RETHINKING INACTIVITY IN RESIDUAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Controlling activation variance does not determine how many features are zero on every input. Such features contribute no signal to a linear readout on that input set. We develop an inverse theory that chooses a residual initialization scale from a prescribed fraction of inactive features. For Gaussian branches $B\phi(Ax)$ followed by a terminal ReLU, we prove convergence of the joint empirical feature law on any fixed input set as state width and depth grow at arbitrary relative rates, allowing branch width to remain finite. The law yields unique inverses for marginal inactivity on positive inputs and, for ReLU or leaky ReLU branches, joint inactivity on positive or centered Gaussian input laws. The marginal inverse explains why input rescaling requires opposite corrections for GELU and SELU, while ReLU and leaky ReLU are invariant. We also construct positive non-Gaussian input laws with identical pairwise terminal feature distributions but different unused-feature fractions and expected readout ranks at every fixed budget of sampled limiting features. Experiments verify calibration and the predicted fitting differences. The results identify the input and activation information needed to control initial feature usage beyond variance.

1 INTRODUCTION

A feature that is zero on every input in a given set leaves an unused column in a linear readout's feature matrix. Its weight receives no gradient from the prediction loss. Enough zero columns can prevent the readout from fitting labels (Steinwart, 2019; Shin & Karniadakis, 2020). Per-input inactivity does not count these columns: different inputs may switch off different features. Figure 1 shows similar per-input inactivity with different numbers of unused columns and fitting errors.

This raises an inverse question for initialization theory: given the input values and an activation function, which residual scale produces a prescribed fraction of unused features? The inverse would turn a tolerance on missing columns into a scale constraint before drawing weights, and let us compare activations at the same initial feature usage.

We study unnormalized residual branches $B\phi(Ax)$ with independent Gaussian matrices and one ReLU after the last block. The established $L^{-1/2}$ scaling keeps accumulated variance finite as depth grows (Hayou et al., 2021; Marion et al., 2025), leaving a coefficient to choose. As the residual state evolves, shared weights couple its values across inputs. We follow their joint distribution to determine which scale realizes the prescribed fraction.

Our contributions are:

- Unique scale inverses for marginal inactivity on positive inputs and joint inactivity on positive or centered Gaussian input laws under ReLU or leaky ReLU. The marginal inverse also gives opposite input-scale corrections for GELU and SELU (Theorems 3.3, 3.4, and 4.2).
- A supporting joint empirical law for the two-matrix residual branch as state width and depth grow at any relative rate, including branch width one. It preserves a possibly non-Gaussian input-coordinate law (Theorem 4.1).
- For every fixed input count $N \geq 3$, positive input laws with identical limiting feature distributions on every proper input subset, but different unused-column fractions and expected ranks at each fixed feature budget (Proposition 4.3; Corollary E.5).

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.

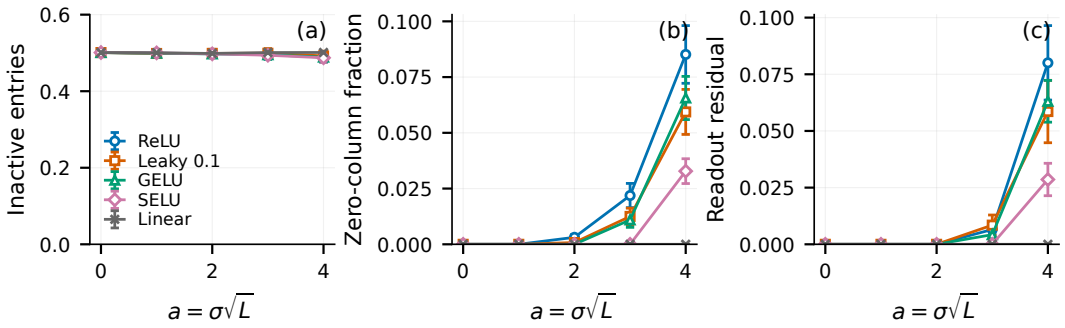


Figure 1: Similar per-input inactivity can conceal unused features. (a) Inactive entries. (b) Features zero on all inputs. (c) Least-squares readout error on balanced random labels. $N = n = 128$, $m = 32$, $L = 64$; means and standard errors over ten initializations. This finite-size illustration has $N = n$; the joint limit fixes N.

The last result identifies a limit of describing finite feature collections by their mean kernel. It also distinguishes the non-Gaussian problem from the Gaussian case, where joint inactivity is determined by correlations. Covariance propagation, joint width–depth limits, and the leaky-ReLU correlation correction are established (Hayou & Yang, 2023; Hayou, 2024; Zhang et al., 2022). Our analysis uses these tools to invert inactivity events and relate them to the features available to a readout. The experiments test those predictions, including the finite-feature fitting difference.

2 Model

Architecture. We study the forward pass at initialization of a residual network of depth L, state width n and branch width m. From a deterministic input vector $u \in \mathbb{R}^n$, the state evolves as

$$x_0 = u, \qquad x_{k+1} = x_k + \sigma\, B_k\, \phi(A_k x_k), \qquad \sigma = \frac{a}{\sqrt{L}}, \qquad k = 0, \dots, L-1, \tag{1}$$

where $a \geq 0$, $A_k \in \mathbb{R}^{m \times n}$ has independent $N(0, 1/n)$ entries, $B_k \in \mathbb{R}^{n \times m}$ has independent $N(0, 1/m)$ entries, layers are independent, and ϕ acts coordinatewise. No nonlinearity acts on the identity path between blocks. A single rectifier $\mathrm{ReLU}(x_L)$ produces the terminal features, and the quantity we control is

$$\widehat{p}_{n,L} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_{L,i} \leq 0\}, \tag{2}$$

the share of terminal coordinates that this rectifier zeroes at initialization. Using the same terminal rectifier makes this measurement comparable across branch activations. Negative outputs inside a GELU or SELU branch are not counted as inactive. Positivity of the input will be required for the single-input inverse, but not for the forward law or the Gaussian-input results below.

Branch activations. The branch activation is ReLU, leaky ReLU with slope $\beta \in [0, 1)$, exact GELU $\phi(z) = z\Phi(z)$ (Hendrycks & Gimpel, 2016), or SELU with the standard constants $\lambda \approx 1.0507$, $\alpha \approx 1.6733$ (Klambauer et al., 2017); Φ is the standard normal distribution function and G a standard normal variable. The single-input calibration law depends on the activation through its Gaussian second-moment profile

$$v_\phi(q) = \mathbb{E}\, \phi(\sqrt{q}\, G)^2, \qquad q \geq 0, \tag{3}$$

the mean-square output of a unit whose preactivation has variance q. Table 1 summarizes the four profiles.

Conditional increment. The centered output weights can move coordinates in either direction, even with a nonnegative branch activation. Given the current state, each coordinate increment has mean zero and variance $a^2 v_\phi(Q_k)/L$, where $Q_k = \|x_k\|^2/n$ is the state energy. Appendix A.1 gives the Gaussian representation. Shared weights also correlate increments across inputs (Section 4).

Branch activation	$k_\phi(0+)$	$k_\phi(\infty)$	Required a for larger inputs
ReLU	$1/2$	$1/2$	unchanged
Leaky ReLU	$(1+\beta^2)/2$	$(1+\beta^2)/2$	unchanged
GELU	$1/4$	$1/2$	decreases
SELU	$\lambda^2(1+\alpha^2)/2$	$\lambda^2/2$	increases

Table 1: Input-scale dependence through $k_\phi(q) = v_\phi(q)/q$. The last column keeps the terminal fraction fixed (Theorem 3.4).

Local input sensitivity. For fixed $u \in (0,\infty)^n$ and σ, residual blocks have invertible Jacobians almost surely, so the final feature map $F(u) = \mathrm{ReLU}(x_L(u))$ satisfies

$$\operatorname{rank} J_F(u) = n(1 - \widehat{p}_{n,L}). \tag{4}$$

Appendix A proves the identity and explains how changing the activation placement changes this interpretation. Table 3 distinguishes the limiting regimes used below.

3 Calibrating inactivity for one input

Standing assumption. Throughout, $\phi(0) = 0$, ϕ is K-Lipschitz, and $c_- q \le v_\phi(q) \le c_+ q$ for all $q \ge 0$ with $0 < c_- \le c_+$; the four branch activations satisfy this (Appendix D). Unless stated otherwise, constants C may depend on ϕ and on upper bounds for a and Q_0, never on n, m, L. We write $k_\phi(q) = v_\phi(q)/q$ and q_0 for the input energy.

3.1 Variance profile

For one input, the branch variance is determined by the energy Q_k. Expanding $\|x_{k+1}\|^2$ in (17) gives

$$Q_{k+1} - Q_k = \frac{a^2}{L}\, v_\phi(Q_k) + D_{k+1}, \qquad \mathbb{E}[D_{k+1} \mid x_0, \dots, x_k] = 0, \tag{5}$$

where D_{k+1} is a centered martingale increment. Its conditional variance separates randomness from the state dimension and randomness from the finite branch (Appendix C). The drift gives the ordinary differential equation

$$\bar{q}'(t) = a^2 v_\phi(\bar{q}(t)), \qquad \bar{q}(0) = q_0, \qquad \text{equivalently} \qquad \int_{q_0}^{\bar{q}(t)} \frac{dq}{v_\phi(q)} = a^2 t, \tag{6}$$

where $q_0 = Q_0$. For ReLU, $\bar{q}(t) = q_0 e^{a^2 t/2}$. For GELU and SELU, the integral determines the curve numerically. Residual addition changes the role of SELU: the energy has positive drift, and saturation reduces its growth rate. The feedforward variance fixed point of SELU (Klambauer et al., 2017) is therefore absent from this unnormalized residual block.

Proposition 3.1. *Under the standing assumption, for all $n, m, L \ge 1$ and $\bar{q}$ solving (6) with the given a and $q_0 = Q_0$,*

$$\mathbb{E} \max_{k \le L} \big| Q_k - \bar{q}(k/L) \big|^2 \le C\Big(\frac{1}{n} + \frac{1}{mL} + \frac{1}{L^2} \Big). \tag{7}$$

The three terms come from state width, branch randomness averaged over depth, and discretization. The bound vanishes as $n, L \to \infty$ at any relative rate, even at $m = 1$. Appendix C proves it by a martingale comparison with the usual residual variance ODE (Hayou et al., 2021; Hayou & Yang, 2023).

3.2 Empirical inactive fraction in the joint limit

For one input vector, its *coordinate law* is the distribution obtained by choosing a coordinate uniformly at random. We denote its limiting law by ν; it records the input values, including their signs and magnitudes. The finite-branch randomness averages over depth. Summing the centered increments

and applying Proposition 3.1, a coordinate's terminal displacement is asymptotically Gaussian with variance

$$w = a^2 \int_0^1 v_\phi(\bar{q}(t))\, dt = \bar{q}(1) - q_0. \tag{8}$$

Thus w is the variance accumulated along the residual stack. A coordinate that enters at u_i leaves approximately distributed as $u_i + \sqrt{w}\, Z$, with Z standard normal. Averaging its negativity probability over the input distribution gives the limiting fraction. Convergence in 2-Wasserstein distance below means that the coordinate distributions converge weakly and their second moments converge as well.

Theorem 3.2. *Let $n, L \to \infty$ at arbitrary relative rates, let $m = m_{n,L} \geq 1$ be arbitrary, let $a_{n,L} \to a \in (0, \infty)$, and let the empirical input laws $\nu_n = n^{-1} \sum_i \delta_{u_i}$ converge in 2-Wasserstein distance to a law ν with $q_0 = \int u^2\, d\nu(u) > 0$. Let w be given by (8) with the limiting a and q_0. Then*

$$\widehat{p}_{n,L} \xrightarrow{L^2} p(w, \nu) := \int \Phi(-u/\sqrt{w})\, d\nu(u).$$

This forward statement also permits signed input coordinates.

The branch width may stay fixed, even at $m = 1$. The proof replaces each conditional characteristic function by the one obtained from the deterministic energy curve. Two randomly selected coordinates then acquire independent Gaussian increments in the limit. Convergence of the first and second moments gives the empirical fraction, without assuming independence of coordinates in a finite network (Appendix C).

3.3 Prescribing the terminal fraction

The inverse separates two questions. How much added variance makes a chosen fraction p of the input coordinates negative? What residual coefficient produces that variance for this activation? Write w_p for the answer to the first question and a_p for the second. Positive coordinates make the first step monotone: each Gaussian tail $\Phi(-u/\sqrt{w})$ increases from 0 to $1/2$ as the added variance w grows.

Theorem 3.3. *Under the standing activation assumption, let positive input arrays have coordinate laws $\nu_n \to \nu$ in 2-Wasserstein distance, with $\nu((0, \infty)) = 1$ and $q_0 = \int u^2\, d\nu(u) > 0$. Let $n, L \to \infty$ at arbitrary relative rates and $m \geq 1$ be arbitrary. Fix a target inactive fraction $p \in (0, 1/2)$.*

(i) There is a unique $w_p \in (0, \infty)$ with $\int \Phi(-u/\sqrt{w_p})\, d\nu(u) = p$; the finite-$n$ equations $\int \Phi(-u/\sqrt{w})\, d\nu_n(u) = p$ have unique solutions $w_{p,n} \to w_p$.

(ii) With

$$a_p^2 = \int_{q_0}^{q_0 + w_p} \frac{dq}{v_\phi(q)}, \qquad a_{p,n}^2 = \int_{Q_0}^{Q_0 + w_{p,n}} \frac{dq}{v_\phi(q)}, \tag{9}$$

we have $a_{p,n} \to a_p \in (0, \infty)$, and the choice $\sigma = a_{p,n}/\sqrt{L}$ in (1) yields $\widehat{p}_{n,L} \to p$ in L^2.

Equation (9) integrates the reciprocal branch variance over the energy needed to reach the target. Using the observed coordinate values gives $a_{p,n}$, so both steps are scalar calculations without network simulation. The proof brackets the monotone root and separates variables in the energy ODE (Appendix C.4); Appendix C.2 gives a finite-size mean-bias bound.

The input distribution determines which targets are possible. For nonnegative inputs with limiting zero mass $\rho < 1$, every target in $(\rho/2, 1/2)$ is attained by a unique $a > 0$. A symmetric signed input gives one half for every $a > 0$ (Appendix C.4). For a fixed collection of positive inputs, averaging their individual probability curves before solving for a gives a common coefficient (Appendix C.5).

For ReLU at $p = 0.25$, a constant input requires $a_p \approx 1.52$ and a half-normal input requires $a_p \approx 1.18$, despite having the same second moment (Example C.4). The input law matters beyond its variance. For a given input law, matching p also matches the limiting alignment with the identity path (Corollary C.3).

3.4 INPUT AMPLITUDE

Scaling the input by $s > 0$ multiplies the target variance and the input energy by s^2, since $\Phi(-su/\sqrt{w}) = \Phi(-u/\sqrt{w/s^2})$. Substituting into (9) and changing variables $q = s^2 t$,

$$a_p(s)^2 = \int_{s^2 q_0}^{s^2(q_0+w_p)} \frac{dq}{v_\phi(q)} = \int_{q_0}^{q_0+w_p} \frac{dt}{t\, k_\phi(s^2 t)}. \tag{10}$$

The input scale enters through k_ϕ over the energies traversed by the state.

Theorem 3.4. *Fix ν with $\nu((0,\infty)) = 1$ and $0 < \int u^2\, d\nu(u) < \infty$. Let $a_p(s)$ be the multiplier (9) for the input law scaled by s.*

(i) If $v_\phi(q) = c_\phi q$, then $a_p(s)^2 = c_\phi^{-1} \log(1 + w_p/q_0)$ for every s and p.

(ii) If $a_p(s)$ does not depend on s for every $p \in (0, 1/2)$, then $v_\phi(q) = c_\phi q$ for some $c_\phi > 0$.

(iii) For GELU, $s \mapsto a_p(s)$ is strictly decreasing for every p, with $a_p(0^+)/a_p(\infty) = \sqrt{2}$. For SELU it is strictly increasing, with $a_p(\infty)/a_p(0^+) = \sqrt{1+\alpha^2} \approx 1.95$.

GELU changes from approximately $z/2$ to ReLU as scale grows, raising its second-moment ratio. SELU's negative branch saturates, lowering its ratio. This explains their opposite corrections (Appendix D), which also hold at fixed width in the depth limit (Appendix D.3). Homogeneity gives finite-network amplitude invariance for ReLU and leaky ReLU.

With ideal root-mean-square (RMS) normalization before A_k, the coefficient instead scales linearly with input amplitude for every activation. Appendix J gives this comparison, its finite-size bounds and the effect of stabilization.

4 FEATURES SHARED ACROSS INPUTS

An inactive coordinate on one input can still be useful on another. For a fixed set of N inputs, call a terminal feature *silent on the input set* if it is zero on every input. Its empirical fraction is

$$\widehat{d}_N = \frac{1}{n} \sum_{i=1}^{n} \prod_{s=1}^{N} \mathbf{1}\{x_{L,i}^{(s)} \le 0\}. \tag{11}$$

Every such feature gives a zero column in the readout feature matrix and zero gradient from a prediction-dependent loss to its readout weight. Changing the residual weights can reopen a feature during training.

Let $\mathcal{F}_{si} = \mathrm{ReLU}(x_{L,i}^{(s)})$ and let $\mathcal{F}_c = (I_N - \mathbf{1}\mathbf{1}^\mathsf{T}/N)\mathcal{F}$ be its column-centered feature matrix. It has at least $n\widehat{d}_N$ zero columns, so

$$\operatorname{rank} \mathcal{F}_c \le \min\{N-1, n(1-\widehat{d}_N)\}. \tag{12}$$

For even $N \ge 2$, choose labels uniformly with $N/2$ entries each of $+1$ and -1, independently of the features. Their expected relative least-squares residual is $1 - \operatorname{rank}(\mathcal{F}_c)/(N-1)$ (Appendix H). Joint inactivity thus constrains readout capacity when $n(1-\widehat{d}_N) < N-1$, so too few usable columns remain to span all centered labels. With N fixed and n growing, this bound becomes slack; the rank comparison below instead fixes the number of sampled features.

Let $U_i = (u_i^{(1)}, \dots, u_i^{(N)})$ be the initial values of coordinate i across inputs, and suppose their empirical law $\nu_n^{(N)}$ converges in 2-Wasserstein distance to $\nu^{(N)}$. Each vector U_i is a row across the input set. Its joint law records which input coordinates tend to be large, small or negative together. Write $C_0 = \mathbb{E}UU^\mathsf{T}$ for its raw second-moment matrix and define

$$\mathsf{K}_\phi(C) = \mathbb{E}[\phi(Z)\phi(Z)^\mathsf{T}], \qquad Z \sim N(0, C). \tag{13}$$

The matrix $\mathsf{K}_\phi(C)$ is the second moment after applying the activation to a Gaussian vector with covariance C. Shared branch weights give the matrix flow

$$\bar{C}_a'(t) = a^2 \mathsf{K}_\phi(\bar{C}_a(t)), \qquad \bar{C}_a(0) = C_0, \qquad W_a = \bar{C}_a(1) - C_0. \tag{14}$$

Its diagonal entries are the scalar energy curves in (6). The off-diagonal entries retain input dependence, and W_a is the covariance added by the full residual stack. The next theorem lifts this covariance calculation to a law for entire coordinate rows.

Theorem 4.1. *Assume the standing activation hypothesis, fix N, and suppose $\nu_n^{(N)} \to \nu^{(N)}$ in 2-Wasserstein distance with $C_0 = \mathbb{E}UU^\mathsf{T}$ positive definite. Let $n, L \to \infty$ at arbitrary relative rates, $m = m_{n,L} \geq 1$ be arbitrary, and $a_{n,L} \to a > 0$. For each bounded continuous $f : \mathbb{R}^N \to \mathbb{R}$,*

$$\frac{1}{n}\sum_{i=1}^{n} f(x_{L,i}^{(1)}, \ldots, x_{L,i}^{(N)}) \xrightarrow{L^2} \mathbb{E}f(U + G_a), \qquad U \sim \nu^{(N)}, \quad G_a \sim N(0, W_a),$$

where U and G_a are independent. In particular,

$$\widehat{d}_N \xrightarrow{L^2} D_N(a) := \mathbb{E}_U \mathbb{P}(G_a \leq -U), \tag{15}$$

with componentwise inequalities. The terminal feature Gram entries $n^{-1}\sum_i \mathcal{F}_{si}\mathcal{F}_{ti}$ also converge in L^2 to

$$\mathbb{E}[\mathrm{ReLU}(U_s + G_{a,s})\,\mathrm{ReLU}(U_t + G_{a,t})].$$

The covariance flow is standard Gaussian-kernel propagation; the empirical law allows finite m and non-Gaussian input rows. Appendix E controls the state Gram and the laws of one and two rows to prove it.

