F = 512$	ratio CI in $[0.9, 1.1]$	1.25 at 256‡	inconclusive
ReLU shift at held-out F	CIs in $\pm 3\%$ of 2.70, 2.50	2.70, 2.48 (CI to -4%)	inconclusive
… powered rerun, $F = 1024$	CI in $\pm 3\%$ of 2.504	2.502 $[-1.7, +1.5]\%$	confirmed

Eighteen predictions, fixed before measuring. Each was committed with its decision rule before the run that tested it; Table 2 records what happened, per-graph numbers in Appendix J.

The failures have three causes, detailed in Appendix J: a depth too shallow for the width, a rule written against $\mathcal{C}_{\mathrm{q}}$ while the experiment measures $\widehat{\mathcal{C}}_{L,S}$, and a rule without the power to detect what it asked. *A post hoc fourth width, pooled, puts the slopes on -1 and $-1/2$, but their difference is -0.63 $[-1.11, +0.83]$. The void test is discussed in Appendix I.

The confirmations carry the mechanism: both predicted exponents appear, and so does the κ_p prefactor ratio. The bipartite verdicts did not survive a pre-registered audit of our estimator (Appendix J): in 15–28 of 32 seed groups per graph the field changes sign more than once. Pooling each graph's fields puts the regular controls on the exact $\mathcal{C}_{\mathrm{q}}$ (5.44, 5.28 against 5.32%) and the bipartite graphs at 1.5–2.9% with intervals reaching 3.7%, which leaves the class and irregular tests inconclusive. The field itself is well-conditioned: a registered field test on fresh graphs finds the tangency of Theorem 3(i) (Table 2, Appendix J).

Graphs. A d-regular graph with uniform input keeps its state node-uniform, so its forward pass is exactly the MLP of gain g. Its tangent adds blocks of smaller gain and, if the graph is bipartite, a block of equal gain with weakened alignment (Theorem 3(i)). Irregular graphs differ from the MLP at small width. At $F \leq 6$ the MLP has no crossing at $\sigma_{\mathrm{s}} \leq 3$, nor does a random 4-regular graph at $g \leq 5$,

while all 17 irregular graphs cross, at 45–91% (mean 63%) beyond $\mathcal{C}_{\mathrm{lin}}(6)$ once averaged until each field changes sign exactly once (Figure 18); in two dimensions the measured level set contains no pixel inside $\mathcal{C}_{\mathrm{lin}}$ on six graphs (Figure 17, Appendix L). At $F = 16$ irregular graphs are *less* displaced than the MLP (Table 1), a random 4-regular graph matches it, and from $F = 48$ to 256 graphs and MLP agree within their intervals. Node heterogeneity therefore creates or pulls in the crossing at small width. Near $g_{\mathrm{lin}}(F)$ it drops out: gapped irregular graphs match the exact chain there (Table 2). At the largest widths no graph effect survives the audit. At $F = 1024$, $L = 160$ the 8-graph mean, 6.7% $[4.9, 8.8]$, exceeds the MLP's 4.2%, mostly through bipartite trees (11.2%), where the exponent only touches zero. Our registered test that more depth removes the excess is refuted: at $L = 640$ all four irregular graphs fall, but only one leaves its $L = 160$ interval. The exception is hub-dominated graphs, displaced 1.2–1.4× as much as Erdős–Rényi graphs from $F = 64$ to 256, the largest width at which that test resolves (Table 2). Otherwise we find no large-width difference between non-bipartite graphs and MLPs (all per-graph values in Appendix J).

Calibration and robustness. Linear controls recover $\mathcal{C}_{\mathrm{lin}}(6)$ to within 1.6%. The displacement survives A, $A + I$ and $D^{-1/2}AD^{-1/2}$ (GCN-type), is lost only for the non-symmetric $D^{-1}A$, and moves under three points with L, S or graph size (Appendix J).

6 Beyond regular graphs: a proved bound and an open problem

Theorem 3 needs node-uniform states. For every symmetric operator what can be proved is a one-sided bound with the *quenched* linear exponent. Let $\bar{m}$ denote the long-run average of $\phi'(Z)^2$ realized along the trajectory, over layers, nodes and channels.

Theorem 4 (Quenched domination for MLPs and regular graphs)**.** *Let ϕ be odd, bounded and C^1 with $|\phi'| \le \phi'(0) = 1$, ϕ'^2 non-increasing in $|x|$ and $|\log \phi'(x)| \le C(1 + x^2)$. Let A be d-regular (or $n = 1$), H^0 node-uniform, and $\sigma_{\mathrm{b}} \ge 0$. Then, almost surely,*

$$\chi_\phi^\star \;\le\; \chi_{\mathrm{lin}}^\star(F) + \tfrac{1}{2}\log \bar{m}. \tag{5}$$

Here $\bar{m}$ is read as a lim sup*. Consequently any crossing satisfies $g\,\bar{m} \ge g_{\mathrm{lin}}(F) > 1$, so none lies inside $\mathcal{C}_{\mathrm{lin}}(F)$. For $\sigma_{\mathrm{b}} > 0$, $\bar{m} < 1$ always, so every crossing lies strictly outside it; at $\sigma_{\mathrm{b}} = 0$ that the sign change is not on $\mathcal{C}_{\mathrm{lin}}(F)$ itself is Theorem 3, for large F.*

The proof block-diagonalizes the tangent as for Theorem 1. Given the past, output channels are i.i.d., so the linear gain in each block is an exact scaled χ_F^2; the nonlinearity enters through pairs (Z_a, u_a) that are jointly Gaussian but correlated, and a total-positivity argument shows that this state–tangent alignment can only *lower* the gain. A martingale strong law and Jensen finish (Appendix C.14). Numerically equation 5 holds in all 48 tested cells, with slack under 0.005 at $F = 256$. The bound cannot fix the magnitude of the displacement, because the alignment cost that sets it (Theorem 3) is exactly what it discards.

General graphs. Two steps of the proof need a node-uniform state: on an irregular graph the channel variance differs across nodes, and the linear gain from a node-spread tangent is no longer a scaled χ_F^2. We therefore state equation 5 for general symmetric A as a conjecture.

Conjecture 1. *Under the hypotheses of Theorem 4 without regularity or uniform input, equation 5 holds.*

It is certified on 199 of 200 graph × activation cells at $F = 6$ with one unresolved and none violated, and an adversarial search over star, wheel, lollipop, barbell, path and tree graphs at $F \in \{2, 6, 24\}$ found no violation beyond three standard errors in 168 cells (Appendix C). What holds for every symmetric A, depth and bias is a weaker bound on the expectation of the estimator we measure, with $\frac{1}{2}\log g$ in place of $\chi_{\mathrm{lin}}^\star(F)$:

Proposition 4 (Finite-depth domination)**.** *Let ϕ be as in Proposition 2 with ϕ'^2 non-increasing in $|x|$, A symmetric, $\sigma_{\mathrm{b}} \ge 0$, $\sigma_{\mathrm{s}} > 0$ (or read $\log 0 = -\infty$), and $q_{\min}^\ell = \min_{i,a} \mathrm{Var}(Z_{ia}^\ell \mid \mathcal{F}_\ell) \ge \sigma_{\mathrm{b}}^2$. If $\mathbb{E}|\log m(q_{\min}^\ell)| < \infty$, then for all $0 \le L_0 < L$, $\mathbb{E}\big[\frac{1}{L-L_0}\log\frac{\|v^L\|}{\|v^{L_0}\|}\big] \le \frac{1}{2}\log g + \frac{1}{2(L-L_0)}\sum_\ell \mathbb{E}\log m(q_{\min}^\ell) \le \frac{1}{2}\log\big(g\,m(\sigma_{\mathrm{b}}^2)\big)$.*

Graph dependence beyond the spectral radius is open. On the ray $\mathcal{C}_{\mathrm{lin}}$ is graph-independent, yet at $F = 6$ the converged crossings spread 1.41–1.85 on the 17 graphs, and $\mu_{\max}$ explains $R^2 = 0.24$ on a 65-graph corpus. Theorem 4 says where to look: a non-bipartite regular graph with uniform input has the MLP's forward pass and top exponent, so the spread must come from node heterogeneity or bipartite structure. The pre-registered heterogeneity statistic, the regularity ratio $\bar{d}/\mu_{\max}$, fails its $R^2 \geq 0.60$ gate, and its apparent partial support did not survive converging the noisy targets (Appendix J). The governing functional is open.

7 Discussion and limitations

Why boundary location matters. Every downstream use of the order/chaos boundary — box-counting its fractal geometry or comparing architectures — inherits its answer from the locator used to find it. At a simple critical point the standard locators are accurate to $O(1/F)$, but at the edge of chaos the mean-field edge is *degenerate* (Theorem 2): a double root, for every smooth saturating odd activation at once ($\kappa_2 = 1/3$; $\pi/32$ for softsign). There an $O(1/F)$ finite-width perturbation becomes $O(F^{-1/2})$ in the boundary's location (§3): a measurement that linearizes at a degenerate edge carries an error larger, in the quantity it reports, than the perturbation that causes it. The practical reading for finite-size studies of neural phase transitions is to check the root's multiplicity before trusting a linearized boundary (for $\tanh$, 28% more gain at $F = 64$).

The crossing is a diagnostic, not an initialization prescription. A pre-registered grid of 4512 training runs finds initializing at the measured crossing no better than at $\mathcal{C}_{\mathrm{lin}}$: -0.0027 $[-0.0072, +0.0020]$ paired accuracy over 96 cells, with 59 unable to separate the arms at all (Appendix H). The theory says why: between the two loci the exponent is within $(1 + o(1))/(2F)$ of zero per layer (Corollary 1), and where the state has decayed the network is effectively linear, so the gap multiplies every layer's gradient by one common factor, which a tuned learning rate removes at any depth. What the sign does control is whether nearby inputs merge. In a registered test at depth $L = 16F$, pairs at initial cosine $1 - 10^{-4}$ merge below, and separate above, $\nu\sqrt{F/3} = 1.23$ and 1.03 at $F = 32$ and 128 (Appendix I), against $\mathcal{C}_{\mathrm{q}}$ at 1.26 and 1.09: the theorem, restated in input space. Deeper networks do care: at $L = 4F$ a $\tanh$ MLP started at $\mathcal{C}_{\mathrm{q}}$ barely leaves chance on MNIST and CIFAR-10 while one at $g_{\mathrm{lin}}(F)$ trains (post hoc; its registered mechanism failed its control), and a registered sweep puts the collapse near half of $\mathcal{C}_{\mathrm{q}}$: training fails before the crossing, not at it (Appendix I).

Related work. Finite width is usually treated perturbatively: Roberts et al. (2022) expand in the depth-to-width ratio (see also Yaida, 2020; Dyer & Gur-Ari, 2020; Hanin & Nica, 2020a) and even re-tune criticality at $O(1/n)$, and Terin (2025) find finite-width MLPs following the mean-field threshold perturbatively. Theorems 2 and 3 give a second, non-perturbative criterion: the mean-field edge is a double root, so the Lyapunov crossing moves by $O(F^{-1/2})$, outside any $1/n$ expansion (Appendix B), and the activation enters (Hayou et al., 2019) only through κ_p. Cohen & Newman (1984) and Newman (1986) computed the Lyapunov spectrum of Gaussian matrix products (see also Crisanti et al., 1993), which Theorem 1 applies blockwise; their log-normal fluctuations, governed by $\sum_\ell 1/n_\ell$ (Hanin, 2018; Hanin & Nica, 2020b; Li et al., 2021), reappear as heavy-tailed gradients at $\mathcal{C}_{\mathrm{q}}$ (Appendix I). Those results correct the exponent at a fixed gain; the window law is a shift of its root. The two loci form a noisy bifurcation in the sense of random dynamics (Baxendale, 1994; Arnold, 1998), with noise of variance $1/F$ (Appendix B). Reservoir computing leaves random recurrent and graph networks untrained, with the weight scale set by an echo-state or stability condition (Jaeger, 2001; Gallicchio & Micheli, 2010; 2020) and often tuned toward the edge of chaos (Bertschinger & Natschläger, 2004); Table 2 tests whether graph reservoirs track the finite-width log-gain. Oversmoothing (Oono & Suzuki, 2020; Rusch et al., 2023; Epping et al., 2024; DeZoort & Hanin, 2023) is a different observable; kernel and signal-propagation analyses of deep GNNs (Du et al., 2019; Huang et al., 2022; Wang et al., 2025) work at infinite width.

Limitations. Theorem 3 covers node-uniform states and gives the leading constant only; the $\sigma_{\mathrm{b}} > 0$ cost is proved for the MLP only; the quenched bound on irregular graphs is certified, not proved; $\widehat{\mathcal{C}}_{L,S}$ is not $\mathcal{C}_{\mathrm{q}}$; and graphs are small ($n \leq 97$; a registered test to $n = 500$ was inconclusive), with odd activations and symmetric (GCN-type, not attention-based) operators.

REPRODUCIBILITY STATEMENT

The code and measured outputs are provided as anonymized supplementary material. Every number in this paper is produced by a command listed in Appendix M, run from that directory against the outputs it ships. Three things are worth naming for anyone attempting to reproduce them. First, every runner that measures the $\chi^\star$ field asserts **bit-identical** agreement with an independently validated reference implementation — $\max|\Delta| = 0$ exactly, not approximately — before it measures anything; reproducing that requires matching the reference's operation order, since an algebraically equivalent matrix product takes a different BLAS path and differs in the last unit in the last place. Second, all measurements use float64: this work hit χ^L-amplified float32 error large enough to invert a conclusion, and float32 is not safe anywhere in this pipeline. Third, the seed set is fixed and stated $(42, 7, 13, 21, 37, 55, 71, 89)$, every interval is a $10{,}000$-resample bootstrap over the units named in its caption, and the benchmark graphs for the displacement measurements are embedded as `graph6` strings, so that pipeline needs no external data. The one experiment with an external dependency is the initialization grid of §7, whose datasets and expected layout are given in Appendix M. Pre-registrations, including the gates this work failed and reports as failed, are included with the code.

REFERENCES

T. W. Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2):170–176, 1955.

Ludwig Arnold. *Random Dynamical Systems*. Springer Monographs in Mathematics. Springer, 1998.

Ludwig Arnold, N. Sri Namachchivaya, and Klaus R. Schenk-Hoppé. Toward an understanding of stochastic Hopf bifurcation: A case study. *International Journal of Bifurcation and Chaos*, 6(11): 1947–1975, 1996.

Jacob Bamberger, Benjamin Gutteridge, Scott le Roux, Michael Bronstein, and Xiaowen Dong. On measuring long-range interactions in graph neural networks. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2506.05971.

Peter H. Baxendale. A stochastic Hopf bifurcation. *Probability Theory and Related Fields*, 99: 581–616, 1994.

Nils Bertschinger and Thomas Natschläger. Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation*, 16(7):1413–1436, 2004.

Joel E. Cohen and Charles M. Newman. The stability of large random matrices and their products. *The Annals of Probability*, 12(2):283–310, 1984.

Hans Crauel and Franco Flandoli. Additive noise destroys a pitchfork bifurcation. *Journal of Dynamics and Differential Equations*, 10:259–274, 1998.

Andrea Crisanti, Giovanni Paladin, and Angelo Vulpiani. *Products of Random Matrices in Statistical Physics*, volume 104 of *Springer Series in Solid-State Sciences*. Springer, 1993.

Gage DeZoort and Boris Hanin. Principles for initialization and architecture selection in graph neural networks with ReLU activations, 2023. arXiv:2306.11668.

Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Liò, and Michael M. Bronstein. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, 2023.

Giuseppe Alessio D'Inverno, Zhiyuan Hu, Leo Davy, Michael Unser, Gianluigi Rozza, and Jonathan Dong. Revisiting deep information propagation: fractal frontier and finite-size effects. *Neural Networks*, 202:109013, 2026. arXiv:2508.03222.

Darshil Doshi, Tianyu He, and Andrey Gromov. Critical initialization of wide and deep neural networks using partial Jacobians: General theory and applications. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Simon S. Du, Kangcheng Hou, Ruslan Salakhutdinov, Barnabás Póczos, Ruosong Wang, and Keyulu Xu. Graph neural tangent kernel: Fusing graph neural networks with graph kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from Feynman diagrams. In *International Conference on Learning Representations (ICLR)*, 2020.

Bastian Epping, Alexandre René, Moritz Helias, and Michael T. Schaub. Graph neural networks do not always oversmooth. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.02269.

Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, 1986.

Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. Wiley, 2nd edition, 2003.

Harry Furstenberg and Harry Kesten. Products of random matrices. *The Annals of Mathematical Statistics*, 31(2):457–469, 1960.

Claudio Gallicchio and Alessio Micheli. Graph echo state networks. In *International Joint Conference on Neural Networks (IJCNN)*, 2010. doi: 10.1109/IJCNN.2010.5596796.

Claudio Gallicchio and Alessio Micheli. Fast and deep graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3898–3905, 2020.

Alessio Gravina. Information propagation dynamics in deep graph networks, 2024. arXiv:2410.10464.

Weiqi Guan and Zihao Shi. Measuring over-smoothing beyond Dirichlet energy, 2025. arXiv:2512.06782.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1706.02216.

Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. arXiv:1801.03744.

Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations (ICLR)*, 2020a. arXiv:1909.05989.

Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376:287–322, 2020b.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2672–2680, 2019.

Wei Huang, Yayong Li, Weitao Du, Jie Yin, Richard Yi Da Xu, Ling Chen, and Miao Zhang. Towards deepening graph neural networks: A GNTK-based optimization perspective. In *International Conference on Learning Representations (ICLR)*, 2022.

Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks. Technical Report GMD Report 148, GMD – German National Research Institute for Computer Science, 2001.

Jonathan Kadmon and Haim Sompolinsky. Transition to chaos in random neuronal networks. *Physical Review X*, 5:041030, 2015.

Rafail Khasminskii. *Stochastic Stability of Differential Equations*, volume 66 of *Stochastic Modelling and Applied Probability*. Springer, 2nd edition, 2012.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Omri Lesser and Debanjan Chowdhury. Competing nonlinearities, criticality, and order-to-chaos transition in deep networks, 2026. arXiv:2605.05294.

Mufan Bill Li, Mihai Nica, and Daniel M. Roy. The future is log-Gaussian: ResNets and their infinite-depth-and-width limit at initialization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. arXiv:2106.04013.

Xuerong Mao. *Stochastic Differential Equations and Applications*. Woodhead Publishing, 2nd edition, 2008.

Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition, 2009.

L. Molgedey, J. Schuchhardt, and H. G. Schuster. Suppressing chaos in neural networks by noise. *Physical Review Letters*, 69(26):3717–3719, 1992.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TUDataset: A collection of benchmark datasets for learning with graphs. In *ICML Workshop on Graph Representation Learning and Beyond*, 2020. arXiv:2007.08663.

Charles M. Newman. The distribution of Lyapunov exponents: exact results for random matrices. *Communications in Mathematical Physics*, 103(1):121–126, 1986.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations (ICLR)*, 2020.

Valery I. Oseledets. A multiplicative ergodic theorem: Lyapunov characteristic numbers for dynamical systems. *Transactions of the Moscow Mathematical Society*, 19:197–231, 1968.

Stefano Peluchetti and Stefano Favaro. Infinitely deep neural networks as diffusion processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Daniel A. Roberts, Sho Yaida, and Boris Hanin. *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press, 2022. arXiv:2106.10165.

T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks, 2023. arXiv:2303.10993.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations (ICLR)*, 2017.

Mariia Seleznova and Gitta Kutyniok. Analyzing finite neural networks: Can we trust neural tangent kernel theory? In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, 2022.

Jascha Sohl-Dickstein. The boundary of neural network trainability is fractal, 2024. arXiv:2402.06184.

Rodrigo Carmo Terin. Gauge-covariant stochastic neural fields: Stability and finite-width effects, 2025. arXiv:2508.18948, revised 2026.

James Theiler. Estimating fractal dimension. *Journal of the Optical Society of America A*, 7(6): 1055–1073, 1990.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations*, 2022.

Senmiao Wang, Yupeng Chen, Yushun Zhang, Ruoyu Sun, and Tian Ding. Exploring and improving initialization for deep graph neural networks: A signal propagation perspective. *Transactions on Machine Learning Research (TMLR)*, 2025. arXiv:2506.16790.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1810.00826.

Sho Yaida. Non-Gaussian processes and neural networks at finite widths. In *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pp. 165–192, 2020.

Greg Yang and Samuel S. Schoenholz. Mean field residual networks: On the edge of chaos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Kaicheng Zhang, Piero Deidda, Desmond Higham, and Francesco Tudisco. Are we measuring oversmoothing in graph neural networks correctly? In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2502.04591.

USE OF AI TOOLS

A large language model (Claude) was used substantially in this work. We disclose its role by task, with the verification procedure for each, in the order of the categories the ICLR AI Policy for Authors lists. The authors are responsible for every claim in this paper.

Implementing methods. All measurement code was written with AI assistance: the boundary-crossing sweeps, the $\sigma_{\mathrm{b}} > 0$ runner, the certificate for Conjecture 1, the activation-family construction, the corpus staging, and the training runner; the figures and the compression to the page limit were produced the same way. *Verification.* Every runner that measures the $\chi^\star$ field asserts bit-identical agreement ($\max|\text{ours} - \text{reference}| = 0.000\text{e}{+00}$) with a pre-existing, independently validated reference implementation before measuring anything. Components that compute no field (corpus staging, training runner) are not covered by that gate, and we do not claim it for them; every runner carries a self-test with hand-checkable invariants.

Mathematical claims and proofs. The AI proposed the $\varepsilon = 0$ form of the quenched bound, identified the joint-Gaussian structure equation 19, and proposed Proposition 4, Theorem 1 (the blockwise application of Newman (1986)), Theorem 2, and Theorem 3 (including the exact two-state reduction, the alignment mechanism and the bipartite exception) and Theorem 4, with their proofs. The proofs of Lemmas 11 and 12 were AI-assisted. *Verification.* Each claim was checked against direct simulation before being written down: the Kronecker lemmas to machine precision, the covariance formula to 0.1% over 2×10^5 draws, Theorem 1 against simulation on three irregular graphs and against an independent random-matrix computation, Theorem 2 by adaptive quadrature for five activations, the reduction behind Theorem 3 against full-matrix simulation and its conclusion by simulating the exact chain to $F = 4096$, the total-positivity inequality of Theorem 4 by Monte Carlo at 36 settings, and the theorem itself in 48 cells, with an adversarial search over 168 cells for the conjecture it does not cover (`theory/`). **No unproved statement appears in a theorem environment**. Conjecture 1 is labelled a conjecture. An earlier AI-produced proof of an alignment-tail bound for Theorem 3 was found in a later audit to have four gaps; it was withdrawn, and the theorem is now proved by a different route (Lemmas 3–7, Proposition 5) that needs no tail bound. That route was itself found by the AI during the audit, and one of its first formulations was wrong and caught by simulation before it entered the paper. The AI also proposed and proved Propositions 6, 12, 13 and 14 (the nonzero-bias cost, first computed as a formal expansion and later proved; a numerical decomposition showed that an apparent missing term was next-order). Each was checked numerically before entering the paper: the trade-off bound exactly, by quadrature, to 10^{-15}. These proofs were produced quickly with AI assistance. A second AI system (ChatGPT), used as an outside referee, read the proofs and, separately, the main text. It found one unproved step (a negative-moment bound

in Proposition 5, now replaced by a truncation argument), one false inequality ($K(q) \le K(0)$ in Proposition 6), hypotheses missing from Propositions 4, 12 and 13, and three compressed arguments (Lemmas 1, 4 and 7), all now corrected or written out. A second pass on the revised proofs, in a fresh session, found:

- a positive-density claim in Lemma 1(b) that was asserted, not proved, and was used again in Proposition 14. It is now a two-step minorization, with a separate no-atom lemma for the corner.
- a bound in Proposition 6 whose error term was additive, not relative. It is withdrawn.
- a sentence that presented the sharp constant as established when it is only numerical.

The same pass re-derived and confirmed the next-order expansion of Appendix C.10. A third pass, on the rewritten Lemma 1 alone, found a Meyn–Tweedie theorem cited for the wrong purpose, which is now corrected. A fourth, on the corrected text, raised no fatal objection and asked for four clarifications, all made.

The authors' reading. An author then read every proof in a theorem environment line by line, one result at a time. Each read used a reading guide prepared with AI assistance, which listed the load-bearing steps and the numerical checks behind them; the AI re-derived each proof before writing its guide. Between them, the reading and the re-derivations found:

- a one-step density claim in Lemma 1(b) that is false at $\rho = \pm 1$ (there ρ' has the sign of ρ);
- a Meyn–Tweedie theorem cited for the wrong purpose;
- a compressed step in Theorem 3 (Step 4(i)) whose AI-written gloss in a reading guide was wrong;
- a sharpness claim in Lemma 3 that its check did not support;
- a consequence of Theorem 4 stated more strongly than its proof gave;
- a hypothesis ($c > 0$) missing from Theorem 2;
- several justifications that were true but not as written.

All are corrected. Every proof on the main line of the paper carries the author's sign-off, including the results the AI later proposed and proved: Lemmas 5, 6, 8 and 9, Propositions 7 and 9, Corollary 1 and Theorem 5. That reading found two statements whose typeset form could be misread, and both were rewritten. The authors have also read Proposition 8 (the critical point $x = \frac{1}{2}$, which the AI proposed and proved after that reading) and four results no earlier reading had covered: the proof of Lemma 10, Proposition 11, and Lemmas 11 and 12. That reading found one error, an overstated lower bound in Step 3 of Proposition 8's proof, and four gaps; all are fixed, with a new numerical check of the corrected step. The ReLU result (Proposition 10), proposed and proved by the AI on the last day, was also read and signed off by the authors. Every proved result in the paper now carries the authors' sign-off. An outside reader, not an author, has also checked the five results the main line rests on (Theorems 3 and 5, Propositions 5 and 8, and Lemma 6). That reading found a citation error, a missing hypothesis and a missing step, all now fixed. The proof map at the start of Appendix C lists each result's dependencies and numerical check.

Hypotheses and experimental design. The AI proposed the spectral-deficit hypothesis of §6 and designed its corpus, the identifiability families, the σ_{b} sweep, the MLP control, the width–depth scaling study, the nonzero-bias falsification test, the softsign prefactor test and the bipartite test of §5, and, in a later round, the depth test of Appendix I, the out-of-sample test of its gradient-tail model, the Perron-weighted bound test, the box-counting test of Appendix F, and audits of the bipartite tests and of the zero-bias width study, and the input-merging test of Appendix I with its exploratory follow-up. Three of the four tests came back falsified and one void; the merging test was supported and its follow-up found nothing. The bipartite audit changed three verdicts, two of them from confirmed to inconclusive. The width audit replaced Table 1's values and withdrew the claim that the measurements lie below $\mathcal{C}_{\mathrm{q}}$; a third audit refuted the registered claim that more depth removes a large-width graph excess and widened the other activations' exponent range. The AI also designed the ReLU control and the $n = 256$ graph check, both supported, and the training test of Appendix I, whose registered mechanism failed its control (its large post hoc effect is labelled as

such). Each is reported as such. *Verification.* Hypotheses, predictors, acceptance thresholds and void conditions were fixed in writing before the corresponding runs (the pre-registrations, and the docstrings of `experiments/scaling_study.py` and `experiments/bias_scaling.py`, whose confidence-interval rule was evaluated only after a later deterministic re-run stored per-group values; the softsign and bipartite predictions were committed to version control before their runs); the spectral-deficit hypothesis **failed** its threshold and is reported as failed. Two AI-set thresholds that proved infeasible are recorded as errors in the pre-registration amendments. The MLP control was run after the width results and overturned the paper's earlier claim that the displacement is specific to graph coupling; that claim was withdrawn.

Synthetic data. The synthetic odd activation families and the widened graph corpus were generated with AI assistance. *Verification.* Each constructed activation is checked odd, C^1, with $\phi'(0) = \sup|\phi'| = 1$ to 10^{-12}, so that Propositions 1–2 apply unchanged.

Interpretation, literature and drafting. The AI drafted the interpretation of the measurements, conducted the related-work search, and drafted the manuscript; which readings the paper adopts and which claims it makes were decided by the authors. *Verification.* Every number has an entry in an internal claim ledger (owning script, output file, interval). Cited near-neighbours were checked against their abstracts or full text. The Gaussian-process prescription of Epping et al. (2024) was implemented from its recursion and validated against its analytic special case to 3×10^{-13}; this surfaced coefficients in the printed manuscript inconsistent with its own equations, while the released code is correct, and we follow the code. Checking procedures that caught real errors during the work are listed in Appendix M. The AI did not select which results to report, did not set or relax any acceptance threshold after the fact, and did not decide what the paper claims.

A Roadmap, notation and the five loci

Roadmap. Appendix B compares with prior work, including a known-versus-new table. Appendix C holds every proof. The window theorem is in §C.9, with its lemmas L1–L3c and Propositions 5–6, and the irregular-graph bounds are in §C.15. Appendices D–E cover side results: the metric half and the depth continuum. Appendix F is the box-counting test, G the finite-width law, H the initialization experiment, and I the depth tests, including the input-merging test. Appendix J collects the per-graph tables and the three audits of our own crossing estimator (E-f, E-g, and the bipartite tests). Appendix K covers nonzero bias, L the supplementary figures, and M the commands that reproduce every number.

Notation.

F, n, L, S	width (channels), nodes, depth, realizations averaged
σ_{s}, σ_{n}, σ_{b}	self, neighbour and bias weight scales; $g = \sigma_{\mathrm{s}}^2 + \mu_{\max}^2 \sigma_{\mathrm{n}}^2$
$\mu_{\max}(A)$	spectral radius of the symmetric operator A
$\chi^\star$	top (quenched) tangent Lyapunov exponent
γ_F, $g_{\mathrm{lin}}(F)$	$\frac{1}{2}(\psi(F/2) - \log(F/2))$ and $e^{-2\gamma_F}$: the exact linear penalty and boundary
ν	$\log g - \log g_{\mathrm{lin}}(F)$, distance above the exact linear boundary
q, ρ	channel variance of the state; state–tangent cosine (the two-state chain)
$G(g)$, κ_p	mean-field tangent gain at the induced fixed point; its quadratic coefficient
$K(q)$, $S(q)$, $\Lambda(q)$	$-\mathbb{E}(\phi'^2)''(\sqrt{q}\,\xi)$; $\mathbb{E}\phi'(\sqrt{q}\,\xi)^2$; the mean-field log-gain

The five loci.

$\mathcal{C}_{\mathrm{lin}}$, $\mathcal{C}_{\mathrm{lin}}(F)$	linear stability boundary at the origin; its exact finite-width version
$\mathcal{C}_{\mathrm{ann}}$	locus where the annealed second moment has unit growth
$\mathcal{C}_{\mathrm{q}}$	locus where the quenched top Lyapunov exponent vanishes
$\widehat{\mathcal{C}}_{L,S}$	sign-change set of the finite-L, S-averaged estimate of $\chi^\star$
$\mathcal{C}_{\mathrm{ind}}$	mean-field boundary, linearized at the *induced* fixed point

For MLPs and regular graphs with uniform input, $\mathcal{C}_{\mathrm{q}}$ is computed from the exact reduction (§3); **elsewhere we do not claim to have located** $\mathcal{C}_{\mathrm{q}}$. On irregular graphs the small-width spread of the crossings is the open problem of §6.

B Related work in detail

Random dynamical systems and noisy bifurcations. Both loci of this paper are sign changes of Lyapunov exponents, which is the language of random bifurcation theory (Arnold, 1998). At $\mathcal{C}_{\mathrm{lin}}(F)$ the exponent of the linearization at the origin changes sign; past it the origin loses stability and is replaced by an invariant measure, here the stationary law of the two-state chain. Baxendale (1994) studies exactly this transition for stochastic differential equations and shows that, as the linearized exponent tends to zero from above, the rescaled invariant measures converge to a nontrivial limit. Our boundary layer is the analogue at finite width: with $\nu = x/F$, the rescaled state Fq has the explicit Gamma limit of Proposition 7. The crossing $\mathcal{C}_{\mathrm{q}}$ is a *second* sign change, of the top exponent of the tangent dynamics along that invariant measure, and it is the one the window law locates. Small noise is known to reshape deterministic bifurcations qualitatively: additive noise destroys a pitchfork (Crauel & Flandoli, 1998), and stochastic Hopf bifurcation splits into distinct phenomenological and dynamical transitions (Arnold et al., 1996). In our setting the noise is intrinsic to finite width, of variance $1/F$, and it acts at a degenerate (double) root. That is why it moves the second crossing by $F^{-1/2}$ rather than by $O(1/F)$.

Effective theory at finite width. Roberts et al. (2022) describe networks at initialization and in training as nearly Gaussian, with the depth-to-width ratio controlling the deviations, and obtain criticality by tuning the infinite-width layer-to-layer recursions; activations fall into universality classes by their behaviour there. This paper agrees with, and specializes, both points. The depth-to-width ratio is the parameter in our consequence bound (Proposition 6: the loss between the two loci is $O(L/F)$) and in the spread of per-example gradients at $\mathcal{C}_{\mathrm{q}}$ (Appendix I). The activation enters the displacement only through κ_p, which is universal $(1/3)$ across smooth saturating odd activations and changes with the order p of the leading correction to $\phi(x)/x$ (softsign, $p = 1$, gives $\pi/32$). The location of criticality at finite width is where the two works meet most closely, and they must be kept apart. Roberts et al. (2022, §5.4.2) re-tune the initialization hyperparameters order by order in $1/n$ and find, for this class, $C_W = \sigma_1^{-2}\big(1 + \frac{2}{3n} + \dots\big)$, chosen so that the next-to-leading-order kernel correction stays marginal under depth: a finite-width critical point, shifted at $O(1/n)$ and analytic in $1/n$. Ours is a different criterion, the sign change of the top quenched Lyapunov exponent of the tangent map, and it is not analytic in $1/n$. At zero bias it lies at $\log g = \log g_{\mathrm{lin}}(F) + \sqrt{3/F}\,(1+o(1))$ (Theorem 3), a shift of order $n^{-1/2}$ that no finite order of a $1/n$ expansion can produce: at $n = 64$ it is 0.22 in $\log g$ against 0.010. So two notions of finite-width criticality coexist, the kernel's, perturbative and $O(1/n)$, and the Lyapunov exponent's, $O(n^{-1/2})$; they disagree at leading order, and a study that tunes by one does not sit at the other. The double root that produces the $n^{-1/2}$ is also latent in their expansions: from the Taylor coefficients of Roberts et al. (2022, §5.3.3), the fixed point just above criticality is $K^\star \approx (C_W - 1)/(2C_W)$ for $\tanh$, and the gain $C_W\langle\sigma'^2\rangle_{K^\star} = 1 + O((C_W - 1)^2)$, the first order cancelling. Theorem 2 makes that cancellation exact for every odd activation with $|\phi'| \le 1$, gives its constant κ_p, and covers non-smooth activations ($p = 1$). A study that places its network at the infinite-width critical point is off by $F^{-1/2}$ at zero bias, and the measured displacements (45–91% at $F = 6$, 3–4% at $F = 1024$) are of that size.

Is the double root already known? We checked the standard signal-propagation sources against the full texts. Poole et al. (2016) describe the bifurcation itself: at $\sigma_{\mathrm{b}} = 0$ the fixed point $q^\star = 0$ loses stability at $\sigma_w = 1$ and a stable nonzero fixed point appears. Schoenholz et al. (2017) study depth scales on the order/chaos boundary; Hayou et al. (2019) characterize the edge-of-chaos curve for $\tanh$-like activations, whose $\sigma_{\mathrm{b}} \to 0$ end is $\sigma_w \to 1$, and note that $\sigma_{\mathrm{b}} = 0$ itself drives q to zero; Doshi et al. (2023) derive power-law critical exponents in depth at critical points, for $\tanh$-class networks among others. None of them states that the tangent gain on the nontrivial branch touches 1 quadratically, or gives κ_p, and all four work at infinite width, where the location of the edge is not in question. The closest statement is the first-order cancellation latent in PDLT's Taylor coefficients, discussed above. Table 3 separates what we take from prior work from what is new.

Table 3: What is known, and what this paper adds.

object	known before	new here
edge at $\sigma_\mathrm{b} = 0$	at $\sigma_w = 1$, infinite width (Poole et al., 2016; Schoenholz et al., 2017)	exact finite-width linear boundary $\mathcal{C}_\mathrm{lin}(F)$ for every symmetric operator (Theorem 1)
fixed point $q^\star$	pitchfork at $\sigma_w = 1$ (Poole et al., 2016)	—
tangent gain G on the branch	first-order cancellation, latent for $\tanh$ (Roberts et al., 2022, §5.3.3)	exact identity; $G-1 = \kappa_p(g-1)^2$ with $\kappa_2 = 1/3$, $\kappa_1 = \pi/32$ (Theorem 2)
finite-width criticality	kernel criterion, shifted at $O(1/n)$, analytic (Roberts et al., 2022, §5.4.2)	Lyapunov crossing at $\sqrt{3/F}$, non-analytic in $1/F$, both branches (Theorem 3)
depth behaviour at criticality	power laws at infinite width (Doshi et al., 2023; Roberts et al., 2022)	—
Gaussian matrix products	Lyapunov spectrum (Newman, 1986); log-normal limits (Hanin & Nica, 2020b)	blockwise application to graph operators; alignment cost exactly $1/F$ (Proposition 5)
activations	edge-of-chaos curve for general ϕ (Hayou et al., 2019)	activation enters the displacement only through κ_p
graphs	infinite-width GNN kernels and initialization (Du et al., 2019; Huang et al., 2022; Wang et al., 2025)	quenched bound for regular graphs (Theorem 4); bipartite branch; small-width graph effects (measured)
fractal boundary	fractal at finite width (D'Inverno et al., 2026)	where the boundary lies; cost of box-counting it (Appendix F)

Finite-width corrections and where they apply. Finite-width effects are usually computed as corrections to an infinite-width description at fixed hyperparameters: non-Gaussian priors organized as a flow of preactivation distributions (Yaida, 2020), $1/n$ asymptotics of training by Feynman diagrams (Dyer & Gur-Ari, 2020), and neural tangent kernel fluctuations whose standard deviation is exponential in the depth-to-width ratio (Hanin & Nica, 2020a). Seleznova & Kutyniok (2022) show that whether a finite network is described by the NTK at all depends on its initialization hyperparameters and depth, and fails for deep networks in the chaotic phase. One might expect the finite-width crossing to mark where the NTK description breaks down. An exploratory check does not support that. Across 96 weight draws of a $\tanh$ MLP at depth $L = 16F$, the log-fluctuation of the NTK diagonal is smallest just above $\mathcal{C}_\mathrm{lin}(F)$ ($g \approx 1.1$ at $F = 16$ and 32) and rises smoothly *through* $\mathcal{C}_\mathrm{q}$, with no feature there. At $\mathcal{C}_\mathrm{q}$ it equals 5.2 at both widths, as a zero exponent with log-normal fluctuations set by L/F predicts (`experiments/sa_ntk_explore.py`). Because the double root makes the exponent flat near $\mathcal{C}_\mathrm{q}$, quantities built from gradient magnitudes vary smoothly across it.

Random networks in physics. In discrete-time random networks, external noise suppresses chaos and moves the transition to larger gain (Molgedey et al., 1992). The finite-width alignment cost acts in the same direction: sampling noise from F channels lowers the exponent by $1/F$, and the double root turns that into an $F^{-1/2}$ shift. Kadmon & Sompolinsky (2015) analyse the continuous-time transition at infinite size and leave finite-size effects open. Our results are a finite-size theory for the discrete-time case with smooth saturating activations.

Fractal boundaries. Sohl-Dickstein (2024) found that the boundary between hyperparameters giving convergent and divergent *training* is fractal over more than ten decades of scale. D'Inverno et al. (2026) study the analogous frontier for signal propagation at initialization, the order/chaos boundary of this paper, and find it fractal at finite width. These are different boundaries: one concerns the gradient-descent dynamics, the other the forward and tangent dynamics at initialization. Our results bear on the second. Any box-counting estimate of it presupposes knowing where it lies, and at finite averaging its estimated level set mixes noise near a tangency with a genuine crossing elsewhere (Appendix F). D'Inverno et al. (2026) locate the frontier as a level set of a divergence between two independent inputs over the (σ_w, σ_b) plane, with erf, width $N = 10^3$ taken as close to mean field, and depth $D = 10^3$. Our results say where that proxy is and is not safe. Away from $\sigma_b = 0$ the mean-field root is simple, so the finite-width shift is $O(1/N)$, a few tenths of a percent at $N = 10^3$

in our bias test. As $\sigma_b \to 0$ the root becomes double and the shift grows to order $\sqrt{3/N}$ in $\log g$, about 5% at $N = 10^3$. Their observable concerns two distant inputs; the one tied to the sign of $\chi^\star$ is the fate of *nearby* inputs, which switches at $\mathcal{C}_{\mathrm{q}}$ (Appendix I).

Products of random matrices and log-normal gradients. Newman (1986) gives the Lyapunov spectrum of products of Gaussian matrices, the linear case of Theorem 1; Crisanti et al. (1993) survey the methods for Lyapunov exponents of random matrix products in statistical physics. Hanin (2018) shows that the variance of squared input–output Jacobian entries is exponential in $\sum_\ell 1/n_\ell$, Hanin & Nica (2020b) proves log-normal limits for such products when depth and width grow together, and Li et al. (2021) find log-Gaussian behaviour, governed by the depth-to-width ratio, in ReLU ResNets. The same mechanism appears at $\mathcal{C}_{\mathrm{q}}$. The *typical* exponent there is zero, but per-input log-norms spread at a rate $\propto 1/F$ per layer, and a batch gradient follows its extreme inputs (Appendix I). A model treating inputs as independent predicted the resulting slope in-sample and was falsified out of sample: inputs share weights.

Dynamical isometry and the edge-of-chaos curve. Pennington et al. (2017) compute the full singular value distribution of the input–output Jacobian and show that sigmoidal networks with orthogonal weights can achieve dynamical isometry, where ReLU networks cannot; Hayou et al. (2019) analyse the edge-of-chaos curve for general activations and show that initializing on it helps training. Both work at infinite width and with the whole spectrum or the full curve. Our results locate where the *top* exponent vanishes at finite width, which is necessary for isometry but not sufficient. They also show that, under a tuned learning rate, the gap between the infinite-width and finite-width edges is a uniform per-layer factor in the linear regime (Appendix I). That is why placing a network exactly at the finite-width edge buys little in our experiments.

Kernels and signal propagation for GNNs. The graph neural tangent kernel (Du et al., 2019) gives the infinite-width limit of GNN training. Huang et al. (2022) analyse the trainability of deep GNNs through it, and Wang et al. (2025) choose initial weight variances for deep GCNs by optimizing forward, backward and graph-embedding-variation metrics of signal propagation. Both work at infinite width, or with metrics other than the tangent exponent. Our initialization experiment, which finds no training benefit from initializing at $\mathcal{C}_{\mathrm{q}}$ rather than $\mathcal{C}_{\mathrm{lin}}$ (Appendix H), does not contradict SPoGInit's improvements. It compares two gains on one ray under a tuned learning rate, while SPoGInit optimizes different objectives and architectures.

Oversmoothing. Oono & Suzuki (2020) show that deep GCNs whose weights satisfy a spectral condition converge exponentially to a subspace carrying only connected-component and degree information, and Rusch et al. (2023) define oversmoothing axiomatically as exponential convergence of node-similarity measures. This concerns the forward state; order and chaos concern the tangent. The two meet in one place here. A non-bipartite regular graph with uniform input keeps the state node-uniform, so it is oversmoothed from the start, and there the message-passing network is exactly an MLP of gain g (Theorem 4). Graph structure affects the crossing through departures from that state: node heterogeneity (Appendix C.15) and bipartite spectra (Theorem 3(i)). We do not claim that the crossing governs oversmoothing. Nor do we address over-squashing, the insensitivity of a node to distant nodes through bottlenecks (Topping et al., 2022). Width, depth and topology all shape it, with commute time the strongest factor (Di Giovanni et al., 2023). The tangent exponent aggregates over the whole graph and is set by $\mu_{\max}$, not by pairwise sensitivities.