Theorem 4.2. *Fix N and $U \sim \nu^{(N)}$ with finite second moment and positive-definite $C_0 = \mathbb{E}UU^\mathsf{T}$. Let $\phi(z) = \max(z, \beta z)$ with $0 \leq \beta < 1$, and let $D_N(a) = \Pr(U + G_a \leq 0)$, where $G_a \sim N(0, W_a)$ is independent of U and W_a is determined by* (14)*.*

(i) If $\nu^{(N)}((0,\infty)^N) = 1$, then $D_N(a)$ is continuous and strictly increasing from 0 to $1/2$. Every $d \in (0, 1/2)$ therefore has a unique coefficient a_d. For strictly positive finite input arrays, the inverse computed from coordinate-row laws $\nu_n^{(N)} \to \nu^{(N)}$ in 2-Wasserstein distance and their raw second moments is consistent. Using this coefficient divided by $\sqrt{L}$ gives $\widehat{d}_N \to d$ in L^2 as $n, L \to \infty$ at any relative rate, with arbitrary $m \geq 1$.

(ii) If $\nu^{(N)} = N(0, C_0)$ and $N \geq 2$, then $D_N(a)$ is continuous and strictly increasing from $\mathbb{P}(U \leq 0)$ to $1/2$, although each marginal of $U + G_a$ is nonpositive with probability $1/2$ for every a. Each target strictly between these endpoints has a unique coefficient.

For positive inputs, increasing a raises both the marginal noise variances and the correlations of the added noise. The first change makes each terminal gate easier to close; the second increases their joint closure probability. The proof uses both effects (Appendix E.5). For signed Gaussian inputs, the inverse instead controls dependence among gates whose marginal probabilities do not change (Appendix E.6). Such coordinate laws arise from an independent Gaussian linear embedding of fixed data with a positive-definite input Gram matrix (Appendix E.11). Among M sampled limiting features, the expected number of zero columns is $MD_N(a)$. A tolerance of b unused columns in expectation gives $a \leq a_{b/M}$ whenever b/M is attainable. The inverse specifies the largest coefficient meeting this constraint; it also lets us compare activations at a common fraction of usable features. For Gaussian inputs with a common pairwise correlation, the silent fraction is a strictly increasing function of the terminal correlation: targeting either quantity gives the same coefficient (Corollary E.1). Arbitrary signed laws need not give a monotone D_N.

Beyond pairwise dependence. This equivalence is specific to the Gaussian family. For ReLU or leaky ReLU, consider three positive inputs whose coordinate rows are $U = \mathbf{1} + \delta S$, $0 < \delta < 1$. Choose S uniformly from sign triples with product $+1$, or from those with product -1. The two laws have identical one- and two-coordinate marginals. Their limiting marginal inactivity and terminal feature Grams therefore agree, but their zero-column fractions differ for every $a > 0$ (Proposition E.4). Joint inactivity can thus distinguish unused columns that pairwise feature statistics cannot identify. For balanced finite input arrays, each Gram entry also agrees in law by a pair-specific coordinate permutation (Appendix E.10).

Proposition 4.3. *Fix a ReLU or leaky-ReLU branch with $0 \leq \beta < 1$. For these two input laws, let $F_\pm^{(M)} \in \mathbb{R}^{3\times M}$ contain M independent limiting feature vectors $\mathrm{ReLU}(U_\pm + G_a)$ as columns. For*

every $a > 0$ and fixed $M \geq 1$,

$$\mathbb{E}[F_+^{(M)} F_+^{(M)\mathsf{T}}] = \mathbb{E}[F_-^{(M)} F_-^{(M)\mathsf{T}}], \qquad \mathbb{E}\,\mathrm{rank}\, F_+^{(M)} > \mathbb{E}\,\mathrm{rank}\, F_-^{(M)}.$$

Consequently, the minus law has larger expected relative least-squares residual for independent labels uniform on $\{y : \|y\|^2 = 3\}$, without an intercept.

Appendix E.10 proves the ordering and connects these columns to uniformly sampled coordinates of the wide network. Corollary E.5 extends the construction to every fixed $N \geq 3$, including a fitted intercept, with agreement on all proper input subsets. The two laws have different inverse curves, so pairwise input statistics alone cannot calibrate the same positive joint target for both. Thus matching the mean kernel can leave different fitting capacity when only finitely many random features are used (Rahimi & Recht, 2007).

For leaky ReLU, the established correlation identity (Zhang et al., 2022) gives a rate proportional to $a^2(1-\beta)^2$, while the energy grows at rate $a^2(1+\beta^2)/2$. These different factors lead to different transfer rules. The energy multipliers h_0 and h_β below are terminal-to-input energy ratios; dividing features by their square roots compares laws at the same amplitude.

Corollary 4.4. *Let $\nu^{(N)} = N(0, C_0)$ with C_0 positive definite. Compare ReLU branches with coefficient $a_0 > 0$ and leaky ReLU branches of slope $\beta \in [0, 1)$ with coefficient*

$$a_\beta = \frac{a_0}{1-\beta}, \qquad h_0 = e^{a_0^2/2}, \qquad h_\beta = e^{(1+\beta^2)a_\beta^2/2}. \tag{16}$$

The limiting terminal coordinate-row laws, divided respectively by $\sqrt{h_0}$ and $\sqrt{h_\beta}$, are identical. Their joint gate probabilities agree, as do their terminal ReLU feature Gram matrices divided by h_0 and h_β, respectively.

Matching the single-input accumulated variance instead uses $a_\beta = a_0/\sqrt{1+\beta^2}$ and does not match this joint geometry. A larger negative slope raises branch variance but slows the growth of correlations. Matching the first therefore requires less residual noise; matching joint inactivity requires more. For these Gaussian inputs, the normalized feature law determines the limiting kernel for a linear readout. With the features held fixed, least-squares readouts therefore have the same limiting gradient-flow dynamics when their initial prediction errors agree (Appendix E.7). At finite depth, the coefficient and energy normalization have distinct corrections, given in Appendix E.8.

5 Related work

Initialization and inactive units. Variance-based methods choose scales to preserve signal magnitudes (He et al., 2015; Klambauer et al., 2017; Defazio & Bottou, 2022; Bingham & Miikkulainen, 2023); LSUV estimates them from data (Mishkin & Matas, 2016). Prescribing marginal sparsity instead inverts a Gaussian tail (Price et al., 2024), and sparse initialization can affect learning (Yang et al., 2024). Work on dying ReLU units identifies the stronger event of inactivity on an entire input set (Steinwart, 2019; Lu et al., 2020; Shin & Karniadakis, 2020). We solve the residual-scale inverse for this event, using the input coordinate law and its evolution through shared weights. The distinction is consequential: equal marginal sparsity can leave different numbers of usable feature columns.

Residual scaling and covariance propagation. Activation placement changes residual signal propagation (He et al., 2016b;a), while normalization and initialization control branch magnitudes (De & Smith, 2020; Zhang et al., 2019; Bachlechner et al., 2021; Brock et al., 2021). The $L^{-1/2}$ scale, diffusion descriptions and covariance limits are established (Hayou et al., 2021; Peluchetti & Favaro, 2020; Marion et al., 2025; Li et al., 2022). In particular, Hayou & Yang (2023); Hayou (2024) prove that width and depth can grow jointly without a prescribed relative rate in related residual models. We treat the two-matrix branch with finite internal width and retain the non-Gaussian coordinate-row law needed for inactivity inversion. The leaky-ReLU correlation factor is also established by Zhang et al. (2022); we use it to transfer joint feature laws and contrast this with matching marginal variance.

Kernels and finite feature collections. Gaussian activation kernels describe pairwise feature statistics (Cho & Saul, 2009) and guide deep-network initialization (Martens et al., 2021). Random-feature methods approximate the resulting kernels with finitely many columns (Rahimi & Recht,

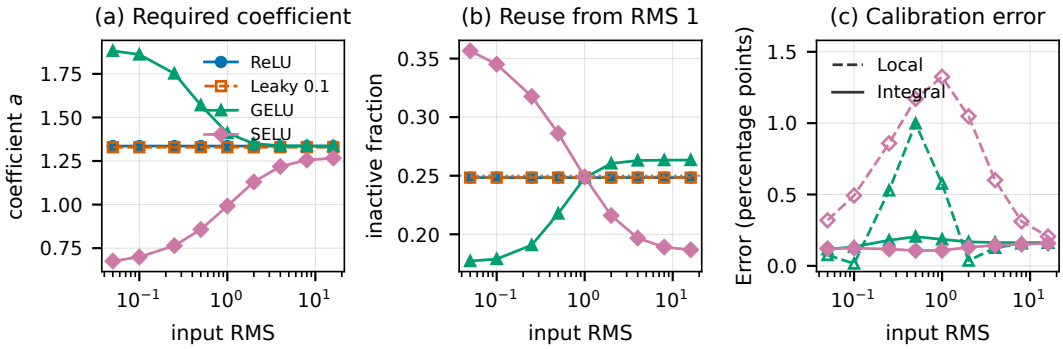


Figure 2: Input scale at target $p = 0.25$. (a) Required coefficient. (b) Reuse from unit RMS. (c) Input-variance correction (dashed) and full integral (solid). $n = 128$, $m = 32$, $L = 512$.

2007). Our construction holds every pairwise feature law fixed but changes the expected rank of such a finite collection. It isolates information about unused columns that the mean kernel does not retain.

6 Experiments

We test the scale inverses, compare activation-transfer rules, and measure the readout capacity lost when features are zero on an entire input set. Leaky ReLU uses slope 0.1 unless another value is specified.

6.1 Single-input calibration

Input scale and activation. Figure 2 compares three ways to prescribe $p = 0.25$: reuse the coefficient calibrated at unit input RMS, correct only the branch variance at the input, or integrate (9) along the energy curve. Inputs span RMS values 0.05 to 16 at $n = 128$, $m = 32$, $L = 512$. For ReLU and leaky ReLU, homogeneity makes all three rules identical. Reusing the unit-scale coefficient misses the target by an average 3.19 percentage points for GELU and 5.76 for SELU, in the opposite directions predicted by Theorem 3.4. Correcting the input variance reduces these errors to 0.30 and 0.70 points; the full integral gives 0.16 and 0.13 points. Following the energy through the stack therefore improves on a correction based only on its initial value.

Width, depth and input shape. Appendix F varies activations, input shapes and scales, and targets 0.1, 0.25, 0.4. The mean absolute calibration error falls from 0.46 percentage points at $n = 32$, $L = 64$ to 0.12 at $n = 128$, $L = 512$. At the latter size, the standard deviation across networks still averages 3.39 points (Table 5). The appendix also checks narrow branches, signed inputs and shared image inputs, separating mean bias from network variation. Appendix J.2 tests normalized branches.

6.2 Features shared across inputs

The same marginal fraction, different feature usage. We compare four activations at matched terminal-to-input energy ratio, with $n = 1024$, $m = 32$, $L = 256$ and ten initializations per setting. Figure 3(a,b) uses $N = 8$ independent, unit-RMS Gaussian inputs; Appendix G varies input count, scale and correlation. At energy ratio 10.47, mean zero-column fractions are 9.75% for ReLU and 0.57% for SELU, against predictions 9.93% and 0.78%. Both have marginal inactivity near 50%; their off-diagonal activation kernels distinguish the joint event that variance matching misses.

Transferring joint activity. For Gaussian inputs, we compare the correction $a_\beta = a_0/(1-\beta)$ with variance matching $a_\beta = a_0/\sqrt{1+\beta^2}$. Targets are 0.025, 0.05 and 0.10 at $N \in \{8, 32\}$, $L \in \{128, 512\}$. At slopes 0.1 and 0.5, the mean absolute errors of ten-network means are 0.43 and 0.85 percentage points under correlation correction, versus 1.78 and 5.18 under variance matching (Figure 3c; Appendix G). At slope 0.5, the same correction raises mean relative Gram error from 0.265 to 0.329. A 720-network follow-up separates coefficient correction from finite-depth energy

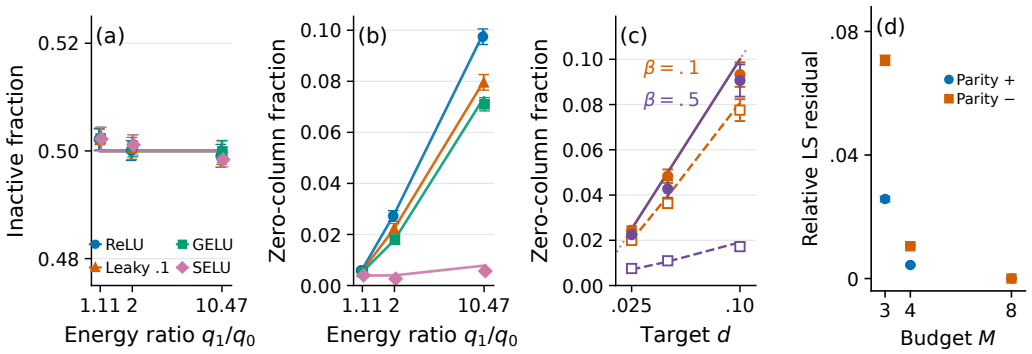


Figure 3: Shared-input calibration. (a,b) Matching single-input variance leaves different zero-column fractions. (c) Correlation correction (solid/filled) versus variance matching (dashed/open). Lines in (a–c) are limiting predictions. (d) No-intercept least-squares residual with M sampled features, analytically averaged over isotropic labels, under the two parity laws. Points show ten-network means with standard errors. Settings are given in Appendix G.

normalization. The latter reduces Gram error; coefficient correction alone slightly increases it (Appendix G.1).

The same pairwise features, different unused columns. The positive sign-triple construction uses $\delta = 0.99$, $n = 1024$, $m = 32$ and depths 128 and 512. At $a = 0.2$, the two limiting zero-column fractions are essentially zero and 6.60%, despite identical marginal inactivity and pairwise feature laws (Appendix E.9). At $a = 0.2$, $L = 512$, $M = 3$, ten independent network–sampling seed pairs give a zero-column difference of 6.46 ± 0.15 percentage points (mean $\pm$ SE). The relative fitting residuals are 2.57% and 7.07%, against predictions 2.66% and 7.19%, with paired difference 4.49 ± 0.14 points (Figure 3d). The predicted gap falls to 0.61 points at $M = 4$ and below 0.0001 at $M = 8$. Appendices E.10 and G.2 give the proof and all settings.

Finite-width readout capacity. Figure 1 tests the rank constraint at $N = n = 128$, $m = 32$, $L = 64$ on centered Gaussian inputs. Ten initializations are evaluated by least squares with an intercept and balanced random labels. Increasing the ReLU coefficient from 0 to 4 leaves marginal inactivity near one half but raises zero-column frequency to 8.5%. The mean centered rank falls from 127 to 117, and fitting residual rises to 8.0%. Appendix H varies widths, correlations and activations, including a linear-branch control.

7 Conclusion

We develop an inverse theory of inactivity for unnormalized Gaussian residual networks with a terminal ReLU. For fixed input sets, the joint empirical limit permits arbitrary relative growth rates of state width and depth, while allowing branch width to remain fixed. Marginal calibration depends on the input-coordinate law and the activation's second-moment profile; joint calibration additionally requires input dependence. Under the stated assumptions, the inverses uniquely determine the scale for an admissible inactivity target.

For every fixed $N \geq 3$, we construct positive input laws with identical limiting feature distributions on every proper input subset, but different unused-feature fractions and expected readout ranks at each fixed finite budget of independently sampled limiting features. The rank gap persists when an intercept is fitted. Thus, lower-order feature laws alone cannot determine finite-feature fitting capacity. These results concern initialization; our short classification experiments show no consistent accuracy gain (Appendix I). The theory identifies the information needed to control initial feature usage and provides a basis for comparing activations at matched inactivity.

AI USE STATEMENT

We used generative AI tools to check grammar, improve the readability of author-written text, assist with literature searches, and provide feedback on experimental design. We reviewed all AI-assisted content and take full responsibility for the content of this paper.

REFERENCES

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. ReZero is all you need: fast convergence at large depth. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1352–1361. PMLR, 2021. URL https://proceedings.mlr.press/v161/bachlechner21a.html.

Garrett Bingham and Risto Miikkulainen. AutoInit: Analytic signal-preserving weight initialization for neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6823–6833, 2023. doi: 10.1609/aaai.v37i6.25836. URL https://ojs.aaai.org/index.php/AAAI/article/view/25836.

Andrew Brock, Soham De, and Samuel L. Smith. Characterizing signal propagation to close the performance gap in unnormalized ResNets. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=IX3Nnir2omJ.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, 2009. URL https://proceedings.neurips.cc/paper/2009/hash/5751ec3e9a4feab575962e78e006250d-Abstract.html.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical Japanese literature, 2018. URL https://arxiv.org/abs/1812.01718.

Soham De and Sam Smith. Batch normalization biases residual blocks towards the identity function in deep networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19964–19975. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e6b738eca0e6792ba8a9cbcba6c1881d-Paper.pdf.

Aaron Defazio and Léon Bottou. A scaling calculus for the design and initialization of ReLU networks. *Neural Computing and Applications*, 34(17):14807–14821, 2022. doi: 10.1007/s00521-022-07308-z. URL https://link.springer.com/article/10.1007/s00521-022-07308-z.

Lester E. Dubins and Gideon Schwarz. On continuous martingales. *Proceedings of the National Academy of Sciences*, 53(5):913–916, 1965. doi: 10.1073/pnas.53.5.913.

Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, 1986. doi: 10.1002/9780470316658. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316658.

Alberto Raúl Galmarino. Representation of an isotropic diffusion as a skew product. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 1:359–378, 1963. doi: 10.1007/BF00533411. URL https://link.springer.com/article/10.1007/BF00533411.

Soufiane Hayou. On the infinite-depth limit of finite-width neural networks. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=RbLsYz1Az9.

Soufiane Hayou. Commutative scaling of width and depth in deep neural networks. *Journal of Machine Learning Research*, 25(299):1–41, 2024. URL http://jmlr.org/papers/v25/23-1163.html.

Soufiane Hayou and Greg Yang. Width and depth limits commute in residual networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12700–12723. PMLR, 2023. URL https://proceedings.mlr.press/v202/hayou23a.html.

Soufiane Hayou, Eugenio Clerico, Bobby He, George Deligiannidis, Arnaud Doucet, and Judith Rousseau. Stable ResNet. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1324–1332. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/hayou21a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. URL https://arxiv.org/abs/1502.01852.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, pp. 630–645. Springer International Publishing, 2016a. ISBN 9783319464930. doi: 10.1007/978-3-319-46493-0_38. URL http://dx.doi.org/10.1007/978-3-319-46493-0_38.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016b. URL https://arxiv.org/abs/1512.03385.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs), 2016. URL https://arxiv.org/abs/1606.08415.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/5d44ee6f2c3f71b73125876103c8f6c4-Paper.pdf.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yann Le Cun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient based learning applied to document recognition. *Proceedings of IEEE*, 86(11):2278–2324, 1998. URL https://bottou.org/papers/lecun-98h.

Mufan Li, Mihai Nica, and Daniel M. Roy. The neural covariance SDE: Shaped infinite depth-and-width networks at initialization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10795–10808. Curran Associates, Inc., 2022. doi: 10.52202/068431-0784. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/45fc4a0da7e7f6fbabaabe2d20a441d1-Paper-Conference.pdf.

Lu Lu, Yeonjong Shin, Yanhui Su, and George Em Karniadakis. Dying ReLU and initialization: Theory and numerical examples. *Communications in Computational Physics*, 28(5):1671–1706, Nov 2020. ISSN 1991-7120. doi: 10.4208/cicp.oa-2020-0165. URL http://dx.doi.org/10.4208/cicp.OA-2020-0165.

Pierre Marion, Adeline Fermanian, Gérard Biau, and Jean-Philippe Vert. Scaling resnets in the large-depth regime. *Journal of Machine Learning Research*, 26(56):1–48, 2025. URL http://jmlr.org/papers/v26/22-0664.html.

James Martens, Andy Ballard, Guillaume Desjardins, Grzegorz Swirszcz, Valentin Dalibard, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping, 2021. URL https://arxiv.org/abs/2110.01765.

Aleksandar Mijatović, Veno Mramor, and Gerónimo Uribe Bravo. Projections of spherical brownian motion. *Electronic Communications in Probability*, 23, Jan 2018. ISSN 1083-589X. doi: 10.1214/18-ecp162. URL http://dx.doi.org/10.1214/18-ECP162.

Aleksandar Mijatović, Veno Mramor, and Gerónimo Uribe Bravo. A note on the exact simulation of spherical brownian motion. *Statistics & Probability Letters*, 165:108836, Oct 2020. ISSN 0167-7152. doi: 10.1016/j.spl.2020.108836. URL http://dx.doi.org/10.1016/j.spl.2020.108836.

Dmytro Mishkin and Jiri Matas. All you need is a good init. In *International Conference on Learning Representations*, 2016. URL https://cmp.felk.cvut.cz/~mishkdmy/papers/mishkin-iclr2016.pdf.

Stefano Peluchetti and Stefano Favaro. Infinitely deep neural networks as diffusion processes. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1126–1136. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/peluchetti20a.html.

R. L. Plackett. A reduction formula for normal multivariate integrals. *Biometrika*, 41(3-4):351–360, 1954. doi: 10.1093/biomet/41.3-4.351. URL https://doi.org/10.1093/biomet/41.3-4.351.

Ilan Price, Nicholas Daultry Ball, Samuel C. H. Lam, Adam C. Jones, and Jared Tanner. Deep neural network initialization with sparsity inducing activations. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2402.16184.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Yeonjong Shin and George Em Karniadakis. Trainability of ReLU networks and data-dependent initialization. *Journal of Machine Learning for Modeling and Computing*, 1(1):39–74, 2020. ISSN 2689-3967. doi: 10.1615/jmachlearnmodelcomput.2020034126. URL http://dx.doi.org/10.1615/JMachLearnModelComput.2020034126.

David Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41 (2):463–501, 1962. doi: 10.1002/j.1538-7305.1962.tb02419.x. URL https://doi.org/10.1002/j.1538-7305.1962.tb02419.x.

Ingo Steinwart. A sober look at neural network initializations, 2019. URL https://arxiv.org/abs/1903.11482.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the Transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533, 2020. URL https://proceedings.mlr.press/v119/xiong20b.html.

Hongru Yang, Ziyu Jiang, Ruizhe Zhang, Yingbin Liang, and Zhangyang Wang. Neural networks with sparse activation induced by large bias: Tighter analysis with bias-generalized NTK. *Journal of Machine Learning Research*, 25(321):1–51, 2024. URL https://jmlr.org/papers/v25/23-0831.html.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

Guodong Zhang, Aleksandar Botev, and James Martens. Deep learning without shortcuts: Shaping the kernel with tailored rectifiers, 2022. URL https://arxiv.org/abs/2203.08120.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization, 2019. URL https://arxiv.org/abs/1901.09321.

A THE RESIDUAL STRUCTURE AND THE ACTIVITY MEASUREMENT

The final rectifier is important to the definition of inactivity. Its output is exactly zero at a nonpositive coordinate of x_L, irrespective of the activation inside a residual branch. Throughout the proofs, the initialization scale is held fixed when differentiating the network with respect to its input.

Symbol	Meaning
L, n, m	Depth, state width, and width inside each residual branch
σ, a	Branch initialization scale and its depth-adjusted coefficient, $\sigma = a/\sqrt{L}$
u, ν_n, ν	Input vector, its empirical coordinate law, and its limit
$Q_k, q_0, \bar{q}(t)$	State energy $\|x_k\|^2/n$, input energy (finite or limiting), and deterministic energy curve
$v_\phi(q), k_\phi(q)$	$\mathbb{E}\phi(\sqrt{q}G)^2$ and its ratio to q
w, w_p	Added coordinate variance, and the variance required for target p
$\widehat{p}_{n,L}, p$	Observed terminal inactive fraction and its prescribed limit
$N, U, \nu^{(N)}$	Number of inputs, a coordinate row across inputs, and its limiting law
$C_0, \bar{C}_a(t), W_a$	Initial raw second-moment matrix, its limiting evolution, and added covariance
$\mathsf{K}_\phi(C)$	Gaussian activation kernel $\mathbb{E}[\phi(Z)\phi(Z)^\mathsf{T}]$ for $Z \sim N(0, C)$
$\widehat{d}_N, D_N(a), d$	Observed zero-column fraction, its limiting probability, and a prescribed target
$J_N(c)$	Zero-orthant probability for a standard Gaussian vector with common correlation c
h_β	Deterministic energy multiplier for the leaky-ReLU transfer in (16)
$M, R_\pm(M)$	Retained feature-column budget and expected relative fitting residual for the parity laws
X_t, R_t, T_t	Fixed-width diffusion, its radius, and the time change for angular motion

Table 2: Principal notation.

A.1 AN EXACT REPRESENTATION OF THE FORWARD PASS

Let $\mathcal{F}_k$ contain the input and the matrices preceding block k. Conditionally on $\mathcal{F}_k$, the coordinates of $A_k x_k$ are independent centered Gaussians of variance $Q_k = \|x_k\|^2/n$. Given $A_k x_k$, independence of the rows of B_k gives

$$B_k\phi(A_k x_k) \overset{d}{=} \left(\frac{\|\phi(A_k x_k)\|^2}{m}\right)^{1/2} g_k, \qquad g_k \sim N(0, I_n).$$

Drawing fresh independent standard Gaussian vectors $z_k \in \mathbb{R}^m$ and $g_k \in \mathbb{R}^n$ at each step therefore gives the entire one-input Markov chain in the form

$$x_{k+1} = x_k + \frac{a}{\sqrt{L}}\sqrt{H_k}\, g_k, \qquad H_k = \frac{1}{m}\sum_{j=1}^{m} \phi\big(\sqrt{Q_k}\, z_{kj}\big)^2, \qquad Q_k = \frac{\|x_k\|^2}{n}. \tag{17}$$

In particular,

$$\mathbb{E}[H_k \mid \mathcal{F}_k] = v_\phi(Q_k), \qquad \mathrm{Var}(H_k \mid \mathcal{F}_k) = \frac{\mathrm{Var}(\phi(\sqrt{Q_k}G)^2)}{m}, \tag{18}$$

$$\mathbb{E}[\Delta x_k \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\Delta x_k \Delta x_k^\mathsf{T} \mid \mathcal{F}_k] = \frac{a^2}{L} v_\phi(Q_k) I_n. \tag{19}$$

The representation integrates out the matrices for one input. Several inputs propagated through the same matrices have correlated increments, which the explicit-matrix networks of Appendices F.1 and I retain.

Correlations between inputs. For two states x, y before a fresh layer, write $Q_x = \|x\|^2/n$, $Q_y = \|y\|^2/n$, and $C_{xy} = \langle x, y\rangle/n$. Let

$$\mathcal{K}_\phi(Q_x, Q_y, C_{xy}) = \mathbb{E}[\phi(G_1)\phi(G_2)], \qquad \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} \sim N\left(0, \begin{pmatrix} Q_x & C_{xy} \\ C_{xy} & Q_y \end{pmatrix}\right).$$

Regime	What is fixed or grows	Quantity described
Fixed state width	n, m fixed; $L \to \infty$	Expected inactive fraction of a diffusion (Appendix B)
Joint width–depth limit	$n, L \to \infty$; $m \geq 1$ arbitrary; N fixed	Empirical coordinate law and inactive fractions (Theorems 3.2, 4.1)
Finite-depth covariance recursion	L fixed; $n, m \to \infty$	Deterministic second moments and their activation correction (Appendix E.8)
Finite feature budget	N, M fixed; M columns sampled after the joint limit	Expected readout rank and fitting residual (Appendix E.10)

Table 3: The limits used in the paper. Finite-network experiments retain all dimensions explicitly.

This is the Gaussian activation kernel familiar from Cho & Saul (2009). Averaging first over B_k, then over A_k, gives the exact conditional cross-covariance

$$\mathbb{E}[\Delta x\, \Delta y^{\mathsf{T}} \mid x, y] = \frac{a^2}{L} \mathcal{K}_\phi(Q_x, Q_y, C_{xy}) I_n. \tag{20}$$

Indeed, the first average is $m^{-1}\langle \phi(A_k x), \phi(A_k y)\rangle I_n$, and each row of A_k produces the stated Gaussian pair.

For leaky ReLU $\phi_\beta(z) = \beta z + (1-\beta)\,\mathrm{ReLU}(z)$, Gaussian symmetry gives $\mathcal{K}_\beta = \beta C_{xy} + (1-\beta)^2 \mathcal{K}_{\mathrm{ReLU}}$. Matching its single-input variance to ReLU uses $a_\beta^2 = a_0^2/(1+\beta^2)$, where a_0 is the ReLU coefficient. After this rescaling the kernel difference, divided by a_0^2, is

$$\frac{\mathcal{K}_\beta}{1+\beta^2} - \mathcal{K}_{\mathrm{ReLU}} = \frac{\beta}{1+\beta^2}(C_{xy} - 2\mathcal{K}_{\mathrm{ReLU}}).$$

For noncollinear states and $\beta > 0$ this is strictly negative: $2\mathcal{K}_{\mathrm{ReLU}} = (C_{xy} + \mathbb{E}|G_1 G_2|)/2 > C_{xy}$. The distinction therefore holds even for strictly positive, noncollinear inputs. As a simple numerical example, at $C_{xy} = 0$ with $Q_x, Q_y > 0$ and $a > 0$, corresponding increment coordinates have correlation $1/\pi \approx 0.318$ for ReLU and $(1-\beta)^2/[\pi(1+\beta^2)] \approx 0.255$ for $\beta = 0.1$. Thus the single-input calibration and alignment can agree while the correlations between inputs depend on the activation.

A.2 Why the terminal fraction measures a local rank loss

We prove (4). For $\sigma = 0$, the stack is the identity and the strictly positive input makes the result immediate. Suppose $\sigma \neq 0$. Every coordinate of x_k is nonzero almost surely, by induction. Indeed, condition on the past and A_k, and put $b = \phi(A_k x_k)$. If $b = 0$, the state is unchanged. Otherwise $B_k b$ is a nondegenerate isotropic Gaussian vector, so each updated coordinate has zero probability of being exactly zero. Taking a finite union over coordinates and layers completes the induction.

Every coordinate of $A_k x_k$ is also a nondegenerate Gaussian, conditionally on the past. Thus the activation is differentiable at every realized preactivation, including for ReLU, leaky ReLU and SELU. Write

$$\Lambda_k = \operatorname{diag} \phi'(A_k x_k), \qquad J_k = I_n + \sigma B_k \Lambda_k A_k.$$

Given the past and A_k, the determinant of J_k is a polynomial in the entries of B_k. It equals one at $B_k = 0$ and hence is not the zero polynomial. Its zero set has Lebesgue measure zero. The Gaussian matrix has a density, so J_k is invertible almost surely. This argument does not require $m \geq n$.

The chain rule now gives an invertible stack Jacobian $J_{\mathrm{stack}} = J_{L-1} \cdots J_0$. Since no terminal coordinate is zero,

$$J_{\mathrm{ReLU\,ostack}}(u) = \operatorname{diag}\big(\mathbf{1}\{x_{L,i} > 0\}\big) J_{\mathrm{stack}}(u).$$

Multiplication on the right by an invertible matrix preserves rank. The rank is therefore the number of positive terminal coordinates, proving the claimed identity. This is a pointwise statement at any predetermined positive input. It describes the local map on the residual state space; a preceding embedding or a following classifier may impose further rank restrictions.

A.3 WHERE THE RECTIFIER IS PLACED

For $x_{k+1} = x_k + \mathrm{ReLU}(\sigma Z_k x_k)$, a positive input remains positive coordinatewise for every depth and every σ. This block cannot tune terminal inactivity. The signed second matrix in (1) removes that restriction.

In contrast, the postactivation stack $x_{k+1} = \mathrm{ReLU}(W_k x_k)$ includes a zero-derivative gate at every layer. Its Jacobian rank cannot exceed the smallest number of active coordinates along the path. Later activity can recover without recovering rank. For example, take

$$u = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad W_0 = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}, \quad W_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

The two states are $(1,0)^\mathsf{T}$ and $(1,1)^\mathsf{T}$, but the Jacobian has rank one. The strict preactivation signs persist on an open set of matrix pairs. Consequently this behavior has positive probability under $W_k = I + \sigma Z_k$ whenever $\sigma \neq 0$. This elementary comparison explains why the final inactive fraction has a different interpretation when the identity path itself is rectified.

B THE DEPTH LIMIT AT FIXED WIDTH

At fixed width, the depth limit is a diffusion. Its expected inactive fraction admits a unique inverse, although the empirical fraction does not concentrate.

Proposition B.1. *Fix $n \geq 2$, $m \geq 1$, and $u \in (0,\infty)^n$. As $L \to \infty$ with $\sigma = a/\sqrt{L}$, the interpolated state $x_{\lfloor tL \rfloor}$ converges in law to the unique strong solution of*

$$dX_t = a\sqrt{v_\phi(\|X_t\|^2/n)}\, dW_t, \qquad X_0 = u, \qquad t \in [0,1], \tag{21}$$

where W is an n-dimensional Brownian motion. The expected fraction converges, $\mathbb{E}\widehat{p}_{n,L} \to \pi_n(a) := n^{-1}\sum_i \mathbb{P}(X_{1,i} \leq 0)$. The function π_n is continuous and strictly increasing in a, from 0 as $a \to 0$ to $1/2$ as $a \to \infty$, so every $p \in (0,1/2)$ is attained by exactly one a. The fraction itself remains random at fixed n; only its expectation is calibrated.

The proof separates the radius $R_t = \|X_t\|$ from an independent Brownian motion on the unit sphere. The elapsed time for the spherical motion is

$$T_1 = a^2 \int_0^1 \frac{v_\phi(R_t^2/n)}{R_t^2}\, dt \in \Big[\frac{c_- a^2}{n}, \frac{c_+ a^2}{n}\Big]. \tag{22}$$

Brownian scaling identifies X_1 at coefficient a with the unit-coefficient diffusion at time a^2. Coupling these unit-coefficient diffusions through the same radial process makes their elapsed spherical times strictly increase with a, which proves monotonicity. When $v_\phi(q) = c_\phi q$, this time is the deterministic value $c_\phi a^2/n$ and the input amplitude cancels. For GELU and SELU, the radial process changes the elapsed time and hence the inverse.

Figure 4 compares depth-16 to depth-1024 networks at $n = m = 32$ with an Euler–Maruyama approximation of (21). At depth 1024, the largest difference of their estimated mean fractions is below 0.0017. Appendix F reports Monte Carlo uncertainty and a two-mesh check. These curves estimate the fixed-width expectation $\pi_{32}(a)$; the rule (9) instead inverts the joint-limit law. The width-two example below distinguishes these inverses at small state width.

B.1 CONVERGENCE AND NONATTAINMENT OF THE ORIGIN

Let K be a Lipschitz constant of ϕ, with $\phi(0) = 0$. The reverse triangle inequality in $L^2(G)$ gives

$$\big|\sqrt{v_\phi(q)} - \sqrt{v_\phi(r)}\big| \leq K|\sqrt{q} - \sqrt{r}|, \tag{23}$$

$$|v_\phi(q) - v_\phi(r)| \leq K^2|q - r|. \tag{24}$$

For the second inequality, bound the difference of squares by $K^2|\sqrt{q} - \sqrt{r}|(\sqrt{q} + \sqrt{r})\mathbb{E}G^2$. It follows that $x \mapsto \sqrt{v_\phi(\|x\|^2/n)}$ is globally Lipschitz and has linear growth. The limiting stochastic differential equation therefore has a unique strong, nonexplosive solution.

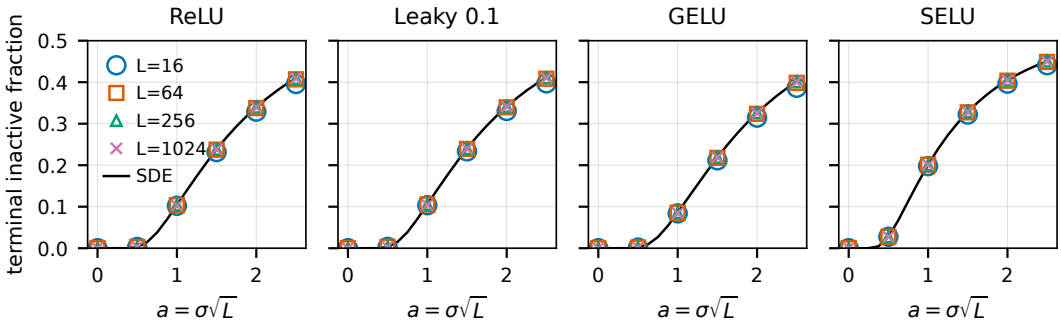


Figure 4: Depth convergence at $n = m = 32$. Curves show the fixed-width diffusion; markers show networks at four depths.

We verify the martingale-problem convergence criterion (Ethier & Kurtz, 1986, Chapters 4 and 7). The transition moments (19) give the drift and covariance exactly. Also, $\mathbb{E}[H_k^2 \mid \mathcal{F}_k] \le K^4 Q_k^2 (1 + 2/m)$, so, at fixed n, $\mathbb{E}[\|\Delta x_k\|^4 \mid x_k = x] \le C\|x\|^4/L^2$. For each fixed radius R and $\epsilon > 0$, this gives

$$\sup_{\|x\| \le R} L\mathbb{E}[\|\Delta x_k\|^2 \mathbf{1}\{\|\Delta x_k\| > \epsilon\} \mid x_k = x] \le \frac{C_R}{L\epsilon^2} \longrightarrow 0.$$

Before the first exit from the radius-R ball, the expected squared increment over a time interval of length h starting at a stopping time is at most $C_R(h+1/L)$. The martingale tightness criterion applies, and the fourth-moment estimate makes the largest jump vanish in probability. Every subsequential limit is therefore continuous.

For $f \in C_c^3(\mathbb{R}^n)$, Taylor expansion gives the limiting generator

$$\mathcal{L}f(x) = \frac{a^2}{2} v_\phi(\|x\|^2/n) \Delta f(x).$$

Indeed, $L\mathbb{E}|f(x + \Delta x) - f(x) - \nabla f(x) \cdot \Delta x - \frac{1}{2}\Delta x^\mathsf{T} \nabla^2 f(x) \Delta x|$ is $O_R(L^{-1/2})$, uniformly on that ball. Passing the discrete test-function martingales to the limit identifies the stopped martingale problem. Finally, the second-moment recursion bounds $\mathbb{E}\|x_L\|^2$ uniformly in L. Since the state is a martingale, the maximal inequality gives $\mathbb{P}(\max_{k \le L} \|x_k\| > R) \le C/R^2$. Letting R grow removes the stopping. The Lipschitz diffusion coefficient makes the martingale problem well posed, so all subsequences have the same limit. This is the usual depth-limit argument for independent residual increments (Peluchetti & Favaro, 2020; Hayou, 2023).

To study signs, we also need to exclude the origin. Set $R_t = \|X_t\|$ and stop before R_t reaches either zero or infinity. With a one-dimensional Brownian motion W^{rad}, Itô's formula gives

$$d \log R_t = \frac{a\sqrt{v_\phi(R_t^2/n)}}{R_t} \, dW_t^{\mathrm{rad}} + \frac{(n-2)a^2 v_\phi(R_t^2/n)}{2R_t^2} \, dt.$$

Both coefficients are bounded because $c_- q \le v_\phi(q) \le c_+ q$. The logarithm cannot reach an infinite value in finite time. Exhausting the stopping intervals proves $0 < R_t < \infty$ for every finite t. For $n = 1$ the same logarithmic argument shows that the sign is preserved.