Further reading. Closely related to the questions here are Lesser & Chowdhury (2026) on competing nonlinearities and the order-to-chaos transition, Gravina (2024) on information propagation in deep graph networks, and, on how oversmoothing is measured, Zhang et al. (2026) and Guan & Shi (2025).

C PROOFS

Proof map. Every proved result, where its proof is, what it uses, and its numerical check in `theory/verify_proofs.py`, named by label. Conjecture 1 is certified numerically, not proved (Appendix C.15).

result	proof	uses	check
Prop. 1	p. 3	—	`---`
Thm. 1	p. 3	—	`Thm 1`
Prop. 2	p. 3	—	`Prop [bound]`
Thm. 2	p. 4	—	`Thm 2`
Prop. 3	p. 4	Thm. 2	`Prop [mfnull]`
Thm. 4	p. 8	—	`Thm 4`
Prop. 4	p. 8	Lem. 10	`Prop [finite]`
Lem. 10	p. 42	—	`(in text)`
Prop. 11	p. 43	Thm. 1	`---`
Prop. 12	p. 44	Lem. 10	`---`
Prop. 13	p. 45	—	`Prop [tradeoff]`
Lem. 11, 12	p. 47	—	`---`
Lem. 1 (L1)	p. 21	—	`L1(b), L1(d)`
Lem. 2 (L2)	p. 22	—	`(in text)`
Lem. 3 (L3a)	p. 23	—	`L3a`
Lem. 4 (L3b)	p. 23	L3a	`L3b`
Lem. 5	p. 24	—	`---`
Lem. 6 (L3b′)	p. 25	L3b, Lem. 5	`L3b′`
Lem. 7 (L3c)	p. 27	L2, L3a	`L3c`
Prop. 5	p. 27	L1, L2, L3a–L3c	`Prop [cost]`
Prop. 6	p. 28	Prop. 5	`Prop [consequence]`
Lem. 8 (L4)	p. 28	Prop. 5, Lem. 5	`Lemma L4`
Cor. 1	p. 29	L3b′, L4	`---`
Thm. 3	p. 4	L1–L3c, Prop. 5, Thm. 2	`Thm 3`
Prop. 7	p. 30	—	`Prop [diffusion]`
Lem. 9	p. 32	Lem. 5, L3b′	`Lemma localize`
Thm. 5	p. 32	Prop. 7, L3b′, Lem. 9, Cor. 1	`Theorem [layerlimit]`
Prop. 8	p. 35	L3b′, Lem. 9, 5, Thm. 5	`Prop [critical]` (2)
Prop. 9	p. 39	L4, Thm. 3 (Step 4)	`Prop [nextorder]`
Prop. 10	p. 40	Lem. 5	`Prop [relu]`
Prop. 14	p. 63	L1, L2	`Prop [bias]`

C.1 Proposition 1: $\mathcal{C}_{\mathrm{lin}}$ in closed form

Oddness gives $\phi(0) = 0$, so $H \equiv 0$ solves equation 1 at $\sigma_{\mathrm{b}} = 0$. Linearizing, $\phi'(0) = 1$ reduces equation 2 to $v \mapsto \sigma_{\mathrm{s}} v W_s + \sigma_{\mathrm{n}} A v W_n$. With $C^\ell = \mathbb{E}[v^\ell (v^\ell)^\top]$, taking expectations over the independent Gaussian weights gives $C^{\ell+1} = \sigma_{\mathrm{s}}^2 C^\ell + \sigma_{\mathrm{n}}^2 A C^\ell A^\top$, whose top growth rate is $\sigma_{\mathrm{s}}^2 + \rho(A)^2 \sigma_{\mathrm{n}}^2$, attained on the leading eigenvector, with $\rho(A) = \mu_{\max}(A)$ for symmetric A. Setting it to 1 gives the conic; a conic arc is rectifiable, hence of box dimension 1. □

C.2 Theorem 1: the exact finite-width linear boundary

Let $A = \sum_k \lambda_k e_k e_k^\top$ with orthonormal e_k, and write the tangent as $v = \sum_k e_k y_k^\top$ with $y_k \in \mathbb{R}^F$, so that $\|v\|^2 = \sum_k \|y_k\|^2$. Since $Av = \sum_k \lambda_k e_k y_k^\top$, the linear recursion acts blockwise, $y_k^{\ell+1} = y_k^\ell M_k^\ell$ with $M_k^\ell = \sigma_{\mathrm{s}} W_s^\ell + \sigma_{\mathrm{n}} \lambda_k W_n^\ell$, whose entries are i.i.d. $\mathcal{N}(0, s_k^2/F)$, $s_k^2 = \sigma_{\mathrm{s}}^2 + \sigma_{\mathrm{n}}^2 \lambda_k^2$, independent across ℓ (the blocks share weights, but each block alone is a product of i.i.d. Gaussian matrices). For fixed $y \neq 0$, $yM_k/\|y\|$ is $\mathcal{N}(0, (s_k^2/F) I_F)$ by rotation invariance, so $\|y_k^{\ell+1}\|^2/\|y_k^\ell\|^2 = (s_k^2/F)\, X_\ell$ with $X_\ell \sim \chi_F^2$ i.i.d. and independent of the past. The strong law gives $L^{-1} \log \|y_k^L\| \to \frac{1}{2} \log s_k^2 + \frac{1}{2} \mathbb{E} \log(X/F)$ a.s., and $\mathbb{E} \log \chi_F^2 = \psi(F/2) + \log 2$ (Newman, 1986). With finitely many blocks, $L^{-1} \log \|v^L\| \to \max_k$ over blocks with $y_k^0 \neq 0$, which is attained at $|\lambda_k| = \mu_{\max}$ for generic v^0. Hence $\chi_{\mathrm{lin}}^\star(F) = \frac{1}{2} \log g + \gamma_F$ with $\gamma_F = \frac{1}{2}(\psi(F/2) - \log(F/2)) = -\frac{1}{2F} - \frac{1}{6F^2} + O(F^{-4})$, and the linear crossing is $g = g_{\mathrm{lin}}(F) = e^{-2\gamma_F} = 1 + F^{-1} + O(F^{-2})$ for every symmetric A. For $n = 1$ this is Newman (1986)'s top exponent. □

C.3 Proposition 2: annealed domination

Condition on the weights of layers before ℓ, which fix H^ℓ and v^ℓ, and write $u^\ell = \sigma_{\mathrm{s}} v^\ell W_s^\ell + \sigma_{\mathrm{n}} A v^\ell W_n^\ell$ for the linear tangent increment, so that $v^{\ell+1} = \phi'(Z^\ell) \odot u^\ell$. Since $|\phi'(x)| \leq \phi'(0) = 1$ pointwise,

$\|v^{\ell+1}\|^2 \le \|u^\ell\|^2$ for every draw of the layer-ℓ weights; Z^ℓ depends on that draw, the bound does not. The matrices W_s^ℓ, W_n^ℓ are independent of v^ℓ and of each other, with mean zero and entry variance $1/F$, so $\mathbb{E}\|xW\|^2 = \|x\|^2$ and the cross term vanishes:

$$\mathbb{E}\big[\|u^\ell\|^2 \mid v^\ell\big] = \sigma_{\mathrm{s}}^2\|v^\ell\|^2 + \sigma_{\mathrm{n}}^2\|Av^\ell\|^2 \le \big(\sigma_{\mathrm{s}}^2 + \mu_{\max}(A)^2\sigma_{\mathrm{n}}^2\big)\|v^\ell\|^2 = g\,\|v^\ell\|^2,$$

using $\|Ax\| \le \|A\|_2\|x\| = \mu_{\max}(A)\|x\|$ for symmetric A. Taking expectations, $Q_\phi^{\ell+1} \le g\,Q_\phi^\ell$, hence $Q_\phi^\ell \le g^\ell Q_\phi^0$ and $\limsup_\ell \ell^{-1}\log Q_\phi^\ell \le \log g$. For $\phi = \mathrm{id}$ the first inequality is an equality and, by the recursion for C^ℓ in the proof of Proposition 1, the rate $\log g$ is attained whenever v^0 has a component on the leading eigenvector. So wherever the linear map is annealed-stable ($g < 1$) so is ϕ, and the unit-growth locus of ϕ lies in $g \ge 1$: on or outside $\mathcal{C}_{\mathrm{lin}}$. □

The comparison is of growth *rates*, not of Q_ϕ^ℓ and Q_{lin}^ℓ layer by layer: $Q_{\mathrm{lin}}^{\ell+1}$ depends on $\|Av_{\mathrm{lin}}^\ell\|$, not on Q_{lin}^ℓ alone, so a scalar induction between the two recursions does not close.

C.4 Proposition 3: mean field predicts no displacement

Write $G(g) = g\,\mathbb{E}[\phi'(Z)^2]$ with $Z \sim \mathcal{N}(0, q^\star)$ and $q^\star = q^\star(g)$ the nontrivial root of $q = g\,\mathbb{E}[\phi(\sqrt{q}\,\xi)^2]$ at $\sigma_{\mathrm{b}} = 0$.

Fix $g > 1$ and suppose $q^\star > 0$ exists. Since ϕ is odd and Z is symmetric, $\mathbb{E}[\phi(Z)] = 0$, so $\mathbb{E}[\phi(Z)^2] = \mathrm{Var}\,\phi(Z)$. Because $|\phi'| \le 1$, ϕ is Lipschitz, and the Gaussian Poincaré inequality gives

$$\mathrm{Var}\,\phi(Z) \;\le\; q^\star\,\mathbb{E}[\phi'(Z)^2]. \tag{6}$$

Multiplying equation 6 by g and substituting the fixed-point equation $q^\star = g\,\mathbb{E}[\phi(Z)^2]$ on the left,

$$q^\star \;=\; g\,\mathbb{E}[\phi(Z)^2] \;\le\; q^\star\,g\,\mathbb{E}[\phi'(Z)^2] \;=\; q^\star\,G(g),$$

and dividing by $q^\star > 0$ gives $G(g) \ge 1$. Equality in equation 6 holds only for affine ϕ; ϕ is odd and nonlinear, so the inequality is strict and $G(g) > 1$. At $g = 1$ the only root is $q^\star = 0$: for $q > 0$ the same inequality, strict, gives $\mathbb{E}\phi(\sqrt{q}\,\xi)^2 < q\,\mathbb{E}\phi'(\sqrt{q}\,\xi)^2 \le q$, and there $Z \equiv 0$ and $G(1) = \phi'(0)^2 = 1$. Hence $G(g) = 1$ has no root above $g = 1$ and $\mathcal{C}_{\mathrm{ind}} = \mathcal{C}_{\mathrm{lin}}$. □

Three remarks. First, only oddness and $|\phi'| \le 1$ are used; $\phi'(0) = 1$ enters only in the last line. Second, the argument never needs $q^\star$ to exist for all $g > 1$ — where the nontrivial branch does not open, $\mathcal{C}_{\mathrm{ind}}$ is simply undefined there and there is still no root, which is why the proposition is stated over the g at which it exists. Third, the identity activation is exactly the affine equality case, so it sits *on* $\mathcal{C}_{\mathrm{lin}}$ rather than outside it; this is the same fact the linear control of §5 measures from the other side.

C.5 Theorem 2: the exact mean-field identity

Let $X \sim \mathcal{N}(0, q)$, $q > 0$. Stein's lemma states $\mathbb{E}[Xh(X)] = q\,\mathbb{E}[h'(X)]$ for absolutely continuous h with $\mathbb{E}|h'(X)| < \infty$. Put $\psi(x) = \phi(x)/x$. Because $|\phi'| \le 1$ and $\phi(0) = 0$, $|\psi| \le 1$ and $|x\psi'(x)| = |\phi'(x) - \psi(x)| \le 2$ for $x \ne 0$, so $h(x) = x\psi(x)^2$ has bounded derivative $h' = \psi^2 + 2x\psi\psi'$ and

$$\mathbb{E}\phi(X)^2 = \mathbb{E}[Xh(X)] = q\,\mathbb{E}[\psi^2 + 2X\psi\psi'].$$

Since $\phi' = \psi + x\psi'$, $q\,\mathbb{E}\phi'(X)^2 = q\,\mathbb{E}[\psi^2 + 2X\psi\psi' + X^2\psi'^2]$, and subtracting,

$$q\,m(q) - \mathbb{E}\phi(X)^2 = q\,\mathbb{E}[X^2\psi'(X)^2]. \tag{7}$$

At $q = q^\star = g\,\mathbb{E}\phi(X)^2 + \sigma_{\mathrm{b}}^2$, dividing by $q^\star/g$ gives $G(g) = g\,m(q^\star) = (q^\star - \sigma_{\mathrm{b}}^2)/q^\star + g\,\mathbb{E}[X^2\psi'(X)^2]$, which is equation 4. At $\sigma_{\mathrm{b}} = 0$ the right side is non-negative and vanishes only if $\psi' = 0$ a.e., i.e. ϕ is linear; this re-proves Proposition 3 without the Poincaré inequality. For the expansion, the fixed-point equation gives $g - 1 = \mathbb{E}[X^2(1-\psi^2)]/\mathbb{E}[X^2\psi^2]$. If $\psi(x) = 1 - c|x|^p + O(|x|^{p+\delta})$ and $\psi'(x) = -cp\,|x|^{p-1}\,\mathrm{sgn}\,x + O(|x|^{p-1+\delta})$ near 0 for some $\delta > 0$ (the remainders are bounded elsewhere, since $|\psi| \le 1$ and $|x\psi'| \le 2$), then with $M_s = \mathbb{E}|\xi|^s$ and $q^\star \downarrow 0$ as $g \downarrow 1$,

$$g - 1 = 2cM_{p+2}\,q^{\star p/2}(1 + o(1)), \qquad G(g) - 1 = g\,\mathbb{E}[X^2\psi'^2] = c^2p^2M_{2p}\,q^{\star p}(1 + o(1)),$$

so $(G-1)/(g-1)^2 \to \kappa_p = p^2M_{2p}/(4M_{p+2}^2)$. For $p = 2$, $M_4 = 3$ gives $\kappa_2 = 1/3$; for $p = 1$, $M_2 = 1$ and $M_3 = 2\sqrt{2/\pi}$ give $\kappa_1 = \pi/32$. Both identities hold to $\le 2 \times 10^{-8}$ by adaptive quadrature for the five activations, and $(G-1)/(g-1)^2$ at $g - 1 = 10^{-4}$ is 0.3333 (tanh, erf, arctan, sublinear) and 0.0982 (softsign) (`theory/mf_identity.py`). □

C.6 THE TWO-STATE CHAIN: LEMMAS L1–L3C

This subsection and the next four prove Theorem 3 and its extensions (the proof map at the start of this appendix shows how they depend on one another):

- this subsection: the lemmas on the two-state chain;
- Appendix C.7: the alignment cost (Proposition 5) and the band;
- Appendix C.8: the boundary layer;
- Appendix C.9: the proof of Theorem 3 itself;
- Appendix C.10: the next order.

The argument runs through stationarity of $\log(1-\rho^2)$, where the corner $\rho^2 = 1$ costs only its stationary mass, so **no tail condition on π is needed**. One routine step in the proof of Theorem 3, the $O(1/F)$ decoupling of self-normalized weighted averages, is L2's delta-method argument applied to $h(u,v) = v/u$.

Lemma 1 (L1: ergodicity and representation)**.** *Let ϕ be bounded, so that $q \le q_{\max} := g \sup \phi^2$, and assume, as holds for* tanh*, erf,* arctan *and softsign, that $\psi(y) = \phi(y)/y$ satisfies $\psi(y) \ge c_\psi/(1+|y|)$ and that $h(y) = y\phi'(y)/\phi(y)$ has Lebesgue-null level sets (h is real-analytic and non-constant on $(0,\infty)$ for the first three and equals $1/(1+|y|)$ for softsign). Fix $g > g_{\mathrm{lin}}(F)$ and work in $w = \log q$ on $(-\infty, w_{\max}]$, $w_{\max} = \log q_{\max}$. One step is $w' = w + \Delta$ with $\Delta = \log g + \log\big(\frac{1}{F}\sum_a \xi_a^2 \psi(\sqrt{q}\,\xi_a)^2\big)$, ξ_a i.i.d. standard normal, and write $D(q) = \mathbb{E}[\Delta \mid q]$. Then the pair (q,ρ) has a unique invariant probability π, with $\mathbb{E}_\pi|\log q| < \infty$, and $2\chi^\star = \mathbb{E}_\pi \log(gB/F)$ almost surely.*

Proof. (a) Two-sided drift. The bound $\psi \ge c_\psi/(1+|y|)$ gives $\big|\log(\frac{1}{F}\sum_a \xi_a^2\psi^2)\big| \le \big|\log(\frac{1}{F}\sum_a \xi_a^2)\big| + \max_a 2\log\big(1+\sqrt{q_{\max}}|\xi_a|\big) + 2|\log c_\psi|$, integrable uniformly in $q \le q_{\max}$; dominated convergence then makes D continuous on $[0, q_{\max}]$ with $D(0^+) = \log g + \mathbb{E}\log(\chi^2_F/F) = \log g + 2\gamma_F = \nu > 0$, so $D \ge \nu/2$ on some $[0, q_2]$. At the other end Jensen gives $D(q) \le \log g + \log(\Phi(q)/q) = \nu + \log g_{\mathrm{lin}}(F) - \ell(q) \le \nu + F^{-1} - c_0 q$, where $\Phi(q) = \mathbb{E}\,\phi(\sqrt{q}\,\xi)^2$ and $\ell(q) = -\log(\Phi(q)/q)$; ℓ is continuous on $(0, q_{\max}]$, positive because $|\phi(y)| < |y|$ for $y \ne 0$, and $\ell(q) = 6|a|q + O(q^2)$, so $\ell(q) \ge c_0 q$ there. This is at most $-c_0 q/2$ once $q \ge q_1 := 2(\nu + F^{-1})/c_0$. Fix $q_3 \in (q_1, q_{\max})$; the interval $I = [\log q_2, \log q_3]$ is thus reached from both sides.

(b) Minorization. Given q, $q' = gA/F$ with A a sum of F i.i.d. absolutely continuous variables, so the kernel has a density, jointly continuous in (q,q') and strictly positive on $I \times I$ (both ends lie inside $(0, q_{\max})$), hence bounded below there: I is a small set. Write $C_I = \{q : \log q \in I\}$. Since Δ has exponential moments of small order uniformly in q (the bound in (a)), the two-sided drift of (a) gives the Foster–Lyapunov condition $PV \le V - 1 + b\,\mathbf{1}_{C_I}$ for $V(w) = 1 + a_+(w - \log q_3)^+ + a_-(\log q_2 - w)^+$, with $a_\pm$ and b large enough, after enlarging I by a fixed margin if necessary. V is everywhere finite and C_I is small, so the q-chain is positive Harris recurrent (Meyn & Tweedie, 2009, Thm. 11.3.4), with a unique invariant probability π_q, and it is aperiodic. *The pair.* The q-coordinate evolves autonomously, so the pair chain on $(0, q_{\max}] \times [-1,1]$ inherits the drift of (a) with the same function of q alone, and it remains to show that $C = C_I \times [-1,1]$ is small; we show a two-step minorization. One step cannot suffice: at $\rho = \pm 1$ every term of $\sum_a \phi(z_a)\phi'(z_a)\zeta_a$ has the sign of ρ, so ρ' does too, and no one-step density is positive on all of $(-1,1)$. Given (q,ρ) and x, write $z = \sqrt{q}\,x$, $w = \phi(z)$, $D = \operatorname{diag}\phi'(z)$, so that $q' = g\|w\|^2/F$ and ρ' is the cosine between w and $D\zeta$, $\zeta = \rho x + \sqrt{1-\rho^2}\,e$. *Step 1: interior states.* Let $K_1 = I_1 \times [-1+\eta, 1-\eta]$ with $I_1 \subset (0, q_{\max})$ compact and $\eta \in (0,1)$, and fix a compact rectangle $K \subset (0, q_{\max}) \times (-1,1)$. Given x, $D\zeta \sim N(\rho Dx, (1-\rho^2)D^2)$ with $(1-\rho^2)D^2 \succeq \eta(2-\eta)\min_a \phi'(z_a)^2\, I \succ 0$. So ρ', the cosine of a nondegenerate Gaussian vector with the fixed direction w, has a conditional density $f(r \mid x, s)$ on $(-1,1)$, continuous in (x,s,r) and strictly positive, off the fixed null set $\{x = 0\}$ (where $w = 0$), which does not depend on s. Write $x = (x_1, \tilde{x})$ and $u_0(\tilde{x}) = \frac{g}{F}\sum_{a \ge 2}\phi(\sqrt{q}\,x_a)^2$. Given $\tilde{x}$, $q' = u_0 + \frac{g}{F}\phi(\sqrt{q}\,x_1)^2$ is a smooth function of $|x_1|$, strictly increasing in $|x_1|$ (its x_1-derivative $2g\sqrt{q}\,\phi\phi'/F$ vanishes only at $x_1 = 0$), with range $[u_0, u_0 + q_{\max}/F)$; the density below is on the open interval. Changing variables from x_1 to u, the pair (q', ρ') has, given $\tilde{x}$, the density

$$p(s; u, r \mid \tilde{x}) = \sum_{\pm} \frac{\gamma(x_1^{\pm}(u))}{|\partial_1 q'|}\, f\big(r \mid (x_1^{\pm}(u), \tilde{x}), s\big)\, \mathbf{1}\{u_0(\tilde{x}) < u < u_0(\tilde{x}) + q_{\max}/F\},$$

with γ the standard normal density and $x_1^{\pm}(u)$ the two roots. It is continuous and positive where the indicator is 1 and zero elsewhere, hence lower semicontinuous in (s, u, r). So $p(s; u, r) = \mathbb{E}_{\tilde{x}}\, p(s; u, r \mid \tilde{x})$ is the transition density, lower semicontinuous by Fatou. It is strictly positive whenever $0 < u < q_{\max}$, since then the event $u_0(\tilde{x}) \in (u - q_{\max}/F, u)$ is open, non-empty and of positive probability. A positive lower semicontinuous function attains a positive minimum c on the compact $K_1 \times K$, so $P(s, A) \geq c\,\mathrm{Leb}(A \cap K)$ for every $s \in K_1$. *Step 2: leaving the corner.* From each $s \in C$ some open set of (x, e) of positive probability maps into the interior of such a K_1. For $|\rho| < 1$, choose e with $D\zeta \perp w$ and x with q' in range. At $\rho = \pm 1$ only x enters, and $|\rho'| < 1$ as soon as the $h(z_a)$ are not all equal (see (d)). The map $(q, \rho, x, e) \mapsto (q', \rho')$ is continuous off a null set, so $s \mapsto P(s, U)$ is lower semicontinuous for open U (Fatou). A finite subcover of the compact C therefore gives one K_1 and one $p_0 > 0$ with $P(s, K_1) \geq p_0$ for every $s \in C$. Together, $P^2(s, A) \geq p_0 c\,\mathrm{Leb}(A \cap K)$ on C: C is small. Every state reaches C with positive probability in one step: from any $q > 0$ the q'-kernel has a density positive on $(0, q_{\max})$, and so on C_I. So the chain is $\mathrm{Leb}(\,\cdot \cap K)$-irreducible. The same V, a function of q alone, satisfies $PV \leq V - 1 + b\,\mathbf{1}_C$ for the pair, with C now shown petite, so the pair chain is positive Harris recurrent (Meyn & Tweedie, 2009, Thm. 11.3.4) and has a unique invariant probability π, whose q-marginal is π_q.

(c) Integrability. For small $\theta > 0$, $V(w) = e^{-\theta w} = q^{-\theta}$ satisfies $\mathbb{E}[V(w') \mid w] = V(w)\,\mathbb{E}[e^{-\theta\Delta}]$, and for $q \leq q_2$, $\mathbb{E}[e^{-\theta\Delta}] = 1 - \theta D(q) + O(\theta^2) \leq 1 - \theta\nu/4 < 1$, the $O(\theta^2)$ uniform because Δ has exponential moments of small order by the bound in (a). Hence $\mathbb{E}_\pi q^{-\theta} < \infty$, so $\mathbb{E}_\pi |\log q| < \infty$. With $|\log \phi'(x)| \leq C(1 + x^2)$ and $q \leq q_{\max}$, every functional averaged below ($\log q$, $\log B$, H, $\rho^2 qK$) is π-integrable. The law of large numbers for positive Harris chains (Meyn & Tweedie, 2009, Thm. 17.0.1) then gives $2\chi^\star = \mathbb{E}_\pi \log(gB/F)$ almost surely from every initial state, and licenses the stationarity identities. (Geometric ergodicity is neither needed nor claimed.)

(d) No atom at the corner. From every state, $\mathbb{P}(\rho' = \pm 1) = 0$ for $F \geq 2$. If $|\rho| < 1$, then given x, $D\zeta$ is a nondegenerate Gaussian vector, which lies on the line through w with probability 0. If $\rho = \pm 1$, then $\zeta = \pm x$, and $\rho' = \pm 1$ forces $Dx \parallel w$, that is, $\phi'(z_a)x_a/\phi(z_a) = h(z_a)/\sqrt{q}$ equal for all a; since z_2 has a density and the level sets of h are null, $\mathbb{P}(h(z_1) = h(z_2)) = 0$. By invariance, $\pi(\rho^2 = 1) = 0$. □

Lemma 2 (L2: moments of the sample cosine)**.** *Write $X_a = \phi(\sqrt{q}\,x_a)/\sqrt{q} = x_a\psi(\sqrt{q}\,x_a)$ and $Y_a = \phi'(\sqrt{q}\,x_a)\zeta_a$, so the alignment update is the uncentered sample cosine of the i.i.d. pairs (X_a, Y_a), scale-free in each argument. With $W = (X^2, Y^2, XY)$, $\mu = \mathbb{E}W$, $C = \mathrm{Cov}(W)$, $\bar{W}$ the sample mean and $h_2(u, v, w) = w^2/(uv)$,*

$$\mathbb{E}\big[\hat{\rho}^{\,2} \;\big|\; q, \rho\big] = h_2(\mu) + \frac{1}{2F}\sum_{i,j} \partial^2_{ij} h_2(\mu)\, C_{ij} + O(F^{-2}) = r^2 + \frac{v_2(q, \rho)}{F} + O(F^{-2}), \tag{8}$$

uniformly for $q \in (0, q_{\max}]$, $\rho \in [-1, 1]$, where $r = \mu_3/\sqrt{\mu_1\mu_2}$ is the population cosine; likewise with $h_4 = w^4/(u^2v^2)$, $\mathbb{E}[\hat{\rho}^{\,4}] = r^4 + v_4/F + O(F^{-2})$ with $v_4 = O(r^2)$.

Proof. Three facts make the second-order delta method exact to this order. (i) $|\phi(y)| \leq |y|$ and $|\phi'| \leq 1$ give $|X_a| \leq |x_a|$, $|Y_a| \leq |\zeta_a|$, so W has moments of every order, bounded uniformly in (q, ρ), and $\mathbb{E}|\bar{W} - \mu|^k \leq C_k F^{-k/2}$. (ii) $\mu_1 = \mathbb{E}[x^2\psi^2]$ and $\mu_2 = \mathbb{E}[\phi'^2\zeta^2]$ are bounded below by a positive constant for $q \leq q_{\max}$, so h_2 and h_4 are C^4 with bounded derivatives on a neighbourhood N of μ. (iii) $\hat{\rho}^{\,2} \leq 1$ by Cauchy–Schwarz, so $\{\bar{W} \notin N\}$ contributes at most $2\,\mathbb{P}(\bar{W} \notin N) = O(F^{-2})$ by Chebyshev at fourth order. On N, Taylor to third order with fourth-order remainder, $\mathbb{E}[\bar{W} - \mu] = 0$, $\mathbb{E}[(\bar{W} - \mu)_i(\bar{W} - \mu)_j] = C_{ij}/F$, $\mathbb{E}[(\bar{W} - \mu)_i(\bar{W} - \mu)_j(\bar{W} - \mu)_k] = O(F^{-2})$ and $\mathbb{E}|\bar{W} - \mu|^4 = O(F^{-2})$ give equation 8 (a third-order remainder alone would give only $O(F^{-3/2})$); restricting the first-order term to N costs $\|\nabla h\|\,\mathbb{E}\big[|\bar{W} - \mu|\mathbf{1}\{\bar{W} \notin N\}\big] = O(F^{-2})$, since off N (a ball of radius δ) $|\bar{W} - \mu| \leq |\bar{W} - \mu|^4/\delta^3$. The same bound handles restricting the second- and third-order terms to N. The moments are expectations of functions even in $\sqrt{q}$, hence C^1 in q (as in Lemma 7), which gives $v_2(q, \rho) = (1 - r^2)(1 - 2r^2) + O(q)$, the $q = 0$ value being the Gaussian one. □

Checks on L2. At $q = 0$ the derivative formulas give exactly $v_2 = (1 - r^2)(1 - 2r^2)$ and $v_4 = 6r^2 - 14r^4 + 8r^6$. For the actual pairs, equation 8 matches Monte Carlo over 12 cells ($q \in \{0.01, 0.1, 0.5\}$, $\rho \in \{0.2, 0.6\}$, $F \in \{256, 1024\}$, 2×10^5 replicates): $|z| \leq 1.9$ at $F = 256$

and $|z| \le 0.7$ at $F = 1024$, the $F = 256$ residual carrying the sign and size of the $O(F^{-2})$ term. The gap $v_2 - (1-r^2)(1-2r^2)$ scales as q (ratio $\to -3.02$ at $\rho = 0.2$, $+1.97$ at $\rho = 0.6$) (`theory/sample_cosine.py`). **The Gaussian sample-correlation value $v_4 = 6r^2(1-r^2)^2$ is not this estimator's coefficient**: Monte Carlo rejects it at 15 standard errors at $r^2 = 1/2$, $F = 256$. The proof uses only $v_4 = O(r^2)$, which the delta computation supplies.

Lemma 3 (L3a: population contraction)**.** *Let (x, ζ) be standard bivariate normal with correlation ρ, $z = \sqrt{q}\,x$, and let r be the cosine of $(\phi(z), \phi'(z)\zeta)$. Write $P = \mathbb{E}\phi(z)^2$, $Q = \mathbb{E}[\phi'(z)^2 x^2]$, $S = \mathbb{E}\phi'(z)^2$, $T = \mathbb{E}[\phi(z)\phi'(z)x]$, and $u = 1 - \rho^2$. Then $S - Q = qK(q)$ with K as in Step 3, and*

$$1 - r^2 \;\ge\; u\Big(1 + \tfrac{K(q)}{S(q)}\, q\, \rho^2\Big), \tag{9}$$

with equality at $\rho = 0$. Moreover $1 - r^2 \ge \rho^2(PQ - T^2)/(PQ) > 0$ at $\rho^2 = 1$ unless ϕ is linear.

Proof. $\mathbb{E}[\phi(z)\phi'(z)\zeta] = \rho T$ and $\mathbb{E}[\phi'(z)^2\zeta^2] = \rho^2 Q + uS$, so $1 - r^2 = N/D$ with $N = \rho^2(PQ - T^2) + uPS$ and $D = \rho^2 PQ + uPS$. Using $1 - u = \rho^2$,

$$N - uD \;=\; \rho^2\big[(PQ - T^2) + uP(S - Q)\big].$$

Cauchy–Schwarz gives $PQ \ge T^2$, with equality only if $\phi(z) \propto \phi'(z)x$ a.s., i.e. ϕ linear. Stein's lemma twice gives $Q - S = \mathbb{E}[(x^2 - 1)\phi'(\sqrt{q}x)^2] = q\,\mathbb{E}[(\phi'^2)''(\sqrt{q}\,\xi)] = -qK(q)$, and $K \ge 0$ because ϕ'^2 is non-increasing in $|x|$ (the association argument of Appendix C.14, Step 2, applied to ϕ'^2 and x^2). Since $Q \le S$ we have $D \le PS$, so $N - uD \ge \rho^2\big[(PQ - T^2) + uPqK\big] \ge \rho^2\,\frac{K}{S}\,q\,u\,D$, which is equation 9. □

Check. The slack in equation 9 is never negative at $q \in \{0.01, \dots, 2\}$ for tanh, erf and softsign (its minimum over ρ is 0, attained at $\rho = 0$ where the bound is an equality). The constant K/S is sharp as $q \to 0$, not at fixed q: the slack relative to the kept term $u(K/S)q\rho^2$, minimized over $\rho \in (0, 1)$, is 0.0062 at $q = 0.01$ and 0.0032 at $q = 0.005$, about $0.64\,q$ (`theory/l3a_sharpness.py`); $K/S \to 12|a|$ as $q \to 0$ (3.81 against 4 for tanh at $q = 0.01$) (`theory/l3a_contraction.py`).

Lemma 4 (L3b: sampling drift)**.** *Under the hypotheses of Theorem 3, $K(q) \ge c_0 S(q)$ on $[0, q_{\max}]$ for some $c_0 > 0$: indeed $qK(q) = S - Q = -\mathrm{Cov}(\phi'(\sqrt{q}\,x)^2, x^2) > 0$ for $q > 0$, strictly because ϕ'^2 is non-increasing and not constant, $K(0) = 12|a| > 0$, K is continuous on the compact $[0, q_{\max}]$, and $S \le 1$. Then there is C, independent of F, of $q \le q_{\max}$ and of $\rho \in [0, 1)$, with*

$$\mathbb{E}\big[\log(1 - \rho'^2) \;\big|\; q, \rho\big] \;\ge\; \log(1 - \rho^2) \;+\; \tfrac{K(q)}{S(q)}\, q\rho^2 \;-\; \frac{C}{F}. \tag{10}$$

Proof. Write $D = \mathrm{diag}(\phi'(z_a))$, $X = \phi(z)$, $\zeta = \rho x + \sqrt{u}\,e$ with $u = 1 - \rho^2$ and x, e independent standard normal, so $Y = D\zeta = \rho Dx + \sqrt{u}\,De$ and $1 - \rho'^2 = \|P_{X^\perp} Y\|^2/\|Y\|^2$ with $P_{X^\perp} = I - XX^\top/\|X\|^2$. Put $A = P_{X^\perp} Dx$, which is fixed given x, and $B = P_{X^\perp} De$, and let $P_{A^\perp}$ project, inside the range of $P_{X^\perp}$, onto the complement of A. Splitting along A and discarding that component,

$$\|\rho A + \sqrt{u}B\|^2 \;=\; \big(\rho\|A\| + \sqrt{u}\,\langle B, A/\|A\|\rangle\big)^2 + u\,\|P_{A^\perp} B\|^2 \;\ge\; u\,\|P_{A^\perp} B\|^2,$$

so $\mathbb{E}\log(1 - \rho'^2) \ge \log u + \mathbb{E}\log\|P_{A^\perp} B\|^2 - \mathbb{E}\log\|Y\|^2$. *This is where the corner stops mattering*: the discarded term is the one that survives at $\rho = 1$, and the term kept is Gaussian however close ρ is to 1.

Denominator. By Jensen, $\mathbb{E}\log\|Y\|^2 \le \log\mathbb{E}\|Y\|^2 = \log\big(F(\rho^2 Q + uS)\big)$.

Numerator. Given x, $T = \|P_{A^\perp} B\|^2 = e^\top M e$ with $M = DP_{X^\perp} P_{A^\perp} P_{X^\perp} D$. The nonzero spectrum of M is that of $\Pi D^2 \Pi$ with $\Pi = P_{X^\perp} P_{A^\perp} = P_{A^\perp}$, a projection of corank at most 2, so by Cauchy interlacing $\lambda_k(M) \ge \lambda_{k+2}(D^2)$, and $\lambda_{\max}(M) \le 1$. *Good event.* Choose t_0 with $\mathbb{P}(|\xi| \le t_0) = 3/4$ and $\delta = \min_{q \le q_{\max}} \phi'(\sqrt{q}\,t_0)^2 > 0$ (as $\phi' > 0$ is continuous and even). Let $G = \{m \ge F/2\}$ with $m = \#\{a : \phi'(z_a)^2 \ge \delta\}$; by Hoeffding $\mathbb{P}(G^c) \le e^{-cF}$, uniformly in $q \le q_{\max}$. Take $F \ge 6$, so that $m - 2 \ge 1$ on G. On G, M has at least $m - 2$ eigenvalues $\ge \delta$, so $T \ge \delta\,\chi^2_{m-2}$ in distribution given x, and $\mathrm{tr}\,M \ge \sum_a \phi'(z_a)^2 - 2 \ge \delta F/2 - 2$. With $R = T/\mathrm{tr}\,M$, $\mathrm{Var}(R \mid x) = 2\,\mathrm{tr}M^2/(\mathrm{tr}M)^2 \le 2/\mathrm{tr}M = O(1/F)$. Since $\log(1 + t) \ge t - t^2$ for $t \ge -1/2$, $\mathbb{E}[\log R\,\mathbf{1}\{R \ge 1/2\} \mid x] \ge -\frac{1}{2}\mathbb{P}(R < 1/2 \mid x) - \mathrm{Var}(R \mid x)$; the probability

is at most $e^{-c\,\mathrm{tr}M}$ (Laurent & Massart, 2000, Lemma 1), and $\mathbb{E}[\log^2 R \mid x] \le C$ on G because $R \ge \delta\chi^2_{m-2}/F$; Cauchy–Schwarz then bounds the contribution of $\{R < 1/2\}$ by e^{-cF}. *Exceptional event.* On G^c every eigenvalue of D^2 is at least $\min_a \phi'(z_a)^2 \ge e^{-2C(1+q_{\max}\max_a x_a^2)}$ by the growth hypothesis on $\log\phi'$, so $T \ge e^{-2C(1+q_{\max}\max_a x_a^2)}\chi^2_{F-2}$ and $\mathbb{E}[\log^2 T] \le \mathrm{poly}(F)$; by Cauchy–Schwarz, $\mathbb{E}[|\log T|\,\mathbf{1}_{G^c}] \le \mathrm{poly}(F)e^{-cF/2}$. Together, $\mathbb{E}\log T \ge \mathbb{E}\log \mathrm{tr}M - C/F$. *The trace.* $\sum_a \phi'(z_a)^2$ is a sum of F i.i.d. terms in $[0,1]$ with mean $S \ge c > 0$; the same three bounds (Hoeffding for the good event, $\log(1+t) \ge t - t^2$ on it, and the same polynomial bound off it) give $\mathbb{E}\log(\sum_a \phi'(z_a)^2 - 2) \ge \log(FS) - C/F$, and hence $\mathbb{E}\log T \ge \log(FS) - C/F$.

Algebra. Combining and using $Q = S - qK$, so that $\rho^2 Q + uS = S\big(1 - \frac{K}{S}q\rho^2\big)$ with $\frac{K}{S}q\rho^2 \le 1$ because $Q \ge 0$,

$$\mathbb{E}\log(1-\rho'^2) \;\ge\; \log u + \log(FS) - \log\big(F(\rho^2 Q + uS)\big) - \tfrac{C}{F} \;=\; \log u - \log\Big(1 - \tfrac{K}{S}q\rho^2\Big) - \tfrac{C}{F},$$

and $-\log(1-t) \ge t$. □

Check. $F\big(\mathbb{E}\log(1-\rho'^2) - \log u - (K/S)q\rho^2\big) \ge -1.0$ over $F = 32, \dots, 4096$, $q \in [0.3, 3]/\sqrt{F}$ and ρ up to 0.9999; the bound is slack by factors of hundreds near the corner and tight at $\rho = 0$, where it reduces to $\mathbb{E}\log(\chi^2_{F-1}/F)$ (`theory/l3b_drift.py`).

Lemma 5 (logarithms of concentrated positive variables)**.** *Let $N > 0$ be a random variable with $s = \mathbb{E}N < \infty$, let $n \ge 1$, and suppose* (i) $\mathbb{P}(|N/s - 1| \ge \tau) \le 2e^{-c_1 n \min(\tau, \tau^2)}$ *for all $\tau \ge 0$.*

- *(a) If also $\mathbb{E}(s/N)^{p_0} \le C_1$ for some $p_0 \in (0,1]$, then $|\mathbb{E}\log N - \log s| \le C/n$, and $\log \mathbb{E}\, e^{\lambda(\log N - \mathbb{E}\log N)} \le C\lambda^2/n$ for $|\lambda| \le p_0/4$.*
- *(b) If instead $\mathbb{P}(N \le \varepsilon s) \le (C_2\varepsilon)^{c_2 n}$ for $0 < \varepsilon \le 1$, the same two bounds hold, the second for $|\lambda| \le cn$.*

The constants depend only on c_1, C_1, p_0, C_2, c_2.

Proof. Put $y = N/s - 1$ and $Z = \log(N/s)$. By (i), $\mathbb{E}|y|^k \le C_k n^{-k/2}$ for every k, and $\mathbb{P}(y < -\frac{1}{2}) \le 2e^{-c_1 n/4}$.

Centring. $Z \le y$ gives $\mathbb{E}Z \le 0$. On $\{y \ge -\frac{1}{2}\}$, $Z \ge y - y^2$, and $\mathbb{E}y = 0$. So $\mathbb{E}Z \ge -\mathbb{E}y^2 - (\mathbb{E}(Z-y)^2)^{1/2}\,\mathbb{P}(y < -\frac{1}{2})^{1/2} \ge -C/n$. Here $\mathbb{E}Z^2 < C$ because both $(N/s)^{p_0}$ and $(s/N)^{p_0}$ have bounded means: the first by (i), the second by assumption, or under (b) by $\mathbb{E}(s/N) \le C$ once $c_2 n > 1$.

(a). Let $X = Z - \mathbb{E}Z$. From $e^v \le 1 + v + v^2 e^{|v|}$, $\mathbb{E}e^{\lambda X} \le 1 + \lambda^2(\mathbb{E}X^4)^{1/2}(\mathbb{E}e^{2|\lambda||X|})^{1/2}$.

- For $2|\lambda| \le p_0/2$: $\mathbb{E}e^{2|\lambda||Z|} \le \mathbb{E}(N/s)^{p_0/2} + \mathbb{E}(s/N)^{p_0/2} \le C$.
- $\mathbb{E}Z^4 \le C\,\mathbb{E}y^4 + (\mathbb{E}Z^8)^{1/2}\,\mathbb{P}(y < -\frac{1}{2})^{1/2} \le C/n^2$, using $|Z| \le 2|y|$ on $\{y \ge -\frac{1}{2}\}$.

Hence $\mathbb{E}e^{\lambda X} \le 1 + C\lambda^2/n$.

(b). Upper tail: $\mathbb{P}(Z \ge t) \le \mathbb{P}(y \ge t) \le 2e^{-c_1 n \min(t,t^2)}$. Lower tail: let $t_0 = 2\log(C_2 \vee e)$ and $\kappa = (1 - e^{-t_0})/t_0$.

- For $t \le t_0$: $1 - e^{-t} \ge \kappa t$, so (i) gives $\mathbb{P}(Z \le -t) \le 2e^{-c_1 n \min(\kappa t, \kappa^2 t^2)}$.
- For $t > t_0$: $\mathbb{P}(Z \le -t) \le (C_2 e^{-t})^{c_2 n} \le e^{-c_2 n t/2}$.