B.2 Angles, hitting zero, and a unique inverse

The polar form of the generator separates a radial diffusion from a Brownian motion U on the unit sphere with generator $\Delta_{\mathbb{S}}/2$. The latter can be chosen independently of the radial process. Their skew product is

$$X_t = R_t U_{T_t}, \qquad T_t = a^2 \int_0^t \frac{v_\phi(R_s^2/n)}{R_s^2} \, ds, \qquad U_0 = \frac{u}{\|u\|}. \tag{25}$$

Here is a direct construction of this classical skew product (Galmarino, 1963). On a product probability space, solve

$$dR_t = a\sqrt{v_\phi(R_t^2/n)} \, dW_t^{\mathrm{rad}} + \frac{(n-1)a^2 v_\phi(R_t^2/n)}{2R_t} \, dt$$

and choose an independent spherical Brownian motion U. On each annulus, the generator of $R_t U_{T_t}$ is

$$\frac{a^2 v_\phi(r^2/n)}{2}\left(\partial_{rr} + \frac{n-1}{r}\partial_r + \frac{1}{r^2}\Delta_{\mathbb{S}}\right) = \mathcal{L}.$$

The logarithmic bound keeps R inside $(0,\infty)$ at every finite time. Uniqueness of the vector martingale problem therefore identifies the constructed process with X in law. The one-coordinate spherical diffusion is described, for example, by Mijatović et al. (2018; 2020). For $a > 0$, the clock at time one is at least $c_- a^2/n > 0$. Conditionally on the radial path, the sphere process at that time has a density, so no coordinate equals zero with positive probability. The finite-network sign indicators and their expectations consequently converge with the vector process. The case $a = 0$ is deterministic.

For a spherical coordinate starting positive, let τ_i be its first visit to the equator $\{U_i = 0\}$. Reflecting that coordinate after τ_i preserves the law. The strong Markov property and absence of an atom at zero therefore imply

$$\mathbb{P}(U_i(t) < 0) = \frac{1}{2}\mathbb{P}(\tau_i \le t). \tag{26}$$

This probability is continuous and strictly increasing from zero to one half. Strictness follows because, from any point in the open hemisphere, the diffusion can reach the equator in any positive time and can also remain inside the hemisphere for any finite time. The equator is reached almost surely. These statements also hold for spherical Brownian motion on the circle, so $n = 2$ is included.

Let Y be the vector diffusion with coefficient $a = 1$. Brownian scaling identifies $X_1^{(a)}$ in law with Y_{a^2}. The corresponding angular diffusion time is

$$T(a) = \int_0^{a^2} \frac{v_\phi(\|Y_s\|^2/n)}{\|Y_s\|^2}\, ds, \qquad \frac{c_- a^2}{n} \le T(a) \le \frac{c_+ a^2}{n}.$$

The clock is pathwise continuous, strictly increasing in a, and unbounded. Let $F_i(t) = \mathbb{P}(\tau_i \le t)$ for the spherical coordinate started at $u_i/\|u\|$. Applying (26) conditionally on the radial process gives $\mathbb{P}(X_{1,i}^{(a)} < 0) = \frac{1}{2}\mathbb{E}F_i(T(a))$. Continuity follows by bounded convergence. Strict increase follows because both F_i and $T(a)$ are strictly increasing. As a goes from zero to infinity, $F_i(T(a))$ goes from zero to one. Averaging over coordinates proves that the expected inactive fraction is continuous and strictly increasing from zero to one half. It therefore has a unique inverse at every target $p \in (0, 1/2)$. This proves Proposition B.1.

When $v_\phi(q) = c_\phi q$, the clock is deterministic, $T(a) = c_\phi a^2/n$. For the same input direction, ReLU and leaky ReLU consequently have the same depth-limit law after rescaling a by the square root of their second-moment coefficients. At finite depth and branch width their laws differ: a ReLU branch can be identically zero with positive probability, whereas a nonzero-slope leaky branch has no such event.

An explicit width-two example. For ReLU branches, $n = 2$ and $u = (1,1)^{\mathsf{T}}$, the angular diffusion time is $T_1 = a^2/4$. The terminal angle has the law $\pi/4 + (a/2)G$ modulo 2π, and the coordinates have equal negativity probabilities. The Fourier series of the indicator of a negative cosine therefore gives, for $a > 0$,

$$\pi_2(a) = \frac{1}{2} - \frac{2}{\pi}\sum_{j=0}^{\infty} \frac{(-1)^j \cos((2j+1)\pi/4)}{2j+1} \exp\left(-\frac{(2j+1)^2 a^2}{8}\right).$$

This absolutely convergent series has the unique quarter-target root $a \approx 2.17454$. The joint-limit rule for equal positive coordinates instead gives $a_\infty = \sqrt{2\log(1 + \Phi^{-1}(1/4)^{-2})} \approx 1.52483$, at which $\pi_2(a_\infty) \approx 0.15247$. Thus state width matters for the inverse even after depth tends to infinity. Here π_2 is an expectation: one network's terminal fraction still takes values in $\{0, 1/2, 1\}$.

C THE EMPIRICAL FRACTION WHEN WIDTH AND DEPTH GROW

The proof has two steps. First, the squared norm follows a deterministic ordinary differential equation. Second, a randomly chosen pair of coordinates acquires independent Gaussian increments in the limit. The pair argument is what turns a one-coordinate probability into a statement about the fraction observed in one large network.

C.1 CONTROLLING THE SQUARED NORM

In this subsection, a denotes the prefactor of the finite array under consideration. Its upper bound and an upper bound on Q_0 are fixed. Expanding the exact transition gives

$$Q_{k+1} - Q_k = \frac{a^2}{L} v_\phi(Q_k) + D_{k+1}, \tag{27}$$

$$D_{k+1} = \frac{2a\sqrt{H_k}}{n\sqrt{L}} \langle x_k, g_k \rangle + \frac{a^2}{L} \left(H_k \frac{\|g_k\|^2}{n} - v_\phi(Q_k) \right). \tag{28}$$

The two terms in D_{k+1} have zero conditional covariance: their product contains an odd Gaussian moment in g_k. Using $\mathbb{E}\|g_k\|^2 = n$ and $\mathrm{Var}(\|g_k\|^2) = 2n$ yields

$$\mathbb{E}[D_{k+1}^2 \mid \mathcal{F}_k] = \frac{4a^2 Q_k v_\phi(Q_k)}{nL} + \frac{a^4}{L^2} \left[\frac{2v_\phi(Q_k)^2}{n} + \left(1 + \frac{2}{n}\right) \frac{\mathrm{Var}(\phi(\sqrt{Q_k}G)^2)}{m} \right]. \tag{29}$$

Since $v_\phi(q) \le K^2 q$ and $\mathbb{E}\phi(\sqrt{q}G)^4 \le 3K^4 q^2$, expansion of Q_{k+1}^2 gives

$$\mathbb{E}[Q_{k+1}^2 \mid \mathcal{F}_k] \le (1 + C/L) Q_k^2.$$

Iteration bounds $\max_{k \le L} \mathbb{E} Q_k^2$ uniformly. Also, Q_k is a nonnegative submartingale, so Doob's inequality bounds $\mathbb{E} \max_{k \le L} Q_k^2$ uniformly.

Put $M_j = \sum_{k<j} D_{k+1}$. Orthogonality of martingale increments and (29) give

$$\mathbb{E} \max_{j \le L} |M_j|^2 \le 4 \sum_{k<L} \mathbb{E} D_{k+1}^2 \le C \left(\frac{1}{n} + \frac{1}{mL} \right). \tag{30}$$

Let $b_0 = Q_0$ and $b_{k+1} = b_k + a^2 v_\phi(b_k)/L$ be the deterministic Euler iteration. From (24),

$$\max_{j \le k} |Q_j - b_j| \le \max_{j \le k} |M_j| + \frac{a^2 K^2}{L} \sum_{j<k} \max_{i \le j} |Q_i - b_i|.$$

Discrete Gronwall transfers (30) to $Q_k - b_k$. The solution of $\bar{q}' = a^2 v_\phi(\bar{q})$ stays in a bounded interval on $[0, 1]$. Its drift is Lipschitz there, so its one-step Euler error is $O(L^{-2})$, and the accumulated error is $O(L^{-1})$. Squaring this deterministic error proves

$$\mathbb{E} \max_{k \le L} |Q_k - \bar{q}(k/L)|^2 \le C \left(\frac{1}{n} + \frac{1}{mL} + \frac{1}{L^2} \right). \tag{31}$$

The ODE here uses the finite array's a and Q_0. Continuous dependence on these parameters gives the limiting ODE when $a \to a_\infty$ and $Q_0 \to q_0$.

C.2 A FINITE-SIZE BOUND FOR THE EXPECTED FRACTION

The squared-norm estimate also bounds the error of the expected inactive fraction. The comparison uses the same finite input vector, so it does not require a rate of convergence for its empirical distribution.

Theorem C.1. *Assume the standing activation hypothesis. Fix $0 < a_- \le A < \infty$ and $0 < q_- \le q_+ < \infty$. For every $n, m, L \ge 1$ and deterministic $u \in \mathbb{R}^n$ with $Q_0 = \|u\|^2/n \in [q_-, q_+]$, let*

$$\bar{q}_a'(t) = a^2 v_\phi(\bar{q}_a(t)), \qquad \bar{q}_a(0) = Q_0, \qquad w_n(a) = \bar{q}_a(1) - Q_0, \qquad \epsilon_{n,m,L} = \frac{1}{n} + \frac{1}{mL} + \frac{1}{L^2}.$$

There is a finite constant C, depending only on $K, c_-, c_+, a_-, A, q_-, q_+$, such that

$$\sup_{a \in [a_-, A]} \left| \mathbb{E}\widehat{p}_{n,L}(a) - \frac{1}{n} \sum_{i=1}^n \Phi\left(-\frac{u_i}{\sqrt{w_n(a)}} \right) \right| \le \min\{1, C\epsilon_{n,m,L}^{1/5}\}. \tag{32}$$

The inputs may be signed or contain zeros. The expectation is over the network initialization; the supremum is outside that expectation.

Proof. Write $v = v_\phi$, fix $a \in [a_-, A]$, and abbreviate $w = w_n(a)$ and $\epsilon = \epsilon_{n,m,L}$. All constants below are uniform over these parameter ranges. The ODE and the lower variance bound give

$$w \ge w_* := q_- \left(e^{c_- a_-^2} - 1 \right) > 0. \tag{33}$$

The realized variance clock. Using the exact representation (17), put

$$\tau = \frac{a^2}{L} \sum_{k=0}^{L-1} H_k.$$

The centered branch-energy sum

$$J_L = \frac{a^2}{L} \sum_{k=0}^{L-1} \big(H_k - v(Q_k)\big)$$

is a sum of orthogonal martingale differences with respect to the discrete past before the fresh branch Gaussians are drawn. Therefore the conditional branch variance and the uniform second-moment bound on Q_k imply

$$\mathbb{E}J_L^2 = \frac{a^4}{mL^2} \sum_{k=0}^{L-1} \mathbb{E}\,\mathrm{Var}\Big(\phi(\sqrt{Q_k}G)^2 \mid Q_k\Big) \le \frac{3A^4K^4}{mL^2} \sum_{k=0}^{L-1} \mathbb{E}Q_k^2 \le \frac{C}{mL}.$$

Decompose the remaining error as

$$\begin{aligned}\tau - w = J_L &+ \frac{a^2}{L} \sum_{k=0}^{L-1} \big[v(Q_k) - v(\bar{q}_a(k/L))\big] \\ &+ a^2 \left[\frac{1}{L} \sum_{k=0}^{L-1} v(\bar{q}_a(k/L)) - \int_0^1 v(\bar{q}_a(t))\,dt\right].\end{aligned}$$

Since v is K^2-Lipschitz, the mean square of the middle term is at most $A^4K^4\mathbb{E}\max_k |Q_k - \bar{q}_a(k/L)|^2$. The last term is $O(L^{-1})$ uniformly: the ODE trajectory is bounded by $q_+e^{A^2c_+}$ and has uniformly bounded derivative. The radial estimate (31) consequently gives

$$\mathbb{E}|\tau - w|^2 \le C_\tau \epsilon, \tag{34}$$

where C_τ depends only on K, c_+, A, q_+.

An exact continuous interpolation and its stopping time. Construct an n-dimensional Brownian motion $\mathcal{W}$, independent of all the branch Gaussians, and use $g_k = \sqrt{L}(\mathcal{W}_{(k+1)/L} - \mathcal{W}_{k/L})$ in (17). Reveal z_k at time k/L. Thus H_k is known at the left endpoint of the interval, while its Brownian increment is still independent of the past. For a fixed coordinate i, define the continuous martingale

$$M_t = \sum_{k=0}^{L-1} a\sqrt{H_k}\big(\mathcal{W}_{i,t\wedge(k+1)/L} - \mathcal{W}_{i,t\wedge k/L}\big), \qquad 0 \le t \le 1.$$

Its terminal value is exactly $x_{L,i} - u_i$ and its bracket at time one is τ. Let $(\mathcal{H}_t)$ be the usual augmentation of the Brownian filtration with the branch Gaussians revealed as just specified. Extend M after time one by $M_t = M_1 + \mathcal{W}_{i,t} - \mathcal{W}_{i,1}$. Then $\Gamma_t := \langle M\rangle_t$ is continuous, $\Gamma_t = \tau + t - 1$ for $t \ge 1$, and $\Gamma_\infty = \infty$.

The Dambis–Dubins–Schwarz theorem (Dubins & Schwarz, 1965) now gives $M_t = W_{\Gamma_t}$, where

$$\mathcal{T}_s = \inf\{t \ge 0 : \Gamma_t > s\}, \qquad W_s = M_{\mathcal{T}_s}$$

is a standard Brownian motion for $\mathcal{G}_s = \mathcal{H}_{\mathcal{T}_s}$. The terminal clock τ is a stopping time for this filtration, because

$$\{\tau \le s\} = \{\mathcal{T}_s \ge 1\} \in \mathcal{H}_{\mathcal{T}_s} = \mathcal{G}_s. \tag{35}$$

This argument permits flat clock intervals and $\tau = 0$, both of which can occur for ReLU branches.

A fourth-moment coupling bound. Since τ and the deterministic time w are stopping times,

$$N_s = W_{s\wedge\tau} - W_{s\wedge w}$$

is a continuous martingale, with $\langle N\rangle_\infty = |\tau - w|$ and $N_\infty = W_\tau - W_w$. The fourth-moment Burkholder–Davis–Gundy inequality yields

$$\mathbb{E}|W_\tau - W_w|^4 \le C_4\mathbb{E}|\tau - w|^2 \le C_4C_\tau\epsilon. \tag{36}$$

For completeness, $C_4 = 1024/9$ suffices. After localization, Itô's formula, Cauchy–Schwarz and Doob's L^4 inequality give

$$\mathbb{E}N_T^4 \le 6\big(\mathbb{E}\sup_{s\le T}|N_s|^4\big)^{1/2}\big(\mathbb{E}\langle N\rangle_T^2\big)^{1/2} \le \frac{32}{3}(\mathbb{E}N_T^4)^{1/2}\big(\mathbb{E}\langle N\rangle_T^2\big)^{1/2}.$$

Divide and square when the first factor is nonzero, and remove the localization by Fatou's lemma to obtain (36).

From the coupling to a sign probability. For every $h > 0$, the coupling $x_{L,i} = u_i + W_\tau$ gives

$$\begin{aligned}\big|\mathbb{P}(x_{L,i} \le 0) - \Phi(-u_i/\sqrt{w})\big| &\le \mathbb{P}(|u_i + W_w| \le h) + \mathbb{P}(|W_\tau - W_w| > h)\\ &\le \sqrt{\frac{2}{\pi w_*}}\,h + C_4 C_\tau \epsilon h^{-4}.\end{aligned}$$

Here W_w has the ordinary $N(0, w)$ law; it need not be independent of τ. Choosing $h = \epsilon^{1/5}$ proves a bound $C\epsilon^{1/5}$ for every coordinate, with the same constant. Averaging proves (32). All estimates are uniform in $a \in [a_-, A]$, and the trivial upper bound of one gives the stated minimum. □

Corollary C.2. *Under the assumptions of Theorem 3.3, use the empirical coefficient $a_{p,n}$ in (9). There are $C < \infty$ and n_0 such that, for every $n \ge n_0$ and every $m, L \ge 1$,*

$$|\mathbb{E}\widehat{p}_{n,L}(a_{p,n}) - p| \le C\left(\frac{1}{n} + \frac{1}{mL} + \frac{1}{L^2}\right)^{1/5}. \tag{37}$$

The constant depends on the activation, limiting input law, target, and chosen eventual input-energy bounds, but not on n, m, L. The threshold n_0 may depend on the input sequence.

Proof. Theorem 3.3 gives $Q_{0,n} \to q_0 > 0$ and $a_{p,n} \to a_p \in (0, \infty)$. For all sufficiently large n, these parameters therefore lie in fixed positive compact intervals, uniformly over m, L. Separation of variables in the finite-input ODE and the definition of the empirical inverse give

$$w_n(a_{p,n}) = w_{p,n}, \qquad \frac{1}{n}\sum_i \Phi(-u_i/\sqrt{w_n(a_{p,n})}) = p$$

exactly. Apply Theorem C.1. □

C.3 Gaussian increments and the empirical fraction

Choose a fixed number d of distinct coordinate indices at time zero, independently of the weights, denote their collection by $\mathcal{I}$, and include this choice in $\mathcal{F}_0$. For $\xi \in \mathbb{R}^d$, the exact conditional characteristic function of their increment is

$$\psi_k(\xi) = \mathbb{E}\left[\exp\left(-\frac{a^2\|\xi\|^2}{2L}H_k\right)\middle|\mathcal{F}_k\right].$$

Set $\chi_k = \exp[-a^2\|\xi\|^2 v_\phi(b_k)/(2L)]$. The inequality $|e^{-t} - 1 + t| \le t^2/2$ for $t \ge 0$, together with (18), implies

$$\sum_{k<L}\mathbb{E}|\psi_k - \chi_k| \le C_\xi\left(\mathbb{E}\max_{k\le L}|Q_k - b_k| + \frac{1}{L}\right) \longrightarrow 0.$$

The increments are also asymptotically independent of the selected initial values $u_\mathcal{I}$, which are random when $\mathcal{I}$ is drawn at random. To see this, take any bounded $\mathcal{F}_0$-measurable random variable Y. Peel off the last increment using conditional expectation, replace its characteristic function by χ_{L-1}, and repeat backwards. At each replacement the other exponential factors have modulus at most one. Consequently,

$$\left|\mathbb{E}\Big[Y e^{i\xi\cdot(x_{L,\mathcal{I}} - u_\mathcal{I})}\Big] - \mathbb{E}[Y]\prod_{k<L}\chi_k\right| \le \|Y\|_\infty \sum_{k<L}\mathbb{E}|\psi_k - \chi_k| \longrightarrow 0.$$

The Euler recursion gives $\prod_{k<L}\chi_k = \exp[-\|\xi\|^2(b_L - Q_0)/2]$ exactly. This converges to $\exp(-w\|\xi\|^2/2)$, where $w = \bar{q}(1) - q_0 > 0$. The selected terminal increments are therefore

asymptotically independent Gaussians, independent of their initial values. At each step H_k is averaged conditionally on $\mathcal{F}_k$, retaining its dependence on earlier increments.

Now draw I, J independently and uniformly from $\{1, \dots, n\}$. Their collision probability is $1/n$. Off that event, the preceding argument applies to two distinct indices. Convergence of the empirical input law gives

$$(x_{L,I}, x_{L,J}) \Rightarrow (U + \sqrt{w}Z, \widetilde{U} + \sqrt{w}\widetilde{Z}),$$

where $U, \widetilde{U}$ are independent with law ν, and $Z, \widetilde{Z}$ are independent standard normals, independent of them. Because I and J are uniform and independent of the network, $\mathbb{E}\widehat{p}_{n,L} = \mathbb{P}(x_{L,I} \le 0)$ and $\mathbb{E}\widehat{p}^2_{n,L} = \mathbb{P}(x_{L,I} \le 0,\, x_{L,J} \le 0)$. Gaussian convolution has no atom at zero, so these probabilities converge to $p = \int \Phi(-u/\sqrt{w})\, \nu(du)$ and p^2, respectively. This proves the claimed L^2 convergence in Theorem 3.2.

C.4 Inverting the limiting law

For a positive input law, the function $w \mapsto \int \Phi(-u/\sqrt{w})\, \nu(du)$ is continuous and strictly increasing from zero to one half. Its root w_p is unique. For any $0 < w_- < w_p < w_+$, weak convergence of the empirical law makes the two empirical mixture values eventually lie on the corresponding sides of p. The empirical root is trapped between w_- and w_+. Letting these endpoints approach w_p proves $w_{p,n} \to w_p$.

Since $Q_{0,n} \to q_0 > 0$ and v_ϕ is continuous and positive on $(0, \infty)$, the integral defining $a^2_{p,n}$ converges to that defining a^2_p. Separating variables in the ODE shows that this coefficient makes $\bar{q}(1) = q_0 + w_p$. The empirical-fraction theorem then gives $\widehat{p}_{n,L} \to p$ in L^2, proving Theorem 3.3. The bounds on the second-moment profile also give the useful brackets

$$\frac{\log(1 + w_p/q_0)}{c_+} \le a^2_p \le \frac{\log(1 + w_p/q_0)}{c_-}.$$

If the input law is nonnegative with an atom $\rho < 1$ at zero, then for every positive a the limiting fraction is

$$p(a) = \frac{\rho}{2} + \int_{(0,\infty)} \Phi(-u/\sqrt{w(a)})\, \nu(du). \tag{38}$$

It increases strictly from $\rho/2$ to one half. The same root-bracketing proof works for any target in that interval. At $a = 0$, the finite-network fraction is the empirical exact-zero mass ρ_n. Its convergence to ρ requires that additional assumption: small positive coordinates can converge to zero without ever being exactly zero.

For a symmetric signed input law, symmetry of Φ makes $p(a) = 1/2$ for every $a > 0$. A nonsymmetric signed law need not be monotone either. For example, $\nu = \frac{1}{3}\delta_{-1} + \frac{2}{3}\delta_2$ gives $p(w) = \frac{1}{3}\Phi(1/\sqrt{w}) + \frac{2}{3}\Phi(-2/\sqrt{w})$, which first decreases and then increases. These examples identify the role of the input assumption in the inverse problem.

C.5 A fixed collection of inputs

For training experiments, the calibration data contain several inputs, propagated through the same weights. Each input has its own initial energy and empirical coordinate distribution. For a fixed number N_{pr} of inputs, let their limiting laws be ν_b and let $\bar{q}'_b = a^2 v_\phi(\bar{q}_b)$ with $\bar{q}_b(0) = q_{0,b} > 0$. Their average limiting inactive fraction is

$$\bar{p}(a) = \frac{1}{N_{\mathrm{pr}}} \sum_{b=1}^{N_{\mathrm{pr}}} \int \Phi\left(-\frac{u}{\sqrt{\bar{q}_b(1) - q_{0,b}}} \right) \nu_b(du).$$

For positive input laws this is again continuous and strictly increasing from zero to one half. A scalar root search in a, evaluating one radial ODE per input, therefore gives a unique coefficient. Applying the single-input theorem to each input proves convergence of their average fraction in L^2: the norm of the average error is at most the average of the individual L^2 norms. Independence between inputs is unnecessary. The same root-bracketing argument proves consistency when the empirical coordinate

laws are used in the calculation. This is the rule used for the image inputs in Appendices F.1 and I; each input keeps its own energy curve. For nonnegative input laws with zero masses $\rho_b < 1$, the same argument gives the target interval $(\frac{1}{2N_{\mathrm{pr}}}\sum_b \rho_b,\ 1/2)$. An independently initialized embedding can be conditioned on throughout; the convergence statements then require the corresponding empirical input laws to satisfy the stated assumptions.

C.6 Alignment with the input

Corollary C.3. *In the setting of Theorem 3.3, with $\sigma = a_{p,n}/\sqrt{L}$, the state before the terminal rectifier satisfies*

$$\frac{\langle u, x_L\rangle}{\|u\|\|x_L\|} \xrightarrow{L^2} \sqrt{\frac{q_0}{q_0 + w_p}}.$$

Thus matching p for the same input distribution also matches the limiting alignment with the identity path across activations. The proof below controls the martingale $\langle u, x_k - u\rangle/n$. For a fixed input law, p, w_p/q_0 and this cosine parameterize the same one-dimensional curve. Correlations between inputs still depend on the activation: ReLU and leaky ReLU have distinct cross-covariances of their increments even after their single-input variances are matched (Appendix A.1).

The calibrated variance also describes how far the residual state has turned away from its input. Put $N_k = \langle u, x_k - u\rangle/n$. Conditional centering makes it a martingale, with

$$\mathbb{E}[(N_{k+1} - N_k)^2 \mid \mathcal{F}_k] = \frac{a^2 Q_0}{nL} v_\phi(Q_k).$$

The uniform energy bound gives $\mathbb{E}N_L^2 \le C/n$. Hence $\langle u, x_L\rangle/n \to q_0$, whereas (31) gives $\|x_L\|^2/n \to q_0 + w$. Their ratio converges in probability, and the cosine is bounded by one in absolute value, so

$$\frac{\langle u, x_L\rangle}{\|u\|\|x_L\|} \xrightarrow{L^2} \sqrt{\frac{q_0}{q_0 + w}}. \tag{39}$$

After calibration, substitute w_p. For the same limiting input law, this alignment is independent of the branch activation. For equal positive coordinates it is $z_p/\sqrt{1 + z_p^2}$, where $z_p = -\Phi^{-1}(p)$. For a half-normal input law it is $\cos(\pi p)$.

Example C.4. For a constant input $u_i = s > 0$, $\Phi(-s/\sqrt{w_p}) = p$ gives $w_p = s^2/\Phi^{-1}(p)^2$, and for $v_\phi(q) = c_\phi q$

$$a_p^2 = \frac{1}{c_\phi} \log\Big(1 + \frac{1}{\Phi^{-1}(p)^2}\Big), \tag{40}$$

which does not involve s; for ReLU, $c_\phi = 1/2$. If ν is the half-normal law of $s|G|$, then $q_0 = s^2$, rotation invariance of a standard Gaussian pair gives $p = \pi^{-1}\arctan\sqrt{w_p/q_0}$, so $w_p = q_0\tan^2(\pi p)$ and $a_p^2 = -(2/c_\phi)\log\cos(\pi p)$. For ReLU at $p = 0.25$, these two input laws require $a_p \approx 1.52$ and 1.18, respectively, despite having the same second moment. The input distribution matters beyond its variance.

D Activation profiles and input scale

D.1 Explicit Gaussian second moments

For ReLU and leaky ReLU with negative slope β,

$$v_{\mathrm{ReLU}}(q) = \frac{q}{2}, \qquad v_{\mathrm{leaky}}(q) = \frac{1 + \beta^2}{2} q.$$

For exact GELU, $\phi(x) = x\Phi(x)$,

$$v_{\mathrm{GELU}}(q) = q\left[\frac{1}{4} + \frac{1}{2\pi}\arcsin\frac{q}{1+q} + \frac{q}{\pi(1+q)\sqrt{1+2q}}\right]. \tag{41}$$

Here is a derivation. Set $F(t) = \mathbb{E}\Phi(tG)^2$, and let φ denote the standard normal density. Gaussian integration by parts gives

$$J(t) := \mathbb{E}[G\Phi(tG)\varphi(tG)] = t\mathbb{E}\varphi(tG)^2 - t^2 J(t) = \frac{t}{2\pi\sqrt{1+2t^2}} - t^2 J(t).$$

Thus $F'(t) = 2J(t) = t/[\pi(1+t^2)\sqrt{1+2t^2}]$. Integrating from $F(0) = 1/4$ yields $F(t) = 1/4 + (2\pi)^{-1}\arcsin(t^2/(1+t^2))$. A second integration by parts gives

$$\mathbb{E}[G^2\Phi(tG)^2] = F(t) + 2t^2\mathbb{E}\varphi(tG)^2 - 2t^3 J(t) = F(t) + \frac{t^2}{\pi(1+t^2)\sqrt{1+2t^2}}.$$

Taking $t = \sqrt{q}$ proves (41).

The SELU profile is the zero-mean second-moment map in Klambauer et al. (2017). Using $\phi(x) = \lambda x$ for $x > 0$ and $\lambda\alpha(e^x - 1)$ for $x \le 0$ gives

$$v_{\mathrm{SELU}}(q) = \lambda^2\left[\frac{q}{2} + \alpha^2\left(e^{2q}\Phi(-2\sqrt{q}) - 2e^{q/2}\Phi(-\sqrt{q}) + \frac{1}{2}\right)\right]. \tag{42}$$

This follows by expanding the square on the negative half-line and using $\mathbb{E}[e^{tG}\mathbf{1}\{G < 0\}] = e^{t^2/2}\Phi(-t)$. All four activations are globally Lipschitz and vanish at zero. The profile limits below give the strictly positive lower and finite upper bounds required in the main text.

D.2 Why GELU and SELU require opposite corrections

Write $k_\phi(q) = v_\phi(q)/q$. Symmetry gives

$$k_{\mathrm{GELU}}(q) = \frac{1}{4} + \mathbb{E}\left[G^2\left(\Phi(\sqrt{q}G) - \frac{1}{2}\right)^2\right].$$

For each $G \neq 0$, the squared term is strictly increasing in q. Dominated convergence shows that k_{GELU} increases strictly from $1/4$ to $1/2$. For SELU, put $h(z) = (1 - e^{-z})/z = \int_0^1 e^{-tz}\,dt$, with $h(0) = 1$. Then

$$k_{\mathrm{SELU}}(q) = \frac{\lambda^2}{2} + \lambda^2\alpha^2\mathbb{E}[G^2 h(\sqrt{q}|G|)^2\mathbf{1}\{G < 0\}].$$

The function h decreases strictly from one to zero. Therefore the SELU ratio decreases strictly from $\lambda^2(1+\alpha^2)/2$ to $\lambda^2/2$.

Multiplying a fixed positive input by s multiplies q_0 and w_p by s^2. A change of variables in the inverse formula gives

$$a_p(s)^2 = \int_{q_0}^{q_0+w_p} \frac{dt}{t\,k_\phi(s^2 t)}. \tag{43}$$

The preceding monotonicities prove the opposite directions of the two calibration corrections. The endpoint limits of k_ϕ give $a_p(0+)/a_p(\infty) = \sqrt{2}$ for GELU and $a_p(\infty)/a_p(0+) = \sqrt{1+\alpha^2}$ for SELU. For either homogeneous rectifier, k_ϕ is constant and the coefficient is independent of s. In fact, positive homogeneity makes the finite-depth state scale exactly by s for every realization of the weights.

To prove the converse in Theorem 3.4, use invariance for every target. For a fixed positive input law, put $x = \log(s^2 q_0)$, $\ell = \log(1 + w_p/q_0)$, and $h_0(x) = 1/k_\phi(e^x)$. As p ranges over $(0, 1/2)$, ℓ ranges over $(0, \infty)$. Scale independence says that $\int_x^{x+\ell} h_0(t)\,dt$ is independent of x for every $\ell > 0$. Divide by ℓ and let $\ell \downarrow 0$. Continuity forces h_0 to be constant, and hence $v_\phi(q) = c_\phi q$.

D.3 The same scale directions at fixed width

The same input-scale directions hold at fixed width. Fix $n \ge 2$, $a > 0$ and a positive input shape y. For input sy, the radial energy of the limiting diffusion satisfies

$$dQ_t = a^2 v_\phi(Q_t)\,dt + \frac{2a}{\sqrt{n}}\sqrt{Q_t v_\phi(Q_t)}\,dW_t^{\mathrm{rad}}, \qquad Q_0 = s^2\|y\|^2/n.$$

For $s_2 > s_1$, use the same radial Brownian motion. On a compact subinterval of $(0, \infty)$ the coefficients are Lipschitz. Their difference $\Delta_t = Q_t^{s_2} - Q_t^{s_1}$ satisfies $d\Delta_t = \Delta_t b_t \, dt + \Delta_t c_t \, dW_t^{\mathrm{rad}}$ with bounded divided-difference coefficients. Its stochastic-exponential solution is strictly positive. Exhaust the compact intervals, using nonattainment of zero and nonexplosion, to obtain $Q_t^{s_2} > Q_t^{s_1}$ at every finite time.

Choose a common spherical Brownian motion, independent of the radial driver, starting at $y/\|y\|$. Its clock is $T_s = (a^2/n) \int_0^1 k_\phi(Q_t^s) \, dt$. This clock increases strictly with s for GELU and decreases strictly for SELU. Conditional on the radial path, the mean hemisphere probability is a strictly increasing function of the clock, by (26). The expected inactive fraction thus increases with input scale for GELU and decreases for SELU. Combining this with its strict increase in a proves the corresponding directions for the exact fixed-width inverse.

The endpoint ratios also agree. The drift and diffusion coefficients of $\log Q_t$ are bounded uniformly in s, so $\sup_{t \le 1} |\log Q_t - \log Q_0|$ is uniformly tight. Therefore Q_t tends uniformly in probability to zero as $s \downarrow 0$, while its minimum tends to infinity as $s \uparrow \infty$. The angular diffusion time converges to $a^2 c_0/n$ or $a^2 c_\infty/n$, where c_0, c_∞ are the two limits of k_ϕ. Bounded continuity gives convergence of the hemisphere probabilities. Bracketing their unique inverse gives a ratio $\sqrt{c_\infty/c_0}$ or its reciprocal, proving the same factors stated above for the fixed-width diffusion inverse.

E JOINT ACTIVITY ON A FIXED INPUT SET

We prove Theorems 4.1 and 4.2, then derive the signed Gaussian inverse and Corollary 4.4. The number N of inputs is fixed throughout. The state width n and depth L grow at arbitrary relative rates; the branch width m need not grow.

E.1 AN EXACT TRANSITION FOR SHARED WEIGHTS

Place the N states in the columns of $X_k \in \mathbb{R}^{n\times N}$ and put $C_k = X_k^{\mathsf{T}} X_k/n$. Let $\mathcal{F}_k$ contain the inputs and all matrices preceding layer k. Conditional on $\mathcal{F}_k$, the rows of $A_k X_k$ are independent $N(0, C_k)$ vectors. Thus the shared-matrix network has the exact transition

$$\begin{aligned} z_{kj} &\sim N(0, C_k) \quad (j = 1, \dots, m), \qquad H_k = \frac{1}{m}\sum_{j=1}^{m} \phi(z_{kj})\phi(z_{kj})^{\mathsf{T}}, \\ X_{k+1} &= X_k + \frac{a}{\sqrt{L}} G_k H_k^{1/2}, \end{aligned} \tag{44}$$

where G_k has independent standard Gaussian entries and is independent of the branch vectors conditional on the past. A symmetric positive semidefinite square root is used. Conditional on H_k, different coordinate rows receive independent increments, each with covariance $a^2 H_k/L$. Dependence across the N inputs is retained in H_k. In particular,

$$\mathbb{E}[H_k \mid \mathcal{F}_k] = \mathsf{K}_\phi(C_k).$$

E.2 CONVERGENCE OF THE MATRIX OF INNER PRODUCTS

In this proof, $\bar{C}$ denotes the solution of (14) with the finite array's initial matrix $C_{0,n}$ and coefficient a. For all sufficiently large arrays there are fixed constants $\lambda, R_0, A > 0$ such that

$$C_{0,n} \succeq \lambda I_N, \qquad \operatorname{tr} C_{0,n} \le R_0, \qquad 0 \le a \le A. \tag{45}$$

This follows from 2-Wasserstein convergence of the input laws, positive definiteness of C_0, and convergence of the coefficients. Constants denoted by K_* below can depend on these quantities, on N, and on the activation, but not on n, m, L.

The map K_ϕ is locally Lipschitz on the positive-definite cone: differentiate the Gaussian density on a compact subset of that cone, and use the linear growth of ϕ to dominate the integrals. Moreover, $\mathsf{K}_\phi(C)$ is positive semidefinite and $\operatorname{tr}\mathsf{K}_\phi(C) \le K^2 \operatorname{tr} C$, where K is a Lipschitz constant of ϕ. Consequently the ODE is well posed,

$$\bar{C}(t) \succeq C_{0,n}, \qquad \operatorname{tr}\bar{C}(t) \le R_0 e^{A^2K^2},$$

and its Euler iteration remains in the same fixed positive-definite region.

Write

$$D_{k+1} = C_{k+1} - C_k - \frac{a^2}{L}\mathsf{K}_\phi(C_k).$$

Expanding (44) shows that $\mathbb{E}[D_{k+1} \mid \mathcal{F}_k] = 0$. The conditional variance has the identity

$$\begin{aligned} \mathbb{E}[\|D_{k+1}\|_{\mathrm{F}}^2 \mid \mathcal{F}_k] = {} & \frac{2a^2}{nL}\{\operatorname{tr} C_k \operatorname{tr}\mathsf{K}_\phi(C_k) + \operatorname{tr}(C_k\mathsf{K}_\phi(C_k))\} \\ & + \frac{a^4}{mL^2}\{\mathbb{E}\|\phi(Z)\|^4 - \|\mathsf{K}_\phi(C_k)\|_{\mathrm{F}}^2\} + \frac{a^4}{nL^2}\mathbb{E}\{(\operatorname{tr} H_k)^2 + \|H_k\|_{\mathrm{F}}^2 \mid \mathcal{F}_k\}, \end{aligned} \tag{46}$$

where $Z \sim N(0, C_k)$ conditional on $\mathcal{F}_k$. To verify the formula, put $S = X_k^{\mathsf{T}} G_k H_k^{1/2}$. The centered increment is the sum of $a(S+S^{\mathsf{T}})/(n\sqrt{L})$, $a^2(H_k - \mathsf{K}_\phi(C_k))/L$, and the centered Gaussian sample-covariance term $a^2 H_k^{1/2}(G_k^{\mathsf{T}}G_k/n - I_N)H_k^{1/2}/L$. Their cross terms have zero expectation. The first term uses

$$\mathbb{E}[\|S\|_{\mathrm{F}}^2 \mid H_k, \mathcal{F}_k] = n \operatorname{tr} C_k \operatorname{tr} H_k, \qquad \mathbb{E}[\operatorname{tr}(S^2) \mid H_k, \mathcal{F}_k] = n \operatorname{tr}(C_k H_k).$$

The other terms are the variance of an average of m independent matrices and the usual Gaussian sample-covariance second moment.