So $\mathbb{P}(|Z| \ge t) \le 4e^{-cn\min(t,t^2)}$ for all t, and the same holds for X after adjusting c, since $|\mathbb{E}Z| \le C/n$. Then $\mathbb{E}|X|^k \le C^k(k^{k/2}n^{-k/2} + k^k n^{-k})$, and summing the exponential series gives $\log \mathbb{E}e^{\lambda X} \le C\lambda^2/n$ for $|\lambda| \le cn$. □

Application to $\log(q'/q)$. $q'/q = (g/F)\,N$ with $N = A/q = \sum_a W_a$ and $W_a = \xi_a^2 \psi(\sqrt{q}\,\xi_a)^2$, which are i.i.d. with $0 \le W_a \le \xi_a^2$, each of mean $\Phi(q)/q \ge c$ on $q \le q_{\max}$. So $s = \mathbb{E}N = F\Phi(q)/q \ge cF$.

- Condition (i), with $n = F$, is Bernstein's inequality.
- Since $\psi(y) \ge c_\psi/(1+|y|)$, $W_a \ge c\min(\xi_a^2, 1)$. With $\eta = \sqrt{\varepsilon}$ and $t = 1/(2C'\varepsilon)$, Chernoff's bound gives $\mathbb{P}(N \le \varepsilon s) \le e^{tC'\varepsilon F}\big(\mathbb{P}(\xi^2 \le \eta) + e^{-t\eta}\big)^F \le (C_2\varepsilon)^{F/4}$.

So Lemma 5(b) applies, uniformly in $q \le q_{\max}$: $\log(q'/q)$ has conditional cumulants bounded by $C\lambda^2/F$ for $|\lambda| \le cF$.

Lemma 6 (L3b′: the sharp drift, up to the corner)**.** *Under the structural hypotheses of Theorem 3 ($\sigma_{\mathrm{b}} = 0$), write $u = 1-\rho^2$ and $c = 6|a|$. There is C such that, uniformly for $q \le q_{\max}$ and $|\rho| < 1$,*

$$\mathbb{E}\big[\log(1-\rho'^2) \,\big|\, q, \rho\big] \;\ge\; \log u \;+\; \psi\big(\tfrac{F-1}{2}\big) - \psi\big(\tfrac{F}{2}\big) \;+\; K(0)\,q\rho^2 \;-\; C\big(q^2 + q/F + F^{-2}\big). \quad (11)$$

More precisely, there is a random variable $\tilde{u}'$ (the centred step*), defined jointly with q', such that:*

- $\mathbb{E}[\Psi(1-\rho'^2, q') \mid q, \rho] \le \mathbb{E}[\Psi(\tilde{u}', q') \mid q, \rho]$ *for every $\Psi \ge 0$ that is non-increasing in its first argument;*
- $\mathbb{E}\log(\tilde{u}'/u)$ *equals the right side of equation 11 minus* $\log u$*, up to* $\pm C(q^2 + q/F + F^{-2})$*;*
- $\mathbb{E}\log(\tilde{u}'/u) \ge (K/S)q\rho^2 - C/F$ *for every $q \le q_{\max}$;*
- $L = \log(\tilde{u}'/u)$ *and* $L = \log(q'/q)$ *satisfy* $\log \mathbb{E}[e^{\lambda(L - \mathbb{E}L)} \mid q, \rho] \le C\lambda^2/F$ *for* $|\lambda| \le \lambda_0$.

Proof. We use the notation of the proof of Lemma 4: D, X, $Y = D\zeta$, and $P = P_{X^\perp}$.

Centring. Given x, $Y \sim \mathcal{N}(\rho Dx, uD^2)$. Let $y_s = \langle \hat{X}, Y\rangle$ with $\hat{X} = X/\|X\|$, and $Y_\perp = PY$, so that $1-\rho'^2 = f(\|Y_\perp\|^2)$ with $f(t) = t/(y_s^2 + t)$ increasing. Given (x, y_s), $Y_\perp$ is Gaussian with mean $m = \rho PDx + \kappa\,(y_s - \rho\langle \hat{X}, Dx\rangle)$, where $\kappa = PD^2\hat{X}/(\hat{X}^\top D^2 \hat{X})$, and covariance $u\Sigma_0$, where $\Sigma_0 = PD^2P - \kappa\kappa^\top(\hat{X}^\top D^2 \hat{X})$. Let $Z_0 \sim \mathcal{N}(0, \Sigma_0)$ be independent of y_s given x, and put $\tilde{u}' = f(u\|Z_0\|^2)$.

Anderson's inequality (Anderson, 1955) gives $\mathbb{P}(\|m + \sqrt{u}Z_0\| \le r) \le \mathbb{P}(\sqrt{u}\|Z_0\| \le r)$ for every r. It applies to a centred Gaussian and a symmetric convex set. For a ball, split m into its components in and orthogonal to the range of Σ_0; the orthogonal part only shrinks the ball's radius. So $1-\rho'^2 \ge \tilde{u}'$ in the stochastic order given (x, y_s). Since $q' = gA/F$ is a function of x, the domination claim follows, and $\Psi = -\log$ gives $\mathbb{E}\log(1-\rho'^2) \ge \mathbb{E}\log\tilde{u}'$.

At $q = 0$ we have $m = 0$ and $\tilde{u}' = 1-\rho'^2$ exactly (the Bartlett law of Lemma 7). Nothing is discarded. Lemma 4 instead drops the component along $A = PDx$, which costs a Gaussian degree of freedom, that is, $1/F$ in the constant.

The drift of the centred step. $\log(\tilde{u}'/u) = \log\|Z_0\|^2 - \log\|\hat{Y}\|^2$, with $\|\hat{Y}\|^2 := y_s^2 + u\|Z_0\|^2$.

- *Numerator.* Given x, $\|Z_0\|^2$ is a Gaussian quadratic form, with $0 \preceq \Sigma_0 \preceq I$ and rank at least $F-2$. Expand log to third order about the mean, using $\mathbb{E}(R-1)^2 = 2\,\mathrm{tr}\Sigma_0^2/(\mathrm{tr}\Sigma_0)^2$ and $\mathbb{E}|R-1|^3 = O(F^{-3/2})$ for $R = \|Z_0\|^2/\mathrm{tr}\Sigma_0$, and bound the lower tail as in Lemma 4. This gives $\mathbb{E}[\log\|Z_0\|^2 \mid x] = \log\mathrm{tr}\Sigma_0 - \mathrm{tr}\Sigma_0^2/(\mathrm{tr}\Sigma_0)^2 + O(F^{-2})$. Here $\mathrm{tr}\Sigma_0 = \sum_a \phi'(z_a)^2 - r_x$ with $r_x = 1 + O(q)$ (and $r_x = 1$ at $q = 0$), and $\mathrm{tr}\Sigma_0^2/(\mathrm{tr}\Sigma_0)^2 = (F-1)^{-1}(1+O(q))$. Averaging over x, with $\sum_a(1-\phi'(z_a)^2)/F$ of mean $1 - S(q) = cq + O(q^2)$ and variance $O(q^2/F)$, gives

  $$\mathbb{E}\log\|Z_0\|^2 = \psi\big(\tfrac{F-1}{2}\big) + \log 2 - cq + O(q^2 + q/F + F^{-2}),$$

  since $\psi(\frac{F-1}{2}) + \log 2 = \log(F-1) - (F-1)^{-1} + O(F^{-2})$.

- *Denominator.* In law given (x, y_s), $\|Y\|^2 = y_s^2 + \|m + \sqrt{u}Z_0\|^2$. Anderson again gives $\mathbb{E}\log\|\hat{Y}\|^2 \le \mathbb{E}\log\|Y\|^2$. Conversely, $\log(1+t) \le t$ and the symmetry $Z_0 \mapsto -Z_0$ (which

kills the cross term $2\sqrt{u}\langle m, Z_0\rangle$) give $\mathbb{E}\log\|Y\|^2 - \mathbb{E}\log\|\hat{Y}\|^2 \le \mathbb{E}[\|m\|^2/\|\hat{Y}\|^2] = O(q^2)$. This uses $\|PDx\| = \|P(\phi'(z)z - \phi(z))\|/\sqrt{q} \le C\|z^{\circ 3}\|/\sqrt{q}$ and $\|\kappa\| = O(q)$.

Finally, $\log\|Y\|^2 = \log\|\zeta\|^2 + \log(1 - t_\zeta)$ with $t_\zeta = \sum_a (1 - \phi'(z_a)^2)\zeta_a^2/\|\zeta\|^2$, and $1 - \phi'(z)^2 = cz^2 + O(z^4)$. Since $\zeta \sim \mathcal{N}(0, I)$ with $\mathbb{E}\log\|\zeta\|^2 = \psi(\frac{F}{2}) + \log 2$, and $\mathbb{E}[x_a^2\zeta_a^2/\|\zeta\|^2] = (1 + 2\rho^2)/F + O(F^{-2})$ (Dirichlet moments), we get $\mathbb{E}\log\|Y\|^2 = \psi(\frac{F}{2}) + \log 2 - cq(1 + 2\rho^2) + O(q^2 + q/F)$. The upper bound uses $\log(1 - t) \le -t$; the lower bound uses $\log(1 - t) \ge -t - t^2/(1 - t)$ with the good and exceptional events of Lemma 4.

Subtracting: $\mathbb{E}\log(\tilde{u}'/u) = \psi(\frac{F-1}{2}) - \psi(\frac{F}{2}) + 2cq\rho^2 + O(q^2 + q/F + F^{-2})$, and $2c = 12|a| = K(0)$. For $q \ge q_0$ the remainder is absorbed into Cq^2. The general-q bound follows as in Lemma 4: $\mathrm{tr}\Sigma_0 \ge FS(q) - 2$ with Jensen on $\mathbb{E}\log\|\hat{Y}\|^2 \le \log\mathbb{E}\|Y\|^2 = \log F(\rho^2 Q + uS)$.

Tails. For $\log(q'/q)$, see the application after Lemma 5. For $L = \log(\tilde{u}'/u)$ the argument has four parts: a good event G_x of probability $1 - e^{-cF}$; on it, conditional cumulant bounds for $\log N_1$ and $\log N_2$ from Lemma 5(a); the fluctuation of $\mathbb{E}[L \mid x]$ over x; and, off G_x, crude exponential moments that the small probability of G_x^c absorbs. Write $L = \log N_1 - \log N_2$, with $N_1 = \|Z_0\|^2$ and $N_2 = \|\hat{Y}\|^2 = y_s^2 + u\|Z_0\|^2$, we claim that $\log\mathbb{E}[e^{\lambda(L - \mathbb{E}L)} \mid q, \rho] \le C\lambda^2/F$ for $|\lambda| \le \lambda_0 := 1/32$. Let G_x be the event that the good event of Lemma 4 holds and that the empirical means $\overline{\phi'^2}$, $\overline{\phi^2}/q$, $\overline{\phi\phi' x}/\sqrt{q}$, $\overline{\phi^2\phi'^2}/q$ and $\overline{\phi^2\phi'^4}/q$ lie within a fixed factor of their expectations. Bernstein's inequality gives $\mathbb{P}(G_x^c) \le e^{-cF}$, uniformly in $q \le q_{\max}$.

Given $x \in G_x$, both N_j satisfy Lemma 5(a) with $n = F$ and $p_0 = \frac{1}{4}$.

- *Condition (i).* N_1 is a Gaussian quadratic form $g^\top\Sigma_0 g$. N_2 is a Gaussian quadratic form plus a linear term $2\mu_s y_s$, with $\mu_s = \rho\langle\hat{X}, Dx\rangle$. Both covariances are $\preceq I$. On G_x, $s_1 = \mathrm{tr}\Sigma_0 \ge cF$ and $s_2 = \mu_s^2 + u(\hat{X}^\top D^2\hat{X} + \mathrm{tr}\Sigma_0) \ge cF$. Condition (i) follows from Laurent & Massart (2000, Lemma 1) for the quadratic part and the Gaussian tail for the linear part, whose variance is at most $4s_2$.
- *Negative moments of* N_1. On G_x, at least $F/2 - 2$ eigenvalues of Σ_0 exceed δ (the interlacing argument of Lemma 4). So $N_1 \ge \delta\chi^2_{F/2-2}$ in law, and $\mathbb{E}(s_1/N_1)^{1/4} \le C$.
- *Negative moments of* N_2. If $\mu_s^2 \ge s_2/2$: for $y \sim \mathcal{N}(\mu, v)$ and $p < 1$, $\mathbb{E}|y|^{-p} \le (|\mu|/2)^{-p} + (2\pi v)^{-1/2}e^{-\mu^2/(8v)}\int_{|y|<|\mu|/2}|y|^{-p}dy \le C|\mu|^{-p}$ uniformly in v, so $\mathbb{E}(s_2/y_s^2)^{1/4} \le C$. Otherwise $u\,\mathrm{tr}\Sigma_0 \ge s_2/2 - u \ge s_2/4$, and $\mathbb{E}(s_2/(uN_1))^{1/4} \le C$ as for N_1. In both cases $\mathbb{E}(s_2/N_2)^{1/4} \le C$, since $N_2 \ge \max(y_s^2, uN_1)$.

By Cauchy–Schwarz, $\log\mathbb{E}[e^{\lambda(L - \mathbb{E}[L|x])} \mid x] \le C\lambda^2/F$ on G_x for $|\lambda| \le 1/32$, and $\mathbb{E}[L \mid x] = \log(s_1/s_2) + O(1/F)$. Moreover s_1/F and s_2/F are smooth functions of the empirical means listed in G_x, bounded away from 0 there, plus $O(1/F)$ terms. Bernstein's inequality for those means then gives sub-exponential fluctuations of $\log(s_1/s_2)$ at scale $F^{-1/2}$.

Off G_x, $|L| \le |\log N_1| + |\log N_2|$ has all exponential moments of order $\le \frac{1}{16}$ bounded by $\mathrm{poly}(F)$, by the exceptional-event bounds of Lemma 4 ($\min_a \phi'(z_a)^2 \ge e^{-2C(1+q_{\max}\max_a x_a^2)}$ and $\mathbb{E}e^{c\max_a x_a^2} \le F\,\mathbb{E}e^{cx^2}$).

Combining. As in the proof of Lemma 5(a), write $X = L - \mathbb{E}L$ and $\mathbb{E}e^{\lambda X} \le 1 + \lambda^2\,\mathbb{E}[X^2 e^{|\lambda X|}]$. The contribution of G_x is $O(1/F)$ by the two bounds just shown. That of G_x^c is at most $\mathbb{P}(G_x^c)^{1/2}(\mathbb{E}[X^4 e^{2|\lambda X|}])^{1/2} \le e^{-cF/2}\,\mathrm{poly}(F)$ by Cauchy–Schwarz, which is $O(1/F)$. So both parts are $O(1/F)$, and the claim follows. The conditional bound becomes unconditional because the x-fluctuation term is itself sub-exponential at scale $F^{-1/2}$. □

Check. On the exact chain at $F = 64, 256, 1024$, $Q = Fq \in \{0.25, 1, 4\}$ and ρ up to $1 - 10^{-8}$ (45 cells), the Anderson gap $F\,\mathbb{E}[\log(1 - \rho'^2) - \log\tilde{u}']$ is never negative beyond one standard error. It is 0.0000 at $\rho = 0$ and grows without bound at the corner, where it is the injection. $F\,\mathbb{E}\log(\tilde{u}'/u)$

matches $F(\psi(\frac{F-1}{2}) - \psi(\frac{F}{2})) + 4Q\rho^2$ within 2.6, 0.8 and 0.2 at $Q = 4$ and the three widths, which is $O(Q^2/F)$, and far closer at smaller Q. $F \operatorname{Var} \log(\tilde{u}'/u) = 4\rho^2 + o(1)$, which is the diffusion's noise (`theory/r5_corner.py`).

Lemma 7 (L3c: the sampling constant away from the corner)**.** *Fix $\eta \in (0, 1)$. Uniformly for $q \le q_{\max}$ and $1 - \rho^2 \ge \eta$,*

$$\mathbb{E}\big[\log(1 - \rho'^2) \mid q, \rho\big] \;=\; \log(1 - r^2) \;-\; \frac{1}{F} \;+\; O_\eta\big(q/F + F^{-3/2}\big). \tag{12}$$

Proof. Put $h(w) = \log(1 - w_3^2/(w_1 w_2))$, so $\log(1 - \rho'^2) = h(\bar{W})$ with $\bar{W}$ the sample mean of $W = (X^2, Y^2, XY)$ as in Lemma 2. By Lemma 3, $1 - r^2 \ge 1 - \rho^2 \ge \eta$ at $w = \mu$, so h is C^4 with bounded derivatives on a neighbourhood N of μ whose size depends only on η, and the argument of Lemma 2 gives $\mathbb{E}[h(\bar{W})\mathbf{1}\{\bar{W} \in N\}] = h(\mu) + c_2(q, \rho)/F + O(F^{-3/2})$ with $c_2 = \frac{1}{2}\sum_{ij} \partial^2_{ij} h(\mu) C_{ij}$. Off N, the numerator bound in the proof of Lemma 4 gives $|h(\bar{W})| \le \log(1/\eta) + |\log T| + |\log \|Y\|^2|$, and there $\mathbb{E} \log^2 T \le \operatorname{poly}(F)$ (shown in that proof). Likewise $\mathbb{E} \log^2 \|Y\|^2 \le \operatorname{poly}(F)$ uniformly in (q, ρ), by the same lower-tail argument: $\|Y\|^2 = \sum_a \phi'(z_a)^2 \zeta_a^2 \ge \min_a \phi'(z_a)^2 \sum_a \zeta_a^2$, with $\min_a \phi'(z_a)^2 \ge e^{-2C(1 + q_{\max} \max_a x_a^2)}$ by the growth hypothesis on $|\log \phi'|$. Since $\bar{W}$ averages F i.i.d. vectors with all moments bounded uniformly in (q, ρ), Markov's inequality at order $2k$ gives $\mathbb{P}(\bar{W} \notin N) = O(F^{-k})$ for every k, uniformly; by Cauchy–Schwarz the event contributes $O(F^{-2})$, uniformly. At $q = 0$ the pairs are Gaussian with correlation ρ and the law is exact: by the Bartlett decomposition of the 2×2 Wishart matrix of the pairs, $1 - \rho'^2 = W_{22\cdot 1}/W_{22}$ with $W_{22\cdot 1} \sim (1 - \rho^2)\chi^2_{F-1}$ and $W_{22} \sim \chi^2_F$, so, by linearity alone,

$$\mathbb{E} \log(1 - \rho'^2) = \log(1 - \rho^2) + \psi\big(\tfrac{F-1}{2}\big) - \psi\big(\tfrac{F}{2}\big) = \log(1 - \rho^2) - \tfrac{1}{F} + O(F^{-2}) \quad \text{for every } \rho.$$

Hence $c_2(0, \rho) = -1$. After the scale-free normalization $X = \phi(\sqrt{q}\, x)/\sqrt{q}$, the moments entering μ and C are expectations of functions even in $\sqrt{q}$ with polynomially bounded derivatives, so they are C^1 in $q \in [0, q_{\max}]$ (differentiation under the integral, dominated by polynomial growth), uniformly in ρ. Moreover c_2 is a smooth function of (μ, C) on the compact set where $\mu_1, \mu_2 \ge c$ and $1 - \mu_3^2/(\mu_1\mu_2) \ge \eta/2$. So c_2 is C^1 in q uniformly on $\{1 - \rho^2 \ge \eta\}$, and $c_2(q, \rho) = -1 + O(q)$ there. □

Check. With the Gaussian pairs as a control variate on the same draws, $F\big(\mathbb{E}\log(1 - \rho'^2) - \log(1 - r^2)\big) = -0.750, -0.832, -0.898, -0.949$ at $\rho = 0$ and $F = 64, 256, 1024, 4096$ ($q = 1/\sqrt{F}$): the gap to -1 (0.25, 0.17, 0.10, 0.05) shrinks roughly in proportion to q, as $O(q)$ requires. At $\rho = 0.9$ it is $-1.36, -1.26, -1.10, -1.15 \pm 0.21$ (`theory/l3c_constant.py`).

C.7 The alignment cost and the band

Proposition 5 (the alignment cost is exactly $1/F$)**.** *Under the hypotheses of Theorem 3, and uniformly for $\nu\sqrt{F} \in [s_0, s_1]$:* (i) $\mathbb{E}_\pi[q\rho^2] \le C/F$*;* (ii) $\mathbb{E}_\pi[\rho^2 q K(q)] = F^{-1} + o(F^{-1})$*. No condition on the tail of π is assumed.*

Proof. Truncation. Write $u = 1 - \rho^2$, P for the chain's one-step kernel, and $D(q, \rho) = \mathbb{E}[\log u' - \log u \mid q, \rho]$, which is finite for $u > 0$ by Lemma 4. For $\varepsilon \in (0, 1)$ put $f_\varepsilon = \log(u \vee \varepsilon)$. It is bounded, so stationarity gives $\mathbb{E}_\pi[Pf_\varepsilon - f_\varepsilon] = 0$ *with no integrability assumption on* $\log u$. Since $\log(u' \vee \varepsilon) \ge \log u'$, on $\{u \ge \varepsilon\}$ we have $Pf_\varepsilon - f_\varepsilon \ge D$, and on $\{u < \varepsilon\}$ we have $Pf_\varepsilon - f_\varepsilon \ge \log\varepsilon - \log\varepsilon = 0$. Hence, for any bounded B with $D \ge B$ on $\{u > 0\}$,

$$0 \;=\; \mathbb{E}_\pi[Pf_\varepsilon - f_\varepsilon] \;\ge\; \mathbb{E}_\pi\big[B\,\mathbf{1}\{u \ge \varepsilon\}\big]. \tag{13}$$

The chain puts no mass on the corner: from any state $\mathbb{P}(\rho' = \pm 1) = 0$ (Lemma 1(d)), so $\pi(u = 0) = 0$. Letting $\varepsilon \downarrow 0$ with dominated convergence (B bounded) gives $\mathbb{E}_\pi B \le 0$.

(i). Lemma 4 gives $D \ge B := (K/S)q\rho^2 - C/F$, which is bounded, so $\mathbb{E}_\pi[(K/S)q\rho^2] \le C/F$, and $K \ge c_0 S$ gives (i). With (i) and Step 4, $\mathbb{E}_\pi\rho^2 = O(F^{-1/2})$: split at $q = \mathbb{E}_\pi q/2$, and use Chebyshev with $\operatorname{Var}_\pi q = O(F^{-3/2})$ and $\mathbb{E}_\pi q \asymp F^{-1/2}$ (Step 4) for the lower part and (i) for the upper, to get $\mathbb{E}_\pi\rho^2 \le \mathbb{P}_\pi(q < \mathbb{E}_\pi q/2) + 2\mathbb{E}_\pi[q\rho^2]/\mathbb{E}_\pi q = O(F^{-1/2})$. Also $\mathbb{E}_\pi[q^2\rho^2] = O(F^{-5/4})$.

Upper bound in (ii). Fix $\eta \in (0,1)$. On $\{u \geq \eta\}$, Lemmas 3 and 7 give $D \geq \log\big(1+\frac{K}{S}q\rho^2\big) - \frac{1}{F} - C_\eta(q/F + F^{-3/2}) \geq \frac{K}{S}q\rho^2 - Cq^2\rho^2 - \frac{1}{F} - C_\eta(q/F+F^{-3/2})$; on $\{0<u<\eta\}$, Lemma 4 gives $D \geq \frac{K}{S}q\rho^2 - \frac{C}{F}$. Both lower bounds are bounded functions of (q,ρ), so equation 13 applies and gives,

$$\mathbb{E}_\pi\big[\tfrac{K}{S}q\rho^2\big] \;\leq\; \frac{1}{F} + \frac{C}{F}\,\mathbb{P}_\pi(u<\eta) + O\big(\mathbb{E}_\pi[q^2\rho^2] + \mathbb{E}_\pi q/F + F^{-3/2}\big),$$

and $\mathbb{P}_\pi(u<\eta) \leq \mathbb{E}_\pi\rho^2/(1-\eta) = O(F^{-1/2})$. *The corner costs only its stationary mass, which (i) already makes small.* As $S = 1-O(q)$, $\mathbb{E}_\pi[qK\rho^2] \leq F^{-1} + o(F^{-1})$.

Lower bound in (ii). In the notation of Lemma 3, $\rho^2 - r^2 = \rho^2[(PQ-T^2) + uPqK]/D'$ with $D' = \rho^2 PQ + uPS \geq PQ$, and $(PQ-T^2)/(PQ) = O(q^2)$, $Q = 1-O(q)$; so $\rho^2 - r^2 \leq qK\rho^2 + O(q^2\rho^2)$. Stationarity of ρ^2 with Lemma 2 gives $\mathbb{E}_\pi[\rho^2 - r^2] = F^{-1}\mathbb{E}_\pi[(1-r^2)(1-2r^2)] + O(\mathbb{E}_\pi q/F + F^{-2}) \geq F^{-1}(1-3\mathbb{E}_\pi\rho^2) - o(F^{-1})$, hence $\mathbb{E}_\pi[qK\rho^2] \geq F^{-1} - o(F^{-1})$. (In detail: $(1-r^2)(1-2r^2) = 1-3r^2+2r^4 \geq 1-3r^2 \geq 1-3\rho^2$, since $r^2 \leq \rho^2$ by Lemma 3; $\mathbb{E}_\pi q/F = O(F^{-3/2})$ by Step 4; and $\mathbb{E}_\pi\rho^2 = O(F^{-1/2})$ by (i), so the bracket is $F^{-1}(1-o(1))$.) □

Check. On the exact chain at the predicted crossing (8 seeds, burn-in $60\sqrt{F}$), $F\,\mathbb{E}_\pi[\rho^2 qK(q)] = 0.835 \pm 0.006,\ 0.934 \pm 0.013,\ 0.975 \pm 0.038$ at $F = 1024, 4096, 16384$; with K frozen at $K(0) = 12|a|$ it is 1.006, 1.026, 1.022. The two differ by $O(\mathbb{E}_\pi[q^2\rho^2])$, since $K(q) \approx K(0)(1-8.5q)$ for tanh, and $F\,\mathbb{E}_\pi[q^2\rho^2] = 0.0073, 0.0036, 0.0018$ (`theory/alignment_cost.py`).

Remark (the tail, empirically). Proposition 5 needs no bound on how often the chain visits the corner, but the chain rarely does: the time spent at $\rho^2 \geq 1/2$ is 3.1×10^{-2}, 4.8×10^{-3}, 5.0×10^{-4}, 6.5×10^{-5} and 7.5×10^{-6} at $F = 8$, 16, 32, 48, 64, fitting $\log\pi = 1.01 - 1.565\sqrt{F}$ ($R^2 = 0.996$; $\log\pi$ against F fits nearly as well), and the largest ρ^2 visited falls from 0.99 to 0.54 (`theory/alignment_tail.py`).

Proposition 6 (the consequence bound)**.** *Under the structural hypotheses of Theorem 3 (those on ϕ; its window condition $\nu\sqrt{F} \geq s_0$ is dropped here) there is C, independent of F and ν, such that uniformly for $0 < \nu\sqrt{F} \leq s_1$,*

$$2F\chi^\star \;=\; F\nu^2/3 \;-\; F\,\mathbb{E}_\pi[\rho^2 qK(q)] \;+\; o(1),$$
$$0 \;\leq\; F\,\mathbb{E}_\pi[\rho^2 qK(q)] \;\leq\; C + o(1).$$

In particular $|2F\chi^\star| \leq \max(1,C) + o(1)$ for every $\nu > 0$ up to $\mathcal{C}_{\rm q}$ ($\nu = \sqrt{3/F}(1+o(1))$), that is, strictly above $\mathcal{C}_{\rm lin}(F)$ ($\nu = 0$). The chain is built for $\nu > 0$, and the endpoint itself is not covered. An initialization anywhere in that band therefore loses at most $O(L/F)$ nats of log-gain over L layers against one at $\mathcal{C}_{\rm q}$.

Proof. Nothing in Step 4, in Lemma 4 or in part (i) of Proposition 5 uses $\nu\sqrt{F} \geq s_0$. Step 4's stationarity identities give, uniformly for $\nu\sqrt{F} \leq s_1$, $\mathbb{E}_\pi q^k \leq C_k F^{-k/2}$, $6|a|\,\mathbb{E}_\pi q = \nu + O(F^{-1})$ and $F\,\mathbb{E}_\pi\Lambda = F\nu^2/3 + O(F^{-1/2})$. Part (i) of Proposition 5, which rests only on Lemma 4 and on stationarity of $\log(1-\rho^2)$ for any $\nu > 0$, gives $\mathbb{E}_\pi[\rho^2 qK] \leq C/F$. Then $\mathbb{E}_\pi[\rho^2 q^2] \leq F^{-1/4}\mathbb{E}_\pi[\rho^2 q] + \mathbb{E}_\pi[q^2\mathbf{1}\{q > F^{-1/4}\}] = O(F^{-5/4})$ (the second term is at most $F^{1/2}\mathbb{E}_\pi q^4 = O(F^{-3/2})$), so the remainder of equation 16 is $o(F^{-1})$ and equation 15 yields the displayed identity. Between the two loci $F\nu^2/3 \leq 1 + o(1)$. □

Lemma 8 (L4: the state rarely falls far below its mean)**.** *Under the structural hypotheses of Theorem 3 there are $c, C, C_0 > 0$ such that, uniformly for $0 < \nu\sqrt{F} \leq s_1$ with $F\nu \geq C_0 \log F$, writing $m_1 = \mathbb{E}_\pi q$: (a) $6|a|\,m_1 = \nu\,(1+O(\nu+F^{-1}))$; (b) $\mathbb{P}_\pi(q \leq m_1/2) \leq e^{-cF\nu}$; (c) $\mathbb{E}_\pi\rho^2 \leq C/(F\nu)$; (d) $F\,\mathbb{E}_\pi[\rho^2 qK(q)] \leq 1 + C/(F\nu) + o(1)$. So Proposition 6 holds with the sharp constant $C = 1$ wherever $F\nu \geq C_0 \log F$.*

Proof. (a) Step 4(i) gives $\mathbb{E}_\pi q^k \leq C_k(\nu + 1/F)^k$ for every $\nu\sqrt{F} \leq s_1$ (Proposition 6), so stationarity of $\log q$ gives $\mathbb{E}_\pi\ell(q) = \nu + O(\mathbb{E}_\pi q^3 + \mathbb{E}_\pi q/F)$. Since $\ell(q) = 6|a|q + O(q^2)$ and $F\nu \geq C_0\log F$, this is (a).

(b) Work in $w = \log q$, with $\Delta = w' - w$ and $D(q) = \mathbb{E}[\Delta \mid q]$ as in Lemma 1. Two facts. (i) $\Delta - D(q)$ is the centred logarithm of $N = A/q$, so the application of Lemma 5(b) after its proof

gives $\log \mathbb{E}[e^{-\theta(\Delta - D(q))} \mid q] \le C\theta^2/F$ for $0 \le \theta \le cF$, uniformly in $q \le q_{\max}$. Its centring bound gives $D(q) \ge \log g + \log(\Phi(q)/q) - C/F$. (ii) By Step 3, $D(q) = \nu - \ell(q) + O(q^3 + q/F)$. With (a) and $\ell(q) \le 6|a|q(1 + Cq)$, this gives $D(q) \ge \nu/5$ for $q \le q_3 := 3m_1/4$ (three quarters of the mean), once F is large: there $\ell(q) \le \frac{3}{4}\nu(1 + o(1))$, so $D(q) \ge \nu/4 - o(\nu)$.

Put $\theta = F\nu/(10C)$, $\lambda = e^{-\theta\nu/10}$ and $V(w) = (m_1/(2q))^\theta$.

- For $q \le q_3$, (i) and (ii) give $PV \le V\, e^{-\theta\nu/5 + C\theta^2/F} \le \lambda V$.
- For $q > q_3$, the lower bound on D in (i) gives $PV \le \exp\{\theta[\log(m_1/2) - \log\Phi(q) - \log g] + C\theta/F + C\theta^2/F\}$. Since Φ is increasing, $\Phi(q) \ge \Phi(q_3) = q_3(1 + O(\nu))$, so the bracket is at most $\log(m_1/2) - \log(3m_1/4) + O(\nu) = \log(2/3) + O(\nu)$, and $\theta^2/F \le s_1^2$. So $PV \le b := C' e^{-\theta/3}$.

Hence $PV \le \lambda V + b$ everywhere. Iterating from a fixed initial state w_0 gives $\mathbb{E}_{w_0} V(w_n) \le \lambda^n V(w_0) + b/(1 - \lambda)$. The chain is positive Harris and aperiodic (Lemma 1), so $P^n(w_0, \cdot) \to \pi$ in total variation (Meyn & Tweedie, 2009, Thm. 13.0.1). For each K, therefore, $\pi(V \wedge K) = \lim_n \mathbb{E}_{w_0}[V(w_n) \wedge K] \le b/(1 - \lambda)$, and $K \to \infty$ gives $\pi(V) \le b/(1 - \lambda)$. Here $\theta\nu/10 = F\nu^2/(100C) \le s_1^2/(100C) \le 1$, after enlarging C if necessary. So $1 - e^{-t} \ge t/2$ with $t = \theta\nu/10$ gives $1 - \lambda \ge \theta\nu/20 \asymp F\nu^2$, and this yields $\mathbb{P}_\pi(q \le m_1/2) = \pi(V \ge 1) \le C'' e^{-\theta/3}/(F\nu^2)$. Finally $1/(F\nu^2) = F/(F\nu)^2 \le F \le e^{F\nu/C_0}$, which gives (b) for suitable C_0.

(c) Split at $q = m_1/2$: $\mathbb{E}_\pi \rho^2 \le \mathbb{P}_\pi(q \le m_1/2) + (2/m_1)\, \mathbb{E}_\pi[q\rho^2]$. Proposition 5(i), valid for every $\nu > 0$ (Proposition 6), gives $\mathbb{E}_\pi[q\rho^2] \le C/F$, and $m_1 \asymp \nu$ by (a).

(d) The upper-bound argument of Proposition 5(ii) uses only Lemmas 3 and 7, the truncation identity, and the estimates $\mathbb{P}_\pi(u < \eta) \le \mathbb{E}_\pi \rho^2/(1 - \eta)$, $\mathbb{E}_\pi[q^2\rho^2] = O(F^{-5/4})$ and $\mathbb{E}_\pi q/F = o(1/F)$. All of these hold here. With (c) it gives $F\, \mathbb{E}_\pi[(K/S) q\rho^2] \le 1 + C/(F\nu) + o(1)$, and $S \le 1$ finishes. □

Corollary 1 (the sharp constant on the whole band)**.** *Under the structural hypotheses of Theorem 3, uniformly for* $0 < \nu\sqrt{F} \le s_1$*,*

$$F\, \mathbb{E}_\pi[\rho^2 q K(q)] \;\le\; 1 + C\big(F\nu^2 + F^{-1}\big), \qquad \textit{hence} \qquad \sup_{0 < \nu\sqrt{F} \le s_1} F\, \mathbb{E}_\pi[\rho^2 q K(q)] \;\le\; 1 + o(1).$$

So Proposition 6 holds with $C = 1$*:* $|2F\chi^\star| \le 1 + o(1)$ *on the whole band, and an initialization anywhere in it loses at most* $(1 + o(1))L/(2F)$ *nats against one at* $\mathcal{C}_{\mathrm{q}}$*.*

Proof. The right side of equation 11, minus $\log u$, is a bounded function B of (q, ρ), so the truncation identity equation 13 gives $\mathbb{E}_\pi B \le 0$:

$$K(0)\, \mathbb{E}_\pi[q\rho^2] \;\le\; \psi\big(\tfrac{F}{2}\big) - \psi\big(\tfrac{F-1}{2}\big) + C\, \mathbb{E}_\pi\big(q^2 + q/F\big) + CF^{-2}.$$

Here $F(\psi(\frac{F}{2}) - \psi(\frac{F-1}{2})) = 1 + O(F^{-1})$, $K(q) \le K(0) + Cq$, and Step 4(i) gives $\mathbb{E}_\pi q^k \le C_k(\nu + 1/F)^k$. With $\nu \le F\nu^2 + F^{-1}$, this is the first bound. For $F\nu \le C_0 \log F$ we have $F\nu^2 \le C_0^2 \log^2 F/F$. For $F\nu \ge C_0 \log F$, Lemma 8(d) gives $1 + C/(F\nu) + o(1) \le 1 + C/(C_0 \log F) + o(1)$. The rest is Proposition 6's identity, with $F\nu^2/3 \le 1 + o(1)$ up to $\mathcal{C}_{\mathrm{q}}$. □

What is proved and what is not. Proposition 6 proves that the exponent in the band is $O(1/F)$, with a constant C inherited from Lemma 4. Corollary 1 proves the sharp constant $C = 1$ on the whole band. It combines Lemma 6 for $F\nu = O(\log F)$ with Lemma 8 for $F\nu \ge C_0 \log F$. Inside the layer $F\nu = O(1)$, Theorem 5 identifies the limit of the cost itself. On the exact chain the cost $F\, \mathbb{E}_\pi[4q\rho^2]$ never exceeds 1.05 ($F = 256$) or 1.01 ($F = 1024$) anywhere in the band, consistent with $1 + C(F\nu^2 + F^{-1})$. It falls to 0.46 and 0.94 at $\nu\sqrt{F} = 0.02$, where $q \to 0$ although state and tangent partly align (`theory/t1_boundary_layer.py`, `theory/window_sweep.py`).

C.8 The boundary layer

The layer. In the layer $x = F\nu = O(1)$, (Fq, ρ) keeps $O(1)$ randomness, and on the time scale of F layers the chain is close to a two-dimensional diffusion. Proposition 7 identifies that diffusion and

solves it: its cost is $c(x) = \min(1, 2x)$. Theorem 5 proves that the chain's cost converges to it for every $x \neq \frac{1}{2}$. On the chain the cost is 0.36–0.44, 0.69–0.76, 0.94–0.97 and 0.97–1.00 at $x = 0.25$, 0.5, 1 and 1.5 for F between 64 and 4096, below $c(x)$, with the gap closing slowly in F; the remark after the proposition says why. For $x \geq 2$ it approaches 1: from above at $F = 256$ (at most 1.04), within 0.004 at $F = 1024$, and from below at $F = 4096$ (at most 0.997). On the chain, $2F\chi^\star \geq -0.993$, -0.984, -0.980 across the band at $F = 256$, 1024, 4096 (`theory/tf_scaling_layer.py`, `theory/r5_diffusion.py`). None of this bears on Theorem 3, which assumes $\nu\sqrt{F} \geq s_0 > 0$.

Proposition 7 (the boundary-layer diffusion)**.** *Under the structural hypotheses of Theorem 3, put $c = 6|a|$ and $K_0 = K(0) = 12|a| = 2c$. In the layer write $x = F\nu > 0$, $Q = Fq$, $Y = \log Q$ and $\eta = \operatorname{artanh}\rho$, and measure time in units of F layers. (a) Uniformly for (Q, ρ) in compact subsets of $(0, \infty) \times (-1, 1)$, the one-step increments of the exact chain satisfy, as $F \to \infty$,*

$$F\,\mathbb{E}[\Delta Y] = x - cQ + o(1), \qquad F\,\mathrm{Var}\,\Delta Y = 2 + o(1),$$
$$F\,\mathbb{E}[\Delta\rho] = -\tfrac{1}{2}\rho(1-\rho^2)(1+K_0Q) + o(1), \qquad F\,\mathrm{Var}\,\Delta\rho = (1-\rho^2)^2 + o(1),$$
$$F\,\mathrm{Cov}(\Delta Y, \Delta\rho) = \rho(1-\rho^2) + o(1),$$

and the tangent grows at $F\,\mathbb{E}[\Delta \log\|w\|^2] = x - cQ - K_0Q\rho^2 + o(1)$. The limiting diffusion is therefore

$$dY = (x - ce^Y)\,dt + \sqrt{2}\,dB_1, \qquad d\eta = \tfrac{1}{2}\tanh\eta\,(1 - K_0e^Y)\,dt + dB_2, \tag{14}$$
$$d\langle B_1, B_2\rangle_t = 2^{-1/2}\tanh\eta_t\,dt.$$

(b) For $x > \frac{1}{2}$, equation 14 is positive recurrent. Its invariant probability is the product of $Q \sim \mathrm{Gamma}(x, c)$ (shape x, rate c) and the law of ρ with density proportional to $(1-\rho^2)^{x-3/2}$ on $(-1, 1)$. Under it $\mathbb{E}\rho^2 = 1/(2x)$ and $\mathbb{E}[K_0Q\rho^2] = 1$. (c) From any start with $|\rho_0| < 1$, almost surely $\frac{1}{T}\int_0^T K_0Q_t\rho_t^2\,dt \to c(x) := \min(1, 2x)$ for every $x \neq \frac{1}{2}$. So the diffusion's analogue of $2F\chi^\star$, the long-run average of the drift of $\log\|w\|^2$ in (a), is $-c(x) \geq -1$, with equality for every $x > \frac{1}{2}$.

Proof. (a) Write $z_a = \sqrt{q}\,x_a$ and expand $\phi(z) = z + az^3 + O(z^5)$ and $\phi'(z) = 1 + 3az^2 + O(z^4)$. The remainders are uniform because $\phi \in C^5$ has bounded derivatives, and the Gaussian moments of the $O(q^2)$ terms are $O(q^2)$ uniformly on compacts. In the notation of Step 1:

- $\Delta Y = \log g + \log(S_x/F) + 2aq\sum_a x_a^4/S_x + O_P(q^2)$. By Theorem 1, $\log g_{\mathrm{lin}} = -\mathbb{E}\log(S_x/F)$, and $\log g = \log g_{\mathrm{lin}} + x/F$. So the first two terms have mean x/F and variance $2/F + O(F^{-2})$, and the third has mean $6aq + O(q/F) = -cQ/F + O(QF^{-2})$.
- Likewise $\Delta\log\|w\|^2 = \log g + \log(S_\zeta/F) + 6aq\sum_a x_a^2\zeta_a^2/S_\zeta + O_P(q^2)$, with $S_\zeta \sim \chi^2_F$ and $\mathbb{E}[x^2\zeta^2] = 1 + 2\rho^2$. This gives $x - cQ - 2cQ\rho^2$, and $2c = K_0$.
- At $q = 0$, ρ' is the uncentred sample correlation of F i.i.d. pairs with correlation ρ. Call it ρ_0'. By the delta method, its mean is $\rho - \rho(1-\rho^2)/(2F)$, its variance $(1-\rho^2)^2/F$, and its covariance with $\log(S_x/F)$ is $\rho(1-\rho^2)/F$, each up to $O(F^{-2})$.
- The $O(q)$ correction to ρ' is $\big[4aq\sum_a x_a^3\zeta_a\big]/\sqrt{S_xS_\zeta} - \rho_0'\big[aq\sum_a x_a^4/S_x + 3aq\sum_a x_a^2\zeta_a^2/S_\zeta\big]$. By the law of large numbers ($\sum x^3\zeta/F \to 3\rho$, $\sum x^4/F \to 3$, $\sum x^2\zeta^2/F \to 1{+}2\rho^2$) its mean is $aq\rho\,[12{-}3{-}3(1{+}2\rho^2)]{+}o(q) = 6aq\,\rho(1{-}\rho^2){+}o(q) = -\frac{1}{2}K_0Q\,\rho(1-\rho^2)/F + o(F^{-1})$.

Its fluctuations enter the second moments only at order $q/\sqrt{F}$. Itô's formula for $\eta = \operatorname{artanh}\rho$ turns $d\rho = -\frac{1}{2}\rho(1-\rho^2)(1+K_0Q)\,dt + (1-\rho^2)\,dW$ into the additive-noise form equation 14. The noise correlation is $\rho(1-\rho^2)/(\sqrt{2}(1-\rho^2)) = \rho/\sqrt{2}$.