The Lipschitz bound gives $\mathbb{E}\|\phi(Z)\|^4 \le 3K^4(\operatorname{tr} C_k)^2$. Thus (46) is at most

$$K_*(\operatorname{tr} C_k)^2 \left(\frac{1}{nL} + \frac{1}{mL^2} + \frac{1}{nL^2}\right). \tag{47}$$

We also need uniform trace moments. For each fixed integer $p \ge 1$, Gaussian moments give $\mathbb{E}[(\operatorname{tr} H_k)^p \mid \mathcal{F}_k] \le K_*(\operatorname{tr} C_k)^p$. Expand the pth power of $\|X_k + aG_kH_k^{1/2}/\sqrt{L}\|_{\mathrm{F}}^2/n$. Its leading term is $(\operatorname{tr} C_k)^p$; odd Gaussian terms vanish and every other term has at least one factor a^2/L. Bounding the remaining Gaussian moments yields

$$\mathbb{E}[(\operatorname{tr} C_{k+1})^p \mid \mathcal{F}_k] \le (1 + K_*/L)(\operatorname{tr} C_k)^p.$$

Iteration, followed by Doob's inequality for the nonnegative submartingale $\operatorname{tr} C_k$, proves

$$\mathbb{E}\max_{k\le L}(\operatorname{tr} C_k)^p \le K_*, \qquad p = 2, 4. \tag{48}$$

Applying the second-moment martingale maximal inequality to (47) gives

$$\mathbb{E}\max_{j\le L}\left\|\sum_{k<j} D_{k+1}\right\|_{\mathrm{F}}^2 \le K_*\left(\frac{1}{n} + \frac{1}{mL}\right). \tag{49}$$

For completeness, local Lipschitz continuity suffices to compare this recursion with the ODE. Stop at the first index τ where $\lambda_{\min}(C_\tau) < \lambda/2$ or $\operatorname{tr} C_\tau > R_0 e^{A^2K^2} + 1$. Let b_k be the deterministic Euler iteration. Before exit, discrete Gronwall and (49) give

$$\mathbb{E}\max_{k\le L}\|C_{k\wedge\tau} - b_{k\wedge\tau}\|_{\mathrm{F}}^2 \le K_*\left(\frac{1}{n} + \frac{1}{mL}\right).$$

This estimate includes the possible exit step, since every drift term in the stopped sum is evaluated before exit. At exit the difference from b_τ has norm bounded below by a fixed positive constant, so $\mathbb{P}(\tau \le L) \to 0$. The Euler error is $O(L^{-1})$. It follows that

$$\max_{k\le L}\|C_k - \bar{C}(k/L)\|_{\mathrm{F}} \xrightarrow{\mathbb{P}} 0. \tag{50}$$

The trace moments in (48) make the maximum matrix difference uniformly integrable. The same argument, using the linear trace bound on K_ϕ, gives

$$\mathbb{E}\max_{k\le L}\|\mathsf{K}_\phi(C_k) - \mathsf{K}_\phi(\bar{C}(k/L))\|_{\mathrm{F}} \longrightarrow 0. \tag{51}$$

In particular, finite branch width contributes a term of order $1/(mL)$ before this limit, which vanishes even for $m = 1$.

E.3 One and two coordinate rows

Choose $r = 1$ or $r = 2$ distinct coordinate indices $i_1, \dots, i_r$, and fix $\theta_1, \dots, \theta_r \in \mathbb{R}^N$. Let $\Delta X_{k,i}$ denote row i of $X_{k+1} - X_k$ as a column vector, and put $\Theta = \sum_{j=1}^r \theta_j\theta_j^{\mathsf{T}}$. Equation (44) gives

$$\mathbb{E}\left[e^{\,\mathrm{i}\sum_j \theta_j^{\mathsf{T}}\Delta X_{k,i_j}} \,\middle|\, \mathcal{F}_k\right] = \mathbb{E}\left[e^{-a^2\operatorname{tr}(H_k\Theta)/(2L)} \,\middle|\, \mathcal{F}_k\right].$$

Since $|e^{-x} - 1 + x| \le x^2/2$ for $x \ge 0$, the difference between this conditional characteristic function and

$$\exp\left\{-\frac{a^2}{2L}\operatorname{tr}(\mathsf{K}_\phi(\bar{C}(k/L))\Theta)\right\}$$

has expected absolute value at most

$$\frac{K_\theta}{L}\mathbb{E}\|\mathsf{K}_\phi(C_k) - \mathsf{K}_\phi(\bar{C}(k/L))\|_{\mathrm{F}} + \frac{K_\theta}{L^2}.$$

The last term uses (48). Summing these errors, telescoping the conditional characteristic functions, and using (51) proves

$$\mathbb{E}e^{\,\mathrm{i}\sum_{j=1}^{r}\theta_j^{\mathsf{T}}(X_{L,i_j}-U_{i_j})}\longrightarrow\exp\left\{-\frac{1}{2}\sum_{j=1}^{r}\theta_j^{\mathsf{T}}W_a\theta_j\right\}. \tag{52}$$

The bound is uniform over the selected distinct coordinate indices. Here continuous dependence of the ODE on its initial matrix and coefficient identifies the limiting W_a. The Gaussian limits for distinct coordinate rows are independent, while the N inputs within one row retain covariance W_a.

To complete the proof of Theorem 4.1, choose a coordinate index I uniformly, independently of the network. Multiplying (52) by a characteristic-function factor of U_I proves joint convergence $(U_I, X_{L,I}-U_I)\Rightarrow(U,G_a)$, with independence. For two distinct uniform indices I,J, the initial joint law differs from $\nu_n^{(N)}\otimes\nu_n^{(N)}$ by at most $O(1/n)$ on bounded tests. The case $r=2$ therefore gives two independent copies of (U,G_a). Expanding the first and second moments of an empirical average proves its L^2 convergence for bounded continuous tests.

For $a>0$, every diagonal entry of W_a is positive because $\bar{C}'_{ss}=a^2v_\phi(\bar{C}_{ss})$ and $\bar{C}_{ss}(0)>0$. Every terminal marginal consequently has a continuous Gaussian convolution. The boundary of the negative orthant has zero mass, so the same first- and second-moment argument proves (15).

Finally the empirical terminal row law converges weakly in probability, and its second moment is $\operatorname{tr}C_L$. By (50), that second moment converges to $\operatorname{tr}\bar{C}_a(1)=\mathbb{E}\|U+G_a\|^2$. The characterization of 2-Wasserstein convergence by weak convergence and convergence of second moments therefore applies, using subsequences if necessary. It yields convergence in probability of the ReLU Gram entries. Each absolute entry is at most $\operatorname{tr}C_L$; the fourth-moment bound in (48) makes their squares uniformly integrable and upgrades the convergence to L^2. This proves the last statement of Theorem 4.1.

E.4 The correlation calculation for homogeneous branches

For $\phi_\beta(z)=\max(z,\beta z)$, $0\le\beta<1$, define

$$\kappa(c)=\frac{\sqrt{1-c^2}+(\pi-\arccos c)c}{\pi},\qquad k_\beta=\frac{1+\beta^2}{2},\qquad \psi_\beta(c)=\frac{2\beta c+(1-\beta)^2\kappa(c)}{1+\beta^2}. \tag{53}$$

Gaussian integration and symmetry give

$$\mathsf{K}_\beta(C)=\beta C+(1-\beta)^2\mathsf{K}_{\mathrm{ReLU}}(C),\qquad (\mathsf{K}_\beta(C))_{st}=k_\beta\sqrt{C_{ss}C_{tt}}\,\psi_\beta(c_{st}),\quad c_{st}=\frac{C_{st}}{\sqrt{C_{ss}C_{tt}}}.$$

Writing $q_{0,s}=(C_0)_{ss}$, differentiation of the diagonal energies and normalized correlations in (14) yields

$$\bar{C}_{ss}(t)=q_{0,s}e^{k_\beta a^2t},\qquad c'_{st}(t)=\frac{a^2(1-\beta)^2}{2}[\kappa(c_{st}(t))-c_{st}(t)]. \tag{54}$$

Thus unequal input energies have the same multiplicative growth. The function $g(c)=\kappa(c)-c$ satisfies $g'(c)=-\arccos(c)/\pi$ and $g(1)=0$, so $g(c)>0$ for $c<1$. Every off-diagonal correlation starting below one increases toward one as a^2t grows. Also, ψ_β is nondecreasing on $[-1,1]$.

The added covariance is positive definite for finite $a>0$. Using $2\,\mathrm{ReLU}(Z)=Z+|Z|$ and Gaussian symmetry gives

$$\mathsf{K}_\beta(C)=\frac{(1+\beta)^2}{4}C+\frac{(1-\beta)^2}{4}\mathbb{E}[|Z||Z|^{\mathsf{T}}]\succeq\frac{(1+\beta)^2}{4}C.$$

Positive definiteness of C_0 then implies $W_a\succ 0$.

E.5 A Gaussian comparison and the positive-input inverse

We use Slepian's Gaussian comparison (Slepian, 1962). For positive-definite correlation matrices S_0,S_1 with $(S_1)_{st}\ge(S_0)_{st}$ whenever $s\ne t$,

$$\mathbb{P}(N(0,S_1)\le b)\ge\mathbb{P}(N(0,S_0)\le b),\qquad b\in\mathbb{R}^N. \tag{55}$$

For completeness, the Gaussian differentiation identity of Plackett (1954) gives a short proof. Set $S(v) = S_0 + v(S_1 - S_0)$ and let p_v be its density. Differentiating its Gaussian characteristic function gives $\partial_v p_v = \frac{1}{2}\sum_{s,t}(S_1 - S_0)_{st}\partial_s\partial_t p_v$. The diagonal differences vanish. Integrating over the lower orthant and then integrating twice by parts expresses the derivative of its probability as

$$\sum_{s<t}(S_1 - S_0)_{st}\int_{z_{-\{s,t\}}\le b_{-\{s,t\}}} p_v(b_s, b_t, z_{-\{s,t\}})\,\mathrm{d}z_{-\{s,t\}} \ge 0.$$

If any off-diagonal difference is positive and $N \ge 2$, this derivative is strictly positive. Gaussian decay justifies the integrations; for $N = 2$ the displayed integral is the bivariate boundary density.

We now prove part (i) of Theorem 4.2. Let $D_0 = \mathrm{diag}(\sqrt{q_{0,1}}, \dots, \sqrt{q_{0,N}})$, $h = e^{k_\beta a^2}$, and let $R(a)$ be the terminal correlation matrix. The covariance of the *added* Gaussian noise is

$$W_a = (h-1)D_0 S(a) D_0, \qquad S_{st}(a) = \frac{hc_{st}(a) - c_{st}(0)}{h-1}. \tag{56}$$

Its correlation S_{st} is distinct from the total-state correlation c_{st}. Parameterize the latter by h and write it as $\widetilde{c}_{st}(h)$. Equation (54) gives $\frac{\mathrm{d}}{\mathrm{d}h}(h\widetilde{c}_{st}(h)) = \psi_\beta(\widetilde{c}_{st}(h))$, hence

$$S_{st}(a) = \frac{1}{h-1}\int_1^h \psi_\beta(\widetilde{c}_{st}(v))\,\mathrm{d}v. \tag{57}$$

The integrand is nondecreasing, so this average is nondecreasing in a.

Conditional on $U = u \in (0,\infty)^N$, the probability in $D_N(a)$ is the Gaussian orthant probability with correlation $S(a)$ and thresholds

$$b_s(a,u) = -\frac{u_s}{\sqrt{q_{0,s}}\sqrt{h-1}}.$$

Every threshold increases strictly with a, while all off-diagonal noise correlations are nondecreasing. Apply (55) at the larger thresholds, and then increase the thresholds at the smaller covariance. The second change is strict because $S(a) \succ 0$ has an everywhere positive Gaussian density. Thus the conditional probability, and hence $D_N(a)$, is strictly increasing. For $N = 1$ only the threshold argument is needed.

Continuity follows from continuity of Gaussian probabilities and dominated convergence. As $a \downarrow 0$, each positive u_s makes its marginal closure probability tend to zero, which also bounds joint closure. As $a \to \infty$, all c_{st} tend to one, and (57) gives $S_{st} \to 1$. The thresholds tend to zero, and the standardized Gaussian vector converges to a common scalar Gaussian in every component. Its all-nonpositive probability is $1/2$; the boundary has zero probability. Dominated convergence over U proves the asserted endpoints and the unique inverse.

For empirical calibration, compute the same curve $D_{N,n}(a)$ with $\nu_n^{(N)}$ and $C_{0,n}$. Strict positivity of every input coordinate and eventual positive definiteness of $C_{0,n}$ give a unique root $a_{d,n}$. At each fixed $a > 0$, ODE continuity and convergence of the input laws imply $D_{N,n}(a) \to D_N(a)$. The Gaussian smoothing makes the limiting orthant a continuity set. For $a_- < a_d < a_+$, strict monotonicity places the empirical root between a_- and a_+ for all sufficiently large n. Shrinking the bracket proves $a_{d,n} \to a_d$. Applying Theorem 4.1 with these coefficients gives $\widehat{d}_N \to d$ in L^2. The coefficients depend on the input array and are independent of the residual matrices.

E.6 Signed Gaussian inputs

If $\nu^{(N)} = N(0, C_0)$, the limiting row $U + G_a$ is $N(0, \bar{C}_a(1))$. The marginal closure probabilities are all $1/2$, while joint closure depends on the correlation matrix $R(a)$. Every off-diagonal entry of $R(a)$ is strictly increasing for $\beta < 1$, by (54) and $C_0 \succ 0$. The strict form of (55), applied at zero thresholds, makes $D_N(a)$ strictly increasing for $N \ge 2$. Its lower endpoint is $\mathbb{P}(N(0, C_0) \le 0)$. Its upper endpoint is $1/2$ because $R(a)$ tends to the all-ones matrix. Continuity proves part (ii) of Theorem 4.2.

Corollary E.1. *Suppose the initial Gaussian correlation matrix is equicorrelated with off-diagonal value $c_0 \in [0, 1)$ and $N \ge 2$. Define*

$$J_N(c) = \int_{\mathbb{R}} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,\Phi\!\left(-\sqrt{\frac{c}{1-c}}\,z\right)^N \mathrm{d}z, \qquad 0 \le c < 1. \tag{58}$$

For any $d \in (J_N(c_0), 1/2)$ there is a unique $c_d \in (c_0, 1)$ with $J_N(c_d) = d$, and the unique branch coefficient is

$$a_d^2 = \frac{2}{(1-\beta)^2} \int_{c_0}^{c_d} \frac{\mathrm{d}c}{\kappa(c) - c}. \tag{59}$$

In particular, $J_N(0) = 2^{-N}$ and $J_2(c) = 1/4 + \arcsin(c)/(2\pi)$.

Proof. An equicorrelated standard Gaussian vector can be written as $\sqrt{c}\, Z + \sqrt{1-c}\,(Z_1, \dots, Z_N)$, with all displayed scalar Gaussians independent. Conditioning on Z gives (58). The Gaussian comparison makes J_N strictly increasing, with upper limit $1/2$. Separating variables in (54) gives (59). The integral is finite because $c_d < 1$. The value at zero follows by independence. For two inputs, write a correlated pair as projections of an isotropic Gaussian vector in the plane; the intersection of their negative half-planes has angular measure $\pi - \arccos c$, giving the stated value. □

Unequal positive coordinate variances do not change this formula: orthant probabilities depend only on the correlation matrix. If a Gaussian input covariance is estimated from coordinates, continuous dependence of the flow and root bracketing also give a consistent coefficient computed from that Gaussian model. This uses the Gaussian model for the joint law, rather than asserting monotonicity of an arbitrary signed empirical mixture.

E.7 Matching normalized terminal feature laws

To prove Corollary 4.4, substitute $a_\beta = a_0/(1-\beta)$ into (54). Every normalized correlation path agrees with its ReLU counterpart. All diagonal entries differ by the same factor h_β/h_0. Therefore

$$\frac{\bar{C}_{\beta, a_\beta}(1)}{h_\beta} = \frac{\bar{C}_{0, a_0}(1)}{h_0}.$$

The limiting terminal row laws are centered Gaussian, so equality of these covariance matrices is equality of their normalized joint laws. Positive homogeneity of the terminal ReLU preserves that equality. Theorem 4.1 then gives the matching normalized feature Gram matrices.

For a precise readout consequence, let $\mathcal{F} \in \mathbb{R}^{N \times n}$ contain the normalized terminal features. Hold $\mathcal{F}$ fixed and train the readout $\widehat{y} = n^{-1/2}\mathcal{F}w$ with loss $\|\widehat{y} - y\|^2/(2N)$. Gradient flow for the prediction error $e = \widehat{y} - y$ satisfies

$$\dot{e} = -\frac{\mathcal{F}\mathcal{F}^{\mathsf{T}}}{nN}\, e.$$

The normalized limiting matrix on the right agrees under the transfer, so its limiting dynamics agree for the same initial error. Here only the readout weights w evolve; the residual stack remains fixed.

E.8 Calibration for the deterministic finite-depth covariance recursion

To account for discretization at large residual scales, retain the individual layers in the deterministic covariance model. Fix N, $L \geq 1$, $C_0 \succ 0$, and $\phi_\beta(z) = \max(z, \beta z)$ with $0 \leq \beta < 1$. Starting from $C_0^{\beta,a} = C_0$, define

$$C_{k+1}^{\beta,a} = C_k^{\beta,a} + \frac{a^2}{L}\mathsf{K}_\beta(C_k^{\beta,a}), \qquad h_{\beta,L}(a) = \left(1 + \frac{k_\beta a^2}{L}\right)^L, \qquad k_\beta = \frac{1+\beta^2}{2}. \tag{60}$$

Proposition E.2. *For every pair of inputs, the normalized correlations in* (60) *satisfy*

$$c_{k+1} = c_k + t_\beta(a, L)[\kappa(c_k) - c_k], \qquad t_\beta(a, L) = \frac{a^2(1-\beta)^2}{2L + (1+\beta^2)a^2}. \tag{61}$$

For a ReLU reference coefficient $a_0 \geq 0$, suppose

$$(1-\beta)^2 - \frac{\beta a_0^2}{L} > 0.$$

Then the coefficient

$$a_{\beta,L} = \frac{a_0}{\sqrt{(1-\beta)^2 - \beta a_0^2/L}} \tag{62}$$

matches every correlation at every layer and gives

$$\frac{C_L^{\beta,a_{\beta,L}}}{h_{\beta,L}(a_{\beta,L})} = \frac{C_L^{0,a_0}}{h_{0,L}(a_0)}. \tag{63}$$

Consequently, centered Gaussian vectors with the two terminal covariances, divided by the square roots of their respective h factors, have identical laws. Their orthant probabilities and normalized terminal ReLU feature Grams agree. For $\beta > 0$ and $a_0 > 0$, the denominator condition is necessary for a finite coefficient to match the one-step correlation parameter.

Proof. Put $r = a^2/L$ and $q_{0,s} = (C_0)_{ss}$. The diagonal entries are $(C_k^{\beta,a})_{ss} = q_{0,s}(1 + k_\beta r)^k$. The kernel identity from (53) gives

$$c_{k+1} = \frac{c_k + r[\beta c_k + (1-\beta)^2\kappa(c_k)/2]}{1 + k_\beta r} = c_k + \frac{r(1-\beta)^2}{2(1+k_\beta r)}[\kappa(c_k) - c_k],$$

which proves (61). Equating its step parameter to $t_0(a_0, L) = a_0^2/(2L + a_0^2)$ gives

$$a_{\beta,L}^2\left((1-\beta)^2 - \frac{\beta a_0^2}{L}\right) = a_0^2.$$

This has the stated finite nonnegative solution exactly under the denominator condition. Matching the correlations inductively, and dividing their diagonal energies by h, proves (63). Equality of normalized Gaussian laws then follows from equality of their covariances. Positive homogeneity of the final ReLU transfers the same equality to feature Grams. □

For $\beta = 0$, (62) returns a_0 exactly. For fixed a_0 and $\beta < 1$, it tends to $a_0/(1-\beta)$ as $L \to \infty$. If the denominator is zero, the matching step parameter is approached only as $a_{\beta,L} \to \infty$; if it is negative, it is unattainable.

The finite-depth attainable range. Write $T_t(c) = c + t[\kappa(c) - c]$ and $\gamma_\beta = (1-\beta)^2/(1+\beta^2)$. For finite $a \geq 0$, the step parameter ranges over $[0, \gamma_\beta)$. In particular, increasing a at a fixed depth does not drive the deterministic terminal correlations all the way to one.