(b) The diffusion matrix of (Y, η) is $\left(\begin{smallmatrix} 2 & t \\ t & 1 \end{smallmatrix}\right)$ with $t = \tanh\eta$. Its determinant is $2 - t^2 \geq 1$, so the diffusion is uniformly elliptic, and its coefficients are smooth. Let $\mathcal{L}$ be the generator. Since $\mathcal{L}$ has no mixed term on sums of functions of Y and of η separately, and $K_0 = 2c$,

$$\mathcal{L}(\log\cosh\eta) = \tfrac{1}{2}(1 - K_0Q\rho^2), \qquad \mathcal{L}(-Y) = cQ - x, \qquad \mathcal{L}(e^Y) = Q(x + 1 - cQ).$$

Take $W = \log\cosh\eta - Y + \beta e^Y$, which tends to ∞ at infinity. Then

$$\mathcal{L}W = -(x - \tfrac{1}{2}) + cQ(1-\rho^2) + \beta Q(x + 1 - cQ).$$

- Choose β with $\beta(x+1)^2/(4c) \le (x - \tfrac{1}{2})/4$.
- Then $\mathcal{L}W \le -(x - \tfrac{1}{2})/2$ outside a compact set. As $Y \to +\infty$ the $-\beta c Q^2$ term dominates. As $Y \to -\infty$, $Q \to 0$. For bounded Y and $|\eta| \to \infty$, $cQ(1-\rho^2) = cQ\,\mathrm{sech}^2\,\eta \to 0$.

Three conclusions follow, from different sources. *No explosion:* $W \to \infty$ at infinity and $\mathcal{L}W \le C$ everywhere, so Itô's formula for W, stopped at the exit times of growing compacts, rules out explosion. *A unique invariant probability:* the diffusion matrix is uniformly nondegenerate (the noises of Y and η have correlation $\tanh\eta/\sqrt{2}$, at most $1/\sqrt{2}$ in modulus), and $\mathcal{L}W \le -(x - \tfrac{1}{2})/2$ outside a compact set. Take for U an open ball containing that compact set, so that U is bounded with a regular (smooth) boundary; Dynkin's formula then gives a finite expected hitting time of U, uniformly for starting points in compacts. This is Khasminskii's condition (B), and it gives an invariant probability μ (Khasminskii, 2012, Thm. 4.1, p. 108), which is unique (Khasminskii, 2012, Cor. 4.4, p. 112). *Ergodic averages:* time averages of μ-integrable functions converge almost surely to their μ-means (Khasminskii, 2012, Thm. 4.2, p. 110, and §4.5).

Identifying μ. Try the stationary Fokker–Planck equation with the product $p = f(Y)h(\eta)$, where $f \propto \exp(xY - ce^Y)$ solves the Y part. Divide the rest by f and use $f'/f = x - ce^Y$: $\frac{1}{2}h'' - \frac{1}{2}(1 - K_0 e^Y)(th)' + (x - ce^Y)(th)' = 0$. The e^Y terms cancel because $K_0 = 2c$, which leaves $\frac{1}{2}h'' + (x - \frac{1}{2})(th)' = 0$. This is solved by $h = \cosh^{1-2x}\eta$, integrable if and only if $x > \frac{1}{2}$. So $\int \mathcal{L}\varphi\, p = 0$ for every $\varphi \in C_c^\infty$. The martingale problem is well posed (locally Lipschitz coefficients, no explosion), so p is invariant (Ethier & Kurtz, 1986, Thm. 4.9.17), and hence $p = \mu$.

In $\rho = \tanh\eta$, $d\eta = d\rho/(1-\rho^2)$ and $\cosh^2\eta = 1/(1-\rho^2)$ give the density $(1-\rho^2)^{x-3/2}$. So $(1+\rho)/2 \sim \mathrm{Beta}(x - \frac{1}{2}, x - \frac{1}{2})$ and $\mathbb{E}\rho^2 = \mathrm{Var}\,\rho = 1/(2x)$. Also f is the $\mathrm{Gamma}(x, c)$ law of $Q = e^Y$, so $\mathbb{E}[K_0 Q \rho^2] = 2c \cdot (x/c) \cdot 1/(2x) = 1$.

(c) Let $V = \log\cosh\eta \ge 0$. Itô's formula gives $V_T - V_0 = \frac{1}{2}T - \frac{1}{2}\int_0^T K_0 Q\rho^2\,dt + M_T$, where $M_T = \int_0^T \rho\,dB_2$ has $\langle M\rangle_T \le T$, so $M_T/T \to 0$ almost surely.

Case $x > \frac{1}{2}$. The time average of $K_0 Q\rho^2$ converges by (b), so V_T/T converges to a constant. That constant is 0, because $V \ge 0$ and V returns to bounded sets at arbitrarily late times (recurrence). Hence the limit is 1, which also re-derives $\mathbb{E}_\mu[K_0 Q\rho^2] = 1$.

Case $x < \frac{1}{2}$. Y alone is a positive recurrent one-dimensional diffusion with invariant law $\mathrm{Gamma}(x, c)$, so $\frac{1}{T}\int_0^T K_0 Q\,dt \to K_0 x/c = 2x$. Since $\rho^2 \le 1$, $\liminf V_T/T \ge \frac{1}{2} - x > 0$. Hence $1 - \rho_t^2 = \mathrm{sech}^2\,\eta_t \to 0$, and $\frac{1}{T}\int_0^T K_0 Q(1-\rho^2)\,dt \le \sup_{t \ge t_0}(1-\rho_t^2) \cdot \frac{1}{T}\int_{t_0}^T K_0 Q\,dt + o(1) \to 0$ on letting $T \to \infty$ and then $t_0 \to \infty$. So the time average of $K_0 Q\rho^2$ tends to $2x$.

The final sentence follows because $\frac{1}{T}\int_0^T (x - cQ)\,dt \to 0$ by the ergodicity of Y. □

The chain against the diffusion. Part (a) matches generators only for $|\rho| < 1$ bounded away from 1. At the corner the chain has a term the diffusion lacks. The nonlinearity bends state and tangent apart, so from $\rho = \pm 1$ one step gives $\mathbb{E}[1 - \rho'^2] = 24a^2q^2(1 + o(1))$. The measured value is $2.63\,Q^2/F^2$ at $F = 1024$, against $24a^2 = 8/3$ for $\tanh$. In η this is a drift $-\mathrm{sgn}(\eta)\,12a^2Q^2\cosh^2\eta/F$, which becomes $O(1)$ at $|\eta| \approx \frac{1}{2}\log F$: a wall that recedes only logarithmically.

Adding it to equation 14 reproduces the chain's alignment statistics. At $F = 1024$ the chain's $\mathbb{E}\rho^2$ at $x = 0.25$, 0.6, 1 and 2 is 0.916, 0.698, 0.491 and 0.249; the corrected diffusion gives 0.918, 0.705, 0.488 and 0.250. The time spent at $\rho^2 > 0.99$ agrees within 0.025 at $F = 64$, 256 and 1024. This is why the chain's cost sits below $c(x)$ for $x \le 1$ at every width tested. Heuristically, for $x > \frac{1}{2}$ the wall removes mass of order $F^{-(x-1/2)}$ from the Beta law, so convergence is slowest near $x = \frac{1}{2}$.

Checks (`theory/r5_diffusion.py`):

- On the exact chain, the coefficients in (a) match at $F = 256$ and 1024. The errors fall roughly as $1/F$: at $Q = 3$ the ΔY drift is off by 0.17 and then 0.04.
- Simulating equation 14 gives the cost within 0.02 of $\min(1, 2x)$ at ten values of $x \in [0.1, 4]$, and $\mathbb{E}\rho^2 = 1/(2x)$ within 0.005 for $x \geq 1$.

Lemma 9 (one-step localization)**.** *Under the structural hypotheses of Theorem 3, fix $x > 0$ and $R \geq 1$, and use the coordinates $(Y, \eta) = (\log Fq, \operatorname{artanh}\rho)$. Let $K = [-R, R]^2$ and $K' = [-R-1, R+1] \times [-R-2, R+2]$. There are $c > 0$ and C_R such that, for $F \geq F_0(R)$, $\sup_{z \notin K'} \mathbb{P}_z$(the next state lies in K) $\leq C_R e^{-c\sqrt{F}}$.*

Proof. Write $\Delta Y = \log(q'/q)$ and $D(q) = \mathbb{E}[\Delta Y \mid q] = x/F - \ell(q) + O(q/F)$. Fix $q_1 > 0$ with $|D(q)| \leq \frac{1}{4}$ for $q \leq q_1$. For $z \notin K'$ there are three cases.

(1) $q \leq q_1$ and $|Y| > R + 1$. Reaching K needs $|\Delta Y| \geq 1$, hence $|\Delta Y - D(q)| \geq \frac{3}{4}$. By the cumulant bound after Lemma 5 and Chernoff at $\lambda = \pm cF$, this has probability at most $2e^{-cF}$.

(2) $q > q_1$. Then $Y > R + 1$ for $F \geq F_0$, and reaching K needs $q' \leq e^R/F$, that is, $N = A/q \leq \varepsilon s$ with $\varepsilon = e^R/(gF\Phi(q)) \leq e^R/(F\Phi(q_1))$. The small-ball bound after Lemma 5 gives probability at most $(C_2 e^R/(F\Phi(q_1)))^{F/4} \leq e^{-F}$.

(3) $|Y| \leq R + 1$ and $|\eta| > R + 2$. Then $q \leq e^{R+1}/F$ and $u < u_1 := \operatorname{sech}^2(R + 2)$, and reaching K needs $u' \geq u_2 := \operatorname{sech}^2 R$, where $u_2/u_1 = \cosh^2(R + 2)/\cosh^2 R \geq \cosh^2 2 > 14$.

In the notation of Lemma 6, in law given (x, y_s),

$$1 - \rho'^2 \;=\; \frac{\|m + \sqrt{u}Z_0\|^2}{y_s^2 + \|m + \sqrt{u}Z_0\|^2} \;\leq\; \frac{2u\|Z_0\|^2 + 2\|m\|^2}{y_s^2},$$

with $\|m\| \leq Cq\|x^{\circ 3}\| + Cq \max_a x_a^2 |y_s - \mathbb{E}[y_s \mid x]|$ (the factor q multiplies both terms). This bound on $\|m\|$ holds because $\|PDx\| = \|P(\phi'(z)z - \phi(z))\|/\sqrt{q} \leq C\|z^{\circ 3}\|/\sqrt{q}$ and $\|\kappa\| \leq C \max_a |1 - \phi'(z_a)^2| \leq Cq \max_a x_a^2$ on G_x.

Let $G'_x = G_x \cap \{0.95F \leq \|x\|^2 \leq 2F,\ \max_a x_a^4 \leq F\}$, which has $\mathbb{P}(G'^c_x) \leq Ce^{-c\sqrt{F}}$ (the maximum by a union bound over a). On G'_x, $\|x^{\circ 3}\|^2 \leq \max_a x_a^4 \|x\|^2 \leq 2F^2$, so $\|m\|^2 \leq Cq^2F^2 = CQ^2 \leq Ce^{2R+2}$ whenever $|y_s - \mathbb{E}[y_s \mid x]| \leq \sqrt{F}$. Also on G'_x:

- $\langle \hat{X}, Dx\rangle^2 = \|x\|^2(1 + O(q \max_a x_a^2)) \geq 0.9F$, so $\mu_s^2 \geq (1 - u_1)0.9F \geq 0.8F$.
- Since $\mathbb{E}[y_s \mid x] = \mu_s$ and $\operatorname{Var}(y_s \mid x) \leq 1$, $|y_s - \mu_s| \leq 0.1\sqrt{F}$ except on a conditional event of probability at most $2e^{-F/200}$. On that event $|y_s| \geq (\sqrt{0.8} - 0.1)\sqrt{F}$, so $y_s^2 \geq 0.63F$, and $|y_s - \mathbb{E}[y_s \mid x]| \leq \sqrt{F}$. Since $\operatorname{tr}\Sigma_0 \leq F$, $\|Z_0\|^2 \leq 1.2F$ except with probability e^{-cF} (Laurent & Massart, 2000, Lemma 1).
- Then $2u\|Z_0\|^2/y_s^2 \leq 4u < 4u_1 < u_2/2$ and $2\|m\|^2/y_s^2 \leq Ce^{2R}/F < u_2/2$ for $F \geq F_0(R)$, so $u' < u_2$.

So in all three cases the probability is at most $C_R e^{-c\sqrt{F}}$. □

Theorem 5 (the chain converges to the diffusion in the layer)**.** *Under the structural hypotheses of Theorem 3, fix $x > 0$ and put $\nu = x/F$. Let π_F be the chain's stationary law and $c_F(x) = F\,\mathbb{E}_{\pi_F}[\rho^2 qK(q)]$. (a) $c_F(x) \leq \min(1, 2x) + C(1 + x^2)/F$ for every F. (b) If $x > \frac{1}{2}$, the law of (Fq, ρ) under π_F converges weakly to the invariant law μ_x of Proposition 7(b), and $c_F(x) \to 1$. (c) If $x < \frac{1}{2}$, $\mathbb{E}_{\pi_F}[Fq(1 - \rho^2)] \to 0$ and $c_F(x) \to 2x$. Hence $c_F(x) \to c(x) = \min(1, 2x)$ and $2F\chi^\star \to -c(x)$ for every $x \neq \frac{1}{2}$; Proposition 8 adds $x = \frac{1}{2}$.*

Proof. Write $Q = Fq$, $u = 1 - \rho^2$, $c = 6|a|$ and $K_0 = 2c$. Four facts are used throughout.

(F1) Step 4(i) gives $\mathbb{E}_{\pi_F} Q^k \leq C_k(1 + x)^k$, hence $\pi_F(Q > \varepsilon F) \leq C_{k,\varepsilon} F^{-k}$ for every k. Stationarity of $\log q$, as in Lemma 8(a), gives $c\,\mathbb{E}_{\pi_F} Q = x + O((1 + x)^2/F)$.

(F2) For $Q \le \varepsilon F$, $\Delta Y = \log(q'/q)$ has $F\,\mathbb{E}[\Delta Y \mid q] = x - cQ + O((1+Q+Q^2)/F)$ (Step 3, as in Lemma 8(b)(ii)), $F\,\mathbb{E}[(\Delta Y)^2 \mid q] = 2 + O(Q/F + F^{-1})$ and $\mathbb{E}[|\Delta Y|^3 \mid q] = O(F^{-3/2})$.

(F3) Lemma 6: the centred step, its drift and its cumulant bounds.

(F4) *Drift lemma.* If $V \ge 0$ and $F(PV - V) \le -\kappa V + b$ with b bounded for each F, then $\pi_F(V) \le \pi_F(b)/\kappa$. Iterate from a fixed state, average over the first n steps, and let $n \to \infty$ using $P^n \to \pi_F$ in total variation (Meyn & Tweedie, 2009, Thm. 13.0.1) applied to $V \wedge K$ and b, then $K \to \infty$, as in Lemma 8(b).

In (b) and (c), "first order is exact under Cauchy–Schwarz" means the following. If $\log A$ and $\log B$ satisfy the cumulant bound of (F3) at scale $\lambda \le 2\theta$, then $\log \mathbb{E}[AB] \le \mathbb{E}\log A + \mathbb{E}\log B + C\theta^2/F$.

(a). Corollary 1 with $F\nu^2 = x^2/F$ gives $1 + C(1+x^2)/F$. Also $\rho^2 \le 1$ and $K(q) \le K_0 + Cq$ give $c_F \le K_0\,\mathbb{E}Q + C\,\mathbb{E}Q^2/F = 2x + O((1+x)^2/F)$ by (F1).

(b), tightness. Let $\theta \in (0,1)$ and $M \ge 1$, put $w = M/(M+Q)$, $h(Q) = (Q^{-1} + M^{-1})^\theta$ and $W = u^{-\theta/2}h(Q)$. W is non-increasing in u, so by (F3) $PW \le \mathbb{E}[\tilde{u}'^{-\theta/2}h(Q')]$. In $Y = \log Q$, $\partial_Y \log h = -\theta w$ and $\partial_Y^2 \log h = \theta w(1-w)$. Cauchy–Schwarz with (F2) and (F3) then gives, for $Q \le \varepsilon F$,

$$F \log \frac{PW}{W} \;\le\; \theta\big[\Gamma_0(Q) + cQu\big] + C\theta^2 + C\theta\varepsilon(1+Q) + \frac{C}{F},$$
$$\Gamma_0(Q) = \tfrac{1}{2} - wx - cQ(1-w) + w(1-w).$$

Here $\frac{1}{2}(1 - K_0 Q\rho^2) = \frac{1}{2} - cQ + cQu$ comes from the centred step, and $-w(x - cQ) + w(1-w)$ from h.

For $M \ge M_0(x,c)$, $\Gamma_0(Q) \le -\kappa_0(1+Q)$ for all $Q \ge 0$. To see this, note $(M+Q)^2\Gamma_0 = (\frac{1}{2} - x)M^2 + (2-x)MQ + (\frac{1}{2} - cM)Q^2 - cQ^3$. Bound the cross term by $(2-x)MQ \le \frac{1}{2}(x - \frac{1}{2})M^2 + \frac{(2-x)^2}{2(x-1/2)}Q^2$, and take $cM \ge 1 + \frac{(2-x)^2}{2(x-1/2)}$; this uses $x > \frac{1}{2}$.

Take $\theta, \varepsilon \le \kappa_0/(4C)$ and $\delta_1 = \kappa_0/(4c)$. Then:

- on $\{u \le \delta_1\} \cup \{Q \le \delta_1\}$, with $Q \le \varepsilon F$: $F(PW - W) \le -\frac{1}{8}\theta\kappa_0 W$ for $F \ge F_0$;
- on $\{u > \delta_1, \delta_1 < Q \le \varepsilon F\}$: $W \le \bar{W} := \delta_1^{-\theta/2}(\delta_1^{-1} + 1)^\theta$, and $F(PW - W) \le C(1+Q)\bar{W}$;
- on $\{Q > \varepsilon F\}$, if $\rho^2 \ge \frac{1}{2}$: the general-q drift of (F3), $\mathbb{E}\log(\tilde{u}'/u) \ge (K/S)q\rho^2 - C/F \ge c'\varepsilon$, with the cumulant bound, contracts $u^{-\theta/2}$ by a fixed factor per step. Meanwhile $h(Q')/h(Q) \le 1 + C/F$ except on an event of probability e^{-cF} (where $Q' < c''Q$). So $F(PW - W) \le -\kappa W$.
- on $\{Q > \varepsilon F\}$, if $\rho^2 < \frac{1}{2}$: $W \le C$ and $F(PW - W) \le CF$.

So $F(PW - W) \le -\kappa W + b$ with $b = C(1+Q)\bar{W} + CF\,\mathbf{1}\{Q > \varepsilon F\}$, and $\sup_F \pi_F(b) < \infty$ by (F1). By (F4), $\sup_F \pi_F(W) < \infty$. Since $W \ge M^{-\theta}u^{-\theta/2}$ and $W \ge Q^{-\theta}$, and $\mathbb{E}_{\pi_F}Q^2$ is bounded, the laws of $(Y, \eta) = (\log Q, \operatorname{artanh}\rho)$ under π_F are tight in $\mathbb{R}^2$, and $Q\rho^2$ is uniformly integrable.

(b), identification. Let ν_∞ be a subsequential weak limit, and let $f \in C_c^\infty(\mathbb{R}^2)$ in the coordinates (Y, η). Stationarity gives $\mathbb{E}_{\pi_F}[F(P_F f - f)] = 0$.

- On compact sets, $F(P_F f - f) \to \mathcal{L}f$ uniformly, by Proposition 7(a) together with third absolute moments $O(F^{-3/2})$ there.
- Take R with $\operatorname{supp} f \subset K = [-R, R]^2$. For $z \notin K'$, $f(z) = \mathcal{L}f(z) = 0$ and $|F(P_F f)(z)| \le F\|f\|_\infty C_R e^{-c\sqrt{F}} \to 0$ uniformly, by Lemma 9. The compact set K' is covered by the previous item.

Hence, splitting $0 = \mathbb{E}_{\pi_F}[F(P_F f - f)]$ over K' and its complement: on K' the integrand is $\mathcal{L}f + e_F$ with $\sup_{K'} |e_F| \to 0$; off K' it is bounded by $F\|f\|_\infty C_R e^{-c\sqrt{F}}$, uniformly in the state. So $\mathbb{E}_{\pi_F}[\mathcal{L}f] \to 0$. Since $\mathcal{L}f$ is bounded and continuous, weak convergence along the subsequence gives $\int \mathcal{L}f \, d\nu_\infty = 0$ for all such f. The martingale problem for $\mathcal{L}$ is well posed (locally Lipschitz coefficients, no explosion; Proposition 7(b)), so ν_∞ is invariant (Ethier & Kurtz, 1986, Thm. 4.9.17). Uniqueness (Proposition 7(b)) gives $\nu_\infty = \mu_x$, so the whole family converges. Finally, $c_F(x) = \mathbb{E}_{\pi_F}[K(Q/F)Q\rho^2]$, with K bounded and continuous, $K(q) \to K_0$ as $q \to 0$, and $Q\rho^2$ uniformly integrable. So $c_F(x) \to \mathbb{E}_{\mu_x}[K_0 Q\rho^2] = 1$.

(c). Put $\varepsilon_F = F^{-1/2}$, choose $0 < \theta < \gamma < \theta/(2x)$ (possible because $2x < 1$), and set $V = (u + \varepsilon_F)^{\theta/2} Q^\gamma$. Now V is non-decreasing in u, so the injection must be bounded rather than dropped. With f as in Lemma 6, $G(t) = (f(t) + \varepsilon_F)^{\theta/2}$ is concave and non-decreasing. Given (x, y_s), $G(\|m + \sqrt{u}Z_0\|^2) \le G(u\|Z_0\|^2) + G'(u\|Z_0\|^2)(2\sqrt{u}\langle m, Z_0\rangle + \|m\|^2)$. The cross term has conditional mean 0 by $Z_0 \mapsto -Z_0$, and $G'(t) \le \frac{\theta}{2}\varepsilon_F^{\theta/2-1}/(y_s^2 + t)$. So the injection adds at most $C\theta\, \varepsilon_F^{\theta/2-1} q^2 Q^\gamma$ to PV, which is $O(\theta Q^{2+\gamma} F^{-1/2-\theta/4})$ after multiplying by F.

For the centred part, $(\tilde{u}' + \varepsilon_F)/(u + \varepsilon_F) = 1 + \lambda(R - 1)$ with $R = \tilde{u}'/u$ and $\lambda = u/(u + \varepsilon_F)$. Since $\log(1+\lambda t) - \lambda \log(1+t) = \frac{1}{2}\lambda(1-\lambda)t^2 + O(|t|^3)$, the cumulant bound gives $F\,\mathbb{E}\log(1+\lambda(R-1)) \le \lambda F\,\mathbb{E}\log R + C(1-\lambda)$.

- On $\{u \ge K\varepsilon_F, Q \le \varepsilon F\}$, where $1 - \lambda \le 1/K$ and $\lambda(K_0 Q\rho^2 - 1) \le K_0 Q\rho^2 - 1 + 1/K$:

$$F \log \frac{PV}{V} \;\le\; \gamma x - \tfrac{\theta}{2} - cQ(\gamma - \theta\rho^2) + \frac{C\theta}{K} + C(\theta + \gamma)^2 + C\varepsilon(\theta + \gamma)(1 + Q) + o(1),$$

  which is at most $-\frac{1}{2}(\frac{\theta}{2} - \gamma x) - \frac{1}{2}c(\gamma - \theta)Q$ once K is large and θ, ε are small (the ratio γ/θ held fixed).

- On $\{u < K\varepsilon_F\}$: $V \le ((K+1)\varepsilon_F)^{\theta/2} Q^\gamma$ and $F(PV - V) \le C(1 + Q)V$ + injection.

- On $\{Q > \varepsilon F\}$: $F(PV - V) \le CFQ^\gamma$.

By (F4) and (F1), $\pi_F(V) \le C(\varepsilon_F^{\theta/2} + F^{-1/2}) \to 0$, and likewise for V^2 (replace (θ, γ) by $(2\theta, 2\gamma)$). As $u \le (u+\varepsilon_F)^{\theta/2}$, $\mathbb{E}_{\pi_F}[Qu] \le (\pi_F V^2)^{1/2}(\mathbb{E}_{\pi_F} Q^{2-2\gamma})^{1/2} \to 0$, and $c_F \ge K_0\,\mathbb{E}Q - K_0\,\mathbb{E}[Qu] - C\,\mathbb{E}Q^2/F \to 2x$ by (F1).

The last sentence follows from Proposition 6's identity $2F\chi^\star = F\nu^2/3 - c_F(x) + o(1)$ with $F\nu^2 = x^2/F$. □

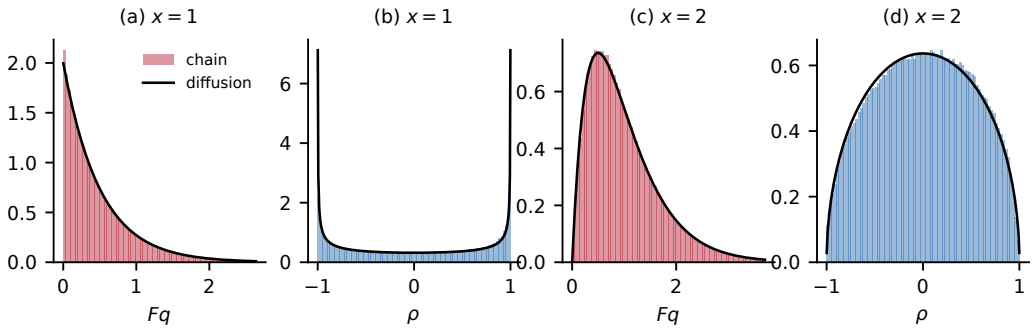


Figure 5: Theorem 5(b) on the exact chain (tanh, $F = 1024$, 16 seeds, 8.2×10^5 stationary samples per panel). Histograms of Fq and ρ in the boundary layer $\nu = x/F$ against the invariant law of the limiting diffusion (Proposition 7(b)): $Fq \sim \mathrm{Gamma}(x, 2)$ (panels a, c) and $(1+\rho)/2 \sim \mathrm{Beta}(x - \frac{1}{2}, x - \frac{1}{2})$ (panels b, d). The sample means are $\mathbb{E}Fq = 0.499, 1.000$ (theory 0.5, 1) and $\mathbb{E}\rho^2 = 0.481, 0.249$ (theory 0.5, 0.25) at $x = 1, 2$ (`theory/fig_chain_vs_diffusion.py`).

Remark (irregular graphs, measured). Near $g_{\mathrm{lin}}(F)$ the state and tangent of a graph network align with the top eigenvector v of A, and to cubic order the network is the MLP with cubic coefficient $a\pi_4$, $\pi_4 = \sum_i v_i^4$: the state variance is rescaled by $1/\pi_4$ and every scale-free statistic is unchanged. A registered replication (E3b; Appendix J) finds the boundary-layer $2F\chi^\star$ and $\mathbb{E}\rho^2$ of six irregular graphs (π_4 varying threefold, $d_{\max}/\bar{d}$ up to 4.7) equal to the exact chain's within ± 0.1 at $F = 256$, and the state variance $\propto 1/\pi_4$. It fails where non-top modes relax slower than F layers. This is a measured regularity, not a theorem.

Proposition 8 (the critical point $x = \frac{1}{2}$)**.** *Under the hypotheses of Theorem 5, at $x = \frac{1}{2}$,*

$$1 - \frac{C}{\log F} \;\le\; c_F(\tfrac{1}{2}) \;\le\; 1 + \frac{C}{F}.$$

Hence $c_F(x) \to \min(1, 2x)$ and $2F\chi^\star \to -\min(1, 2x)$ for every $x > 0$, and at $x = \frac{1}{2}$ the convergence is at least logarithmic.

Proof. Write $u = 1 - \rho^2$, $Q = Fq$, $Y = \log Q$, $c = 6|a|$, $K_0 = 2c$, and $W = \frac{1}{2}\log(1/u) - Y$, the chain's analogue of $\log\cosh\eta - Y$ in Proposition 7. Let $\tilde{u}'$ be the centred step of Lemma 6, and define the *injection*

$$\iota(q, \rho) = \mathbb{E}\big[\log(1 - \rho'^2) - \log\tilde{u}' \;\big|\; q, \rho\big] \;\ge\; 0,$$

which is non-negative by Anderson's inequality. (F1)–(F4) are those of the proof of Theorem 5.

Step 1: the identity. By Lemma 6, $\mathbb{E}\log(\tilde{u}'/u) = \psi(\frac{F-1}{2}) - \psi(\frac{F}{2}) + K_0 q\rho^2 + e_u$ with $|e_u| \le C(q^2 + q/F + F^{-2})$. By Step 3, $\log u$ is π_F-integrable, so stationarity gives $\mathbb{E}_{\pi_F}[\log u' - \log u] = 0$. Hence

$$FK_0\,\mathbb{E}_{\pi_F}[q\rho^2] = F\big(\psi(\tfrac{F}{2}) - \psi(\tfrac{F-1}{2})\big) - F\,\mathbb{E}_{\pi_F}\iota - F\,\mathbb{E}_{\pi_F}e_u.$$

With $K(q) = K_0 + O(q)$ and (F1) this gives $c_F(\frac{1}{2}) = 1 - F\,\mathbb{E}_{\pi_F}\iota + O(1/F)$. The upper bound follows from $\iota \ge 0$. It remains to show $F\,\mathbb{E}_{\pi_F}\iota \le C/\log F$.

Step 2: the injection is local. Couple $1 - \rho'^2 = f(\|m + \sqrt{u}Z_0\|^2)$ and $\tilde{u}' = f(u\|Z_0\|^2)$ through the same Z_0, with $f(t) = t/(y_s^2 + t)$ (Lemma 6). The function $\log f$ is concave with $(\log f)'(t) \le 1/t$, and the cross term $2\sqrt{u}\langle m, Z_0\rangle$ is odd in Z_0. So $\iota \le \mathbb{E}[\|m\|^2/(u\|Z_0\|^2)] \le Cq^2/u$, using the bound on $\|m\|$ in the proof of Lemma 9 and the negative moments of $\|Z_0\|^2$. This holds up to a term that is superpolynomially small in F times a polynomial in $\log(1/u)$, from the complement of the good event, which Step 3 makes π_F-negligible. Therefore

$$F\iota \;\le\; \frac{CQ^2}{Fu} \;=\; \frac{CQ^4 e^{2W}}{F}.$$

Step 3: a floor for u. Let $G_x'' = G_x' \cap \{|\frac{1}{F}\sum_a x_a^6 - 15| \le 1,\ |\frac{1}{F}\sum_a x_a^4 - 3| \le \frac{1}{4}\}$, with G_x' from Lemma 9. Truncating each x_a^6 at $F^{2/3}$ and applying Bernstein's inequality gives $\mathbb{P}(G_x''^c) \le Ce^{-cF^\alpha}$ with $\alpha = 2/9$; every use below needs only superpolynomial decay. From any state, one step gives $\log(1/u') \le 2\log(1/q) + C + R$, except on an event E of probability at most $\varepsilon_F := Ce^{-cF^\alpha}$, on which $\log(1/u') \le \log(1/u) + R$. Here R has exponential moments uniformly in the state. There are two cases.

- If $u \ge c_1 q^2$ (with $c_1 > 0$ small, fixed below): Anderson gives $1 - \rho'^2 \ge_{\mathrm{st}} \tilde{u}' \ge u\|Z_0\|^2/\|Y\|^2$, with $\|Z_0\|^2 \ge cF$ and $\|Y\|^2 \le CF$ outside exponentially small events.
- If $u < c_1 q^2$: $\rho^2 \ge \frac{1}{2}$. By Lemma 6, $m = \rho PDx + \kappa(y_s - \mu_s)$ with $\mu_s = \mathbb{E}[y_s \mid x] = \rho\langle\hat{X}, Dx\rangle$. Three ingredients, each with its source. (i) *The main term, from the moment event G_x'':* $\|PDx\| = \|P(\phi'(z)z - \phi(z))\|/\sqrt{q} \ge 3c_m q\sqrt{F}$, because $\|P(\phi'(z)z - \phi(z))\| \ge c\|P_{z^\perp}z^{\circ 3}\|$ and $\|P_{z^\perp}z^{\circ 3}\|^2/(q^3F) = \frac{1}{F}\sum x^6 - (\frac{1}{F}\sum x^4)^2/(\frac{1}{F}\sum x^2) \ge 14 - 3.25^2/0.95 > 2$. This is for $q \le q_0$, where $\phi'(z)z - \phi(z) = 2az^3 + O(z^5)$. For $q_0 < q \le q_{\max}$, $\|P(\phi'(z)z)\|^2/F$ concentrates, outside an exponentially small event, at $\mathbb{E}(\phi'(z)z)^2 - (\mathbb{E}\phi'(z)z\phi(z))^2/\mathbb{E}\phi(z)^2$. That is positive (Cauchy–Schwarz is strict because $\phi'(z)z$ is not a multiple of $\phi(z)$) and continuous in q, so it is bounded below on $[q_0, q_{\max}]$ and the same bound holds with a smaller c_m. (ii) *The coefficient of the y_s-term, from the same moment event:* $\|\kappa\| \le C_\kappa q$, because $PD^2\hat{X} = P(D^2 - I)\hat{X}$, $|1 - \phi'(z_a)^2| \le Cz_a^2$ and

$\|\phi(z)\| \ge c\sqrt{q}\|x\|$ give $\|\kappa\| \le Cq\|x^{\circ 3}\|/\|x\| \le Cq\sqrt{16/0.95}$. (The bound $Cq\max_a x_a^2$ used in Lemma 9 is too weak here.) (iii) *The y_s-term itself, from Gaussian concentration given x:* since $\mathrm{Var}(y_s \mid x) \le u \le 1$, $|y_s - \mu_s| \le c_*\sqrt{F}$ with $c_* = c_m/C_\kappa$, except with conditional probability $2e^{-c_*^2F/2}$. On that event $\|m\| \ge 3c_mq\sqrt{F}/\sqrt{2} - c_mq\sqrt{F} \ge c_mq\sqrt{F}$. With $c_1 = c_m^2/16$, $\sqrt{u}\|Z_0\| \le \|m\|/2$ whenever $\|Z_0\|^2 \le 2F$, which fails only with exponentially small probability. So $\|m + \sqrt{u}Z_0\| \ge \|m\|/2$, and $1 - \rho'^2 \ge \|m\|^2/(4\|Y\|^2) \ge cq^2$.

Also $\log(1/q') \le \log(1/q) + R$ (Lemma 5).

The negative moment of Q, for every $x > 0$. Theorem 5(b) proves tightness with the joint function $u^{-\theta/2}h(Q)$ under $x > \frac{1}{2}$; here only its Q-part is needed, and it needs only $x > 0$. Let $h(Q) = (Q^{-1} + M^{-1})^\theta$ and $w = M/(M+Q)$, with $M \ge 1$. The computation there, without the factor $u^{-\theta/2}$ (which contributes the $\frac{1}{2} - cQ + cQu$ in Γ_0), gives for $Q \le \varepsilon F$

$$F\log\frac{Ph}{h} \;\le\; \theta\big[-w(x - cQ) + w(1-w)\big] + C\theta^2 + C\theta\varepsilon(1+Q) + \frac{C}{F}.$$

Since $w \ge 1 - Q/M$, the bracket is at most $-x + C_1Q$. So for $Q \le Q_0 := x/(4C_1)$, $\theta \le x/(4C)$, $\varepsilon \le x/(8C)$ and $F \ge F_0$, $F(Ph - h) \le -\frac{1}{4}\theta x h$. On $\{Q_0 < Q \le \varepsilon F\}$ the bracket is at most $cM + 1$ and $h \le (Q_0^{-1} + 1)^\theta$; the right side above is then at most $C(1+Q) \le C(1+\varepsilon F)$, so exponentiating gives $F(Ph - h) \le C(1+Q)$. On $\{Q > \varepsilon F\}$, $h \le 2$ and $Ph \le C$, because $\mathbb{E}[(q'/q)^{-\theta} \mid q] \le C$ by Lemma 5(b); so $F(Ph - h) \le CF$. Hence $F(Ph - h) \le -\kappa h + b$ with $b = C(1+Q) + CF\,\mathbf{1}\{Q > \varepsilon F\}$ and $\sup_F \pi_F(b) < \infty$ by (F1), and (F4) gives $\mathbb{E}_{\pi_F}Q^{-\theta} \le C$ for some $\theta = \theta(x) > 0$. With (F1), $\mathbb{E}_{\pi_F}|Y|^k \le C_k$ for every k.

The moments of $\log(1/u)$. Put $X = \log(1/u) \ge 0$ and $X_M = \min(X, M)$ for $M \ge 1$. By the one-step bound, $X'_M \le (2\log F - 2Y + C + R)_+ + \mathbf{1}_E(X_M + R_+)$. Raise to the power k, and use $(a+b)^k \le 2^{k-1}(a^k + b^k)$, $\mathbb{E}[\mathbf{1}_E X_M^k \mid \cdot] \le \varepsilon_F X_M^k$ (X_M is a function of the state), and Cauchy–Schwarz for $\mathbf{1}_E R_+^k$:

$$\mathbb{E}[X_M'^k \mid \cdot] \;\le\; C_k(1 + \log F + |Y|)^k + 4^k\varepsilon_F X_M^k + C_k.$$

X_M is bounded, so stationarity applies: $(1 - 4^k\varepsilon_F)\,\mathbb{E}_{\pi_F}X_M^k \le C_k(\log F)^k$, using $\mathbb{E}_{\pi_F}|Y|^k \le C_k$. For $F \ge F_k$, $4^k\varepsilon_F \le \frac{1}{2}$; letting $M \to \infty$ gives $\mathbb{E}_{\pi_F}[\log^k(1/u)] \le C_k(\log F)^k$. Since $|W| \le \frac{1}{2}X + |Y|$, also $\mathbb{E}_{\pi_F}|W|^k \le C_k(\log F)^k$; in particular W_+ has all moments at fixed F.

Step 4: a Huber function. Fix $\delta \in (0,1)$ and put $H = \frac{1}{2}(1-\delta)\log F$. Let $V(w) = \frac{1}{2}w^2$ on $[0, H]$, $V(w) = Hw - \frac{1}{2}H^2$ for $w > H$, and $V(w) = 0$ for $w < 0$. Then V is convex, $V' = \min(w_+, H)$, $V'' \le 1$, and $V \le HW_+$ is π_F-integrable by Step 3, so $\mathbb{E}_{\pi_F}[PV - V] = 0$. For every step,

$$V(W') - V(W) = V'(W)\,\Delta W + \int_W^{W'} \big(V'(t) - V'(W)\big)\,dt,$$

and the integral lies between 0 and $\frac{1}{2}(\Delta W)^2$. It vanishes unless the segment between W and W' meets $[0, H]$.

- *Drift.* For $q \le \varepsilon$, $F\,\mathbb{E}[\Delta W] = -\frac{1}{2}F\,\mathbb{E}[\Delta\log u] - F\,\mathbb{E}[\Delta Y]$. By Lemma 6 and (F2), with $x = \frac{1}{2}$ and $K_0 = 2c$,

$$F\,\mathbb{E}[\Delta W] \;=\; \tfrac{1}{2}(1 - K_0Q\rho^2) - (\tfrac{1}{2} - cQ) - \tfrac{1}{2}F\iota + e \;=\; cQu - \tfrac{1}{2}F\iota + e,$$
$$|e| \le C(1 + Q + Q^2)/F.$$

- *Second moments below the wall.* On $\{W \le H+1,\ Q \le F^{\delta/8}\}$ we have $1/u \le e^2F^{1-\delta}Q^2$, so $u/q^2 \ge e^{-2}F^{1+\delta/2}$. The injection there is negligible ($F\iota \le CF^{-\delta/2}$, and $|\log(1 - \rho'^2) - \log\tilde{u}'| \le Cq/\sqrt{u}$ outside exponentially small events). So $F\,\mathbb{E}[(\Delta W)^2] \le C(1+Q^2)$, by the cumulant bounds of Lemmas 5 and 6.

- *No crossing from above.* From $W > H+1$, the segment meets $[0,H]$ only if $\Delta W \le -1$. One step gives $1-\rho'^2 \le C(u+q^2)$ outside exponentially small events (from $\|m+\sqrt{u}Z_0\|^2 \le 2\|m\|^2+2u\|Z_0\|^2$ and $y_s^2 \ge cF$ once $\rho^2 \ge \frac{1}{2}$). So an injection lands at $W' \ge \log F - 2Y - C > H+1$ unless $Q > F^{\delta/8}$. Away from the corner, a drop of 1 is a large deviation, of probability e^{-cF}.

Hence $F(PV-V) \le V'(W)(cQu+|e|-\frac{1}{2}F\iota)+b$, where b is $C(1+Q^2)$ on $\{W \le H+1\}$ plus two kinds of term: $b_1 = F\,\mathbb{E}[|V(W')-V(W)| \mid \cdot]\mathbf{1}_B$ on the set of states $B = \{q > \varepsilon\} \cup \{Q > F^{\delta/8}\}$, and $b_2 = F\,\mathbb{E}[|V(W')-V(W)|\mathbf{1}_{E'} \mid \cdot]$ for a within-step large deviation E' of probability at most ε_F. Since $V \le HW_+$, Cauchy–Schwarz, Jensen and stationarity give $\mathbb{E}_{\pi_F} b_1 \le 2FH\,(\mathbb{E}_{\pi_F} W_+^2)^{1/2}\pi_F(B)^{1/2} \le CF(\log F)^2\,C_k F^{-k\delta/16}$, by (F1) and Markov's inequality for $\pi_F(Q > F^{\delta/8})$, and $\mathbb{E}_{\pi_F} b_2 \le 2FH\,(\mathbb{E}_{\pi_F} W_+^2)^{1/2}\varepsilon_F^{1/2}$. Both are bounded once k is taken large, so $\mathbb{E}_{\pi_F} b \le C$. Taking π_F-expectations,

$$\tfrac{1}{2}\,\mathbb{E}_{\pi_F}\big[V'(W)\,F\iota\big] \;\le\; \mathbb{E}_{\pi_F}[W_+\,cQu] + H\,\mathbb{E}_{\pi_F}|e| + C \;\le\; C,$$

since $W_+ cQu \le cQ\big(\frac{1}{2}u\log(1/u) + |Y|u\big) \le cQ(1+|\log Q|)$ and $H\,\mathbb{E}|e| \le C\log F/F$.

Step 5: conclusion. On $\{W \ge H\}$, $V'(W) = H$, so $F\,\mathbb{E}_{\pi_F}[\iota;\, W \ge H] \le 2C/H$. On $\{W < H\}$, Step 2 gives $F\iota \le CQ^4 e^{2H}/F = CQ^4 F^{-\delta}$. Together, $F\,\mathbb{E}_{\pi_F}\iota \le C/\log F$, and Step 1 finishes. □

Reading. At $x = \frac{1}{2}$ the drift of W away from the wall is $cQu \ge 0$: the diffusion is null recurrent in η, and the chain spreads up to the corner wall at height $\approx \frac{1}{2}\log F$. The cost deficit $1-c_F$ is the wall's share of time, which the Huber function shows is $O(1/\log F)$. On the chain (`theory/cf_half.py`; 64, 64, 32 and 32 seeds), $c_F(\frac{1}{2}) = 0.716 \pm 0.005$, 0.737 ± 0.007, 0.749 ± 0.011 and 0.782 ± 0.013 at $F = 64$, 256, 1024 and 4096, so $(1-c_F)\log F = 1.18$, 1.46, 1.74 and 1.82 ± 0.11. It rises and then levels off, as an $O(1/\log F)$ deficit would; a fit with a nonzero limit through the first three widths overshoots the $F = 4096$ deficit by about two standard errors. At $F = 256$ a four-times-longer run agrees (0.731 ± 0.007), and an earlier 8-seed value of 0.79 was a fluctuation. These numbers are consistent with the proposition but do not establish the rate; the proof does not use them. The exploratory monotonicity in x recorded there is no longer needed. For $x \ne \frac{1}{2}$ no rate is proved; heuristically it is $F^{-|x-1/2|}$ from the wall. At small F the $O(1/F)$ terms in (a) push the chain's cost slightly above 1 (1.06 at $F = 64$, $x = 2$), within the bound of (a).