Corollary E.3. *Suppose $N \geq 2$ and C_0 has an equicorrelation $c_0 \in [0, 1)$. Let J_N be defined by* (58) *and put*

$$c_{\beta,L}^{\max} = T_{\gamma_\beta}^{\circ L}(c_0) < 1.$$

The deterministic Gaussian closure probability $D_{N,\beta,L}^{\mathrm{disc}}(a) = J_N(T_{t_\beta(a,L)}^{\circ L}(c_0))$ is continuous and strictly increasing, with range

$$[J_N(c_0),\ J_N(c_{\beta,L}^{\max})).$$

For any interior target d in this range, first solve $J_N(c_d) = d$, then solve $T_{t_d}^{\circ L}(c_0) = c_d$ for the unique $t_d \in (0, \gamma_\beta)$. Its unique coefficient is

$$a_{\beta,L,d}^2 = \frac{2Lt_d}{(1-\beta)^2 - (1+\beta^2)t_d}. \tag{64}$$

For ReLU this reduces to $a_{0,L,d}^2 = 2Lt_d/(1-t_d)$. Using this ReLU coefficient in (62) *gives the same leaky-ReLU inverse whenever its denominator is positive.*

Proof. For $-1 < c < 1$, $g(c) = \kappa(c) - c > 0$ and $g'(c) = -\arccos(c)/\pi$. Therefore, for $0 \leq t \leq 1$, $\partial_c T_t(c) = 1 - t\arccos(c)/\pi > 0$ and $\partial_t T_t(c) = g(c) > 0$. Moreover $T_t(c) = (1-t)c + t\kappa(c) < 1$ when $c < 1$. Induction proves continuity and strict monotonicity of the L-step terminal correlation in t, and that its upper endpoint is below one. Strict monotonicity of J_N gives the stated range and unique roots. Solving (61) for a^2 proves (64). □

At fixed L the upper endpoint is strictly below $1/2$, even for ReLU. For example, at $L = 1$, $N = 2$ and $c_0 = 0$, the maximal ReLU closure probability is $J_2(1/\pi) \simeq 0.30156$; for $\beta = 1/2$ it is $J_2(1/(5\pi)) \simeq 0.26014$.

Relation to finite networks. At fixed L, the deterministic recursion is the limit as both $n, m \to \infty$, provided the initial empirical coordinate law converges in 2-Wasserstein distance to $N(0, C_0)$. Indeed, in (44), H_k concentrates at $\mathsf{K}_\beta(C_k)$ as $m \to \infty$, and conditional Gaussian row averaging controls the state Gram as $n \to \infty$. Induction over finitely many layers gives (60) and independent limiting row increments; the Gaussian initial law remains Gaussian. With m fixed, branch-covariance randomness persists at fixed L, and $\mathbb{E}\mathsf{K}_\beta(C_k)$ need not equal $\mathsf{K}_\beta(\mathbb{E}C_k)$. The finite-branch theorem also sends L to infinity, averaging this randomness over layers. The finite-depth correction removes the discretization error of the deterministic recursion.

Interpretation as a correlation target. For centered equicorrelated Gaussian inputs, targeting d is equivalent to targeting $c_d = J_N^{-1}(d)$, in both the continuous and deterministic discrete models. The two targets return the same coefficient, with d expressing the constraint as a fraction of zero feature columns. A general Gaussian input law requires the whole correlation matrix, while positive non-Gaussian inputs also require the mixture thresholds and the correlations of the added noise.

For a consistent equicorrelated Gaussian family at a fixed finite coefficient, its terminal $c < 1$ does not depend on N. Conditioning on the shared Gaussian in (58) gives $\Phi(-\sqrt{c/(1-c)}Z)^N \to 0$ almost surely; bounded convergence therefore gives $J_N(c) \to 0$ as $N \to \infty$. This calculation applies to the stated equicorrelated family. The network limit remains a fixed-N result, calibrating a specified finite input set.

E.9 Joint silence is not determined by pairwise geometry

The non-Gaussian joint law contains information beyond all pairwise feature distributions. The following example isolates this distinction.

Proposition E.4. *Fix $0 < \delta < 1$ and $N = 3$. Let ν_+ and ν_- be the laws of $U = \mathbf{1} + \delta S$, where S is uniform on the four vectors in $\{-1, 1\}^3$ with coordinate product $+1$ or -1, respectively. For the same ReLU or leaky-ReLU branch with $0 \le \beta < 1$, and every $a > 0$, these two laws have identical limiting marginal inactivity and terminal ReLU feature Gram matrices, but*

$$D_-(a) - D_+(a) = \frac{1}{4}\mathbb{P}(-(1+\delta) < G_{a,j} \le -(1-\delta),\ j = 1, 2, 3) > 0. \tag{65}$$

Here $G_a \sim N(0, W_a)$ is the common added Gaussian in Theorem 4.1.

Proof. Every one-coordinate sign marginal is uniform and every two-coordinate sign marginal is uniform on $\{-1, 1\}^2$, under either parity. Thus both strictly positive input laws have $C_0 = \delta^2 I_3 + \mathbf{1}\mathbf{1}^\mathsf{T} \succ 0$, and their covariance flows and added covariances W_a agree. The pair marginals of $U + G_a$ consequently agree, proving equality of marginal inactivity and every terminal ReLU Gram entry. Apply inclusion–exclusion to the Gaussian probability of the box in (65). Its four positive terms are exactly the four negative-parity orthant probabilities, and its four negative terms are the positive-parity probabilities. Each input atom has mass $1/4$. Finally, $W_a \succ 0$ gives an everywhere positive Gaussian density, so the box probability is strictly positive. □

For ReLU branches and $\delta = 0.99$, independent one-dimensional Gaussian integration gives the following limiting values:

a	Common marginal p	$D_+(a)$	$D_-(a)$
0.2	0.240031	9.46×10^{-24}	0.066044
0.5	0.246142	0.000040	0.069119
1.0	0.268020	0.028738	0.080319

The shared covariance is equicorrelated. Writing $G_{a,j} = \sqrt{w}(\sqrt{r}Z + \sqrt{1-r}Z_j)$ reduces the difference to $\frac{1}{4}\mathbb{E}[(f_\ell(Z) - f_h(Z))^3]$, where $\ell = 1 - \delta$, $h = 1 + \delta$, and $f_u(z) = \Phi((-u/\sqrt{w} - \sqrt{r}\,z)/\sqrt{1-r})$. This also gives a direct numerical check of (65).

Joint silence therefore distinguishes unused feature columns even when all pairwise terminal feature laws agree. An unrestricted readout has the same limiting kernel dynamics for the two laws when the features are held fixed and the initial prediction errors agree. A restricted feature budget gives a different comparison, proved next.

E.10 A FINITE-FEATURE CONSEQUENCE OF THE PARITY EXAMPLE

The equal limiting Grams in Proposition E.4 do not determine the rank of a readout retaining only a fixed number of randomly chosen feature columns. This distinction is relevant to a restricted feature budget, not to an unrestricted infinite-width readout. For $M \geq 1$, let $F_{\pm}^{(M)} \in \mathbb{R}^{3\times M}$ have independent columns distributed as $\mathrm{ReLU}(U_{\pm} + G_a)$ from that proposition. For labels y independent of the features and uniform on the sphere $\{y : \|y\|^2 = 3\}$, define

$$R_{\pm}(M) = \mathbb{E} \min_b \frac{\|F_{\pm}^{(M)} b - y\|^2}{3} = 1 - \frac{\mathbb{E}\,\mathrm{rank}\, F_{\pm}^{(M)}}{3}.$$

There is no intercept. The identity follows from $\mathbb{E}yy^{\mathsf{T}} = I_3$ and orthogonal projection onto the column span of $F_{\pm}^{(M)}$. The identity also holds for independent Rademacher labels, whose squared norm is three.

We prove Proposition 4.3 in a slightly more general form, with a fixed M and independent, identically distributed columns within each matrix. Its strict ordering holds for two three-input feature laws with the same one- and two-input gate marginals, a positive continuous density on every nonempty support face, and silence probabilities $d_- > d_+$. Quantitatively,

$$R_-(M) - R_+(M) \geq \frac{d_-^M - d_+^M}{3} > 0.$$

In the exchangeable case, put $p = \Pr(F_{1i} = 0)$ and $q = \Pr(F_{1i} = F_{2i} = 0)$. For $M = 3$, the exact difference is

$$\begin{aligned} R_-(3) - R_+(3) = \big[(1-3q)^2 - 3(p-2q)^2\big](d_- - d_+) \\ + 3(1-p-q)(d_-^2 - d_+^2) + 2(d_-^3 - d_+^3). \end{aligned}$$

Proof of Proposition 4.3. Since $W_a \succ 0$, the limiting preactivation law has a positive density, so its positive entries have a continuous density conditional on each support. The expected Grams agree by the equality of pairwise feature laws in Proposition E.4. For a column support $S \subseteq \{1,2,3\}$, fixing all pairwise gate marginals leaves one free probability $d = \Pr(S = \varnothing)$, with

$$\pi_d(S) = \pi_{d_+}(S) + (-1)^{|S|}(d - d_+).$$

The segment $d \in [d_+, d_-]$ is feasible by convexity. The support pattern specifies which matrix entries are zero. Its generic rank is the largest rank attainable by choosing positive values for the other entries; this is the structural rank of the support pattern. Conditional on that pattern, the sampled matrix has this rank almost surely: if its structural rank is r_*, some $r_* \times r_*$ minor is a nonzero polynomial, so its zero set has Lebesgue measure zero. Thus rank depends only on the supports, not on the positive continuous density of the entries within each support. We may use one fixed conditional density for each support throughout the segment.

Fix the other $M-1$ columns, denoted A, and let $T(A) = \{j : e_j \in \mathrm{span}(A)\}$. A continuously distributed column on face S lies in this span with probability one when $S \subseteq T(A)$, and with probability zero otherwise: in the latter case the intersection is a proper linear subspace of the face's linear hull. The rank increment is consequently $\mathbf{1}\{S \not\subseteq T(A)\}$ almost surely. Differentiating the finite polynomial giving the expected rank of M independent columns yields

$$\frac{d}{dd}\mathbb{E}_d\,\mathrm{rank}\, F = -M\ \Pr_d(T(A) = \varnothing),$$

because $\sum_{S\subseteq T}(-1)^{|S|}$ is one for $T = \varnothing$ and zero otherwise. The event that all other columns vanish has probability d^{M-1} and implies $T(A) = \varnothing$. Integration gives the strict ordering and the lower bound.

For the exchangeable law, each particular support of size $0, 1, 2, 3$ has probability $d, q-d, p-2q+d, 1-3p+3q-d$, respectively. When $M = 3$, enumerating the two other columns gives

$$\Pr_d(T(A) = \varnothing) = (1 - 3q + 3d)^2 - 3(p - 2q + d)^2.$$

Indeed, the contributing cases are two empty columns; one empty and one support of size at least two; and two nonempty supports of size at least two whose union is all three coordinates. Integrating this quadratic, using $dR(3)/dd = \Pr_d(T(A) = \varnothing)$, gives the formula. □

The same argument extends beyond three inputs. For a fixed larger input set, even the feature laws on every proper subset can agree while the expected rank of a finite feature collection differs.

Corollary E.5. *Fix $N \geq 3$, $0 < \delta < 1$, $a > 0$, and a common ReLU or leaky-ReLU branch with $0 \leq \beta < 1$. Let $U_{\mathrm{hi}} = \mathbf{1} + \delta S_{\mathrm{hi}}$ and $U_{\mathrm{lo}} = \mathbf{1} + \delta S_{\mathrm{lo}}$, where the sign vectors are uniform subject to coordinate products $(-1)^N$ and $-(-1)^N$, respectively. Their limiting terminal feature laws agree on every proper subset of the N inputs. Write d_{hi} and d_{lo} for their probabilities of a zero feature column. Then*

$$d_{\mathrm{hi}} - d_{\mathrm{lo}} = 2^{1-N}\mathbb{P}\big(-(1+\delta) < G_{a,j} \leq -(1-\delta),\ 1 \leq j \leq N\big) > 0,$$

where $G_a \sim N(0, W_a)$ is their common added Gaussian displacement. Let $F_{\mathrm{hi}}^{(M)}$ and $F_{\mathrm{lo}}^{(M)}$ consist of M independent columns from the respective limiting feature laws. For every fixed $M \geq 1$, their expected Gram matrices agree, whereas

$$\mathbb{E}\,\mathrm{rank}\, F_{\mathrm{lo}}^{(M)} - \mathbb{E}\,\mathrm{rank}\, F_{\mathrm{hi}}^{(M)} \geq d_{\mathrm{hi}}^M - d_{\mathrm{lo}}^M > 0.$$

The same rank bound holds after adjoining a constant column $\mathbf{1}_N$ to both feature matrices. With or without a fitted intercept, for labels independent of the features and uniform on $\{y : \|y\|^2 = N\}$, the hi-minus-lo difference in expected relative least-squares residuals is at least $(d_{\mathrm{hi}}^M - d_{\mathrm{lo}}^M)/N$. Here the residual is $\min_{b,c} \|y - F^{(M)}b - c\mathbf{1}_N\|^2/\|y\|^2$, with $c = 0$ when no intercept is fitted.

Proof. Every assignment on $k < N$ sign coordinates has 2^{N-k-1} completions under either parity. Thus all proper input marginals agree. In particular, $C_0 = \delta^2 I_N + \mathbf{1}\mathbf{1}^{\mathsf{T}}$ is common, as are the covariance flow and the positive-definite added covariance W_a. Adding the same independent Gaussian therefore preserves equality of all proper terminal marginals, including every pair and hence the expected Gram. Writing the difference of the two parity sums as a product gives

$$d_{\mathrm{hi}} - d_{\mathrm{lo}} = 2^{1-N}\mathbb{E}\prod_{j=1}^{N}\left(\mathbf{1}\{G_{a,j} \leq -(1-\delta)\} - \mathbf{1}\{G_{a,j} \leq -(1+\delta)\}\right).$$

The sign vector with all entries -1 has parity $(-1)^N$, which sets the sign of this expansion. For $N = 3$, the hi law is the minus law of Proposition 4.3. The product is the Gaussian box indicator, and positivity of the density gives the strict inequality.

It remains to compare ranks. Write $[N] = \{1, \ldots, N\}$ and $\pi_d(S)$ for the probability of nonzero support $S \subseteq [N]$. Equal proper gate marginals imply $\Delta\pi(S) + \Delta\pi(S \cup \{j\}) = 0$ whenever $j \notin S$. Starting from $\Delta\pi(\varnothing) = \Delta d$ gives the feasible segment

$$\pi_d(S) = \pi_{d_{\mathrm{lo}}}(S) + (-1)^{|S|}(d - d_{\mathrm{lo}}), \qquad d_{\mathrm{lo}} \leq d \leq d_{\mathrm{hi}}.$$

The Gaussian-smoothed feature laws have continuous densities on each nonempty support face. As in the preceding proof, rank therefore equals structural rank almost surely, so the conditional face densities can be held fixed along this segment. For the other $M - 1$ columns A, put $T(A) = \{j : e_j \in \mathrm{span}(A)\}$. A new face-$S$ column raises rank precisely when $S \not\subseteq T(A)$, almost surely. Differentiating the product law of the M independent columns gives

$$\frac{d}{dd}\mathbb{E}_d\,\mathrm{rank}\, F^{(M)} = -M\mathbb{E}_d \sum_{S \subseteq T(A)} (-1)^{|S|} = -M\mathbb{P}_d\{T(A) = \varnothing\}.$$

The event that the other columns are zero has probability d^{M-1} and implies $T(A) = \varnothing$. Integrating gives the rank bound, including $M = 1$, for which this event has probability one. Finally, uniform sphere labels satisfy $\mathbb{E}yy^{\mathsf{T}} = I_N$ and $\|y\|^2 = N$, so the expected relative residual is $1 - \mathbb{E}\,\mathrm{rank}\, F^{(M)}/N$.

For a fitted intercept, replace $\mathrm{span}(A)$ by $\mathrm{span}[\mathbf{1}_N, A]$. Conditional on the column supports, the augmented rank equals its maximum attainable value almost surely: a minor attaining that value is a nonzero polynomial in the active entries, with the intercept held fixed, and its zero set has Lebesgue measure zero. Thus the face densities can again be held fixed, and the same differentiation gives $-M\mathbb{P}_d\{T(A) = \varnothing\}$. When the other columns vanish, their augmented span is $\mathrm{span}(\mathbf{1}_N)$, which contains no coordinate axis for $N \geq 2$. The lower bound and the residual identity therefore also hold for $[\mathbf{1}_N, F^{(M)}]$. □

This construction uses the standard parity example of signs whose every proper subset is independent. The Gaussian box formula bounds the silence gap by 2^{1-N}. The derivative in the proof also bounds the rank gap above by $M(d_{\mathrm{hi}} - d_{\mathrm{lo}}) \le M2^{1-N}$. For fixed M, the gap in this construction therefore tends to zero with N.

Relation to the finite network. Uniformly sample M distinct feature coordinates from a network in Theorem 4.1, independently of its weights. Conditional sampling with replacement has law $\mu_n^{\otimes M}$, where μ_n is the empirical terminal preactivation law. Since this law converges weakly in probability to the deterministic law of $U + G_a$, the sampled columns converge jointly to M independent draws. Sampling without replacement differs in total variation by at most $M(M-1)/(2n)$. To justify the weak convergence in probability, apply the theorem to a countable convergence-determining family of bounded Lipschitz functions. The corresponding summable metric converges in probability by truncating its tail and using the L^2 convergence of each of its finitely many terms. The Gaussian-smoothed limit has a positive density. Within each open sign chamber the rank after ReLU is locally constant except where nontrivial minors vanish; these exceptional algebraic sets and the coordinate hyperplanes have measure zero. The bounded rank, with or without a constant column, therefore also converges in expectation. This is a random-coordinate, fixed-M consequence of the empirical law, not an assertion that deterministic coordinate indices or finite-network columns are independent.

For $N = 3$, in the balanced four-atom input arrays, equality of expected feature Grams holds even at finite n, m, L. For each selected pair of inputs, the two initial pair arrays differ only by a coordinate permutation P because their empirical pair laws agree exactly. Under the coupling $A'_k = A_k P^{\mathsf{T}}$ and $B'_k = PB_k$, the Gaussian weight law is unchanged and the pair trajectories satisfy $X'_k = PX_k$ at every layer. Consequently, each pair's whole empirical terminal law and corresponding Gram entry agree in distribution. The permutation may depend on the pair: this proves equality of the expected full Gram, not equality of its joint distribution across all entries. Uniformly subsampling coordinates preserves the pairwise distributional equality. Finite measured pairwise differences between the two laws are therefore sampling fluctuations, not a systematic finite-depth discrepancy.

For $\delta = 0.99$ and $a = 0.2$, exact support enumeration gives $R_+(3) = 0.026556$ and $R_-(3) = 0.071946$. The corresponding difference falls to 0.006085 at $M = 4$ and 5.49×10^{-7} at $M = 8$. The general lower bound proves strictness but is conservative: at $M = 3$ it gives approximately 9.6×10^{-5}, whereas the exact difference is 0.045390. The rank ordering holds along this family with fixed pairwise gate marginals. It gives a fitting-risk difference at a restricted feature budget; the unrestricted readout has the same limiting kernel. General feature laws also vary their pairwise marginals, so their ranks require more information than the silent fraction.

E.11 Scope and boundary cases

The Gaussian coordinate law can be obtained from fixed, non-Gaussian raw data. Given $y^{(1)}, \dots, y^{(N)} \in \mathbb{R}^d$, let $u^{(s)} = Ey^{(s)}$, where $E \in \mathbb{R}^{n \times d}$ has independent $N(0, 1/d)$ entries and is independent of the residual weights. Then the coordinate rows U_i are independent $N(0, C_0)$ vectors with $(C_0)_{st} = \langle y^{(s)}, y^{(t)} \rangle / d$. Their empirical law converges almost surely in 2-Wasserstein distance. When this raw-data Gram matrix is positive definite and N is fixed, conditioning on the embedding gives the input assumption of Theorem 4.1.

The positive-definite initial Gram assumption and fixed N are used in the covariance localization. No estimate here is uniform when N grows with n. The finite-network bound $\operatorname{rank} \mathcal{F} \le \min\{N, n(1 - \widehat{d}_N)\}$ follows from zero columns, but an interpolation threshold at comparable N and n is not a corollary of the joint limit.

At $a = 0$ the Gaussian smoothing is absent. Strictly positive finite input arrays have $\widehat{d}_N = 0$ exactly. For other laws, weak convergence identifies the closure probability when the limiting law puts no mass on coordinate hyperplanes. It does not by itself identify the finite exact-zero count when such mass is present. The calibration theorems use interior targets with $a_d > 0$.

The restriction $\beta < 1$ also matters. A linear branch, $\beta = 1$, leaves normalized correlations fixed. For centered Gaussian input laws its joint closure probability is constant. For positive inputs its thresholds still change, but its upper endpoint is the Gaussian orthant probability with the initial correlation matrix, generally below $1/2$.