C.9 Proof of Theorem 3

Step 1: an exact two-state chain. Take $n = 1$ first. Let $y^\ell = \phi(z^{\ell-1})$ be the state and w^ℓ the tangent after its nonlinearity. The next pre-activation and tangent increment are $z = M^\top y$ and $u = M^\top w$ for one matrix $M = \sigma_{\mathrm{s}} W_s + \sigma_{\mathrm{n}} d\,W_n$ with i.i.d. $\mathcal{N}(0, g/F)$ entries ($d = 2$ on the MLP ray). Given the past, the channel pairs (z_a, u_a) are therefore i.i.d. bivariate Gaussian with $\mathrm{Var}\, z_a = q := g\|y\|^2/F$ and correlation $\rho := \langle y, w\rangle/(\|y\|\|w\|)$. Write $z_a = \sqrt{q}\,x_a$ and $u_a \propto \zeta_a$ with (x_a, ζ_a) standard with correlation ρ, and set $A = \sum_a \phi(z_a)^2$, $B = \sum_a \phi'(z_a)^2\zeta_a^2$, $S_x = \sum_a x_a^2$, $S_\zeta = \sum_a \zeta_a^2$. Then $q' = gA/F$, $\|w'\|^2/\|w\|^2 = gB/F$, and $\rho' = \sum_a \phi(z_a)\phi'(z_a)\zeta_a/\sqrt{AB}$. So (q_ℓ) is an autonomous Markov chain, (q_ℓ, ρ_ℓ) is a Markov chain, and the tangent log-gain is a function of one transition. For a d-regular graph with node-uniform input, Step 1 of Appendix C.14 gives the same chain for the eigen-block $\lambda_k = d$. For $\lambda_k = -d$ (bipartite graphs) the increment uses $M_- = \sigma_{\mathrm{s}} W_s - \sigma_{\mathrm{n}} d\,W_n$, which has the same law as M but correlation $\beta = (\sigma_{\mathrm{s}}^2 - \sigma_{\mathrm{n}}^2 d^2)/g \in (-1, 1)$ with it; the chain is the same except $\rho' = \beta \sum_a \phi\phi'\zeta_a/\sqrt{AB}$. Blocks with $|\lambda_k| < d$ have gain $s_k^2 < g$ and are handled in Step 6.

Step 2: ergodicity and representation. Lemma 1 gives a unique stationary law π for the pair (q, ρ), by a two-step minorization; one step does not suffice, since at $\rho = \pm 1$ the sign of ρ' is forced. By the law of large numbers for positive Harris chains (Meyn & Tweedie, 2009, Thm. 17.0.1) and the integrability of $\log B$ (from $|\log \phi'(x)| \le C(1+x^2)$), $2\chi^\star = \mathbb{E}_\pi \log(gB/F)$ almost surely. Two terms of zero mean can be subtracted. $\mathbb{E}_\pi \log(q'/q) = 0$ by stationarity, since $\mathbb{E}_\pi|\log q| < \infty$; and $\mathbb{E}[\log(S_\zeta/S_x) \mid q, \rho] = 0$ because (x_a, ζ_a) is exchangeable. Hence

$$2\chi^\star = \mathbb{E}_\pi H(q,\rho), \qquad H = \mathbb{E}\big[\log(1-\bar{U}) - \log(1-\bar{V}) \;\big|\; q, \rho\big], \tag{15}$$

with $\bar{U} = \sum_a (\zeta_a^2/S_\zeta)(1-\phi'(z_a)^2)$ and $\bar{V} = \sum_a (x_a^2/S_x)(1-\psi(z_a)^2)$, two weighted averages of quantities in $[0,1]$ bounded by Cqx_a^2.

Step 3: one-step expansion. Exchangeability and independence of (x_1, ζ_1) from $\sum_{a\geq 2}\zeta_a^2 \sim \chi^2_{F-1}$ give $\mathbb{E}\bar{U} = \mathbb{E}[\zeta^2 f(x)] + O(q/F)$ for $f = 1-\phi'(\sqrt{q}\,\cdot)^2$, and $\mathbb{E}[\zeta^2 f(x)] = \mathbb{E}f + \rho^2\,\mathbb{E}[(x^2-1)f] = 1 - m(q) + \rho^2 qK(q)$, where Stein's lemma twice gives $\mathbb{E}[(x^2-1)f(x)] = \mathbb{E}f''(x) = qK(q)$, $K(q) = -\mathbb{E}[(\phi'^2)''(\sqrt{q}\,\xi)] = 12|a| + O(q)$. Likewise $\mathbb{E}\bar{V} = 1 - \Phi(q)/q + O(q/F)$, and both averages have conditional variance $O(q^2/F)$. Using $|\log(1-u) + u + u^2/2| \leq Cu^3$ for $u \leq 1/2$, and, for $u > 1/2$, the lower bound $\log(B/S_\zeta) \geq -2C\sum_a(\zeta_a^2/S_\zeta)(1+qx_a^2)$ from Jensen and the growth condition,

$$H(q,\rho) = \Lambda(q) - \rho^2 qK(q) + R, \qquad |R| \leq C\big(q^3 + q/F + \rho^2 q^2\big). \tag{16}$$

Here $\Lambda(q) = \log m(q) - \log(\Phi(q)/q)$, the ρ-free part; by equation 7 at variance q, $m(q) - \Phi(q)/q = \mathbb{E}[\xi^2\, q\,\psi'(\sqrt{q}\,\xi)^2] = 12a^2q^2 + O(q^3)$, so $\Lambda(q) = 12a^2q^2 + O(q^3)$. The same expansion applied to $\log(q'/q) = \log g + \log(A/(qS_x)) + \log(S_x/F)$, with $\mathbb{E}\log(S_x/F) = 2\gamma_F$, gives $\mathbb{E}[\log(q'/q) \mid q] = \nu - \ell(q) + O(q^3 + q/F)$, and directly $\mathbb{E}[q' \mid q] = g\Phi(q) = q\, g_{\rm lin}(F)\, e^{\nu-\ell(q)}$ and $\mathbb{E}[q'^k \mid q] = q^k g_{\rm lin}(F)^k e^{k(\nu-\ell(q))}(1 + k(k-1)/F + O(q/F + F^{-2}))$.

Step 4: stationary moments of q. Put $\nu = s/\sqrt{F}$ with $s \in [s_0, s_1]$. (i) *A priori bounds.* Since $\ell \geq c_0 q$, stationarity of q^k gives $\mathbb{E}_\pi[q^k(e^{k(\nu - c_0 q) + C_k/F} - 1)] \geq 0$. Split at $q_* = 2(\nu + C_k/F)/c_0$. For $q \geq q_*$, $k(\nu - c_0 q) + C_k/F \leq -kc_0 q/2$, so by convexity on $[0, q_{\max}]$ the bracket is at most $-cq$ with $c = (1 - e^{-kc_0 q_{\max}/2})/q_{\max} > 0$. (It is not bounded by a negative constant: at q_* the exponent is only $O(F^{-1/2})$.) For $q < q_*$ the bracket is at most $C(\nu + 1/F)$, and $\mathbb{E}_\pi[q^{k+1}\mathbf{1}\{q < q_*\}] \leq q_*\,\mathbb{E}_\pi q^k$. Together, $\mathbb{E}_\pi q^{k+1} \leq C(\nu + 1/F)\,\mathbb{E}_\pi q^k$, and induction from $\mathbb{E}_\pi q^0 = 1$ gives $\mathbb{E}_\pi q^k \leq C_k(\nu + 1/F)^k \leq C_k F^{-k/2}$, uniformly for $\nu\sqrt{F} \leq s_1$. (ii) *First moment.* Stationarity of $\log q$: $\mathbb{E}_\pi \ell(q) = \nu + O(F^{-3/2})$, so $6|a|\,\mathbb{E}_\pi q = \nu + O(F^{-1})$. (iii) *Second moment.* Stationarity of q: $\mathbb{E}_\pi[q(g_{\rm lin}e^{\nu-\ell} - 1)] = 0$ with $g_{\rm lin} = 1 + 1/F + O(F^{-2})$ gives $\mathbb{E}_\pi[q\ell] = (\nu + 1/F)\,\mathbb{E}_\pi q + O(F^{-3/2})$, so $36a^2\,\mathbb{E}_\pi q^2 = \nu^2 + O(F^{-3/2})$ (a ν/F term is of that order, and its coefficient is not fixed at this precision) and

$$\mathbb{E}_\pi \Lambda(q) = 12a^2\,\mathbb{E}_\pi q^2 + O(F^{-3/2}) = \nu^2/3 + O(F^{-3/2}). \tag{17}$$

Also $\mathrm{Var}_\pi q = O(F^{-3/2})$ while $(\mathbb{E}_\pi q)^2 \asymp F^{-1}$, so $\mathbb{P}_\pi(q < \mathbb{E}_\pi q/2) = O(F^{-1/2})$ by Chebyshev.

Step 5: stationary alignment. Let $r(q,\rho)$ be the population correlation of $(\phi(z), \phi'(z)\zeta)$. Stein's lemma gives $r = \rho\,[1 - \lambda(q)(1-\rho^2) + O(q^2)]$ with $\lambda(q) = 6|a|q + O(q^2)$, so $2\lambda(q) = qK(q) + O(q^2)$. The sample correlation of F i.i.d. zero-mean pairs satisfies $\mathbb{E}[\rho'^2 \mid q, \rho] = r^2 + (1-r^2)(1-2r^2)/F + O(q/F + F^{-2})$. The ratio is controlled because A/F and B/F are bounded below outside an event of probability e^{-cF} and $\rho'^2 \leq 1$. Stationarity of ρ^2 therefore gives

$$\mathbb{E}_\pi\big[\rho^2 q\,\big(\tfrac{2\lambda(q)}{q}(1-\rho^2) + O(q)\big)\big] = \tfrac{1}{F}\,\mathbb{E}_\pi[(1-\rho^2)(1-2\rho^2)] + O(F^{-3/2}).$$

The contraction vanishes at $\rho^2 = 1$, which is why this identity alone does not bound $\mathbb{E}_\pi[q\rho^2]$. The bound, and the exact constant, come from stationarity of $\log(1-\rho^2)$ instead, where the corner costs only its stationary mass (Proposition 5). Hence

$$\mathbb{E}_\pi[\rho^2 qK(q)] = F^{-1} + o(F^{-1}). \tag{18}$$

Step 6: conclusion. By equation 15–equation 18, with $\mathbb{E}_\pi|R| = O(\mathbb{E}_\pi q^3 + \mathbb{E}_\pi q/F + \mathbb{E}_\pi[\rho^2 q^2]) = o(F^{-1})$, $2\chi^\star = \nu^2/3 - 1/F + o(1/F)$ uniformly for $\nu\sqrt{F} \in [s_0, s_1]$. This is the non-bipartite claim. Blocks with $|\lambda_k| < d$ satisfy $2\chi_k^\star \leq \log(s_k^2/g) + \nu + \log\bar{m}$ by Theorem 4. Here $\log\bar{m} = \log\mathbb{E}_\pi m(q) = -\nu + O(F^{-1})$ by Step 4, so these blocks are negative and below the top block for large F. For the $\lambda = -d$ block of a bipartite graph, $\rho' = \beta \times$ (sample correlation) with $|\beta| \leq \bar{\beta} < 1$, so $\mathbb{E}_\pi\rho^2 \leq C/F$ without appeal to L3, and $\mathbb{E}_\pi[q\rho^2] \leq F^{-1/4}\mathbb{E}_\pi\rho^2 + \mathbb{E}_\pi[q\,\mathbf{1}\{q > F^{-1/4}\}] = O(F^{-5/4})$: the first term is $F^{-1/4}\cdot O(F^{-1})$, and $q\,\mathbf{1}\{q > F^{-1/4}\} \leq F^{3/4}q^4$ with $\mathbb{E}_\pi q^4 \leq C_4 F^{-2}$ (Step 4(i)) bounds the second by $O(F^{-5/4})$. That block therefore has $2\chi_-^\star = \nu^2/3 + o(1/F)$. The $\lambda = +d$ block satisfies $2\chi_+^\star \leq \nu^2/3 + O(F^{-3/2})$ without appeal to L3 either. In equation 16 the ρ-dependent part is $-\rho^2 q\,(K(q) - Cq) \leq 0$ for $q \leq q_0$, and $\mathbb{E}_\pi[q\,\mathbf{1}\{q > q_0\}] \leq q_0^{-3}\mathbb{E}_\pi q^4 = O(F^{-2})$ by Step 4(i). The maximum over blocks is $\nu^2/3 + o(1/F)$. □

Relaxation time. Near the crossing both the drift of $\log q$ (linearized, $-\ell'(q)q \approx -6|a|q$) and the contraction of ρ ($\lambda(q) \approx 6|a|q$) have rate $\approx \nu \approx \sqrt{3/F}$. The chain therefore relaxes over about $\sqrt{F/3}$ layers, and a finite-depth estimate that averages layers $L/2$ to L is biased unless $L \gg \sqrt{F}$.

Retrodiction: the slack of Theorem 4. To leading order, the slack $\chi^\star - \chi^\star_{\rm lin}(F) - \frac{1}{2}\log \bar{m}$ of Theorem 4 is half the alignment cost, $-1/(2F)$, near the crossing. This was noticed after the checks of Appendix C.14 had been run. For the MLP cells at $g = 1.2$, $-2F\times$slack is 0.98 ± 0.13 at $F = 32$ and 0.56 ± 0.36 at $F = 256$. Cells far from the window ($g \geq 1.6$) are not covered by the expansion.

Checks. The chain is exact in law. At $F \in \{16, 64, 256\}$ and four values of $\sigma_{\rm s}$ its finite-depth estimator reproduces the mean of the full-matrix estimator (12 comparisons, $\sum z^2 = 11.0$ on 12 degrees of freedom). Its stationary exponent gives $(\log g_\times - \log g_{\rm lin})\sqrt{F/3} = 1.150, 1.042, 1.018$ at $F = 64, 256, 1024$, with $F\,\mathbb{E}_\pi[\rho^2 q K(0)]$ at the predicted crossing $1.17, 1.06, 1.02$ (K frozen at $K(0)$; with $K(q)$ the cost is lower, see the check after Proposition 5). It gives the same law for softsign with $\kappa_1 = \pi/32$, although softsign is not C^5 and lies outside the theorem (`theory/reduced_chain.py`). The time the chain spends at $\rho^2 \geq 1/2$ is reported in `theory_out/reduced_chain_true.json`.

C.10 The next order, and a direct test

Direct test of the window equation. The crossing-level checks above test where $2\chi^\star$ changes sign; the residual $Y_F = 2F\chi^\star - (F\nu^2/3 - 1)$ tests the equation itself. On a grid of $x = \nu\sqrt{F} \in [0.4, 2.4]$ at $F = 64, 256, 1024$ (8 seeds), $\max_x |Y_F| = 0.53,\ 0.24,\ 0.12$: the equation holds uniformly over the window, and the residual halves with each fourfold increase in width. Scaled by $\sqrt{F}$ the curves collapse for $F \geq 256$ (Figure 6b; $\sqrt{F}\,Y_F = -3.03$ and -3.06 at $x = 2.25$), so the first correction is $O(F^{-1/2})$ with an x-dependent shape, largest at the upper edge of the window (`experiments/window_residual.py`). Its cubic part is Proposition 9 below: $F\,\mathbb{E}_\pi\Lambda = x^2/3 + (x/3 - 7x^3/18)F^{-1/2} + O(F^{-3/4})$ for $\tanh$. On the chain the measured moments reproduce $F\,\mathbb{E}_\pi\Lambda$ to 3×10^{-4}, and the formula's error falls from 0.022 to 0.003 between $F = 256$ and 2048 (`theory/tg_lambda_check.py`). This term tracks the curve across the window (-3.66 of the measured -3.05 at $x = 2.25$). What is left rises smoothly from about 0 to $+0.8$ across the window, the same at $F = 256$ and 1024 within their errors (± 0.2 and ± 0.4). It is the net of the alignment and remainder terms of equation 16. Each converges slowly ($1 - F\,\mathbb{E}_\pi[\rho^2 q K(q)] = 0.28, 0.15, 0.12$ at the crossing for $F = 256$, 1024, 2048), with coefficients near $\mp 5F^{-1/2}$ that nearly cancel, and their sum is not derived (`theory/tg_next_order.py`). At the crossing, $s_x = 1 + c_1 F^{-1/2} + o(F^{-1/2})$ with $c_1 = 0.72$ from Λ less 0.14–0.18 from the rest, so $c_1 \approx 0.56$. The chain gives $\sqrt{F}(s_x^2 - 1)/2 = 0.57 \pm 0.10$ at $F = 1024$ and 2048; a pure F^{-1} correction to s_x is rejected ($\chi^2 = 24.8$ on 7 degrees of freedom, $F \geq 48$).

Proposition 9 (the next-order Λ term, $\tanh$)**.** *In the setting of Theorem 3 with $\phi = \tanh$, uniformly for $\nu\sqrt{F} \in [s_0, s_1]$ and with $x = \nu\sqrt{F}$,*

$$\mathbb{E}_\pi q = \tfrac{\nu}{2} + \tfrac{11}{24}\nu^2 + O(F^{-3/2}), \qquad \mathrm{Var}_\pi q = \tfrac{\nu}{4F}\big(1 + O(F^{-1/4})\big),$$
$$F\,\mathbb{E}_\pi\Lambda(q) = \tfrac{x^2}{3} + \Big(\tfrac{x}{3} - \tfrac{7x^3}{18}\Big)F^{-1/2} + O(F^{-3/4}).$$

Proof. For $\tanh$, the Taylor series of $\tanh^2$ and sech^4 with Gaussian moments give $\ell(q) = 2q - \frac{11}{3}q^2 + 12q^3 + O(q^4)$ and $\Lambda(q) = \frac{4}{3}q^2 - 8q^3 + O(q^4)$. Write $m_1 = \mathbb{E}_\pi q$, $\delta = q - m_1$, $V = \mathrm{Var}_\pi q$, $M_4 = \mathbb{E}_\pi\delta^4$, and $G = \{q > m_1/2\}$. In the window $F\nu \geq \sqrt{F}s_0 \geq C_0 \log F$, so Lemma 8(b) gives $\pi(G^c) \leq e^{-cF\nu}$. Every term carrying $\mathbf{1}_{G^c}$ below is therefore $O(F^{-k})$ for every k and is dropped.

(i) *Mean.* Step 4(ii) gives $\mathbb{E}_\pi\ell(q) = \nu + O(F^{-3/2})$, that is, $2m_1 - \frac{11}{3}(m_1^2 + V) + O(\mathbb{E}_\pi q^3) = \nu + O(F^{-3/2})$. With $V = O(F^{-3/2})$ (Step 4), this gives $m_1 = \nu/2 + 11\nu^2/24 + O(F^{-3/2})$.

(ii) *Fourth moment.* The mean map $f(q) = \mathbb{E}[q' \mid q] = q\,g_{\rm lin}e^{\nu - \ell(q)}$ is exact. So $\delta' = f(q) - m_1 + \varepsilon$ with $\mathbb{E}[\varepsilon \mid q] = 0$, $\mathbb{E}[\varepsilon^2 \mid q] = (g^2/F)\mathrm{Var}_\xi \tanh^2(\sqrt{q}\,\xi) = (2q^2/F)(1 + O(q))$ and $\mathbb{E}[\varepsilon^4 \mid q] \leq Cq^4/F^2$.

- $f'(t) = 1 + \nu - 4t + O(t^2 + \nu^2 + F^{-1})$ for small t, and $f' \le 1 - c$ away from 0. So on G the average slope of f between q and m_1 is at most $1 - \nu/2 + O(\nu^2 + F^{-1})$.
- $\mathbb{E}_\pi f(q) = m_1$ and f'' is bounded, so $|f(m_1) - m_1| = O(V)$.

Hence on G, $|f(q) - m_1| \le \alpha|\delta| + CV$ with $\alpha = 1 - \nu/2 + O(\nu^2 + F^{-1})$. Stationarity of δ^4 gives $M_4 = \mathbb{E}(f - m_1)^4 + 6\,\mathbb{E}[(f - m_1)^2\varepsilon^2] + 4\,\mathbb{E}[(f - m_1)\varepsilon^3] + \mathbb{E}\varepsilon^4$, since the term linear in ε vanishes. The remaining terms are bounded using $\mathbb{E}|\delta|^k \le M_4^{k/4}$ for $k \le 4$, $\mathbb{E}\delta^2 = V$, $\mathbb{E}q^4 = O(F^{-2})$ and $\mathbb{E}\varepsilon^4 = O(F^{-4})$:

- $\mathbb{E}(f - m_1)^4 \le \mathbb{E}(\alpha|\delta| + CV)^4 \le \alpha^4 M_4 + C\big(M_4^{3/4}V + V^3 + M_4^{1/4}V^3 + V^4\big)$;
- $6\,\mathbb{E}[(f - m_1)^2\varepsilon^2] \le (C/F)(\mathbb{E}(f - m_1)^4)^{1/2}(\mathbb{E}q^4)^{1/2} \le CF^{-2}(M_4 + V^4)^{1/2}$;
- $4\,|\mathbb{E}[(f - m_1)\varepsilon^3]| \le 4(\mathbb{E}(f - m_1)^4)^{1/4}(\mathbb{E}\varepsilon^4)^{3/4} \le CF^{-3}(M_4 + V^4)^{1/4}$.

Since $1 - \alpha^4 \ge \nu$ for large F, $V = O(F^{-3/2})$ (Step 4) and $\nu \asymp F^{-1/2}$, compare νM_4 with each term in turn:

- $M_4^{3/4}V$ gives $M_4^{1/4} \le CV/\nu$, that is, $M_4 = O(F^{-4})$;
- V^3 and V^4 give $O(F^{-4})$, and $\mathbb{E}\varepsilon^4$ gives $O(F^{-7/2})$;
- $M_4^{1/4}V^3$ gives $O(F^{-16/3})$;
- $F^{-2}M_4^{1/2}$ gives $M_4 = O(F^{-4}/\nu^2) = O(F^{-3})$;
- $F^{-3}M_4^{1/4}$ gives $O(F^{-10/3})$.

So $M_4 = O(F^{-3})$, and the dominant constraint is the ε^2 cross term.

(iii) *Variance.* Write $f(q) - m_1 = f'(m_1)\delta + R$. Then $\mathbb{E}_\pi R = 0$ and $|R| \le C(\delta^2 + V)$. Stationarity of δ^2 gives $V = f'(m_1)^2 V + 2f'(m_1)\mathbb{E}[\delta R] + \mathbb{E}R^2 + \mathbb{E}\varepsilon^2$, with $|\mathbb{E}[\delta R]| \le C(VM_4)^{1/2} = O(F^{-9/4})$ and $\mathbb{E}R^2 = O(F^{-3})$. Also $\mathbb{E}\varepsilon^2 = (2/F)(m_1^2 + V)(1 + O(\nu)) = (\nu^2/(2F))(1 + O(F^{-1/2}))$. By (i), $\nu - \ell(m_1) = O(F^{-3/2})$, so $f'(m_1) = (1 + O(F^{-1}))(1 - m_1\ell'(m_1)) = 1 - \nu + O(\nu^2 + F^{-1})$ and $1 - f'(m_1)^2 = 2\nu(1 + O(F^{-1/2}))$. So $V = \nu^2/(2F)/(2\nu)\,(1 + O(F^{-1/2})) + O(F^{-9/4}/\nu) = (\nu/(4F))(1 + O(F^{-1/4}))$.

(iv) *The Λ term.* $\mathbb{E}_\pi\Lambda = \frac{4}{3}(m_1^2 + V) - 8(m_1^3 + 3m_1V + \mathbb{E}\delta^3) + O(\mathbb{E}q^4)$, with $\mathbb{E}\delta^3 = O((VM_4)^{1/2}) = O(F^{-9/4})$ and $\mathbb{E}q^4 = O(F^{-2})$. By (i)–(iii),

$$\mathbb{E}_\pi\Lambda = \tfrac{4}{3}\Big(\tfrac{\nu^2}{4} + \tfrac{11}{24}\nu^3 + \tfrac{\nu}{4F}\Big) - \nu^3 + O(F^{-7/4}) = \tfrac{\nu^2}{3} - \tfrac{7}{18}\nu^3 + \tfrac{\nu}{3F} + O(F^{-7/4}).$$

Multiplying by F with $\nu = x/\sqrt{F}$ gives the claim. □

Proposition 10 (ReLU: a simple root and an $O(1/F)$ shift)**.** *Let $\phi(x) = \max(x, 0)$, $\sigma_{\mathrm{b}} = 0$, and let the weights $W^\ell \in \mathbb{R}^{F\times F}$ have i.i.d. $\mathcal{N}(0, g/F)$ entries (an MLP). Condition on survival, i.e. at every layer at least one unit is active. Then there is an absolute constant C such that, almost surely, for every initial state and tangent,*

$$\log\frac{g}{2} - \frac{5}{2F} - \frac{C}{F^2} \;\le\; \liminf_{L\to\infty}\frac{1}{L}\log\frac{\|v^L\|^2}{\|v^0\|^2} \;\le\; \limsup_{L\to\infty}\frac{1}{L}\log\frac{\|v^L\|^2}{\|v^0\|^2} \;\le\; \log\frac{g}{2} - \frac{5}{2F} + \frac{C}{F^2}.$$

Hence the crossing sits at $\log g_\times = \log 2 + 5/(2F) + O(F^{-2})$. The mean-field gain $G(g) = g\,\mathbb{E}[\mathbf{1}\{\xi > 0\}] = g/2$ has a simple root at $g = 2$, and the shift is of order $1/F$, as the root-shift principle requires.

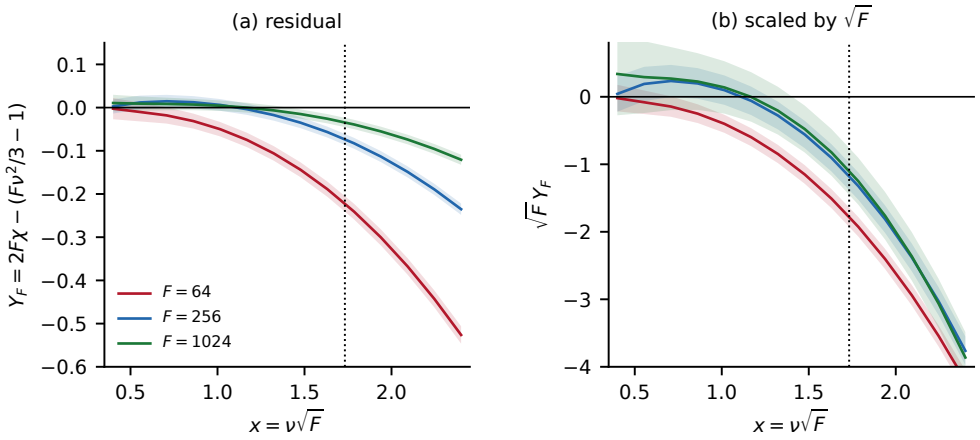


Figure 6: Direct test of the critical-window equation for the exact chain (tanh, ± 1 SE over 8 seeds). (a) The residual $Y_F = 2F\chi^\star - (F\nu^2/3 - 1)$ tends to zero uniformly over the window. (b) Scaled by $\sqrt{F}$, the curves collapse for $F \geq 256$: the first correction is $O(F^{-1/2})$. Dotted line: the predicted crossing $x = \sqrt{3}$.

Proof. Survival. Given the past, the pre-activations $W^\ell h^\ell$ are i.i.d. $\mathcal{N}(0, g\|h^\ell\|^2/F)$, so the number K of active units is $\mathrm{Bin}(F, \frac{1}{2})$, independent of the past, and $\mathbb{P}(K = 0) = 2^{-F}$. Conditioning on survival therefore conditions each layer, independently, on $K \geq 1$, and leaves the rest of the construction below unchanged.

One step. Fix the past, and write $\hat{v} = \rho\hat{h} + \sigma u$ with $u \perp \hat{h}$, $\sigma = (1-\rho^2)^{1/2}$. By rotation invariance $a = (F/g)^{1/2} W\hat{h}$ and $b = (F/g)^{1/2} Wu$ are independent $\mathcal{N}(0, I_F)$, and the active set is $\{i : a_i > 0\}$. So $\|v'\|^2/\|v\|^2 = (g/F)\,X$ with $X = \sum_{i:\,a_i>0} Y_i$ and $Y_i = (\rho|a_i| + \sigma b_i)^2$. Given $K = k$, the Y_i are i.i.d.; the odd powers of b vanish in expectation, so

$$\mathbb{E}Y = \rho^2 + \sigma^2 = 1, \qquad \mathbb{E}Y^2 = 3\rho^4 + 6\rho^2\sigma^2 + 3\sigma^4 = 3,$$

for *every* ρ: Y has the mean and variance of χ^2_1, whatever the direction. All moments of Y are bounded uniformly in ρ. The density of $\rho|a| + \sigma b$ is at most $\min(2\varphi(0)/|\rho|,\ \varphi(0)/\sigma) \leq 1.13$, since $\max(|\rho|, \sigma) \geq 2^{-1/2}$; so $\mathbb{P}(Y \leq t) \leq 2.3\sqrt{t}$, and $\mathbb{P}(X \leq \varepsilon k \mid K = k) \leq (C\varepsilon)^{k/2}$.

Expansion. Write $X/k = 1 + \Delta$ with $\mathbb{E}\Delta = 0$, $\mathbb{E}\Delta^2 = 2/k$ and $\mathbb{E}|\Delta|^p \leq C_p k^{-p/2}$. With the small-ball bound, as in Lemma 5, $\mathbb{E}[\log(X/k) \mid K = k] = -\frac{1}{2}\mathbb{E}\Delta^2 + O(k^{-2}) = -1/k + O(k^{-2})$, uniformly in ρ. For $K \sim \mathrm{Bin}(F, \frac{1}{2})$ conditioned on $K \geq 1$, $\mathbb{E}\log(K/F) = -\log 2 - 1/(2F) + O(F^{-2})$ and $\mathbb{E}[1/K] = 2/F + O(F^{-2})$ (the event $K < F/4$ has probability e^{-cF}). Hence, uniformly in the past,

$$\mathbb{E}\big[\log(\|v'\|^2/\|v\|^2) \,\big|\, \text{past}\big] \;=\; \log\frac{g}{2} - \frac{1}{2F} - \frac{2}{F} + e \;=\; \log\frac{g}{2} - \frac{5}{2F} + e, \qquad |e| \leq \frac{C}{F^2}.$$

Time average. The one-step log-gains have conditional variances bounded uniformly in the past, by the moment and small-ball bounds. Their centred partial sums are therefore a martingale with bounded increment variances, and divided by L they tend to 0 almost surely (the martingale strong law). The time average of the conditional means lies within C/F^2 of $\log(g/2) - 5/(2F)$ at every step, which gives the claim. The direction of the tangent drifts, but no ergodicity of that drift is needed: the one-step bound holds for every direction. □

ReLU: the constant, measured. ReLU is outside the class of Theorem 2 (it is not odd), but it isolates the root-multiplicity principle cleanly. Positive homogeneity gives the mean-field gain $G(g) = g\,\mathbb{E}[\mathbf{1}\{\xi > 0\}] = g/2$ exactly: a *simple* root at $g = 2$. It also makes the finite-width exponent satisfy $2\chi^\star(g) = \log g + 2\chi^\star(1)$ at every width, so the crossing is exactly $\log g_\times = -2\chi^\star(1)$. For a tangent direction held fixed, one layer multiplies the squared

norm by $(g/F)\chi^2_K$, where $K \sim \mathrm{Bin}(F, \frac{1}{2})$ counts the active units. Expanding $\mathbb{E}\log(\chi^2_K/F) = -\log 2 - 5/(2F) + O(F^{-2})$ gives the heuristic $\log g_\times = \log 2 + 5/(2F) + O(F^{-2})$: a shift of order $1/F$, as a simple root requires, not $F^{-1/2}$. On a ReLU MLP ($\sigma_\mathrm{b} = 0$, 16 seeds of 10^5–4×10^5 layers), $F(\log g_\times - \log 2) = 3.16$, 2.82, 2.64, 2.56, 2.53, 2.51 ± 0.02 and 2.48 ± 0.04 at $F = 8, 16, \ldots, 512$, approaching $5/2$ (`experiments/relu_edge.py`). The exact fixed-direction value $-F\,(\mathbb{E}\log(\chi^2_K/F) + \log 2)$, summed over the binomial, is 3.19, 2.82, 2.57 and 2.52 at $F = 8$, 16, 64 and 256: within 1% of the network for $F \geq 16$. A registered check at two held-out widths (`experiments/PREREG_relu_confirm.md`) found 2.704 ± 0.005 at $F = 24$ (predicted 2.696, $+0.3\%$) and 2.475 ± 0.037 at $F = 1024$ (predicted 2.504, -1.2%). Its rule asked both 95% intervals to lie within $\pm 3\%$; the $F = 1024$ interval reaches -4%, so the registered verdict is INCONCLUSIVE, for lack of power at the larger width, not a refutation. A power follow-up, registered after that result and reported with it (`experiments/PREREG_relu_power.md`; 200 fresh seeds, $3.5\times$ the samples), measured 2.502 ± 0.020 at $F = 1024$: -0.1% from the prediction, 95% interval $[-1.7\%, +1.5\%]$, SUPPORTED. With $\tanh$ at zero bias (a double root, $F^{-1/2}$) and nonzero bias (a simple root, F^{-1}), this is a third instance of the same principle. Proposition 10 proves the constant: along the trajectory the tangent direction drifts, but it changes the expected log-gain only at order F^{-2}. Corrections of order $1/n$ to log-norms of random ReLU products are known in general (Hanin & Nica, 2020b).

C.11 THE ALIGNMENT CORRECTION, AND ITS SIGN

Conditional on the state and tangent, Z^ℓ and the linear increment $u^\ell = \sigma_\mathrm{s} v^\ell W_s + \sigma_\mathrm{n} A v^\ell W_n$ are jointly Gaussian, both linear in the same weight draw, with

$$\mathrm{Cov}(Z_{ia}, u_{ia}) = \tfrac{1}{F}\left[\sigma_\mathrm{s}^2 \langle H_i, v_i\rangle + \sigma_\mathrm{n}^2 \langle (AH)_i, (Av)_i\rangle\right]. \tag{19}$$

This state–tangent alignment was long taken to be the obstruction to equation 5. It is not: it acts in the favourable direction, per step (below) and, for node-uniform states, all the way to the exponent (Theorem 4, whose total-positivity step is the many-channel version of this lemma).

Per layer, $v^{\ell+1} = \phi'(Z^\ell) \odot u^\ell$, so the gain of the nonlinear step over the linear one is

$$G_\ell \;=\; \frac{\|v^{\ell+1}\|^2}{\|u^\ell\|^2} \;=\; \frac{\sum_{ia} \phi'(Z_{ia})^2 u_{ia}^2}{\sum_{ia} u_{ia}^2},$$

a *weighted* average of ϕ'^2 with weights $u_{ia}^2/\|u\|^2$. Inequality equation 5 needs that weighted average to sit below the unweighted $\mathbb{E}[\phi'^2]$. Independence would give it immediately; instead we compute the discrepancy exactly.

Lemma 10 (Alignment correction)**.** *Let (Z, u) be jointly Gaussian with mean zero,* $\mathrm{Var}\, Z = q$*,* $\mathrm{Var}\, u = s^2$ *and correlation ρ. Then*

$$\mathbb{E}\big[\phi'(Z)^2 u^2\big] \;=\; s^2\, \mathbb{E}[\phi'(Z)^2]\,\Big(1 - \rho^2\big(1 - \kappa\big)\Big), \qquad \kappa \;=\; \frac{\mathbb{E}[\phi'(Z)^2 Z^2]}{q\, \mathbb{E}[\phi'(Z)^2]}. \tag{20}$$

If moreover ϕ'^2 is non-increasing in $|Z|$, then $\kappa \leq 1$, so $\mathbb{E}[\phi'(Z)^2 u^2] \leq s^2\, \mathbb{E}[\phi'(Z)^2]$, with equality iff $\rho = 0$ or ϕ' is constant.

Proof. Write $u = (\rho s/\sqrt{q})\, Z + \xi$ with ξ independent of Z and $\mathrm{Var}\, \xi = s^2(1-\rho^2)$. Expanding u^2, the cross term has zero mean, so $\mathbb{E}[\phi'(Z)^2 u^2] = \rho^2 (s^2/q)\, \mathbb{E}[\phi'(Z)^2 Z^2] + s^2(1-\rho^2)\mathbb{E}[\phi'(Z)^2]$, which is equation 20. For the second claim, $\kappa \leq 1$ is exactly $\mathrm{Cov}(\phi'(Z)^2, Z^2) \leq 0$. Oddness of ϕ makes ϕ' even, so both $\phi'(Z)^2$ and Z^2 are functions of $|Z|$ alone — the first non-increasing, the second increasing — and oppositely monotone functions of a single random variable have non-positive covariance: with Z' an independent copy, $\mathrm{Cov}(\phi'(Z)^2, Z^2) = \frac{1}{2}\mathbb{E}[(\phi'(Z)^2 - \phi'(Z')^2)(Z^2 - Z'^2)]$, and the integrand is ≤ 0. For equality: if ϕ' is non-constant there are $r_1 < r_2$ with $\phi'(r_1)^2 > \phi'(r_2)^2$, and on the event $\{|Z| \leq r_1, |Z'| \geq r_2\}$, which has positive probability, the integrand is strictly negative. So the covariance is negative, and $\kappa < 1$, unless ϕ' is constant. □

The alignment tightens the bound. The correction factor $1 - \rho^2(1-\kappa)$ is strictly below 1 whenever $\rho \neq 0$ and ϕ' is non-constant, and decreases in ρ^2. So a larger alignment buys *more* room in

equation 5, not less — which is the sign and the monotonicity Figure 9 measures independently, at $r = -0.81$.

Verified against direct simulation over five activations and five (q, ρ) settings, the identity equation 20 agrees to 1.4×10^{-3}, consistent with Monte-Carlo error at 4×10^6 draws; κ attains at most 0.813 over $q \in [0.1, 10]$ (`theory/corrected_quenched_bound.py`). The monotonicity hypothesis on ϕ'^2 is checkable per activation and holds for all five we use.

From one step to the exponent. Lemma 10 is a one-step statement about expectations of products, while the exponent needs expectations of logarithms of channel sums, at a direction that depends on the past. Theorem 4 closes that gap when the state is node-uniform, because the channel pairs are then i.i.d. and the linear gain is an exact χ_F^2. On irregular graphs the channel variance differs across nodes and the gain from a node-spread tangent is not a scaled χ_F^2; that node heterogeneity is what remains between this lemma and Conjecture 1.

C.12 Proposition 11: quenched domination under homogeneous damping

Proposition 11 (Quenched domination under homogeneous damping)**.** *Suppose at every layer $\phi'(Z^\ell) = d_\ell \mathbf{1}$ for a scalar $d_\ell \in [0, 1]$, i.e. the derivative is constant across coordinates. Then,* pointwise in the weight realization, *$\chi^\star = \chi^\star_{\mathrm{lin}} + \limsup_{L\to\infty} \frac{1}{L} \sum_\ell \log d_\ell \le \chi^\star_{\mathrm{lin}}$, with equality of the limits whenever the average of $\log d_\ell$ converges.*

Proof. With $\phi'(Z^\ell)$ a scalar, the tangent map at layer ℓ is $M_\ell = d_\ell J_\ell$ where J_ℓ is the linear tangent map, so the product factors, $M_L \cdots M_1 = (\prod_\ell d_\ell)\, J_L \cdots J_1$. Taking $\frac{1}{L}\log$ of the operator norm and letting $L \to \infty$ separates the two factors, and $d_\ell \le 1$ gives the inequality. No expectation is taken. □

This is the $\mathrm{CV}[\phi'] = 0$ case, and it is exactly where Conjecture 1 is tight: $\frac{1}{2}\log \mathbb{E}[\phi'^2] = \log d$ when $\phi' \equiv d$.

C.13 Proposition 4: finite-depth domination

Let $\mathcal{F}_\ell$ be the σ-field of $(H^0, v^0, \ldots, H^\ell, v^\ell)$; the layer-$\ell$ draws $(W_s^\ell, W_n^\ell, b^\ell)$ are independent of it. Write $u^\ell = \sigma_{\mathrm{s}} v^\ell W_s^\ell + \sigma_{\mathrm{n}} A v^\ell W_n^\ell$, so $v^{\ell+1} = \phi'(Z^\ell) \odot u^\ell$ in the unnormalized recursion. Given $\mathcal{F}_\ell$, both Z_{ia}^ℓ and u_{ia}^ℓ are linear in the Gaussian draws, so each pair is jointly Gaussian with mean zero, with variances q_{ia} and s_{ia}^2; $q_{ia} \ge \sigma_{\mathrm{b}}^2$ because the bias enters Z additively and independently of the weights. Lemma 10 applies to each pair (if $q_{ia} = 0$ then $Z_{ia} \equiv 0$ and the bound holds with equality since $\phi'(0) = 1$): $\mathbb{E}[\phi'(Z_{ia})^2 u_{ia}^2 \mid \mathcal{F}_\ell] \le s_{ia}^2\, m(q_{ia})$. Because ϕ'^2 is non-increasing in $|x|$, m is non-increasing in q, so $m(q_{ia}) \le m(q_{\min}^\ell) =: M_\ell$, which is $\mathcal{F}_\ell$-measurable. Summing over (i, a) and using $\sum_{ia} s_{ia}^2 = \mathbb{E}[\|u^\ell\|^2 \mid \mathcal{F}_\ell] \le g\, \|v^\ell\|^2$ from the proof of Proposition 2,

$$\mathbb{E}\big[\|v^{\ell+1}\|^2 \mid \mathcal{F}_\ell\big] \;\le\; g\, M_\ell\, \|v^\ell\|^2.$$

Hence $Y_\ell = \|v^\ell\|^2 / \prod_{k=L_0}^{\ell-1} (g M_k)$, $\ell \ge L_0$, is a non-negative supermartingale: $\mathbb{E}[Y_{\ell+1} \mid \mathcal{F}_\ell] \le Y_\ell$, the denominator being $\mathcal{F}_\ell$-measurable. So $\mathbb{E}[Y_L \mid \mathcal{F}_{L_0}] \le \|v^{L_0}\|^2$, and by Jensen's inequality for the concave $\log$, $\mathbb{E}[\log Y_L \mid \mathcal{F}_{L_0}] \le \log \|v^{L_0}\|^2$. Expanding $\log Y_L$ and taking expectations, which the integrability assumption permits,

$$\mathbb{E}\big[\log \|v^L\|^2 - \log \|v^{L_0}\|^2\big] \;\le\; (L - L_0) \log g + \sum_{\ell=L_0}^{L-1} \mathbb{E}[\log M_\ell],$$

which is the first inequality after dividing by $2(L - L_0)$. The second follows from $q_{\min}^\ell \ge \sigma_{\mathrm{b}}^2$ and monotonicity of m. For bounded ϕ the state is bounded, so $q_{\min}^\ell$ is bounded and $\log M_\ell$ is bounded below. The renormalized estimator of equation 2 sums $\log(\|v^{\ell+1}\|/\|v^\ell\|)$, which telescopes to the left-hand side, so the bound applies to it exactly. □

Three limits. The bound replaces the quenched linear exponent by the annealed rate $\frac{1}{2}\log g$, which is why it fixes a direction but not a magnitude: at $\sigma_{\mathrm{b}} = 0$ the finite-width linear boundary 0.9729 already lies outside what it certifies. Taking the maximum of $m(q_{ia})$ over coordinates discards the heterogeneity that Conjecture 1 uses. And it bounds an expectation, not the almost-sure exponent, so $\mathcal{C}_{\mathrm{q}}$ is not

reached. Numerically (`theory/supermartingale_bound.py`): the one-step inequality holds with ratio 0.12–0.48 at six fixed states, including tangents aligned with the state; the deterministic corollary certifies order below $\sigma_s = 0.8689, 0.8773, 0.9547, 1.1318$ at $\sigma_b = 0.05, 0.1, 0.3, 0.6$ for $\tanh$, below all 32 measured crossings; and evaluated along 16 trajectories on `pair-close`, the bound exceeds the measured estimator at all 42 points tested, with its own zero crossing at 0.867 ($\sigma_b = 0$) and 1.144 ($\sigma_b = 0.3$) against measured 1.663 and 2.271.

C.14 Theorem 4: the quenched bound for MLPs and regular graphs

Hypotheses. ϕ odd, bounded, C^1, $|\phi'| \le \phi'(0) = 1$, ϕ'^2 non-increasing in $|x|$, $|\log \phi'(x)| \le C(1 + x^2)$; $A\mathbf{1} = d\mathbf{1}$ (d-regular, or $n = 1$); $H^0 = \mathbf{1}\, h^{0\top}$; $\sigma_b \ge 0$ with $b^\ell \sim \mathcal{N}(0, I_F)$ added to every node.

Step 1: the state stays node-uniform. If $H^\ell = \mathbf{1}h^{\ell\top}$ then $AH^\ell = d\,\mathbf{1}h^{\ell\top}$, so $Z^\ell = \mathbf{1}z^{\ell\top}$ with $z^\ell = \sigma_s h^\ell W_s^\ell + \sigma_n d\, h^\ell W_n^\ell + \sigma_b b^\ell$, and $H^{\ell+1} = \mathbf{1}\phi(z^\ell)^\top$. Hence $\phi'(Z^\ell)$ is constant across nodes, and the tangent update acts on each block of Appendix C.2 as $y_k^{\ell+1} = (y_k^\ell M_k^\ell) \odot \phi'(z^\ell)$. Here $d = \mu_{\max}$ and $|\lambda_k| \le d$.

Step 2: one step, one block. Fix k and condition on $\mathcal{F}_\ell$. For each channel a the pair (z_a, u_a), $u = y_k M_k$, is a linear function of column a of (W_s, W_n) and of b_a, so the pairs are i.i.d. across a, jointly Gaussian, mean zero, with $\operatorname{Var} z_a = q^\ell$ and $\operatorname{Var} u_a = s_k^2 \|y_k\|^2/F$. Write $u_a = (s_k\|y_k\|/\sqrt{F})\zeta_a$ and $c_a = \phi'(z_a)^2$. Then $\|y_k^{\ell+1}\|^2/\|y_k^\ell\|^2 = (s_k^2/F)\sum_a c_a\zeta_a^2$, and $\sum_a \zeta_a^2 \sim \chi_F^2$ independently of $\mathcal{F}_\ell$. By Jensen,

$$\mathbb{E}\Big[\log \textstyle\sum_a c_a\zeta_a^2 \;\Big|\; \mathcal{F}_\ell\Big] \le \mathbb{E}\log\chi_F^2 + \log \mathbb{E}\Big[\tfrac{\sum_a c_a\zeta_a^2}{\sum_b \zeta_b^2} \;\Big|\; \mathcal{F}_\ell\Big].$$

With $R_a = \sum_{b\neq a}\zeta_b^2 \sim \chi^2_{F-1}$ independent of (c_a, ζ_a) and $f(t) = \mathbb{E}[t/(t+R)]$ non-decreasing, each summand is $\mathbb{E}[c_a f(\zeta_a^2)]$. The density of $(|z_a|, |\zeta_a|)$ on $[0,\infty)^2$ is proportional to $e^{-\alpha x^2 - \beta y^2}\cosh(\kappa xy)$, which is TP_2 because $\partial_x\partial_y \log\cosh(\kappa xy) = \kappa\tanh(\kappa xy) + \kappa^2 xy\,\operatorname{sech}^2(\kappa xy) \ge 0$ for $\kappa xy \ge 0$ (with κ replaced by $|\kappa|$, as $\cosh$ is even). A TP_2 pair is positively associated, so $\operatorname{Cov}(c_a, f(\zeta_a^2)) \le 0$: c_a is non-increasing and $f(\zeta_a^2)$ non-decreasing in the respective absolute values (for $|\rho| = 1$ this is Chebyshev's inequality). Hence $\mathbb{E}[c_a f(\zeta_a^2)] \le \mathbb{E}c_a\,\mathbb{E}f(\zeta_a^2) = m(q^\ell)/F$, and summing over a,

$$\mathbb{E}\big[\log\|y_k^{\ell+1}\|^2 - \log\|y_k^\ell\|^2 \;\big|\; \mathcal{F}_\ell\big] \le \log s_k^2 + 2\gamma_F + \log m(q^\ell) \le 2\chi_{\mathrm{lin}}^\star(F) + \log m(q^\ell).$$

Step 3: to the exponent. Boundedness of ϕ gives $q^\ell \le g\|\phi\|_\infty^2 + \sigma_b^2$, and the growth bound on $\log\phi'$ makes the conditional second moments of the one-step log-ratio bounded by a constant C_F, uniformly over the state and the layer at each fixed F (it grows with F, e.g. through $\max_a z_a^2$). The martingale strong law then removes the conditional expectations, and Jensen over layers gives $\limsup_L L^{-1}\log\|y_k^L\| \le \chi_{\mathrm{lin}}^\star(F) + \frac{1}{2}\log\bar m$ a.s., where $\bar m$ is the long-run average of $m(q^\ell)$, equal a.s. to the long-run average of $F^{-1}\sum_a \phi'(z_a^\ell)^2$ by the same strong law. The maximum over finitely many blocks gives the theorem. For $n = 1$, taking expectations in Step 2 and summing gives the finite-depth version stated in the text. □

Where it stops. Two steps use node uniformity. On an irregular graph $\operatorname{Var}(Z_{ia})$ differs across nodes, so no single $m(q)$ factors out; and the linear gain from a node-spread tangent is no longer a scaled χ_F^2, so its finite-width penalty is not γ_F. Node heterogeneity of the state, not state–tangent alignment, is therefore the obstruction to Conjecture 1 on general graphs: alignment only tightens the bound. An adversarial search over hub-dominated and bipartite graphs (168 cells, $F \in \{2, 6, 24\}$, $g \in \{1.1, 1.5, 3, 8\}$, $\tanh$ and softsign) found no violation beyond 3 standard errors (`theory/conjecture_adversarial.py`).

C.15 Irregular graphs: what can be proved, and what stops the rest

Proposition 12 (tangent-weighted quenched bound)**.** *Let ϕ satisfy the hypotheses of Theorem 4, A be any symmetric matrix, and $\sigma_b \ge 0$. Given the past $\mathcal{F}_\ell$, let q_i^ℓ and $(s_i^\ell)^2$ be the variances of Z_{ia}^ℓ and u_{ia}^ℓ (they do not depend on the channel a), $w_i^\ell = (s_i^\ell)^2/\sum_j (s_j^\ell)^2$ the expected share of node i in the linear increment, and $m_w^\ell = \sum_i w_i^\ell\, m(q_i^\ell)$. Assume $\sigma_s > 0$, so that $v^\ell \neq 0$ for all ℓ almost surely*

(without it, read $\log 0 = -\infty$*; the bound then holds trivially on the event that the tangent dies). Then, almost surely,*

$$\limsup_{L\to\infty} \frac{1}{L} \log \|v^L\| \;\le\; \tfrac{1}{2}\log g \;+\; \tfrac{1}{2}\limsup_{L\to\infty} \frac{1}{L}\sum_{\ell<L} \log m_w^\ell$$
$$\le\; \tfrac{1}{2}\log g + \tfrac{1}{2}\log\Big(\limsup_{L\to\infty} \frac{1}{L}\sum_{\ell<L} m_w^\ell\Big).$$

Proof. Given $\mathcal{F}_\ell$, each pair (Z_{ia}, u_{ia}) is linear in the layer's Gaussian draws, hence jointly Gaussian with mean zero, and the pairs are i.i.d. across a. Lemma 10 applies to every pair, so $\mathbb{E}[\|v^{\ell+1}\|^2 \mid \mathcal{F}_\ell] = \sum_{ia} \mathbb{E}[\phi'(Z_{ia})^2 u_{ia}^2 \mid \mathcal{F}_\ell] \le F\sum_i s_i^2\, m(q_i) = \mathbb{E}[\|u\|^2 \mid \mathcal{F}_\ell]\; m_w^\ell$, and $\mathbb{E}[\|u\|^2 \mid \mathcal{F}_\ell] \le g\|v^\ell\|^2$ as in Proposition 2. By Jensen, $\mathbb{E}[\log(\|v^{\ell+1}\|^2/\|v^\ell\|^2) \mid \mathcal{F}_\ell] \le \log g + \log m_w^\ell$. The one-step log-ratio has conditional second moments bounded by a constant C_F, uniformly over the state and the layer at each fixed F (the constant grows with F, e.g. through $\max_a Z_{ia}^2$, which the strong law at fixed F does not mind): above, because $|\phi'| \le 1$; below, because each column of u is Gaussian with covariance of trace at least $\sigma_{\mathrm{s}}^2\|v\|^2/F$, hence largest eigenvalue at least $\sigma_{\mathrm{s}}^2\|v\|^2/(nF)$, and $\log\phi'(x) \ge -C(1+x^2)$ with q_i bounded (boundedness of ϕ). The martingale strong law then removes the conditional expectation almost surely, and the second inequality is Jensen over layers. □

Since $m_w^\ell \le \max_i m(q_i^\ell) = m(q_{\min}^\ell)$, this strengthens Proposition 4 on every graph in two ways: it holds almost surely rather than in expectation, and it weights nodes by where the tangent is rather than taking the worst node. It is still *annealed* in its linear part: it has $\frac{1}{2}\log g$ where Conjecture 1 has $\chi_{\mathrm{lin}}^\star(F) = \frac{1}{2}\log g + \gamma_F$, so it is looser than the conjecture by $|\gamma_F|$ (0.088 at $F = 6$).

Why the finite-width penalty does not come for free. On a regular graph with uniform input the tangent lies in one eigen-block, and the linear gain is exactly $(s^2/F)\chi_F^2$, which is where γ_F comes from. On an irregular graph the tangent spreads over node directions, and one linear step gives a Gaussian quadratic form $\sum_k \lambda_k \chi_{F,k}^2$ whose weights are the eigenvalues of the tangent's node covariance. For a concave function of an exchangeable sum, the expectation is Schur-concave in the weights, so the penalty $\mathbb{E}\log\sum_k \lambda_k\chi_{F,k}^2 - \log(F\sum_k\lambda_k)$ is *largest in magnitude at rank one*, where it equals $2\gamma_F$, and shrinks as the tangent spreads. At $F = 6$ it is -0.176 at rank one, -0.086 for two equal weights and -0.021 for eight. A spread tangent therefore escapes part of the finite-width penalty, but it pays for leaving the top eigenvector through a smaller Rayleigh quotient, $(\sigma_{\mathrm{s}}^2\|v\|^2 + \sigma_{\mathrm{n}}^2\|Av\|^2)/\|v\|^2 < g$. Conjecture 1 is precisely the statement that the second cost always outweighs the first. A proof has to control the joint law of the tangent's direction and the state, not one step at a time; that is what remains open.

Proposition 13 (the trade-off, one step at a time)**.** *Let $F > 2$ and let A be symmetric with a simple eigenvalue of largest modulus $\mu_{\max}$, let μ_2 be the largest modulus among the others, and put $s_2^2 = \sigma_{\mathrm{s}}^2 + \sigma_{\mathrm{n}}^2\mu_2^2$. For a tangent $v \in \mathbb{R}^{n\times F}$ let $u = \sigma_{\mathrm{s}} v W_s + \sigma_{\mathrm{n}}(Av)W_n$ be one linear step, and $D(v) = \mathbb{E}\log(\|u\|^2/\|v\|^2)$. Then, for every v,*

$$D(v) \;\le\; \log R(v) + 2\gamma_F + \frac{2b(v)}{F-2},$$

where $R(v) = (\sigma_{\mathrm{s}}^2\|v\|^2 + \sigma_{\mathrm{n}}^2\|Av\|^2)/\|v\|^2$ and $b(v)$ is the share of $\mathbb{E}\|u\|^2$ carried off the top eigenvector of A. If moreover $\log(g/s_2^2) \ge 2/(F-2)$, then $D(v) \le \log g + 2\gamma_F = 2\chi_{\mathrm{lin}}^\star(F)$ for every v.

Proof. Given v, the columns of u are i.i.d. $\mathcal{N}(0, C)$ with $C = (\sigma_{\mathrm{s}}^2 vv^\top + \sigma_{\mathrm{n}}^2(Av)(Av)^\top)/F$, so $\|u\|^2 = \sum_k \lambda_k Y_k$ with λ_k the eigenvalues of C and Y_k i.i.d. χ_F^2, and $F\,\mathrm{tr}\,C = R\|v\|^2$. Write $\theta = \lambda_{\max}/\mathrm{tr}\,C$. Conditioning on Y_1 and applying Jensen to the rest, $\mathbb{E}\log\sum_k\lambda_k Y_k \le \log(F\,\mathrm{tr}\,C) + \Phi(\theta)$ with $\Phi(\theta) = \mathbb{E}\log(\theta Y + (1-\theta)F) - \log F$. Φ is concave and decreasing, $\Phi(1) = 2\gamma_F$ and $\Phi'(1) = 1 - F\,\mathbb{E}[1/Y] = -2/(F-2)$, so $\Phi(\theta) \le 2\gamma_F + 2(1-\theta)/(F-2)$. Expanding $v = \sum_k e_k y_k^\top$ in the eigenvectors of A, the diagonal entry of C along e_1 is $g\|y_1\|^2/F$, so $\theta \ge 1 - b$. This proves the first bound. For the second, $F\,\mathrm{tr}\,C = g\|y_1\|^2 + \sum_{k\ge2}(\sigma_{\mathrm{s}}^2 + \sigma_{\mathrm{n}}^2\lambda_k^2)\|y_k\|^2$ and $\lambda_k^2 \le \mu_2^2$ for $k \ge 2$ give $g/R \ge 1 + b\kappa$ with $\kappa = (g - s_2^2)/s_2^2$. So $D - \log g - 2\gamma_F \le 2b/(F-2) - \log(1+b\kappa)$, which is convex in $b \in [0,1]$ and 0 at $b = 0$, hence nonpositive on $[0,1]$ as soon as it is at $b = 1$, which is the stated condition. □

It is sharp, and it fails where the conjecture is hardest. Near a rank-one tangent the derivative of $D - 2\chi^\star_{\mathrm{lin}}(F)$ in the second-eigenvector share is $2/(F-2) - \kappa$, so the condition cannot be weakened below $\kappa \geq 2/(F-2)$. Evaluated exactly, via $\log x = \int_0^\infty (e^{-t} - e^{-tx})\, dt/t$ (no sampling error), the first bound holds to 10^{-15} on every tangent tried and is attained at rank one; where the condition holds, $D \leq 2\chi^\star_{\mathrm{lin}}(F)$ to 10^{-15}; where it fails, D exceeds it (by $+0.020$ on a lollipop graph at $F = 6$). On a bipartite graph $\mu_2 = \mu_{\max}$ and the condition fails at every width: splitting a tangent between the eigenvectors of $\pm\mu_{\max}$ costs nothing in R and gains $+0.090$ at $F = 6$. At the certificate's operating point ($\sigma_{\mathrm{s}} = 1.5$, so $g = 2.5$ on the graph-active ray) $\log(g/s_2^2) \leq \log(2.5/2.25) = 0.105$, so the condition needs $F \geq 21$ on every graph and holds at $F = 6$ on none (`theory/t3_tradeoff.py`).

What the conjecture now reduces to. Splitting one nonlinear step as $\log \|v^{\ell+1}\|^2 = \log \|u\|^2 + \log(\|v^{\ell+1}\|^2/\|u\|^2)$, bounding the first term by Proposition 13 and the second by Jensen, the martingale argument of Proposition 12 gives the tangent-weighted form of Conjecture 1 at every width where the condition of Proposition 13 holds, provided that, almost surely,

$$\limsup_{L\to\infty} \frac{1}{L} \sum_{\ell<L} \Big(\log \mathbb{E}\big[\|v^{\ell+1}\|^2/\|u^\ell\|^2 \;\big|\; \mathcal{F}_\ell \big] - \log m_w^\ell \Big) \;\leq\; 0. \tag{21}$$

On regular graphs with uniform input every summand is nonpositive: that is the total-positivity step of Theorem 4. **On irregular graphs the summands are not all nonpositive**, and the one-step version of equation 21 is false. By Price's theorem the alignment correction to $\mathbb{E}[\phi'(Z_{ia})^2 u_{ia}^2/\|u\|^2]$ vanishes at first order (since ϕ'^2 is even) and at second order is $\mathbb{E}[(\phi'^2)'']$ times a quadratic form in the cross-covariances $\mathrm{Cov}(Z_{ia}, u_{ja})$. Alignment with the node's own tangent ($j = i$) lowers the ratio, while alignment with another node's tangent ($j \neq i$) raises it. A two-node state with only cross-node alignment, within the $\tanh$ range, exceeds the tangent-weighted bound by 6 standard errors (and the expected-share bound by 25), random bounded states on a path exceed it by up to 8, and even states visited by the dynamics occasionally do: 3 of 80 on NCI1-g2 and a barbell, by up to $+0.011$ (`theory/t3_ratio_step.py`, `theory/t3_ratio_visited.py`). *On average* the summands are strongly negative (-0.03 at visited states; -0.025 to -0.038 along trajectories on eight graphs at $F = 6$, -0.006 to -0.009 at $F = 24$), because along the dynamics own-node alignment dominates. So equation 21 is an ergodic statement about the joint law of the state and the tangent's node distribution, and no argument conditional on one step can prove it. Two gaps therefore remain: equation 21 on irregular graphs, and, at widths below the condition of Proposition 13 (including the certificate's $F = 6$), a multi-layer replacement for that proposition too.

A sharper form that fails. Weighting nodes by the squared Perron vector instead of uniformly, $\chi^\star \leq \chi^\star_{\mathrm{lin}} + \frac{1}{2}\log \sum_i e_i^2 m_i$, is tighter than the conjecture on every graph we probed. A registered test on the eight graphs where it was tightest (`experiments/PREREG_t3_perron.md`, 512 realizations, Bonferroni intervals) **falsified** it: on a star with 64 nodes the slack is $+0.025$ $[+0.017, +0.033]$, and across stars it rises with size (-0.026, -0.000, $+0.025$ at $n = 16, 40, 64$). The uniform and tangent-weighted forms hold with wide margins on the same runs (-0.151 and -0.128 on the 64-node star). The Perron vector puts half its mass on the hub, whose ϕ'^2 is smallest; the tangent does not, so hub-weighting overstates the damping. Of the three forms, the tangent-weighted one is both the natural proof target and, on these data, valid.

C.16 Conjecture 1: the certificate

On irregular graphs equation 5 is *not proved*. It is certified numerically on 40 graphs $\times$ 5 activations, 8 seed-groups, with a 95% bootstrap CI on the slack $\delta = \chi^\star_\phi - \chi^\star_{\mathrm{lin}} - \frac{1}{2}\log \mathbb{E}[\phi'(Z)^2]$ per cell. $\chi^\star_{\mathrm{lin}}$ is measured on the *same* graph at the *same* operating point (it agrees with Theorem 1), and $\mathbb{E}[\phi'(Z)^2]$ under the *measured* late-layer state law rather than the mean-field one — mean field is biased at $F = 6$. A cell certifies only if its CI lies entirely below zero.

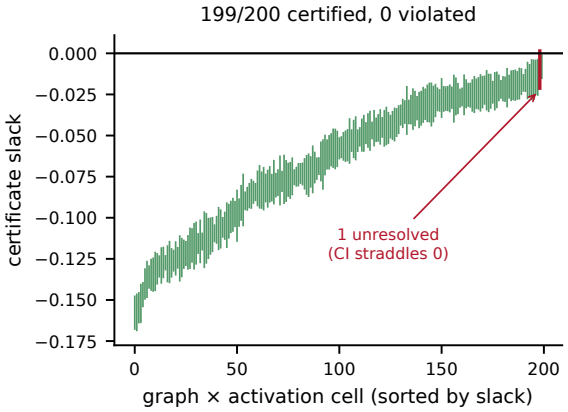


Figure 7: Certificate for Conjecture 1 on 40 irregular graphs $\times$ 5 activations at $F = 6$: slack $\chi^\star_\phi - \chi^\star_{\mathrm{lin}} - \frac{1}{2}\log \bar{m}$ with 95% bootstrap CIs, sorted. $199/200$ certify (interval below zero), one is unresolved, none is violated.

ϕ	cells	certified	mean slack
erf	40	40	-0.0689
tanh	40	39	-0.0671
arctan	40	40	-0.0668
sublinear	40	40	-0.0655
softsign	40	40	-0.0607
total	**200**	**199**	-0.0658

Zero violations. The sublinear rows ($\phi = x(1+x^2)^{-1/4}$, unbounded) lie outside the conjecture's hypotheses, which inherit boundedness from Theorem 4; they are reported as a check beyond its scope. Slack ranges -0.1574 to -0.0091; $148/200$ cells clear zero by more than 0.02. The one unresolved cell straddles zero by 0.0025 and is reported as unresolved. **Scope limit:** the slack correlates with $\mu_{\max}$ at -0.401 over $\mu_{\max} \in [2.45, 31.0]$, so the margin thins as $\mu_{\max}$ grows and the near-zero cells concentrate at the top of that range. We claim nothing beyond $\mu_{\max} = 31$.

C.17 The vectorized form, and the obstruction

This subsection records the vectorized layer map and two second-moment statements about it. Writing $h = \mathrm{vec}(H)$ with $H \in \mathbb{R}^{N\times F}$ and using $\mathrm{vec}(HW) = (W^\top \otimes I_N)\,\mathrm{vec}(H)$, the layer map is $L_\ell = \sigma_{\mathrm{s}}(W_s^{\ell\top} \otimes I_N) + \sigma_{\mathrm{n}}(W_n^{\ell\top} \otimes A)$. Two consequences follow; neither is used by the results above.

Lemma 11. *For W entries i.i.d. $\mathcal{N}(0, 1/F)$ and symmetric A, $\mathbb{E}[L^\top L] = \sigma_{\mathrm{s}}^2 I + \sigma_{\mathrm{n}}^2(I_F \otimes A^2)$.*

Proof. Expanding $L^\top L$ for fixed weights gives four Kronecker terms, $\sigma_{\mathrm{s}}^2(W_s W_s^\top \otimes I) + \sigma_{\mathrm{s}}\sigma_{\mathrm{n}}(W_s W_n^\top \otimes A) + \sigma_{\mathrm{n}}\sigma_{\mathrm{s}}(W_n W_s^\top \otimes A) + \sigma_{\mathrm{n}}^2(W_n W_n^\top \otimes A^2)$ — an algebraic identity, verified to 2.6×10^{-16}. Taking expectations needs only $\mathbb{E}[W_s W_s^\top] = I_F$ and $\mathbb{E}[W_s W_n^\top] = 0$. □

Lemma 12. *For the linear layer map $h \mapsto L_\ell h$ on two graphs A_1, A_2 driven by shared weights, with $\Delta_k = h_1^k - h_2^k$ and $u_k = \mathrm{vec}(A_1 H_1^k) - \mathrm{vec}(A_2 H_2^k)$, $\mathbb{E}\big[\|\Delta_{k+1}\|^2 \mid h_1^k, h_2^k\big] = \sigma_{\mathrm{s}}^2\|\Delta_k\|^2 + \sigma_{\mathrm{n}}^2\|u_k\|^2$, i.e. with $d_k = \|\Delta_k\|$, $\mathbb{E}\big[d_{k+1}^2 \mid h_1^k, h_2^k\big] = d_k^2\,(\sigma_{\mathrm{s}}^2 + \sigma_{\mathrm{n}}^2 Q_k)$ where $Q_k = \|u_k\|^2/\|\Delta_k\|^2$.*

Proof. Condition on (h_1^k, h_2^k), which are independent of the layer-k weights. Then $\Delta_{k+1} = \sigma_{\mathrm{s}}(W_s^\top \otimes I)\Delta_k + \sigma_{\mathrm{n}}(W_n^\top \otimes I)u_k$, and since $W_s \perp W_n$ with both mean zero the cross term vanishes in expectation. With a 1-Lipschitz ϕ applied after the linear map, the same argument gives $\le$ in place of $=$. □

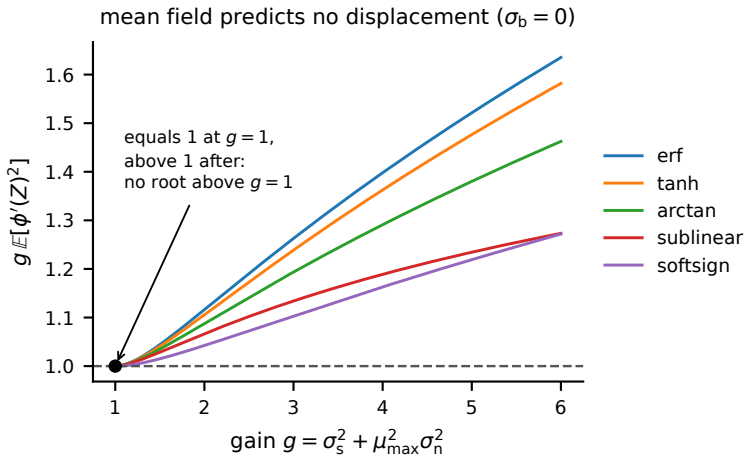


Figure 8: Numerical verification of Proposition 3. The mean-field tangent gain $g\,\mathbb{E}[\phi'(Z)^2]$ at the induced fixed point, five odd activations at $\sigma_b = 0$: it equals 1 at $g = 1$ and rises monotonically, so $g\,\mathbb{E}[\phi'(Z)^2] = 1$ has no root above $g = 1$ and $\mathcal{C}_{ind}$ collapses onto $\mathcal{C}_{lin}$. Mean field predicts zero displacement. The proposition is proved in §C above; this plot is not the evidence for it, but a check that the hypotheses are non-vacuous for the five activations we actually measure.

Convention. In the layer-ordered power series $H^{(k)} = \sum_m A^m X C_m^{(k)}$, the weight product runs layer-increasing left to right under the right-multiplied convention, which is the one used throughout this paper; it reverses under a left-multiplied convention.

Alignment, quantified. Conditional on (H^ℓ, v^ℓ), the pair (Z^ℓ, u^ℓ) is jointly Gaussian with covariance equation 19; verified against 2×10^5 draws to 0.1%. The per-step factorization is exact when that covariance vanishes (relative deviation 0.0097 on a numerically orthogonalized tangent) and fails under alignment (0.87). Measured along trajectories on the certificate's own 40-graph corpus, the alignment has mean 0.281 over $[0.251, 0.338]$ — leading-order, not a remainder.

Alignment predicts the margin. Figure 9 joins the two measurements graph by graph. The alignment tracks the slack at $r = -0.81$ ($[-0.88, -0.73]$, bootstrapped over graphs), and at -0.76 with μ_{max} partialled out, so it is not a restatement of the μ_{max} dependence. *More alignment means more margin*: where the factorization is furthest from exact, the inequality holds most comfortably. Both quantities were measured at the same operating point ($\sigma_s = 1.5$, $F = 6$, $L = 160$, the same 8 seed groups) but the alignment is per-graph while the slack is per-(graph, activation), so the per-cell correlation (-0.80) has an inflated n and the per-graph figure is the one to read.

D THE METRIC HALF

A companion line of work asks a different question — not *where* the boundary is but *with respect to which input metric* the phase classification is defined. On $\mathbb{R}^d$ all norms are equivalent, so the classification is metric-invariant; on graphs that fails. A related but distinct question — how far interactions reach, rather than which metric the classification depends on — is taken up by Bamberger et al. (2025). That question is outside the scope of this paper, and no result in the main text depends on it; the full treatment is deferred to separate work.

E THE DEPTH-CONTINUUM LIMIT

Under a residual parameterization with per-layer variances shrinking as $O(1/L)$, the depth-continuum limit of equation 1 is a diffusion. We use it for one structural fact and one scope condition.

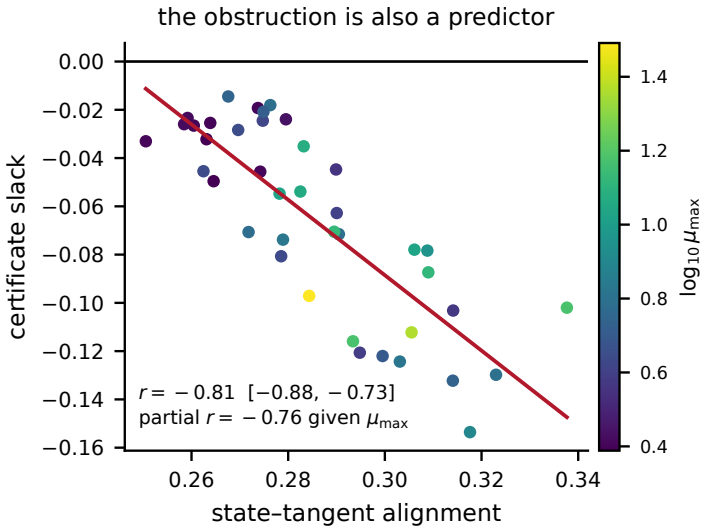


Figure 9: State–tangent alignment against certificate slack, 40 graphs, coloured by $\mu_{\max}$. More alignment means more room in the certificate, the favourable sign that Theorem 4 proves for node-uniform states.

Why ϕ must be odd. This necessity, and the bias cancellation below, are derived here; the depth-continuum limit itself is the construction of Peluchetti & Favaro (2020). The state map $\psi(H) = [H, SH]$ is linear, so a general activation contributes a quadratic Itô drift that violates the linear-growth condition required for non-explosion. Imposing $\phi''(0) = 0$ removes it, and oddness delivers exactly that. Under it the difference SDE collapses to a closed, autonomous linear equation whose drift and diffusion coefficients are bounded linear maps, hence globally Lipschitz, giving existence and pathwise uniqueness together with the fourth-moment bound $\mathbb{E}\|Z_t - Z_0\|^4 \le \kappa t^2(1 + \|Z_0\|^4)$ by standard Itô theory (Mao, 2008).

The bias cancels in node differences. Individual node embeddings accumulate variance σ_b^2 from the bias, but the same b^ℓ is added to every node, so the term vanishes in the difference dynamics and in their quadratic variation. Physically the random bias translates the whole graph through embedding space and plays no role in the relative separation of nodes. §K measures this at 1.1×10^{-16} and reconciles it with the σ_b-dependence of $\mathcal{C}_{\text{ind}}$.

The surviving channel is multiplicative and is measurable as such. Sweeping $\sigma_b : 0 \to 0.6$ on the graph-active ray over the eight graphs of §K, the mean state magnitude $\mathbb{E}\|Z\|$ rises by $+0.2688 \pm 0.0233$ — 11.5 standard errors — while $\mathbb{E}[\phi'(Z)^2]$ falls by -0.1201 ± 0.0122, or 9.8 standard errors, with the mean-field $q^\star$ rising $0.974 \to 1.544$ over the same range. So the bias never enters the tangent recursion additively; it enters only by moving the state at which ϕ' is evaluated, which is exactly the channel $\mathcal{C}_{\text{ind}}$ tracks and the node-difference dynamics do not.

Three defects in the inherited derivation, corrected here. (i) An assumption of $\rho(S) \le 1$ excludes GIN ($\rho = 1 + \bar{d} = 3.5$ on MUTAG) and every validation graph used ($\mu_{\max}$ up to 7.0 for K_8); we widen the scope and state it. (ii) A global Lipschitz constant was asserted from an inequality valid only on bounded sets; the linear difference SDE above supplies the global statement legitimately. (iii) The residual layer used in the derivation was never defined; it is defined above.

F Box-counting near the conic

Dimensions here are box-counting dimensions in the standard sense (Falconer, 2003), estimated by the standard grid method with the standard caveats about scaling range and saturation (Theiler, 1990); the exponent $\chi^\star$ is the top Lyapunov exponent of a random matrix product (Furstenberg &

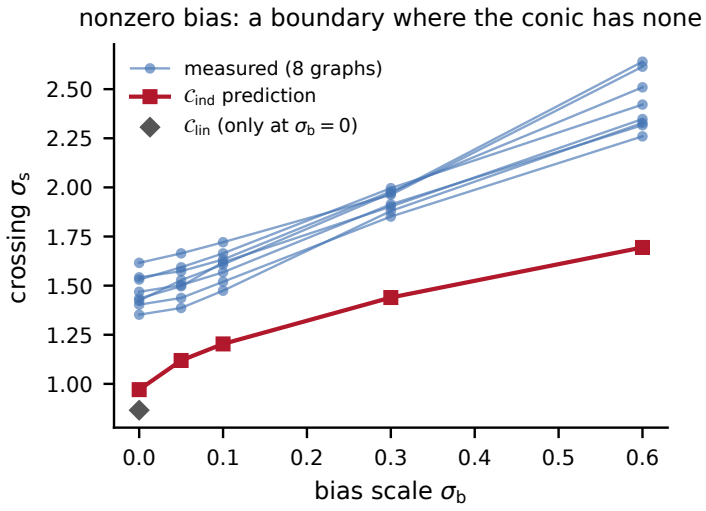


Figure 10: Measured crossings across eight graphs (blue) against the $\mathcal{C}_{\text{ind}}$ prediction (red), both on the finite-width convention of Appendix G. $\mathcal{C}_{\text{lin}}$ exists only at $\sigma_b = 0$ (grey diamond): nonzero bias destroys the origin fixed point, so Proposition 1's hypothesis is gone; at $\sigma_b = 0$, $\mathcal{C}_{\text{ind}} = \mathcal{C}_{\text{lin}}$ by Proposition 3. The prediction has the right direction and ordering on every graph and under-predicts every cell, by about the same margin and with the same sign as the finite-width displacement of §5 — which is what Proposition 3 implies for any mean-field object here.

Kesten, 1960; Oseledets, 1968). The window here is the one box-counted in an earlier version of this work: graph `pair-close` ($\mu_{\max} = 4.48$), $F = 6$, $L = 160$, $\sigma_s \in [0.02, 1.8]$, $\sigma_n \in [0, 1.6/\mu_{\max}]$, 96×96 cells. Its level set $\{\hat{\chi} = 0\}$ is a mixture of two populations that behave oppositely under averaging, and reading it correctly requires separating them (`experiments/e2_step1_field.py`, `e2_step2_field2d.py`).

Near $\mathcal{C}_{\text{lin}}$ there is no boundary. Along the graph-active ray the field averaged over 64 realizations is negative for every $g \in [0.8, 3.0]$ and reaches its maximum, -0.031, at $g \approx 1.18$, beside $\mathcal{C}_{\text{lin}}(6)$'s $g_{\text{lin}}(6) = 1.19$: the exponent *touches* zero near the finite-width linear boundary without crossing it, as the window analysis predicts (Appendix C.9). One realization's standard error there, about 0.04, exceeds $|\chi^\star|$, so single-realization level sets are dense with sign flips. That is the speckle: 1765, 609 and 323 boundary pixels in 60, 11 and 4 components at $S = 1, 8, 32$ (4-connectivity; 44, 7, 2 under 8-connectivity), falling as $S^{-0.655}$ in excess of the rasterized conic's 159 pixels in 1 component (Figure 11).

A genuine crossing lies elsewhere. Averaged over 1024 realizations, the field is significantly positive in 70 of the 9216 cells, all near the window's far corner, and negative elsewhere. The $S = 1024$ level set — 89 pixels, $D = 1.194$, reproducing the 92 pixels and $D = 1.167$ reported earlier — lies *entirely* at $g \in [3.32, 4.75]$, with no pixel near the conic ($g \approx 1$) or near $\mathcal{C}_{\text{lin}}(6)$. As S grows, the speckle near the conic disappears and what remains is this crossing: a short arc clipped by the window's corner.

What this corrects. An earlier version read the falling dimensions as convergence toward the conic, scoring them against the conic rasterized on the same grid ($D_{\text{analytic}} = 1.073$, $R^2 = 0.994$; not 1, because of finite-grid discretization):

S	$D(\widehat{\mathcal{C}}_{L,S})$	95% CI	boundary px	where the pixels are
32	1.466	$[1.426, 1.512]$	368	speckle near the conic, plus the crossing
128	1.289	$[1.213, 1.364]$	204	mixed
512	1.170	$[1.161, 1.180]$	123	mixed
1024	1.167	$[1.080, 1.263]$	92	the crossing only, $g \in [3.32, 4.75]$

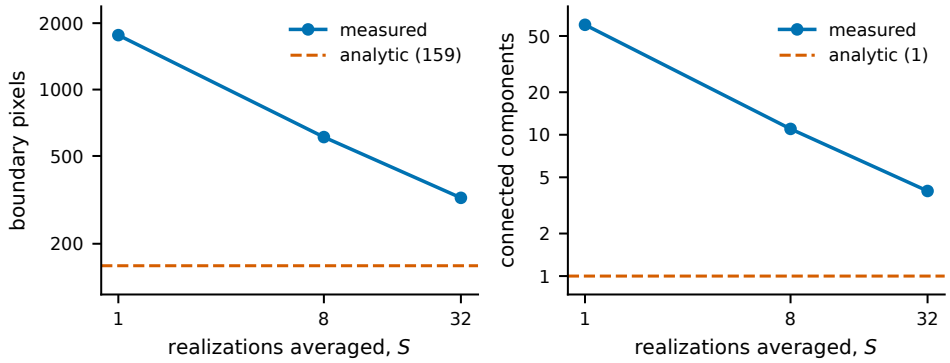


Figure 11: Boundary pixels and connected components against the averaging budget S. Both fall log-linearly: this is speckle disappearing. The dashed lines mark the rasterized conic, which is *not* the limit; the $S = 1024$ level set lies elsewhere in the window (text).

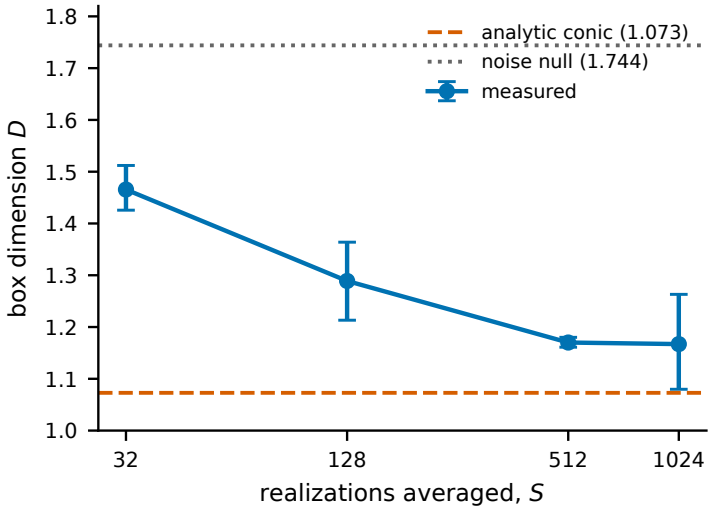


Figure 12: Box dimension against averaging budget, out to $S = 1024$, with the rasterized conic (1.073) and the structureless-noise null (1.744) marked. The decline is noise disappearing; the $S = 1024$ value belongs to a different curve from the conic (text), so the two are not comparable.

That reading was wrong. The sequence measures the *disappearance* of noise, not convergence to a curve: the $S = 1024$ set is a different curve from the conic, so $D = 1.167$ is not comparable with 1.073, and the conic's 159 pixels are not the reference for the speckle counts. A structureless null of independent per-cell draws gives $D = 1.744$, which a single realization matches (1.748 $[1.732, 1.765]$ from a small random start): at $S = 1$ the level set is indistinguishable from noise.

Gating. Requiring $|\chi^\star| > \mathrm{SE}$ before a cell may contribute a boundary — a weak, one-standard-error demand — removes, on `pair-close` at $L = 60$, $S = 6$ and a 72×72 grid, 95.0% of the level set at $F = 6$ (659 pixels to 33), 97.1% at $F = 16$ (859 to 25) and 96.4% at $F = 48$ (414 to 15). This is what the two populations predict: almost all of a small-S level set is noise.

It is not the initialization. Varying the initialization and depth at each budget:

init	L	$S = 1$	$S = 8$	$S = 32$
ones	40	1.565	1.433	1.278
ones	160	1.519	1.493	1.445
small	40	1.748	1.563	1.444
small	160	1.565	1.417	1.347

The averaging budget, not the start, controls the estimate.

Aimed at the finite-width crossing instead. A registered test (`experiments/PREREG_e2_boxcount.md`) compared this window with one of the same size centred on the converged crossing, over 512 fresh realizations. It was **falsified**: the second window's level set did not settle into a stable curve either (pixels $4527 \to 318$ and D $1.77 \to 1.23$ as S rose from 1 to 512). Its criterion for a resolved pixel, both neighbours more than 2 standard errors from zero, turned out to be unsatisfiable at a gentle crossing. A post-hoc classification of cells at $S = 512$ shows what the two windows contain. Around the linearized edge, 97% of cells are clearly negative and the only positive ones are the far corner, so there is no boundary to measure. Around the finite-width crossing, a genuine boundary separates clearly signed cells in 76 of 96 rows, but it lies in an ambiguous band about 5 cells wide. That band narrows as $S^{-1/2}$, so resolving it to one cell would need about 1.3×10^4 realizations. At $F = 6$, locating the crossing correctly is necessary for box-counting and not sufficient (`experiments/RESULT_e2_boxcount.md`).

The lesson for box-counting. A box dimension of $\{\hat{\chi} = 0\}$ at finite width measures two things at once: noise near a tangency that is not a boundary, and a boundary that lies elsewhere. It is interpretable only after averaging has removed the first, and only against the correct locus — here the finite-width crossing, not the conic. The corpus crossings of Figure 18 carry the same noise in a milder form (Appendix J).

G THE FINITE-WIDTH LAW

Theorem 1 gives the critical self-scale at $\sigma_{\mathrm{n}} = 0$ in closed form, $\sigma_{\mathrm{crit}}(F) = g_{\mathrm{lin}}(F)^{1/2} = \exp\big(\frac{1}{2}(\log(F/2) - \psi(F/2))\big)$. An independent random-matrix Lyapunov computation that never touches the tangent pipeline ($L = 1500$, 32 seeds; `boundary_extensions.py`) agrees with it to within Monte-Carlo error:

F	4	8	16	32	64
measured $\sigma_{\mathrm{s}}^{\star}(F)$	1.14304	1.06674	1.03430	1.01621	1.00834
exact (Theorem 1)	1.14474	1.06725	1.03241	1.01591	1.00788

The residuals change sign and have root mean square 0.0012, against 0.0020 for the best one-parameter interpolation $1 + c/F$ ($c = 0.5632$, fitted). On the graph-active ray the linear boundary is $\sqrt{g_{\mathrm{lin}}(F) - 1/4}$: 0.9707 at $F = 6$, 0.8683 at $F = 256$, 0.8666 at $F = 1024$. Three linear controls track it: the 17 graphs of §5 at $F = 6$ (-1.0% to $+1.6\%$, mean $+0.3\%$, on a grid of step 1.3%), the MLP at $F = 48$–1024 on an interpolated grid (within 0.44% at every width), and direct simulation on three irregular graphs at $F = 4, 16$ (within one standard error; `theory/regular_graph_theorem.py`).

H THE INITIALIZATION EXPERIMENT

The grid is 4512 runs: 8 datasets $\times$ 3 architectures $\times$ depths $L \in \{8, 16, 32, 64\} \times 8$ seeds, over 6 arms. The sixth, the GP criterion of Epping et al. (2024), is computable on 7 of the 8 datasets — their condition diagonalises an $N^2 \times N^2$ operator and one dataset has no graph small enough — so it covers 84 of the 96 cells. Contrasts are paired within matched (dataset, architecture, depth) cells and bootstrapped over cells at 10,000 resamples.

The residual margin is concentrated. Over all 96 cells the residual aggregation operator beats Kaiming by $+0.0246$, $[+0.0115, +0.0390]$. One dataset (DD, whose graphs are by far the largest,

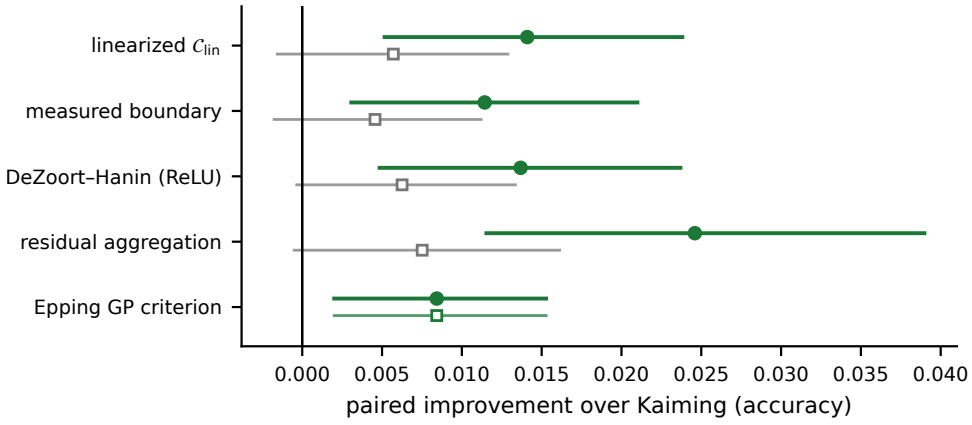


Figure 13: Paired improvement over Kaiming per arm, 95% bootstrap intervals over matched cells; green where the interval excludes zero. Filled circles use each arm's own denominator (96 cells, 84 for the GP criterion); open squares restrict every arm to the 84 cells the GP criterion can reach. The restriction reverses the ranking: on the common cells the GP criterion is the only arm whose margin resolves, and the residual operator no longer does. In 59 of the 96 cells either every arm or no arm trains above the majority-class rate, so those cells cannot separate the arms at all — which is why every resolved difference is small.

median 240 nodes, so depth collapses hardest there) supplies 73% of that effect from 12.5% of the cells: its mean margin is $+0.1442$ against $+0.0075$ across the other seven. Restricted to the remaining cells the margin is $+0.0075$, $[-0.0007, +0.0161]$ — no longer resolved — and on both IMDB datasets the residual operator is *worse* than Kaiming (-0.0242 and -0.0103). The 96-cell number is correct; it is not evenly earned.

A pre-registered void condition fired, and we did not evaluate it. The pre-registration fixes three void conditions. One asks that if default initialization already trains above the majority-class rate at the shallowest depth, that dataset's primary cell is void and must be re-sited deeper. Recomputed from the executed grid, that condition **fires on** 18 **of** 24 **(dataset, architecture) cells**, including the pre-registered primary one, and the runner implements the other two conditions but never implements this one. We report it as a lapse. It moves no number here: every estimand in this experiment is a paired arm-versus-arm contrast within a matched cell, and the majority-class rate is a per-cell constant that cancels identically in all of them. Nor is its remedy outstanding — the remedy is to re-site deeper, and the reported analysis is not the primary cell but the full grid, which already spans four depths. What it costs is narrower and worth stating plainly: a condition we fixed in advance went unchecked, on an experiment whose pre-registration we claim as a contribution. The second condition (numerical infeasibility) is cleanly satisfied — 0 infeasible runs in 4512 — so no null reported here is a disguised numerical failure. The third fixes float64 accuracy and was met. Two secondary diagnostics named in the pre-registration, effective rank and Dirichlet energy, were never recorded on a real run; we do not backfill them, because they were meant to support an oversmoothing identification that §7 explicitly declines to make.

A third of the grid is outside the theory's hypothesis, and it does not matter. The three architectures aggregate with $I + A$ (GIN, Xu et al., 2019), $D^{-1/2}(A + I)D^{-1/2}$ (GCN, Kipf & Welling, 2017) and $I + D^{-1}A$ (GraphSAGE, Hamilton et al., 2017). The last is not symmetric, so 32 of the 96 cells test Proposition 1's prescription outside Proposition 1's scope — the same out-of-derivation transfer we flag for the DeZoort–Hanin and GP arms, and worth checking rather than assuming benign. Splitting the pre-registered contrast along that boundary leaves it unchanged:

$$\text{measured} - \mathcal{C}_{\text{lin}}\text{-init} \;=\; \begin{cases} -0.0039\ [-0.0099, +0.0025] & \text{symmetric (GIN, GCN), 64 cells,} \\ -0.0002\ [-0.0062, +0.0065] & \text{non-symmetric (GraphSAGE), 32 cells.} \end{cases}$$

Both intervals contain zero and overlap each other, so the null is not an artifact of including architectures the theory does not cover. The two symmetric architectures also supply the null on their

own, which is the version of the claim that lies entirely inside the scope of what we prove. Had the split come out the other way it would have been evidence that the symmetry hypothesis reaches into training; it did not, and we report that rather than the hypothetical.

I DEPTH TESTS: A VOID TEST, AND NEARBY INPUTS MERGING

Registered (`experiments/PREREG_e1_depth.md` and its Amendment 1, committed before any run): an MLP ($\tanh$, no biases) on 64 PCA features of MNIST, widths $F \in \{16, 32, 64\}$, depths $L = 2Fr$ with $r \in \{0.5, 1, 2, 4, 8\}$, plain SGD for 3000 steps in float64, the learning rate tuned per arm on a separate seed, and 8 evaluation seeds with shared weight draws. The prediction was that initializing at $\mathcal{C}_{\mathrm{q}}$ ($g_\times = 2.091$, 1.520, 1.303 from the chain) would train better than at $\mathcal{C}_{\mathrm{lin}}$ ($g = 1$), with an advantage growing with L/F. A linear network at $g = 1$ against $g = g_{\mathrm{lin}}(F)$, whose exponent gap is exactly the same $1/(2F)$ per layer, was the positive control, and it had to detect its gap at $r = 8$ in at least two widths before the $\tanh$ arms could be run.

Outcome. It detected the gap in one: the log-loss differences at $r = 8$ are -0.081 $[-0.191, +0.042]$, $+0.108$ $[+0.047, +0.167]$ and -0.072 $[-0.190, +0.057]$ at $F = 16$, 32, 64. The gate is void, and the $\tanh$ test was not run. Across the grid the control's difference runs against the registered expectation: at $r \le 1$ it favours $g_{\mathrm{lin}}(F)$ in 5 of 6 cells, by 0.05–0.22; at $r = 2$, where both arms still train well clear of chance (log-loss 1.3–1.5), it is indistinguishable from zero in all three widths; at $r \ge 4$ both arms end near chance, so those cells say little either way.

Why (found after the result). Per-layer weight-gradient norms at initialization show the linear control's gap to be uniform across depth: 9.4–11.0 nats at $F = 16$, $L = 256$, and 8.3–8.9 at $F = 64$, $L = 1024$. The gradient at layer ℓ is the forward norm up to ℓ times the backward norm from ℓ, roughly $e^{-\ell/(2F)} e^{-(L-\ell)/(2F)} = e^{-L/(2F)}$, the same at every layer, so the gap is one global factor and per-arm tuning of the learning rate removes it. The registered mechanism, a spread across layers that no single learning rate can absorb, is wrong in the linear regime. Two lesser problems compounded it: the learning-rate grid stopped too high for the deepest cells (at $F = 64$, $L = 1024$ one linear arm diverged at every rate), and for $r \ge 4$ both linear arms ended near chance.

Open, and partly explained. The $\tanh$ network at $\mathcal{C}_{\mathrm{q}}$ does *not* have a uniform batch-gradient profile: its gradients spread by 8.0 nats across layers at $F = 16$ and 4.4 at $F = 64$, largest near the input, while at $g = 1$ the profile is flat. Decomposing it (exploratory) shows that the *typical* per-example forward and backward norms are flat, as $\chi^\star = 0$ requires. The spread is a tail effect. Per-example backward norms are log-normally dispersed across inputs (standard deviation about 3 nats at the first layer), the batch sum is dominated by the extreme inputs, and that tail grows at a rate between the quenched rate 0 and the annealed mean-field rate $\frac{1}{2} \log G(g_\times)$: 0.031, 0.005, 0.0015 per layer measured, against 0.056, 0.021, 0.009 at $F = 16$, 32, 64. Since $\mathcal{C}_{\mathrm{q}}$ lies above the mean-field edge, the network there is critical in the typical sense but chaotic in mean square, and heavy-tailed per-example gradients may be the displacement's practical signature. A parameter-free model — a log-normal sum over independent inputs, with the across-input variance computed at the mean-field $q^\star$ — matched the slope at $F = 16$, 32, 64, but its correction was chosen after a gap had been seen. We therefore registered it at five settings it had not seen (width 128, $L/F \in \{8, 24\}$, batch sizes 32 and 512), and it was **falsified** ($\chi^2 = 43.3$ on 5 against 11.07). It gets the directions right — the slope falls with depth ratio and rises with batch size — but it underpredicts in all five settings, by 20–40%, while matching the across-input variance where that was measured. The error is in treating the inputs as independent; the tail rate remains open (`experiments/PREREG_e1b_oos.md`, `experiments/RESULT_e1b_oos.md`).

Do nearby inputs merge? (S-g, pre-registered in `PREREG_sg_merge.md`.) Every consequence above was built from gradient magnitudes, which the double root makes flat near $\mathcal{C}_{\mathrm{q}}$. The sign of $\chi^\star$ does control one thing sharply at depth $L \gtrsim F$: whether two nearby inputs merge. In the tangent limit $1 - c_L = (1 - \rho_L^2)\|v_L\|^2/(2\|h_L\|^2)\,(1 + o(1))$, with $v = h' - h$ and ρ the chain's state–tangent cosine, so $\log(1 - c_L) - \log(1 - c_0) = 2L\chi^\star + O_P(1)$. The test took 256 MNIST (PCA-64) inputs, each paired with an orthogonal perturbation at $1 - c_0 = 10^{-4}$, through a $\tanh$ MLP at $\sigma_{\mathrm{b}} = 0$ with $L = 16F$, 16 weight draws and a 32-point grid in $\nu\sqrt{F/3} \in [-0.5, 2.5]$. The statistic is the median over pairs and draws of $\log(1 - c_L) - \log(1 - c_0)$, and the difference is propagated exactly through $\tanh(z + d) - \tanh(z) = \tanh(d)(1 - \tanh z \, \tanh(z + d))$, because merging goes below

float64's resolution of $h' - h$ (`experiments/sg_merge.py`). The median switches sign at 1.226 $[1.158, 1.298]$ ($F = 32$) and 1.030 $[0.965, 1.129]$ ($F = 128$), against the exact-chain $\mathcal{C}_{\mathrm{q}}$ at 1.264 and 1.087. The median drift runs from about -18 nats at the lowest gains to $+9$ above $\mathcal{C}_{\mathrm{q}}$, and a linear control merges at every grid point. The registered rule returns SUPPORTED. The first run of this test was invalid: a normalization step in the runner changed the $\tanh$ network. It was caught before any analysis and is recorded as an amendment to the pre-registration. Stage 2, exploratory, asked whether merging makes lazy-regime training ill-conditioned in a way a learning rate cannot fix. It used the 64 nearest cross-class pairs in MNIST (median $1 - c_0 = 0.32$) at $F = 32$, $L = 512$ (`experiments/sg_ntk_explore.py`). At this depth the empirical NTK is dominated by its top eigenvalue at every gain, in the $\tanh$ network and the linear control alike. Each pair's separating direction sits 10^{-3} to 10^{-8} below it, and the $\tanh$ curve shows no feature at $\mathcal{C}_{\mathrm{q}}$. The registered go/no-go fails, and this consequence is not pursued.

Does the initialization gap survive training that cannot rescale it away? (S-e, pre-registered in `PREREG_se_training.md` with one amendment.) E1's linear control failed because the linear gap is a single global factor. The $\tanh$ network at $\mathcal{C}_{\mathrm{q}}$ is different: its batch gradients spread across layers. The registered Stage 0 gate confirmed this at every width. At $L = 16F$ the spread is 7.2, 6.1 and 7.5 nats at $F = 16$, 32, 64, against 1.7, 2.3 and 1.8 at $g_{\mathrm{lin}}(F)$, with input-side gradients about e^7 larger. Stage 1 trained E1's MLP, with its exact weight draws, on MNIST and CIFAR-10 (PCA-64) at $L = 4F$ and $16F$, from $g_{\mathrm{lin}}(F)$, $\mathcal{C}_{\mathrm{q}}$ and a point beyond it. It used four optimizers: SGD with a tuned learning rate, SGD with one fixed rate, per-layer-normalized SGD (LARS-style) and Adam. Each arm had 3000 steps and 8 shared seeds (1800 runs, `experiments/se_train.py`). The registered prediction, from the spread's sign, was that $\mathcal{C}_{\mathrm{q}}$ trains worse under tuned SGD at $L = 16F$ and that per-layer normalization removes the gap. The first half holds (MNIST $2/3$ widths, CIFAR-10 $3/3$), but the second fails: per-layer SGD shows the same penalty. The verdict is MIXED/INCONCLUSIVE, and the cross-layer spread is not the mechanism. Test cross-entropy (chance = 2.303):

task, depth	optimizer	$F = 16$: g_{lin} / $\mathcal{C}_{\mathrm{q}}$	$F = 32$	$F = 64$
MNIST, $4F$	SGD / per-layer / Adam	1.70/2.19, 1.35/2.06, 1.68/2.09	1.77/2.04, 1.51/1.59, 1.38/1.95	1.76/2.29, 1.59/1.61, 2.28/2.29
CIFAR-10, $4F$	SGD / per-layer / Adam	2.02/2.29, 1.99/2.19, 2.01/2.28	2.06/2.29, 2.03/2.17, 2.07/2.30	2.07/2.30, 2.08/2.16, 2.30/2.30
MNIST, $16F$	SGD / per-layer / Adam	2.25/2.29, 1.96/2.29, 2.18/2.29	2.18/2.29, 2.09/2.30, 2.28/2.29	2.28/2.29, 2.29/2.29, 2.28/2.29
CIFAR-10, $16F$	SGD / per-layer / Adam	2.26/2.30, 2.19/2.30, 2.26/2.30	2.21/2.30, 2.23/2.30, 2.30/2.30	2.29/2.30, 2.30/2.30, 2.30/2.30

Two things the registration did not anticipate. First, at $L = 16F$ most arms end at chance within 3000 steps, whatever the initialization, so the primary depth has little power: a floor effect. Second, at $L = 4F$ the post hoc contrast is large. From $g_{\mathrm{lin}}(F)$ the network trains, while from $\mathcal{C}_{\mathrm{q}}$ it stays near chance in most cells, under every optimizer. Per-layer SGD at $F \geq 32$ on MNIST and Adam at $F = 64$ are the exceptions, where the arms are close or both at chance. The point beyond $\mathcal{C}_{\mathrm{q}}$ is worse still. The mean-field edge calls both starting points chaotic, while finite width puts $g_{\mathrm{lin}}(F)$ on the ordered side of $\mathcal{C}_{\mathrm{q}}$; training strongly prefers the ordered side. Two initializations cannot tell whether trainability collapses *at* $\mathcal{C}_{\mathrm{q}}$ or simply falls with the gain, and we do not claim the former.

Where does trainability collapse? (S-e2, pre-registered in `PREREG_se2_sweep.md`.) The same MLP at $L = 4F$, $F \in \{16, 32, 64, 128\}$, was trained with SGD from 15 gains per width on a grid in $s = \nu\sqrt{F/3}$ from -0.5 to 3 (learning rate tuned per gain, then 8 seeds; MNIST, test cross-entropy capped at chance). The collapse point s_c is where the seed-mean loss crosses halfway from its best value to chance. If the collapse tracks $\mathcal{C}_{\mathrm{q}}$, $\Delta_c = s_c\sqrt{3/F}$ falls as $F^{-1/2}$; if it sits at a fixed gain, Δ_c is flat. The slope is -0.88 $[-1.05, -0.79]$, which excludes both, so the registered verdict is INCONCLUSIVE. The collapse points are $s_c = 1.07$, 1.26, 0.63, 0.57 against $s_q = 1.56$, 1.26, 1.15, 1.09 ($s_c/s_q = 0.69$, 1.00, 0.55, 0.52; the reported secondary range $[0.5, 2]$ holds). At $F = 64$ and 128 the loss jumps to chance between adjacent grid points $s = 0.5$ and 0.75, both below $\mathcal{C}_{\mathrm{q}}$, so their bootstrap intervals are narrower than the grid can resolve, but the conclusion that collapse precedes the crossing does not depend on that. At $F = 16$ and 32 the loss curves are not monotone in s. The collapse is on the crossing's scale, within a factor of two, but at large width it comes before it, and it moves toward $\mathcal{C}_{\mathrm{lin}}(F)$ faster than $\mathcal{C}_{\mathrm{q}}$ does. The $F = 128$ cells ran on eight, then four, then two GPUs, not the registered eight, because of shared-cluster limits; the code, precision and seeds were unchanged (`experiments/se2_sweep.py`). A pre-registered re-measurement on a finer grid ($\Delta s = 1/16$, 16 fresh seeds) gives $s_c = 1.02$ $[0.81, 1.07]$, 1.19 $[1.14, 1.27]$, 0.72 $[0.70, 0.72]$ and 0.51 $[0.23, 0.53]$ ($s_c/s_q = 0.66$, 0.94, 0.62, 0.47). This confirms collapse before $\mathcal{C}_{\mathrm{q}}$ at large width

and shows $s_c(F)$ is not monotone. None of six statistics computed at initialization predicts s_c to within one grid step at every width, the best (input merging) reaching a leave-one-width-out RMSE of 0.23 against 0.35 for a constant. By its registered rule the search for a predictor stopped there (`experiments/s1_stageA*.py`).

J ADDITIONAL TABLES

Audit of the remaining per-group results (E-h, pre-registered in `PREREG_eh_audit.md`; amendments move arm D to the CPUs and add arm F). The pooled re-analysis was extended to every other result that used per-group first sign changes. Arm C reuses E-g's pooled $L = 160$ values. Linear controls stay within 0.15% of $\mathcal{C}_{\mathrm{lin}}(F)$.

arm	per group (first reported)	pooled	registered verdict
A: F^{-1} at $\sigma_b = 0.3$	-1.07 $[-1.57, -0.85]$	-1.09 $[-1.36, -0.92]$	confirmed → confirmed
B: softsign ratio, $F = 256, 512$	1.97, 2.07	2.24 [1.77, 2.46], 1.71 [1.40, 2.12]	supported → supported
C: more depth removes the large-width excess	all four fall	only NCI1 leaves its $L = 160$ interval	held → **refuted**
D: exponents, erf / arctan / softsign / sublinear	-0.53, -0.56, -0.53, -0.61	-0.57, -0.61, -0.54, -0.66	range changed
E: bipartite rate (v2)	-1.16 vs -0.60	-1.95 $[-2.39, +0.78]$ vs -0.63 $[-0.75, -0.50]$	refuted → refuted
F: rate v3, adds $F = 2048$ (post hoc)	difference -0.51 $[-0.86, -0.15]$	-0.63 $[-1.11, +0.83]$	supported → **inconclusive**

In arm C, from $L = 160$ to $L = 640$ at $F = 1024$: NCI1 $7.0 \to 3.5\%$, PROTEINS $6.1 \to 5.0\%$, ENZYMES $3.3 \to 2.7\%$ and IMDB-MULTI $3.1 \to 2.4\%$. In arm D, arctan ($[-0.71, -0.52]$) and sublinear ($[-0.76, -0.56]$) exclude $-1/2$ on the steep side. The exact chain's own slope over a finite range is steeper than its asymptote (-0.61 over $F \in [16, 2048]$), so this is the pre-asymptotic regime, not a different law. All four exclude -1. The bipartite slope in arm E has a wide interval because a tangency makes each width's crossing poorly defined, which is the E-f finding again. Arm D ran on the CPUs with the original executor, and its per-group means reproduce all 384 stored corpus crossings bit-exactly. Arm F (the $F = 2048$ cells of rate v3, fresh CPU draws) pools to 0.49% $[0.14, 0.85]$ for Q_4 and 2.87% $[1.94, 3.24]$ for rr4-n16, both containing the chain's 0.23 and 2.60%; the four-width slopes are -1.12 $[-1.62, +0.28]$ and -0.50 $[-0.67, -0.44]$, on the predicted -1 and $-1/2$, but the difference interval contains 0.

Controls beyond the theory's scope (E-c, E-b; pre-registered). *ReLU* is positively homogeneous. Given the past, the surviving channels number $N \sim \mathrm{Bin}(F, 1/2)$, and for any alignment the square of a skew-normal variable is χ^2_1, so $\|v'\|^2/\|v\|^2 = (g/F)\chi^2_N$ exactly. The crossing, $\log g_\times = \log F - \mathbb{E}_N[\psi(N/2) + \log 2]$, is a simple root at He's $g = 2$, moved by $O(1/F)$: $+10.48$, $+2.31$, $+0.56$, $+0.14\%$ at $F = 16$, 64, 256, 1024 (slope -1.04). The pooled estimator recovers it at three of four widths: $+2.34$ $[2.01, 2.60]$, $+0.56$ $[0.45, 0.65]$, $+0.18$ $[0.11, 0.23]$. At $F = 16$ one of 64 realizations hit a layer with every channel negative, probability 2^{-16} per layer. The layer dies, the pooled point is undefined, and dropping that realization gives $+11.03\%$, 0.55 points above the exact value. Post hoc, the gap is noise that the interval understated: 1008 fresh fields give $+10.46\%$ $[10.33, 10.59]$, within 0.3 SE of the exact field at every gain, and disjoint 64-field batches scatter with SD 0.20 points, twice what a bootstrap over 8 groups implied (`experiments/ec_relu16_posthoc.py`). The registered verdict is SUPPORTED. *GELU* (exploratory; not odd, and its mean-field edge is not at $g = 2$) crosses $+23.4$, $+10.0$, $+7.0$, $+5.5\%$ above $g = 2$, flattening toward an offset: the displacement law needs the activation's own edge. *Graphs with* $n = 256$ (`experiments/eb_large_graphs.py`): the random 4-regular control matches the MLP (45.3 and 6.6% against 47.0 and 8.3%). Barabási–Albert, Erdős–Rényi and Watts–Strogatz graphs overlap the MLP's interval at $F = 256$ (10.0, 7.9, 8.0%) and lie below it at $F = 16$ (23.6, 37.2, 37.9%). The registered verdict is SUPPORTED, but the $F = 256$ intervals are wide (lower bounds 1.9–3.7%), so the overlap test has little power.

Audit of the zero-bias width study (E-g, pre-registered in `PREREG_eg_width_audit.md` before any run). Every tanh and linear cell of the MLP study, the 8-graph study and the 24-graph corpus was re-run on CUDA at its stored F, L, S, grid and seed groups, with all fields stored (`experiments/eg_width_audit.py`; the code path reproduces a stored CUDA crossing bit-exactly). All three studies were CPU rows, so this is a fresh replication. The per-group statistic (the mean over groups of each group's first sign change) is compared with the pooled one (the crossing of the mean of all fields). In the MLP cells, 1–8 of 8 groups change sign more than once (most at $L = 160$), and in the 8-graph cells 2–8 of 8. Pooled estimates leave the per-group interval in 45 of the 142 tanh cells. The four registered claims: (i) the corpus width exponent is -0.56

$[-0.66, -0.46]$ pooled (-0.49 $[-0.57, -0.41]$ per group in this replication; -0.54 as first reported), so it still contains $-1/2$ and excludes -1; (ii) the statement that the $L = 2560$ estimates lie below $\mathcal{C}_{\mathrm{q}}$ **fails**: pooled, only 3 of 6 lie below, and the pooled interval contains $\mathcal{C}_{\mathrm{q}}$ in 5 of 6 widths (per group, 3 of 6); (iii) no width shows the $L = 2560$ interval entirely below the $L = 160$ one, so the effect is not a finite-depth transient; (iv) $cF^{-1/2}$ still beats cF^{-1}, $cF^{-1/3}$ and $c/\log F$ by more than a factor of 9 in weighted squared error. Linear controls stay within 0.15% of $\mathcal{C}_{\mathrm{lin}}(F)$. The outcome is CLAIM CHANGED, through (ii), and the paper now reports pooled values.

F	$\mathcal{C}_{\mathrm{q}}$	MLP $L = 160$ per group $\to$ pooled	MLP $L = 640$ per group $\to$ pooled	MLP $L = 2560$ per group $\to$ pooled	corpus $L = 160$ per group $\to$ pooled	8 graphs $L = 160$ per group $\to$ pooled
16	50.2	$43.3 \to 50.8$	$44.6 \to 47.0$	$48.6 \to 51.1$	$29.5 \to 37.3$	$22.0 \to 28.2$
48	20.7	$17.8 \to 17.8$	$17.9 \to 19.8$	$20.8 \to 21.1$	$16.7 \to 22.0$	—
128	11.4	$6.4 \to 10.1$	$9.4 \to 10.7$	$10.9 \to 10.8$	$9.1 \to 11.2$	—
256	7.7	$4.6 \to 5.4$	$7.2 \to 8.3$	$6.5 \to 6.5$	$8.1 \to 8.2$	$6.4 \to 6.6$
512	5.3	$4.9 \to 7.1$	$3.7 \to 4.5$	$5.1 \to 5.4$	—	—
1024	3.7	$3.8 \to 4.2$	$2.9 \to 4.0$	$2.5 \to 3.1$	—	$6.2 \to 6.7$
2048	2.6	$2.7 \to 2.8$	—	—	—	—

The per-group statistic scatters by several points between replications (MLP $L = 160$, $F = 16$: 39.4 first reported, 43.3 here), and it is biased low where many groups change sign more than once. Pooling removes most of the shortfall that the paper had attributed to finite depth. As an exploratory cross-check, the exact chain run through the same estimator with common random numbers across the grid (`experiments/eg_chain_crosscheck.py`) gives smoother fields, with at most 1 of 8 groups changing sign more than once, so it does not reproduce the estimator's noise. Its crossings scatter about $\mathcal{C}_{\mathrm{q}}$ by one to two points. Exploratory as well: on the corpus at $F = 256$ the twelve bipartite graphs (trees, stars, complete bipartite graphs, NCI1-g2 and PTC_MR-g1) average 6.8% and the twelve non-bipartite ones 9.7%, the ordering of Theorem 3(i).

The $F = 6$ corpus crossings, converged. The crossings behind Figure 18 were first measured as the first sign change of a field averaged over $S = 32$ realizations. Near the crossing that field sits at about -0.01 while one cell's standard error at $S = 32$ is about 0.009, so a single noisy cell can fire the "first sign change" early, and a noisy negative cell can delay it. Rerunning the identical estimator with more realizations — the first 32 reproducing the stored crossings exactly — and doubling S until the averaged field changes sign *exactly once* (a rule fixed before the run; `experiments/e2_step4_converge.py`) settles every graph by $S \leq 4096$. The $S = 32$ crossing fired early on 11 graphs, late on 3, and within 0.02 on 3. Displacement beyond $\mathcal{C}_{\mathrm{lin}}(6) = 0.9707$ moves from 26–72% (mean 56%) to **45–91**$\%$ (**mean 63**$\%$); requiring the field to exceed $+2$ standard errors gives 50–99% (mean 68%).

graph	$S = 32$ crossing	%	S used	converged	%
MUTAG-1	1.385	43	512	1.423	47
MUTAG-0	1.223	26	512	1.409	45
repo-far	1.560	61	1024	1.721	77
PROTEINS-0	1.672	72	512	1.599	65
repo-middle	1.610	66	512	1.680	73
WS-n16	1.560	61	512	1.684	73
WS-n32	1.423	47	512	1.443	49
repo-close	1.635	68	1024	1.844	90
PROTEINS-1	1.435	48	1024	1.433	48
BA-n16	1.498	54	1024	1.446	49
ENZYMES-1	1.635	68	512	1.560	61
ENZYMES-0	1.485	53	512	1.504	55
ER-n16	1.522	57	512	1.571	62
BA-n32	1.335	38	512	1.454	50
ER-n32	1.410	45	512	1.524	57
IMDB-BINARY-0	1.672	72	1024	1.771	82
IMDB-BINARY-1	1.660	71	4096	1.852	91

The 65-graph spread corpus, converged (post hoc). The pre-registered spread test regressed crossings measured as the mean of 8 first sign changes, each from a field averaged over only $S = 8$ realizations. Its verdict, **FAIL** ($R^2 = 0.179$ for the regularity ratio against a 0.60

gate), was fixed in advance and stands. Rerunning the same estimator — the stored crossings reproduce exactly for all 65 graphs — and applying the same convergence rule as above (`experiments/e2_step6_e8_converge.py`) shows the targets were noise-driven. Every one of the 45 graphs that cross had fired early, by $+0.006$ to $+0.351$ (mean $+0.105$); 14 graphs that had registered a crossing have *none* within $\sigma_s \leq 3$ once averaged (the field stays negative at $S = 4096$); and the 6 recorded as non-crossing remain so. Re-running the test's own regression on the 45 converged crossings, *post hoc*: the regularity ratio falls to $R^2 = 0.009$ with a slope of the wrong sign (CI $[-1.88, +2.67]$), its increment over $\mu_{\max}$ to $\Delta R^2 = +0.061$ ($p = 0.10$, from $+0.104$, $p = 0.014$), while $\mu_{\max}$ alone rises to $R^2 = 0.243$. The partial support once reported for the heterogeneity mechanism was an artifact of noisy targets.

Bipartite test, per-pair numbers (§5, test 3). Displacement (%) over $\mathcal{C}_{\text{lin}}(F)$, tanh, node-uniform input, 8 seed-groups. Primary rule (pre-registered, 0051930): falsified if any bipartite mean lies inside its partner's 95% interval at $F = 1024$; all three do, and the partners read 43–68% of their exact $\mathcal{C}_q$ (3.71%) at $L = 640$. Follow-up rule (pre-registered, f835ee2): at $F = 512$, $L = 2560$, against the chain's predictions 0.92% (bipartite) and 5.32% (partners); inconclusive because both control lower bounds fall below the 2.66% threshold. Its pre-registered power fix (PREREG_bipartite_v2, 32 groups $\times$ S=4, controls only) re-ran the controls: rr4-n16 3.92% $[3.25, 4.58]$, C15 3.60% $[2.98, 4.21]$ — both lower bounds above the threshold, so the v2 rule returns **SUPPORTED**: both bipartite upper bounds (1.84, 2.01) lie below 2.66 and match the chain's 0.92%. The controls still read 68–74% of their exact $\mathcal{C}_q$ at $L = 2560$ (pooled, 99–102%; audit below). The bipartite arm was then re-run at the same 32-group power (PREREG_bipartite_v3 and its amendment): the class rule is supported and the chain's quantitative 0.92% refuted, and two irregular pairs matched in $\mu_{\max}$ (PREREG_bipartite_irregular) separate the same way. At $F = 64$ the bipartite graphs read 1.3, 1.8, 5.5% against 15.2, 15.4, 14.9%; at $F = 256$, 1.7, 1.9, 2.6% against 7.4, 6.3, 4.7%.

Audit of the bipartite tests (E-f, pre-registered in `PREREG_ef_bipartite_audit.md` before any run). The eight graphs behind the two confirmed rows were re-run on CUDA at the registered settings with all 32×4 fields stored (`experiments/ef_bipartite_audit.py`). The six CUDA-origin rows reproduce their stored group crossings exactly; rr4-n16 and C15, originally CPU rows, are fresh replications (4.10 and 3.96% paper-style, against 3.92 and 3.60). Most groups change sign more than once, so the paper-style statistic, the mean over groups of each group's *first* sign change, is biased low; averaging each group's last sign change instead gives 2.9–3.9% (bipartite) and 5.8–7.3% (non-bipartite). The registered primary statistic of the audit pools all fields, with a bootstrap over groups:

graph	paper-style	groups with > 1 sign change	pooled	pooled field sign changes
Q4 (bip.)	1.69 [1.10, 2.32]	15/32	1.46 [0.40, 3.51]	1
C16 (bip.)	1.51 [0.95, 2.13]	17/32	2.07 [0.16, 3.74]	1
rr4-n16	4.10 [3.55, 4.59]	22/32	5.44 [4.73, 5.70]	1
C15	3.96 [3.35, 4.53]	20/32	5.28 [4.64, 5.61]	1
NCI1-g2 (bip.)	1.59 [1.13, 2.07]	24/32	2.54 [1.01, 3.73]	3
NCI1-g1	3.45 [2.74, 4.13]	28/32	6.11 [5.36, 6.78]	1
PTC_MR-g1 (bip.)	1.42 [0.96, 1.92]	24/32	2.87 [0.42, 3.74]	1
MUTAG-g2	3.73 [3.10, 4.33]	28/32	5.52 [4.85, 5.81]	1

Re-applying the registered rules to the pooled values: the v3 class rule becomes *inconclusive* (bipartite upper bounds 3.51 and 3.74 exceed the 2.66% threshold, although no bipartite interval overlaps its partner's); the v3 quantitative rule no longer refutes the chain's 0.92%, but only because the intervals are wide; the irregular rule is supported for NCI1 and inconclusive for PTC_MR against MUTAG (2.87 exceeds half of 5.52 by 0.11). The audit's outcome is VERDICT CHANGED. The reason is visible in the pooled fields ($\times 10^4$, standard error about 1):

displacement over $\mathcal{C}_{\text{lin}}(F)$	0%	1%	2%	3%	4%	6%
Q4, C16 (bipartite)	$-7.3, -7.1$	$-0.8, -1.1$	$+0.5, -0.1$	$+1.7, +1.7$	$+2.8, +2.7$	$+9.5, +9.2$
rr4-n16, C15	$-12.4, -12.3$	$-8.9, -8.9$	$-7.7, -7.5$	$-6.8, -7.0$	$-4.3, -4.8$	$+3.5, +3.3$
NCI1-g2, PTC_MR-g1 (bipartite)	$-7.0, -5.9$	$-1.6, -0.6$	$-1.2, -0.4$	$-0.1, +0.1$	$+1.5, +1.9$	$+5.8, +6.9$
NCI1-g1, MUTAG-g2	$-9.7, -9.9$	$-8.0, -7.7$	$-8.3, -7.8$	$-6.2, -6.0$	$-5.2, -4.5$	$-0.3, +3.2$

Bipartite fields lie on zero across roughly 1–3% and non-bipartite fields well below it, as the tangency of Theorem 3(i) predicts. The gap between the classes, 6–9 $\times 10^{-4}$ over 1–4%, is of the size of the predicted alignment cost $1/(2F) = 9.8 \times 10^{-4}$ (post hoc). A first sign change is the wrong statistic for a tangency.

pair (bip. / partner)	F=1024, L=640	F=512, L=2560
Q_4 / rr4-n16	1.85 [0.70, 3.65] / 2.53 [1.35, 3.87]	0.80 [0.16, 1.84] / 3.82 [2.43, 5.18]
C_{16} / C_{15}	1.25 [0.68, 1.94] / 1.59 [0.86, 2.56]	0.84 [0.15, 2.01] / 2.93 [1.78, 4.06]
$K_{4,4}$ / K_5	1.08 [0.45, 1.74] / 1.93 [0.61, 3.54]	—
at 32 seed groups $\times$ $S = 4$		
Q_4 / rr4-n16	—	1.69 [1.10, 2.32] / 3.92 [3.25, 4.58]
C_{16} / C_{15}	—	1.51 [0.95, 2.13] / 3.60 [2.98, 4.21]
NCI1-g2 / NCI1-g1[‡]	—	1.59 [1.13, 2.07] / 3.45 [2.74, 4.13]
PTC_MR-g1 / MUTAG-g2[‡]	—	1.42 [0.96, 1.92] / 3.73 [3.10, 4.33]

Why the magnitude rule failed. It compared $\widehat{\mathcal{C}}_{L,S}$, a finite-$L$ grid-based first sign change, with $\mathcal{C}_{\mathrm{q}}$, the chain's true crossing. Through the paper's own estimator the chain gives 1.20% [0.44, 2.06] (bipartite) and 3.96% [2.91, 4.92] (standard), both containing the measurements, against 0.41 and 1.59% when grid draws are independent: the protocol moves the same chain by more than the effect the rule asked about (`theory/bipartite_estimator.py`).

The rate test, and why it had no power. Pre-registered across $F = 256, 512, 1024$ at $L = 2560$ with 32 seed groups and one protocol (so grid coupling cancels in the slope), the bipartite arm (Q_4) decays with slope -0.51 $[-0.91, -0.09]$ — 1.87, 1.60, 0.93% — against -0.65 $[-0.82, -0.51]$ for rr4-n16 (6.43, 4.13, 2.60%). The rule asked the bipartite CI to exclude $-1/2$ and sit below its partner's; neither holds, so it is refuted as recorded. It could not have succeeded: passing the chain through the paper's own estimator at the same three widths gives 3.20, 1.35, 1.58%, slope -0.51, while that chain's own crossing $\mathcal{C}_{\mathrm{q}}$ decays at -1.03 (1.78, 0.87, 0.43% pooled over 32 seeds; the 8-seed values registered earlier were 1.67, 0.92, 0.42%, `theory/bip_chain_pooled.py`). A grid-based sign-change locator cannot resolve a displacement falling towards its own resolution, so the measurement is what the chain's -1 law would produce under this estimator.

The diagnosis is resolution, not the locator: v1's step was 0.51% against a 0.4% signal at $F = 1024$. A pre-registered re-run (PREREG_bipartite_rate_v2) scaling each cell's grid half-width to the displacement that cell expects — step $\approx 0.34\times$ the signal everywhere, 28–32 of 32 groups resolving — gives bipartite 1.09, 0.76, 0.22% (slope -1.16 $[-1.70, -0.63]$) against rr4-n16 5.86, 4.17, 2.54% (slope -0.60 $[-0.81, -0.41]$). The point estimates land on the predicted -1 and $-1/2$ and the bipartite interval excludes $-1/2$ ($\mathbb{P} = 0.991$ over resamples), but the intervals overlap, so the rule records a second refutation; a post-hoc difference of slopes is -0.56 $[-1.14, +0.01]$, negative in 97.2% of resamples. Three widths at 32 groups cannot separate two slopes at 95%. A third run (PREREG_bipartite_rate_v3) added $F = 2048$ — bipartite 0.13% [0.07, 0.20], 26 of 32 groups resolving; non-bipartite 1.75% [1.34, 2.18] — giving four-width slopes -1.11 $[-1.43, -0.77]$ and -0.59 $[-0.73, -0.48]$ and a difference of -0.51 $[-0.86, -0.15]$, negative in 99.6% of resamples: by that pre-registration's rule, supported, but only per group. Pooled (E-h arm F, below), the difference is -0.63 $[-1.11, +0.83]$ and the verdict is inconclusive. Its own first line records that the extension was chosen after v2's direction was known, so it is reported as an exploratory power extension and the earlier verdicts stand as measured. One analysis choice deviated from the v2 pre-registration and is recorded in the code: it specified resampling groups jointly across widths, which a narrow grid makes undefined when a cell resolves fewer than 32, so each width is resampled independently — slightly wider intervals, unchanged point estimates (`experiments/bipartite_rate.py`, `bipartite_rate_v2.py`, `theory/bipartite_rate_estimator.py`).

[‡]irregular pairs matched in $\mu_{\max}$ (2.53; 2.55/2.60), measured with the float64 CUDA estimator. Passing the chain through this same estimator (16 groups) gives 1.20 [0.44, 2.06] bipartite and 3.96 [2.91, 4.92] non-bipartite when grid draws are shared as the network shares weights, and 0.41/1.59 when they are independent (`theory/bipartite_estimator.py`, post hoc): the protocol moves the same chain by more than the class gap it is being asked to resolve. Chain reference at $F = 512$: 0.92% (bipartite branch, as registered; 0.87% [0.83, 0.90] over 32 seeds), 5.32% (standard).

The bipartite branch at field level (E1, pre-registered in `PREREG_e1_bipartite_field.md`). Locating a root on a flat field is ill-conditioned (E-f, and the E-d post hoc above), so E1 tests the field. The statistic is B, the mean of $F \cdot 2\chi^\star$ over $s = \nu\sqrt{F/3} \in \{0.30, 0.45, \ldots, 0.90\}$ (below $\mathcal{C}_{\mathrm{q}}$). It was measured with the independent Benettin estimator of E-d, at $L = 4 \times 10^4$ with 192 groups. The pairs are fresh: Q_5 against a random 5-regular graph; the bipartite double covers of random 3- and 4-regular graphs against random 3- and 4-regular graphs, all with $n = 32$; and a random tree against

the same tree with one triangle. C33 was replaced before any run, because its smallest eigenvalue, -1.991, makes a second block compete at small F. The exact chain's predictions were fixed first: $B = +0.162, +0.230, +0.277$ (bipartite) and $-0.625, -0.611, -0.599$ at $F = 128, 256, 512$, so the gap is $0.787, 0.841, 0.876$ (leading order 1). Measured, the gap is 0.783 $[0.765, 0.801]$, 0.842 $[0.819, 0.867]$ and 0.883 $[0.846, 0.921]$. Every network B lies within about one standard error of the chain, so both registered rules hold: SUPPORTED. With node-uniform input and $\sigma_{\mathrm{n}} d = 1/2$, every d-regular graph reduces to the same two blocks, so with common seeds the three regular pairs replicate one test; the registration should have said so. The tree pair, outside the theorem, separates by 0.50, 0.56 and 0.56 (CIs exclude 0). The bipartite tree's field is negative, -0.17 to -0.11, unlike the regular $+s^2$, but it stays about 0.5 above its partner (`experiments/e1_field.py`, `RESULT_e1_bipartite_field.md`).

The width law on a larger corpus (E2, pre-registered in `PREREG_e2_corpus.md`). The corpus has 61 non-bipartite, non-regular graphs: the 40 of the existing corpus ($n \leq 97$) and 21 new ones, seven families at $n \in \{128, 256, 500\}$ (BA with $m = 2$ and 3, Erdős–Rényi, Watts–Strogatz, two-block SBM, random geometric, power-law cluster). The run used $F \in \{16, \ldots, 256\}$, $L = 400$, 64 fields per cell and the pooled crossing. Rule H1 required each family's slope of $\log \Delta_\times$ on $\log F$ to lie in $[-0.8, -0.4]$ with its CI inside $(-1, 0)$; rule H2 required the between-graph spread of $s_\times$ to fall with F.

- *H1.* It holds for the old corpus (-0.485 $[-0.505, -0.478]$), Erdős–Rényi (-0.54) and SBM (-0.59). It fails narrowly for Watts–Strogatz (-0.40) and geometric graphs (-0.34). For the hub-dominated families it fails outright: BA $m = 2$ gives $+0.00$ $[-0.23, +0.01]$, BA $m = 3$ gives -0.16, and power-law cluster gives -0.15 (maximum degree 6–15$\times$ the mean).
- *H2.* The spread rises with F (slope $+0.37$ $[+0.17, +0.50]$).

No family's CI lies outside $(-1, 0)$, so the verdict is INCONCLUSIVE, and the $n \leq 97$ limitation stays.

Two post hoc diagnostics, not registered, locate the failure.

- Split by hubness, the 45 graphs whose maximum degree is below $3\times$ the mean give slope -0.477. The 16 others give -0.27.
- Depth matters on hub graphs. At $F = 256$ the registered $L = 400$ gives $\Delta_\times = 0.34$ on BA $m = 2$ ($n = 256$). At $L = 1600$ and 6400 it gives 0.16 and 0.17, while Erdős–Rényi gives 0.12 at both depths and 0.14 at $L = 400$.

The depth-converged values still separate: 0.16–0.17 on the two hub graphs against 0.12 at $F = 256$, and 0.27–0.29 against 0.19 at $F = 128$, with disjoint CIs. So hub-dominated graphs keep a displacement about $1.4\times$ larger than the MLP-like graphs at these widths, on one graph per family; E4 (below) tests larger F (`experiments/e2_corpus.py`, `e2_posthoc.py`, `RESULT_e2_corpus.md`).

The hub residual at larger width (E4, pre-registered). On 18 hub-dominated graphs (BA $m = 2, 3$ and power-law cluster, $n \in \{128, 256\}$, three generator seeds) against 6 Erdős–Rényi controls, at $L = 1600$, the hub/ER displacement ratio is 1.20 $[1.15, 1.25]$, 1.39 $[1.29, 1.51]$ and 1.25 $[1.05, 1.35]$ at $F = 64$, 128 and 256. This confirms E2's post hoc finding on many graphs. At $F = 512$ the registered depth check failed: doubling the depth moved one control by 17% but the hub graphs by at most 10%. The $F = 512$ ratio (0.99) is biased low and was withdrawn as registered, so whether the gap closes remains open: INCONCLUSIVE. The graph factor $r_6 = \pi_6/\pi_4^2$ correlates with the excess in the registered direction (Spearman $+0.46$ at $F = 256$) (`experiments/e4_hub.py`, `RESULT_e4_hub_residual.md`).

E4b: the same test with deeper runs on new graphs (pre-registered). On 24 new instances (generator seeds 21–23) at $L = 3200$, the ratio at $F = 256$ is 1.35 $[1.23, 1.46]$, which replicates E4. At $F = 512$ the registered depth check failed again: doubling the depth to 6400 moved the ER control and one BA graph by 15% each. The ratio at $F = 512$ (1.09 $[0.97, 1.26]$) and the registered trend test were withdrawn: DEPTH-LIMITED. At $F = 512$ the control fields converge

more slowly than the hub fields, so whether the hub excess closes at large width remains open (`experiments/e4b_hub.py`, `RESULT_e4b_hub_residual.md`).

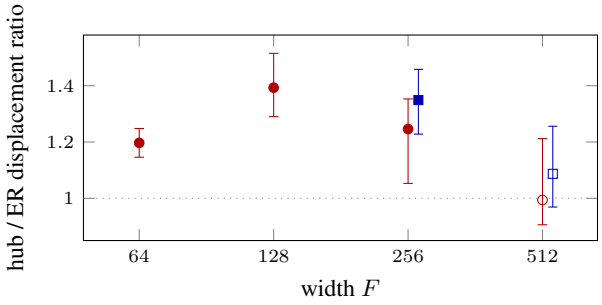


Figure 14: The hub residual (E4, E4b; pre-registered). The ratio of the depth-converged displacement of 18 hub-dominated graphs to that of 6 Erdős–Rényi controls, on two independent sets of graph instances, with 95% bootstrap intervals. Red circles: E4 ($L = 1600$, generator seeds 11–13); blue squares: E4b ($L = 3200$, seeds 21–23). Filled: registered estimates. Open: $F = 512$, withdrawn in both runs because the depth check failed (the controls had not converged). E4b's points are offset slightly in F for legibility.

Small-width graph details (§5). The only complete graphs in the 65-graph corpus are among its six that never cross. A random 4-regular graph is displaced like the MLP: $+49.7\%$ against $+50.8\%$ at $F = 16$, and $+7.0\%$ against $+5.4\%$ at $F = 256$.

The boundary-layer law on irregular graphs (E3, E3b, pre-registered). Theorem 5 is proved where the chain is exact. The cubic reduction (the remark after it) predicts the same boundary-layer statistics on any graph whose non-top modes relax within F layers. E3 compared seven graphs ($n = 48$) with a 16-seed MLP network arm and was INCONCLUSIVE: its intervals (± 0.08) were too wide. It confirmed the mechanism, since the state variance scaled as $1/\pi_4$ within 0.94–1.06 while π_4 varied fourfold. E3b replaced everything that had been looked at: new graph instances, independent seeds per graph, and the exact chain (64 seeds) as reference. SUPPORTED: all six in-scope graphs agree with the chain at $F = 256$, every interval of the difference in $2F\chi^\star$ inside $[-0.08, +0.06]$ at $x \in \{0.25, 0.6, 1, 2, 4\}$ and $\mathbb{E}\rho^2$ within ± 0.05. At $F = 64$ three of six do not. Post hoc, the operative condition is relaxation: a geometric graph outside the registered gap rule (non-top relaxation about 43 layers) passes at $F = 256$ and fails at $F = 64$, and a star with one triangle (about 300 layers) fails at both (`experiments/e3_layer.py`, `e3b_layer.py`, `RESULT_e3b_layer.md`).

Deep random reservoirs (S2′, pre-registered). An exploratory pilot found no accuracy optimum at $\mathcal{C}_{\mathrm{q}}$ for a deep random message-passing reservoir with a trained linear readout, but a plateau between $\mathcal{C}_{\mathrm{lin}}(F)$ and roughly $\mathcal{C}_{\mathrm{q}}$ at depth $4F$. The registered test (six TU datasets, $F \le 64$, $L \le 8F$) asked whether accuracy is organized by the reservoir's measured nonlinear log-gain better than by the linear-theory log-gain $L\nu/2$ (H1), and whether the measured band $[\mathcal{C}_{\mathrm{lin}}(F), \hat{C}]$ is the usable band (H3). H1 held on five datasets (residual ratios 0.72–0.92) and failed on IMDB-MULTI, whose curve is flat. H3 held on four, and the rule needed five: INCONCLUSIVE (`experiments/s2p_reservoir.py`, `RESULT_s2p_loggain.md`).

K NONZERO BIAS ON GRAPHS (§5, TEST 1)

Bias destroys the origin fixed point, so $\mathcal{C}_{\mathrm{lin}}$ is undefined; $\mathcal{C}_{\mathrm{ind}}$ is well-posed and its root is simple. On eight graphs at $F = 6$, all 32 cells at $\sigma_{\mathrm{b}} \in \{0.05, 0.1, 0.3, 0.6\}$ cross, rising monotonically with σ_{b}; $\mathcal{C}_{\mathrm{ind}}$ under-predicts every cell (median relative error 26.7%), and all 32 lie above the proved lower bound of Proposition 4. On the MLP the finite-width displacement from $\mathcal{C}_{\mathrm{ind}}$ decays as $F^{-1.09}$ (§5), as the simple root requires.

Robustness detail (§5). Restricting the 24 graphs to the 11 with $n \ge 48$ changes the $F = 16$ and $F = 256$ displacement values by under a point, and over 576 cells varying L and S the displacement

moves by 2.7 points across a 16× change of S at $L = 640$. $D^{-1}A$ is similar to $D^{-1/2}AD^{-1/2}$ but not symmetric, and is the only operator choice that loses the displacement.

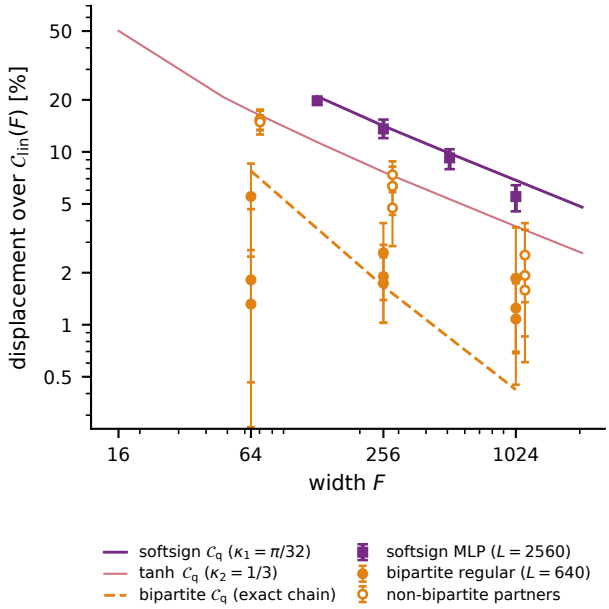


Figure 15: Tests 2–3 of §5 in displacement units: softsign against $\tanh$ $\mathcal{C}_q$ (exact chain) with the full-network softsign estimates at $L = 2560$, and the bipartite chain branch against the $L = 640$ full-network bipartite/partner points (the falsified primary comparison is at $F = 1024$; per-pair numbers above). Figure 4(b) shows the same chain data in the critical-window normalization $\sqrt{F}\,(\log g_\times - \log g_{\text{lin}}(F))$.

Two candidate statistics for the activation ordering are not identifiable. The mean-square derivative under the induced state law, $\mathbb{E}[\phi'(Z)^2]$, orders the five activations exactly (Spearman $+1.000$). So does $\mathrm{CV}[\phi'(Z)]$. These are not separable: constructing 200 admissible odd activations spanning a $5\times$ range of CV and measuring both at $F = 6$ gives $\mathrm{corr}(\mathbb{E}[\phi'^2], \mathrm{CV}[\phi']) = +0.959$. Matching one while spreading the other is infeasible on this class: the widest $\mathbb{E}[\phi'^2]$ spread attainable inside a matched-CV band is 0.034, against a *total reachable range* of 0.054 (both under the mean-field state law, over 4562 candidates). The self-consistency equation 3 absorbs activation shape — a more saturating ϕ shrinks $q^\star$, pushing the state back toward the origin where $\phi' \approx 1$. **The Spearman $+1.000$ stands; only the attribution changes.** We report this as a measured non-identifiability on this class rather than a preference between two candidates. Relatedly, $\mathbb{E}[\phi'^2]$ is a low-contrast predictor: its per-seed standard deviation at $F = 6$ is 0.027, comparable to its entire reachable range.

Operators. On the graph-active ray at $F = 6$, 15 distinct values of $\max \chi^\star$ across 16 (graph, operator) cells confirm the graph participates.

operator	symmetric	$\mu_{\max}$	crosses	displacement (finite-width)
raw A	yes	4.15–5.63	4/4	+54% to +68%
self-loop $A + I$	yes	5.15–6.63	4/4	+55% to +110%
symmetric norm. $D^{-1/2}AD^{-1/2}$	yes	1.000	4/4	+45% to +80%
random walk $D^{-1}A$	**no**	1.000	**0/4**	—

Four graphs, $\tanh$, $F = 6$, $L = 160$, $S = 32$; each operator's ray is set by its own $\mu_{\max}$, so $\mathcal{C}_{\text{lin}}$ predicts 0.8660 and the finite-width law of Appendix G predicts 0.9729 throughout (the law is applied, not re-measured per operator; against 0.8660 the three rows read $+73$–89%, $+74$–136% and $+63$–102%). Four graphs and no confidence intervals — no per-seed spread was retained for

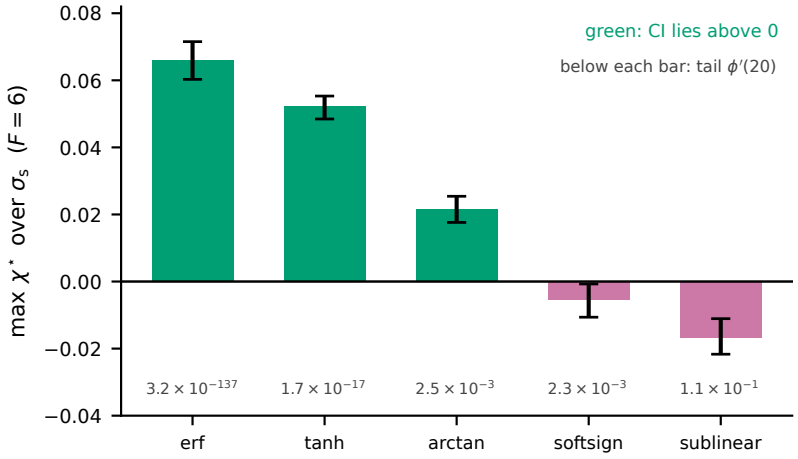


Figure 16: Activation family on `pair-close`, $F = 6$, $L = 160$, 8 seed-groups of $S = 32$: $\max \chi^\star$ over σ_s with 95% bootstrap CIs, green where the CI lies above zero. Activation identity shifts $\chi^\star$ significantly. Printed below each bar is the derivative tail $\phi'(20)$; its order matches the bar order except for arctan and softsign, whose tails differ by 10%, and it cannot be separated from $\mathbb{E}[\phi'(Z)^2]$ as the explanation (text below).

this sweep — so this is a robustness check, not a second measurement of the displacement. The random-walk operator is the sole non-crossing case and the sole non-symmetric one. The last two rows are the isolating comparison: $D^{-1}A = D^{-1/2}(D^{-1/2}AD^{-1/2})D^{1/2}$, so the two are *similar* matrices with identical spectra — the shared $\mu_{\max} = 1.000$ in the table understates how tightly they are matched — and they differ only in normality. Only the symmetric one crosses, which is why we read this row as evidence that Proposition 1's symmetry hypothesis is load-bearing rather than as a case the theory merely fails to cover.

Activations. Five odd activations with $\phi'(0) = \sup|\phi'| = 1$, on the single graph `pair-close` at $F = 6$, $L = 160$, 8 seed-groups of $S = 32$, $\max_{\sigma_s} \chi^\star$ with 95% CI: erf $+0.0661$ $[+0.0603, +0.0715]$ $(8/8$ groups crossing), tanh $+0.0522$ $[+0.0485, +0.0553]$ $(8/8)$, arctan $+0.0216$ $[+0.0176, +0.0254]$ $(8/8)$, softsign -0.0056 $[-0.0106, -0.0007]$ $(2/8)$, sublinear -0.0168 $[-0.0217, -0.0111]$ $(1/8)$. All pairwise comparisons have disjoint intervals. So on this graph softsign and sublinear, although admissible, do not cross at $F = 6$; on the 24-graph width corpus they cross on $18/24$ graphs at $F = 6$ and on every graph from $F = 16$ (Figure 19). The identity, as the linear control, gives $+1.0231$ $[+1.0196, +1.0272]$ $(8/8)$. The two non-odd activations we tried fall outside Propositions 1–2 and behave accordingly: GELU gives -0.185, and ReLU collapses the tangent to numerical zero, so its reported exponent is a log-floor artifact rather than a growth rate. We list them for scope, not comparison.

The $O(1/F)$ cost at $\sigma_b > 0$. At $\sigma_b > 0$ the mean-field root is simple, and the chain of Appendix C.9 fluctuates by $O(F^{-1/2})$ about a fixed point $q^\star$ that is bounded away from zero. That makes the $O(1/F)$ cost provable by plain moment arguments: there is no corner at $q = 0$, and both coordinates contract at rates bounded away from 1. With ξ standard normal and every quantity evaluated at $q^\star$, write $P = \mathbb{E}\phi(\sqrt{q}\,\xi)^2$, $S = \mathbb{E}\phi'(\sqrt{q}\,\xi)^2$, $Q = \mathbb{E}[\phi'(\sqrt{q}\,\xi)^2\xi^2]$, $T = \mathbb{E}[\phi\phi'(\sqrt{q}\,\xi)\xi]$ and $qK = S - Q$; also $f(q) = gP(q) + \sigma_b^2$, $P_b = P + \sigma_b^2/g$, $\kappa = T^2/(SP_b)$, $v = \mathbb{E}[\phi^2\phi'^2]/(SP_b)$, $c = \phi'(\sqrt{q^\star}\xi)^2$ and $V = g^2\mathrm{Var}\,\phi(\sqrt{q^\star}\xi)^2$.

Proposition 14 (the $O(1/F)$ cost at nonzero bias)**.** *Let ϕ satisfy the hypotheses of Theorem 3, let $\sigma_b > 0$, take the MLP, and let J be a compact interval of gains. Uniformly for $g \in J$,*

$$2\chi^\star = \log G(g) + \frac{a(g)}{F} + O(F^{-3/2}), \tag{22}$$

$$a = -\frac{3\mathbb{E}c^2/S^2 - 1}{2} - \frac{qK}{S}\frac{v}{1-\kappa} + \frac{(\log S)' f'' V}{2(1-f')(1-f'^2)} + \frac{(\log S)'' V}{2(1-f'^2)},$$

with f', f'', $(\log S)'$ and $(\log S)''$ taken at $q^\star$; the three groups are a_{samp}, a_{align} and a_{state}. If $g_{\mathrm{mf}} \in \operatorname{int} J$ is the only root of $\log G$ in J and $\partial_g \log G(g_{\mathrm{mf}}) \neq 0$, every sign change of $\chi^\star$ in J lies at $g_\times = g_{\mathrm{mf}} - a(g_{\mathrm{mf}})/(F\,\partial_g \log G) + O(F^{-3/2})$.

Proof. (1) State space and contractions. As in Step 1 of Theorem 3, (q, ρ) is a Markov chain with $q' = gA/F + \sigma_{\mathrm{b}}^2 \in [\sigma_{\mathrm{b}}^2, q_{\max}]$, where here $q_{\max} := g \sup \phi^2 + \sigma_{\mathrm{b}}^2$. The lower end $\sigma_{\mathrm{b}}^2 > 0$ keeps every denominator below bounded away from 0, so all constants are uniform in $g \in J$. Because ϕ is odd with ϕ' non-increasing on $(0, \infty)$, $\psi(y) = \phi(y)/y$ is non-increasing in $|y|$. So $P(q)/q = \mathbb{E}[\xi^2 \psi(\sqrt{q}\,\xi)^2]$ is non-increasing, and $f(q)/q$ is strictly decreasing. Hence $f(q) = q$ has exactly one root $q^\star$, with $f(q) > q$ below it and $f(q) < q$ above. Since f is increasing, $|f(q) - q^\star| < |q - q^\star|$ for every $q \neq q^\star$. The ratio is continuous on the compact state space, and at $q^\star$ it tends to $f'(q^\star) < f(q^\star)/q^\star = 1$, so $\lambda_q := \sup_{q \neq q^\star} |f(q) - q^\star|/|q - q^\star| < 1$, uniformly in $g \in J$. For the alignment, the population cosine obeys $r^2 = \rho^2 T^2/(P_b(\rho^2 Q + (1 - \rho^2)S))$. Since $S \geq Q$ we have $\rho^2 Q + (1 - \rho^2)S \geq Q$, and Cauchy–Schwarz gives $T^2 \leq PQ$, so $r^2 \leq \lambda_\rho \rho^2$ with $\lambda_\rho := \sup_q P/P_b < 1$, because $\sigma_{\mathrm{b}} > 0$.

(2) Ergodicity. After any one step the chain is away from the corner. Here ρ' is the chain variable, the correlation of the next layer's channel pair given the past, and its denominator carries the bias variance: $\rho' = C/\sqrt{B(A + F\sigma_{\mathrm{b}}^2/g)}$ with A, B, C the sums over the current draws. It is not the cosine of two Gaussian vectors, so the cap below is deterministic. By Cauchy–Schwarz, $\rho'^2 \leq A/(A + F\sigma_{\mathrm{b}}^2/g) \leq \sup \phi^2/(\sup \phi^2 + \sigma_{\mathrm{b}}^2/g) =: (1 - \eta_b)^2 < 1$, surely, so $(q', \rho') \in K_1 := [\sigma_{\mathrm{b}}^2, q_{\max}] \times [-1 + \eta_b, 1 - \eta_b]$. From K_1, Step 1 of Lemma 1(b) applies with q' shifted by σ_{b}^2 and with the target rectangle K taken inside the interior of the reachable support, $\{(u, r) : r^2 < 1 - \sigma_{\mathrm{b}}^2/u\}$: when $q' = u$, Cauchy–Schwarz gives $\rho'^2 \leq 1 - \sigma_{\mathrm{b}}^2/u$ (with equality only on a null set), and given x, $D\zeta$ is a nondegenerate Gaussian vector, so every such r carries positive density. The one-step density is jointly continuous and positive, so $P(s, A) \geq c\,\mathrm{Leb}(A \cap K)$ for $s \in K_1$ and a fixed compact K inside the interior. Since every state enters K_1 in one step with probability 1, the corner-escape Step 2 of Lemma 1(b) is not needed here. Hence $P^2(s, A) \geq c\,\mathrm{Leb}(A \cap K)$ for *every* state s, a two-step Doeblin condition, and the chain has a unique stationary law π. The lower bound $\min_a \phi'(z_a)^2 \geq e^{-2C(1 + q_{\max} \max_a x_a^2)}$ used in Lemma 4 makes $\log B$ integrable, so $2\chi^\star = \mathbb{E}_\pi \log(gB/F)$ almost surely. Every function below is bounded on the state space, so $\mathbb{E}_\pi[Ph - h] = 0$ holds with no integrability conditions.

(3) Moments. Write $\delta = q - q^\star$. The noise $q' - f(q) = g(A/F - P(q))$ is an average of F i.i.d. bounded, centred terms, so its $2k$-th moment is at most $C_k F^{-k}$. Stationarity of δ^{2k}, with (1) and Young's inequality, gives $\mathbb{E}_\pi \delta^{2k} \leq C_k F^{-k}$. Likewise $\hat\rho = h(\bar W)$ as in Lemma 2, with the state average shifted by σ_{b}^2/g, which keeps it bounded away from 0. So $\mathbb{E}|\hat\rho - r|^{2k} \leq C_k F^{-k}$, and $\mathbb{E}[\rho'^{2k} \mid q, \rho] \leq (1 + \epsilon)\lambda_\rho^k \rho^{2k} + C_{k,\epsilon} F^{-k}$. Choose ϵ with $(1 + \epsilon)\lambda_\rho^k < 1$; stationarity then gives $\mathbb{E}_\pi \rho^{2k} \leq C_k F^{-k}$.

(4) One step of the exponent. Given (q, ρ), $\mathbb{E}[B/F] = m := S(q) - \rho^2 q K(q)$ exactly, since $\zeta = \rho x + \sqrt{1 - \rho^2}\,e$. Put $Z = B/(Fm) - 1$, an average of F i.i.d. centred terms. Expanding $\log(1 + Z)$ to fourth order on the event $\{|Z| \leq 1/2\}$, and bounding the complement with the good-event and log-moment argument of Lemma 4, gives $\mathbb{E}\log(1 + Z) = -\mathrm{Var}(\phi'^2\zeta^2)/(2Fm^2) + O(F^{-2})$. At $\rho = 0$ this variance is $3\mathbb{E}c^2 - S^2$, and it changes by $O(\rho^2)$. With $\log m = \log S - \rho^2 qK/S + O(\rho^4)$, $\mathbb{E}[\log(gB/F) \mid q, \rho] = \log(gS(q)) - \rho^2\, qK/S + a_{\mathrm{samp}}(q)/F + O(\rho^4 + \rho^2/F + F^{-2})$, uniformly.

(5) Stationary averages. Take π-expectations and Taylor-expand in δ; by (3) the remainders are $O(F^{-3/2})$, using $\mathbb{E}|\delta|\rho^2 \leq (\mathbb{E}\delta^2\,\mathbb{E}\rho^4)^{1/2}$.

- Stationarity of q gives $0 = \mathbb{E}_\pi[f(q) - q] = (f' - 1)\mathbb{E}_\pi \delta + \frac{1}{2} f''\,\mathbb{E}_\pi \delta^2 + O(F^{-3/2})$.
- Stationarity of δ^2 gives $\mathbb{E}_\pi \delta^2 = f'^2\,\mathbb{E}_\pi \delta^2 + V/F + O(F^{-3/2})$.
- Stationarity of ρ^2, with Lemma 2 applied to the bias-shifted cosine $\hat\rho = \bar W_3/\sqrt{\bar W_2(\bar W_1 + \sigma_{\mathrm{b}}^2/g)}$, $\mathbb{E}[\hat\rho^{\,2}] = r^2 + v_2/F + O(F^{-2})$, $r^2 = \kappa(q)\rho^2 + O(\rho^4)$ and $v_2(q, 0) = v(q)$, gives $\mathbb{E}_\pi \rho^2 = v/(F(1 - \kappa)) + O(F^{-3/2})$.

Substituting into $\mathbb{E}_\pi \log S(q) = \log S + (\log S)'\,\mathbb{E}\delta + \frac{1}{2}(\log S)''\,\mathbb{E}\delta^2 + O(F^{-3/2})$ and into (4) gives equation 22, with $\log G(g) = \log(gS(q^\star(g)))$. The constants are continuous in g, so the bounds

are uniform on J. *(6) Crossing.* $\log G$ is bounded away from 0 on J outside any neighbourhood of g_{mf}, so for large F every sign change lies near g_{mf}. There $\log G$ is strictly monotone with slope $\partial_g \log G(g_{\mathrm{mf}}) \neq 0$, and a is continuous. At a sign change, equation 22 gives $|\log G(g) + a(g)/F| = O(F^{-3/2})$, so g lies within $O(F^{-3/2})$ of the root of $\log G + a/F$, which is $g_{\mathrm{mf}} - a(g_{\mathrm{mf}})/(F\,\partial_g \log G) + O(F^{-2})$. No differentiability of the remainder in g is needed. □

For the non-bipartite regular graph with node-uniform input, the top tangent block is the MLP's, so the proposition carries over (Step 1 of Theorem 3). *Checks* (`theory/t2_bias_expansion.py`, `theory/tc_diagnose.py`). At $\sigma_{\mathrm{b}} = 0.3$, $a = -1.854$ ($a_{\mathrm{samp}} = -1.683$, $a_{\mathrm{align}} = -0.352$, $a_{\mathrm{state}} = +0.181$). At $\sigma_{\mathrm{b}} = 0.6$, $a = -2.492$. The hypotheses hold with room to spare: $\partial_g \log G(g_{\mathrm{mf}}) = 0.217$ and 0.206, $f'(q^\star) = 0.45$ and 0.29, $\kappa = 0.20$ and 0.085, and $\lambda_\rho = 0.92$ and 0.81 at $\sigma_{\mathrm{b}} = 0.3$ and 0.6. The chain splits $F \cdot 2\chi^\star$ exactly into the three parts, through $\mathbb{E}[B/F \mid q, \rho] = S - \rho^2 qK$. At $\sigma_{\mathrm{b}} = 0.6$ the state and sampling parts match their coefficients within noise at $F = 64$, 128, 256. The alignment part exceeds a_{align} by 0.012, 0.006, 0.003 and 0.0014 at $F = 64$ to 512 (0.016, 0.008 and 0.004 at $\sigma_{\mathrm{b}} = 0.3$), halving with each doubling of F; at $F = 512$ the total is -2.490 ± 0.013 against $a = -2.492$: a next-order term, as the $O(F^{-3/2})$ remainder allows. Resolving the gap by width shows that it is not a missing $O(1/F)$ term. The crossing $g_\times = g_{\mathrm{mf}} - a/(F\,\partial_g \log G)$ gives displacements of 5.12, 1.95, 0.98, 0.49 and 0.25% at $F = 48$, 128, 256, 512, 1024. The pooled measurements are 6.09, 2.16, 0.87, 0.35 and 0.26% (per group, first reported: 5.79, 1.96, 0.78, 0.47, 0.21); four of five intervals contain the prediction ($F = 512$: 0.49 against $[0.13, 0.48]$). The earlier prediction, which kept only the linear term γ_F, gives 2.92% at $F = 48$, below that cell's interval $[4.33, 6.76]$. These measurements predate the expansion, so this comparison is a retrodiction, not a registered test.

L Supplementary figures

Figure 17 reports a transition band rather than a crossing because the field oscillates inside its own error bar there. For the same reason Figure 18's first crossings, from a separate sweep whose weight draws differ, fall up to 0.09 outside these bands. Gating on $|\chi^\star| > \mathrm{SE}$ removes 95.5–97.7% of these level sets (Appendix F).

M Reproduction

Every headline number derives from a command below against the fixed seed set $[42, 7, 13, 21, 37, 55, 71, 89]$, with 95% bootstrap confidence intervals from $10{,}000$ resamples at a fixed RNG. The bipartite and irregular-graph runs of §5 used a float64 CUDA port of the estimator (`experiments/gpu_field.py`); on identical weights it agrees with the CPU reference to within 4×10^{-15} in $\chi^\star$, and CPU and CUDA rows are never pooled within one measurement.

Checking the proofs numerically. `python theory/verify_proofs.py --full` runs one check per proved result (Theorems 1–4 and 5(a), Lemmas L1(d), L3a–L3c (including the interlacing and trace steps of L3b), L3b′, L4 and 9, Proposition 8, and Propositions 2–9, 13 and 14; check names carry the result's label), each against a stated pass rule, on instances the proof covers and independently of the scripts that produced the paper's numbers. All 43 pass (about 35 s on 8 CPU threads, a few minutes with `--full`). A pass shows that a displayed statement holds where it was tested, not that its proof is correct. Two checks use thresholds chosen after seeing data, as their names say; both passed again on fresh seeds in a registered re-run (`theory/PREREG_confirm_checks.md`).

An independent reimplementation (E-d; targets and pass rules fixed in `experiments/ed_reimpl.py` before any run). Written from the paper's equations alone, it imports none of our simulation code and uses a different estimator: Benettin averaging after a burn-in, common random numbers across a fine bracket of gains, and the root of the realization-averaged exponent by interpolation. It also uses a different, equal-in-law weight parametrization for the MLP. Each target passes if it lies within two standard errors (16 realization groups). Six of seven pass. $\chi^\star_{\mathrm{lin}}(g_{\mathrm{lin}}) = 0$ at $F = 6$ on an irregular 7-node graph and at $F = 64$ ($-5 \times 10^{-5} \pm 5.5 \times 10^{-4}$ and $-2.2 \times 10^{-4} \pm 1.5 \times 10^{-4}$); the MLP's $\mathcal{C}_{\mathrm{q}}$ at $s = 1.147 \pm 0.019$ ($F = 64$; chain 1.150) and 1.055 ± 0.008 ($F = 256$; 1.042); rr4-n16 at $F = 256$, $7.41 \pm 1.05\%$

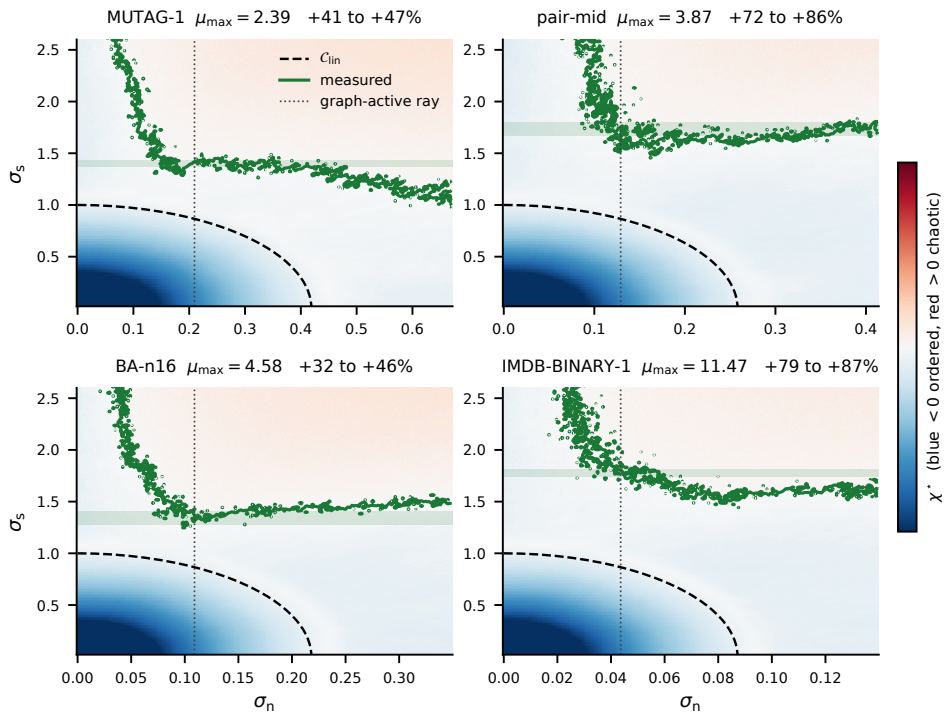


Figure 17: Small width, $F = 6$: $\chi^\star$ over (σ_s, σ_n) for four irregular graphs ($L = 160$, $S = 32$). Dashed: the linearized conic; solid: the measured $\tanh$ locus; dotted: the graph-active ray, with the transition band shaded (last significantly-negative to first significantly-positive $\chi^\star$) and given in each title over $\mathcal{C}_{lin}(6) = 0.9707$. No pixel of the measured level set lies inside $\mathcal{C}_{lin}$. A band is reported because the first sign change is grid-dependent by $\sim 15\%$ where the field oscillates inside its error bar.

(chain 7.65%); and Proposition 14 at $\sigma_b = 0.3$, $F = 128$: $2.00 \pm 0.06\%$ against 1.95%. The bipartite target failed: Q_4 at $F = 256$ gave $0.29 \pm 0.14\%$ against the chain's registered 1.67%. Post hoc, the failure is the statistic, not the chain (`experiments/ed_q4_posthoc.py`). With 384 groups at $L = 60{,}000$ the network's $F \cdot 2\chi^\star$ agrees with the chain's at every gain above 0.25% ($|z| \leq 1.9$, differences ≤ 0.013). Both fields lie within 0.02 of zero from 0.25% to about 1.6%, as Theorem 3(i) predicts for a bipartite graph. A root located on so flat a field is ill-conditioned: the network's first sign change is at 0.25% and its last near 1.6%, with a bootstrap interval of $[0.19, 1.95]$. A jackknife, which the registered rule used, cannot see that instability, because dropping one group rarely moves a first sign change. The chain's variance-reduced field is smooth per seed (no seed changes sign twice), and over 32 seeds its crossing is 1.78% $[1.71, 1.86]$ at $L = 2 \times 10^4$ and 1.86% $[1.81, 1.92]$ at $L = 6 \times 10^4$, against the 1.67% from 8 seeds that the target used (`theory/bip_chain_pooled.py`).

The six checks, with their diagnostics. **P1. Use a graph-active ray** — never sweep $\sigma_n = 0$; diagnostic: $\max \chi^\star$ should differ across (graph, operator) cells, and on the degenerate ray it does not. **P2. Locate before you count** — a dimension can agree while the location does not. **P3. Run the linear control** — replace ϕ by the identity, where $\mathcal{C}_{lin}$ is exact, and require the pipeline to recover it; one that fails this is broken independently of graphs. **P4. Put error bars on every binary call** — this reversed two of our own conclusions, because small magnitude is not the same as unresolved. **P5. Gate on resolvability and report the noise null** — below roughly 50% of the level set resolved, a dimension is not a measurement; and a per-cell gate applied *at* a zero set rejects almost everything by construction, so test that the field is significantly negative on one side and positive on the other instead. **P6. Report the averaging budget against a convergence curve** — achievable accuracy is set by S.

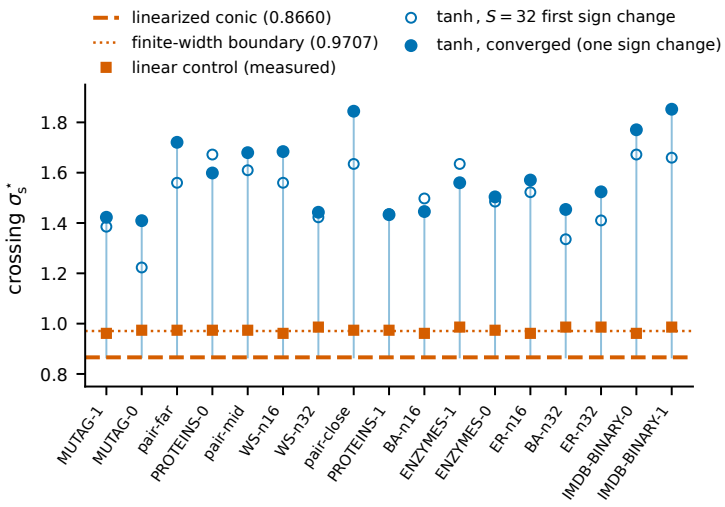


Figure 18: Measured crossing on the graph-active ray at $F = 6$, 17 irregular graphs ordered by $\mu_{\max}$. The linearized conic (dashed) predicts 0.866; the exact finite-width linear boundary $\mathcal{C}_{\text{lin}}(6)$ (dotted) is 0.9707; the $\tanh$ crossing sits 45–91% above it once converged (filled circles: S doubled until the averaged field changes sign once); the $S = 32$ first sign changes (hollow) fired early on noise in 11 of 17 graphs (Appendix J). The linear control (squares) lands on the dotted line to within -1.0% to $+1.6\%$. An MLP of the same gain, and hence every regular graph with uniform input, has no crossing at this width.

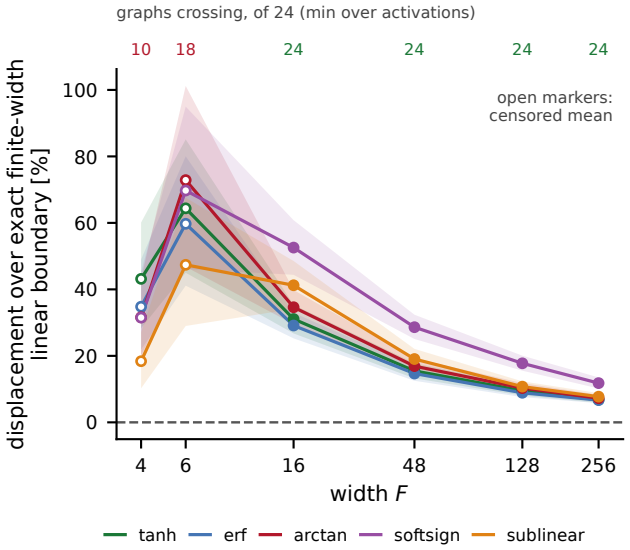


Figure 19: Displacement over $\mathcal{C}_{\text{lin}}(F)$ against width on the 24-graph corpus, five activations, $L = 160$, $S = 8$, first sign change on a grid of step 0.027. Bands are 95% bootstrap intervals over graphs. Numbers above the panel count the graphs that cross, minimum over activations. Open markers are censored means: at $F = 4$ and 6 not every graph crosses (as few as 10 of 24), so those points average only the graphs that do and are biased upward.

Errors the checks caught. Among others: a float32 accumulation artifact of 2.4 that would have read as signal against a float64 truth of 10^{-9}; a selection effect that inflated a fit by a factor of two on re-measurement with independent seeds; a confound that produced an apparently strong trend which vanished when the confounder was held fixed; a plumbing failure being silently recorded as

a scientific outcome; a resolvability gate that tested whether $\chi^\star$ differed from zero *at* the crossing, where it is near zero by construction, caught by a pilot before the run; and the grid-dependence of the first sign change along a ray, which is why Figure 17 reports a transition band. Each was found by a check written before the result it examined.

The fidelity gate. Every measurement runner asserts $\max|\text{ours} - \text{reference}| = 0.000\text{e}{+00}$ — bit-identical, not merely close — against the validated reference field implementation before measuring anything. Reproducing this requires matching the reference's operation order: an otherwise-equivalent matrix product takes a different BLAS path and differs in the last unit in the last place, which fails the gate.

```
python fractal_frontier.py    --selftest  # counter 0.994/1.196/1.982
python induced_boundary.py    --selftest  # 27 checks, mean-field null
python induced_boundary.py    --table2    # C_ind vs C_lin
python boundary_extensions.py --quick     # finite-width law, F=4..64
python theory/certify_quenched_bound.py --full   # the certificate
python theory/state_tangent_alignment.py --measure  # the obstruction
python theory/supermartingale_bound.py --all  # Prop 4 checks
python experiments/e2_sigma_b.py --full  # sigma_b > 0
python experiments/e8_spread.py  --full  # the spread regression
python experiments/e1_width.py --full --graphs 24  # width sweep
python sweeps/sweep.py --only E0,E1,E3  # Prop 2, families
python sweeps/control.py  # init control (App D)
python sweeps/sweep3.py --only G1,G2,G3,G4  # 17 graphs
python theory/epping_gp.py --selftest   # the GP criterion, validated
python experiments/e11_init.py --full  # the init null (s8)
python experiments/r1_convergence.py --full  # the (F,L,S) study
python experiments/make_field2d.py --graph BA-n16  # 2-D field
python experiments/redesign_f6.py --screen  # 200 activations at F=6
```

The analyses below read those outputs rather than re-running the sweeps, and are listed separately because the split matters for reproduction — the measurement half runs on a cluster, the analysis half runs anywhere:

```
python experiments/analyze_e11_merged.py      # init null, all arms
python experiments/analyze_init_by_arch.py    # same, by symmetry
python experiments/analyze_alignment_slack.py # r = -0.81
python experiments/analyze_r1.py              # the (F,L,S) tables
python experiments/analyze_by_size.py         # large-graph subsets
python experiments/redesign_f6.py --select    # identifiability
python experiments/width_exponent.py          # per-group fit, -0.54
python experiments/eg_width_audit.py analyze  # pooled: -0.56
python experiments/sigma_b_reanalyze.py       # sigma_b > 0 errors
python experiments/fig1_fig2_reconcile.py     # 1-D vs 2-D crossings
python experiments/wide_width.py              # F = 48..256 persistence
python experiments/noise_null.py              # D = 1.744 noise null
python experiments/mlp_control.py             # MLP, width-sweep protocol
python experiments/scaling_study.py mlp    # MLP width x depth
python experiments/scaling_study.py graphs # 8 graphs, F<=1024
python experiments/bias_scaling.py         # bias falsification
python experiments/width_models.py            # competing width laws
python theory/regular_graph_theorem.py        # Theorems 1-2 checks
python theory/degenerate_crossing.py          # Lemma, closure prediction
python theory/conjecture_adversarial.py       # adversarial search
```

A layout note, so the commands above are runnable rather than merely nominal: `e11_init.py` was executed in three shards, whose records are collected into one directory that the two `analyze_*` scripts above read; a single `--full` invocation writes one shard, not that collected layout.

Data. The benchmark graphs used for the displacement measurements are embedded as `graph6` strings, so that pipeline requires no data files. The initialization experiment of §7 is the exception: it trains on eight TU benchmark collections (Morris et al., 2020), which must be present locally

in the standard TU layout (`DS_A.txt`, `DS_graph_indicator.txt`, `DS_graph_labels.txt` per collection); datasets that are absent are skipped with a printed notice rather than silently.

Figures. All twelve figures regenerate from scripts; none is hand-edited. The width-law figure comes from `paper/make_figure_widthlaw.py`. Four — the width decay, the initialization forest plot, the four-graph field and the alignment scatter — come from `paper/make_figures_revision.py`. Three — the mean-field null, the certificate and the σ_{b} panel — come from `paper/make_figures_2027.py`. The remaining four are generated under `sweeps/` and *copied* into `paper/figures/` unchanged (byte-identical in both trees): the speckle panel from `make_figures.py`, the averaging panel from `make_sweep_figures.py`, and the displacement and activation panels from `make_final_figures.py`. Figures are emitted with TrueType (Type 42) font embedding — the matplotlib default of Type 3 renders axis labels as `/uniXXXX` glyph codes and makes the text neither extractable nor searchable, which affected six figures in an earlier build. Every shipped figure is verified to embed a TrueType font program with no Type 3 font and no `/uniXXXX` glyph names.