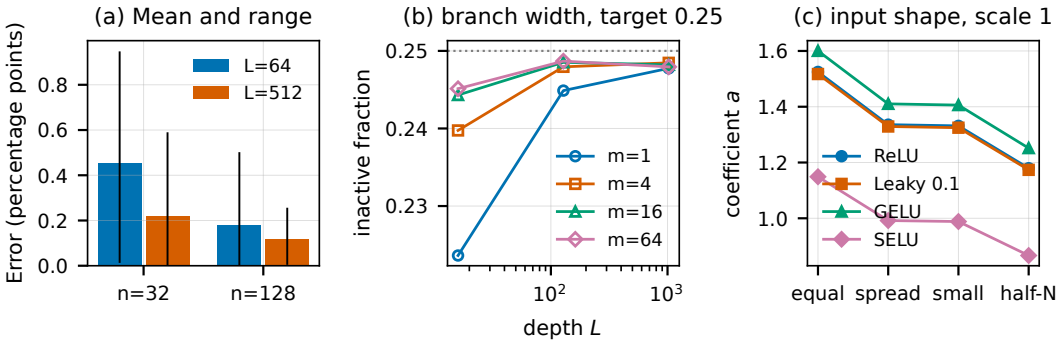


Figure 5: Single-input calibration across architectures and inputs. (a) Error of the mean fraction over 144 target points. (b) Mean over four activations at target 0.25, $n = 128$. (c) Required coefficient for four input shapes of unit RMS.

Experiment	$n; m$	L	Traj./seed	Input and parameter
Depth curves	32; 32	16–1024	4096	Equal; $a = 0, 0.5, \ldots, 2.5$
Diffusion	32; –	2048, 8192 steps	4096	Equal; $a = 0, 0.125, \ldots, 2.5$
Calibration	32, 128; 32	64, 512	512	Four shapes; $p = 0.1, 0.25, 0.4$
Branch width	128; 1–64	16, 128, 1024	1024	Spread; $p = 0.25$
Input scale	128; 32	512	2048	Spread; $p = 0.25$, RMS 0.05–16
Input support	128 (126); 32	512	1024	Zeros or signs; nine $a \in [0, 3]$
Alignment	128, 512; 32	512	1024	Equal, half-normal; three targets

Table 4: Simulation grids for the four branch activations. Each setting uses three independent seed batches.

An arbitrary signed input law can violate monotonicity even with a positive-definite Gram matrix. Let $N = 2$ and take independent U_1, U_2, each equal to -1 with probability $1/3$ and 2 with probability $2/3$. Then $C_0 = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 3 \end{smallmatrix}\right) \succ 0$ and $D_2(0) = 1/9$. Let w be either added marginal variance for small $a > 0$. Loss of closure from the atom $(-1, -1)$ is at least $\Phi(-1/\sqrt{w})/9$, whereas the total possible gain from the other atoms is at most $8\Phi(-2/\sqrt{w})/9$. The latter is negligible relative to the former as $w \downarrow 0$. Thus $D_2(a) < 1/9$ for sufficiently small positive a, although $D_2(a) \to 1/2$ as $a \to \infty$. The centered-Gaussian assumption in the signed inverse cannot be replaced by an unrestricted signed law.

Finally, the feature-law transfer uses Gaussianity, not just a second-moment calculation. For a non-Gaussian $\nu^{(N)}$, the terminal law is the mixture $U + G_a$; dividing by a different energy factor also rescales its unchanged identity component U. Matching normalized Gram matrices before the terminal ReLU then need not match the terminal feature law.

F Numerical verification

Simulator. For one deterministic input, recursion (17) has the network's forward law. We simulate it by drawing m standard normals for H_k and n for g_k at each layer. The shared-input training networks in Appendix I use explicit weight matrices. We use four deterministic positive input shapes: *equal*, all coordinates one; *spread*, equally spaced values from 0.05 to 1.95; *small*, one quarter of the coordinates equal to 0.001 and the rest one; and *half-normal*, midpoint quantiles of the half-normal law. Each is rescaled to a prescribed root-mean-square value, so Q_0 is the squared RMS. Where a target p is prescribed, $\sigma = a_{p,n}/\sqrt{L}$ with $a_{p,n}$ from (9) and the finite-n root $w_{p,n}$. Table 4 lists the grids.

Statistical reporting. Each independently initialized trajectory contributes one terminal fraction (2). Monte Carlo intervals are 1.96 times the standard error of the trajectory mean, pooling three independent seed batches. The coordinates within a trajectory are coupled. Comparisons between coefficients on the same input share random draws.

n	L	Mean error	Mean SE	Network SD	Network RMSE
32	64	0.46	0.17	6.70	6.72
32	512	0.22	0.17	6.58	6.58
128	64	0.18	0.09	3.45	3.45
128	512	0.12	0.09	3.39	3.39

Table 5: Calibration error and network variation, in percentage points. Mean error is the absolute difference between the estimated mean fraction and its target; SE is the Monte Carlo standard error of that mean. Each entry averages the corresponding statistic over 144 target settings.

Activation	$L = 16$	$L = 128$	$L = 1024$
ReLU	22.05	24.44	24.74
Leaky ReLU	22.13	24.46	24.73
GELU	21.44	24.32	24.71
SELU	23.83	24.74	24.91

Table 6: Mean inactive percentage at target 25%, with state width 128 and branch width one; 3072 independent trajectories per entry.

Depth curves and the diffusion limit. Proposition B.1 says that at fixed width the mean fraction $\mathbb{E}\,\widehat{p}_{n,L}$ converges with depth to $\pi_n(a)$ of the diffusion (21). We run the network at $n = m = 32$ for depths 16 to 1024 and integrate (21) by the Euler–Maruyama scheme with 2048 and 8192 steps; apart from the step count, the scheme differs from the network recursion only in replacing H_k by its conditional mean $v_\phi(Q_k)$. The two meshes agree to within 0.0014 in mean fraction, and the $L = 1024$ network lies within 0.0017 of the finer mesh; both gaps contain Monte Carlo and time-discretization error (Figure 4).

Calibration. The calibration grid asks whether $\sigma = a_{p,n}/\sqrt{L}$ reaches its target at finite size, using input RMS values 0.1, 1 and 5. Averaged over activations, shapes, scales and targets, the mean absolute error between the mean achieved fraction and its target is 0.46 percentage points at $n = 32$, $L = 64$; 0.22 at $n = 32$, $L = 512$; 0.18 at $n = 128$, $L = 64$; and 0.12 at $n = 128$, $L = 512$. It falls with both width and depth, consistent with the joint limit of Theorem 3.3 (Figures 5 and 6).

Table 5 distinguishes error in the mean fraction from variation between individual networks. At each target we use 1536 independent trajectories to estimate the standard deviation of $\widehat{p}_{n,L}$ and its root-mean-square error relative to p. The table averages each statistic over the 144 target settings for that width and depth. Quadrupling the state width approximately halves the standard deviation, whereas increasing depth mainly improves the mean. This is the distinction between calibrating an expectation at fixed width and concentration of the fraction in the joint limit.

Branch width. Theorem 3.2 holds for any branch width, which enters (7) only through $1/(mL)$. Figure 7 shows the achieved fraction at target 0.25 for m from 1 to 64 and depths 16, 128 and 1024. Table 6 gives the narrowest-branch values explicitly. At $m = 1$, the error falls from 1.17–3.56 percentage points at $L = 16$ to 0.09–0.29 points at $L = 1024$.

Input scale. Theorem 3.4 makes the calibrated coefficient decrease with the input scale for GELU and increase for SELU. Does matching the branch variance at the input capture this, or is the integral (10) needed? At target 0.25 and nine input scales from RMS 0.05 to 16 we compare, with paired draws, the activation's own coefficient calibrated at RMS 1 and reused unchanged, the ReLU coefficient multiplied by $\sqrt{(1/2)/k_\phi(Q_0)}$, which matches the branch variance at the input energy only, and the full integral (9). Table 7 gives the mean absolute error over the nine scales. For the homogeneous activations the three coefficients coincide, so their identical errors are a check. For GELU and SELU a reused coefficient misses by several points, the local correction removes most of the gap, and the integral brings the error to the homogeneous level (Figure 2).

Input support. Theorem 3.3 needs a positive input; the forward Theorem 3.2 does not. At fixed $a \in \{0, 0.01, 0.1, 0.3, 0.6, 1, 1.5, 2, 3\}$ we compare the achieved fraction with the mixture prediction

Branch activation	Reused coefficient	Local correction	Variance integral
ReLU	0.16	0.16	0.16
Leaky ReLU	0.16	0.16	0.16
GELU	3.19	0.30	0.16
SELU	5.76	0.70	0.13

Table 7: Input-scale comparison: mean absolute error of the terminal fraction in percentage points over nine input scales ($n = 128$, $m = 32$, $L = 512$, target 0.25).

Input	ReLU	Leaky ReLU	GELU	SELU
Positive	0.10 / 0.25	0.10 / 0.26	0.11 / 0.28	0.08 / 0.19
25% zeros	0.09 / 0.21	0.09 / 0.20	0.10 / 0.24	0.07 / 0.13
50% zeros	0.07 / 0.12	0.07 / 0.13	0.08 / 0.14	0.07 / 0.10
75% zeros	0.08 / 0.11	0.08 / 0.11	0.10 / 0.12	0.09 / 0.13
Symmetric signs	0.02 / 0.06	0.02 / 0.05	0.02 / 0.08	0.02 / 0.06
Mixed signs	0.03 / 0.09	0.03 / 0.09	0.04 / 0.13	0.02 / 0.04

Table 8: Input-support check: mean / maximum error in percentage points over the eight positive coefficients. Errors compare the estimated mean fraction with the limiting prediction; Figure 8 also shows the exact zero-coefficient endpoint.

$\int \Phi(-u/\sqrt{w})\, d\nu_n(u)$ for RMS-1 inputs of three kinds: a fraction $\rho \in \{0, 1/4, 1/2, 3/4\}$ of exact zeros with equal positive remaining coordinates, equal numbers of -1 and $+1$, and one third -1 with two thirds $+2$ at $n = 126$. In the limiting formula, the zero coordinates contribute ρ at $a = 0$ and $\rho/2$ for every $a > 0$; the symmetric input gives a constant one half, and the mixed-sign law can be nonmonotone (Figure 8). Table 8 summarizes the discrepancies over the eight positive coefficients.

Alignment. Corollary C.3 gives the cosine between u and x_L as $\sqrt{Q_0/(Q_0 + w_{p,n})}$, whose limits by Example C.4 are $|\Phi^{-1}(p)|/\sqrt{1 + \Phi^{-1}(p)^2}$ for an equal input and $\cos(\pi p)$ for the half-normal law. For both inputs at RMS 1, $n \in \{128, 512\}$ and the three targets, we store the cosine and the fraction of every trajectory and compare their means with the finite-n formula and the two limiting curves. The mean absolute error of the cosine is 0.00358 at $n = 128$ and 0.000756 at $n = 512$ (Figure 9).

F.1 Image-derived inputs with exact zeros

To check the nonnegative-input extension in Appendix C.5, we replace the positive input by a ReLU applied to a Xavier-normal affine embedding of an image, with zero bias. The resulting coordinates have an average zero fraction close to one half. Consequently the limiting range under positive noise begins near 0.25, and we choose targets 0.30, 0.40 and 0.45.

For each of the four datasets in Appendix I, we use a nearly class-balanced subset of 512 training images, state width $n = 256$, branch width $m = 64$ and depth $L = 256$. Each image has its own input distribution and energy curve. We invert the mean of their predicted fractions, then propagate the images through a common set of explicit Gaussian matrices drawn independently of the embedding. Table 9 reports means and sample standard deviations over three network initializations. Averaged over the 48 dataset–activation–target combinations, the absolute error of the three-seed mean is 0.55 percentage points, with a maximum of 2.04. This checks calibration with a conventional ReLU input, including its exact zeros and the dependence between images caused by shared weights.

G Joint-input numerical validation

Finite-network simulation. We simulate the Gaussian-matrix residual model with width $n = 1024$ and branch width $m = 32$, followed by one terminal ReLU. For the $n \times N$ matrix X whose columns contain the input features, each layer draws $Z \in \mathbb{R}^{m \times N}$ with independent rows of covariance $C = X^\top X/n$, forms $H = \phi(Z)^\top \phi(Z)/m$, and updates $X \leftarrow X + (a/\sqrt{L})EH^{1/2}$, where E has

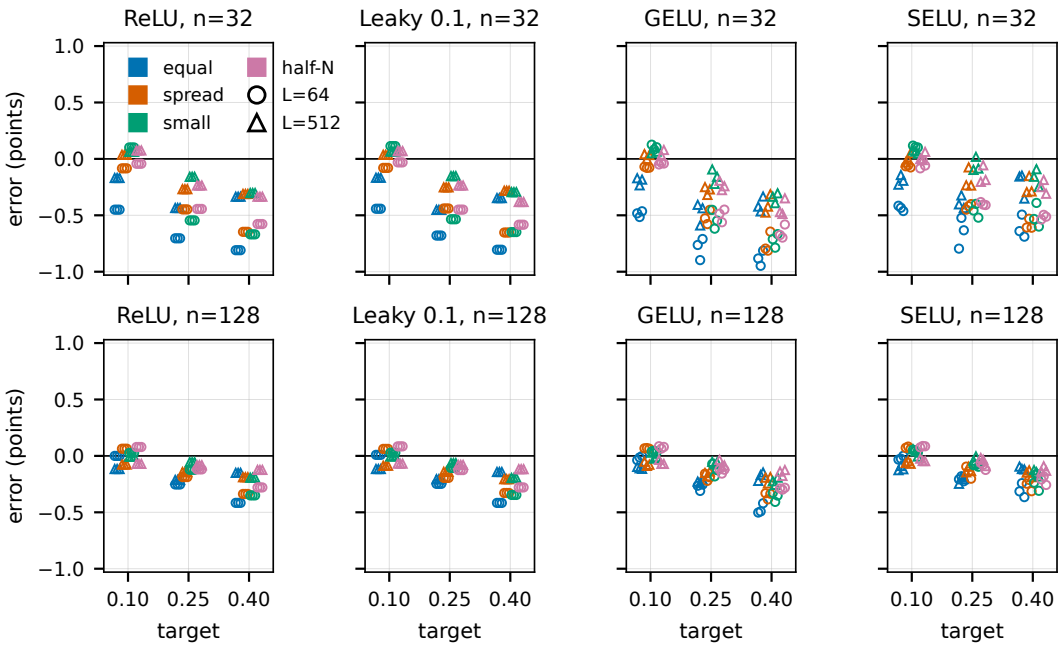


Figure 6: Calibration error for each activation, width and target. Colors denote input shapes, with circles for $L = 64$ and triangles for $L = 512$. Horizontal offsets separate shapes and scales within a target.

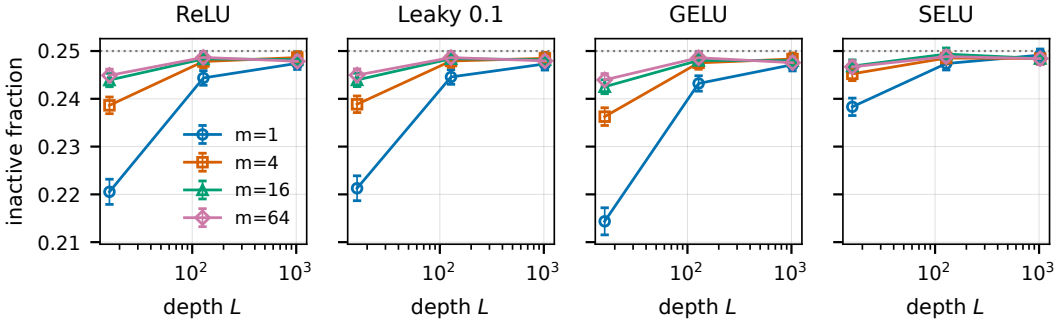


Figure 7: Achieved fraction at target 0.25 for branch widths 1 to 64.

independent standard Gaussian entries. This samples the exact conditional shared-input law of the Gaussian A, B residual branch, including finite branch width (Appendix E.1). The initial coordinate rows are independent centered Gaussian vectors with covariance $s^2[(1 - c_0)I + c_0\mathbf{1}\mathbf{1}^\top]$.

Predictions and calibration. We solve $q' = a^2 v_\phi(q)$ and $c' = a^2[K_\phi(q, c) - c v_\phi(q)]/q$, using $\log q$ as the integration variable. Here $K_\phi(q, c) = \mathbb{E}[\phi(Z_1)\phi(Z_2)]$ for centered jointly Gaussian Z_1, Z_2 with variance q and correlation c. ReLU and leaky-ReLU use closed kernels; GELU and SELU use Gaussian quadrature of order 128, with an order-64 sensitivity check. Predicted zero-column fractions are obtained from the one-dimensional equicorrelated Gaussian orthant integral $J_N(c)$ in (58). Across the matched-energy grid, increasing quadrature order from 64 to 128 changes the predicted fraction by at most 0.000858 for GELU and 0.000420 for SELU.

The matched-energy study uses $N \in \{8, 32\}$, $c_0 \in \{0, 1/2\}$, $s \in \{0.1, 1, 10\}$, four branch activations (ReLU, leaky-ReLU with slope 0.1, GELU, SELU), and $L = 256$. For each activation, a is chosen for the limiting terminal/input energy ratio $q_1/q_0 \in \{1.105573, 2, 10.472136\}$.

The inverse-transfer study uses $N \in \{8, 32\}$, $c_0 = 0$, $s = 1$, $d \in \{0.025, 0.05, 0.1\}$, $L \in \{128, 512\}$, and leaky slopes $\beta \in \{0, 0.1, 0.5\}$. For each (N, d) we first invert the ReLU prediction. Correlation

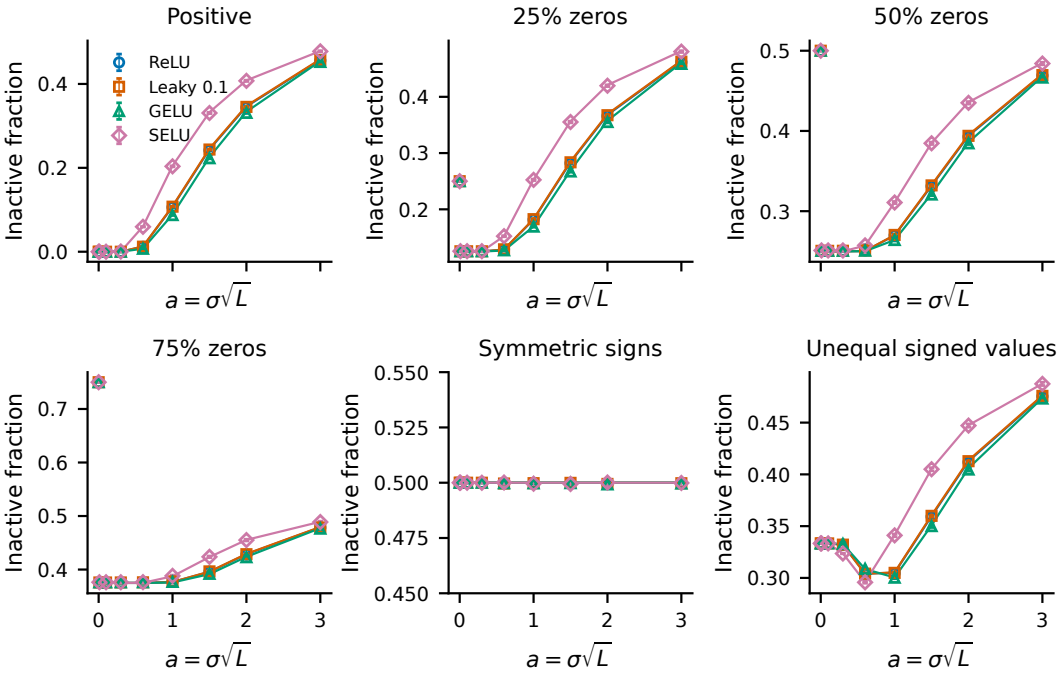


Figure 8: Inputs with zeros or signed coordinates against the mixture prediction.

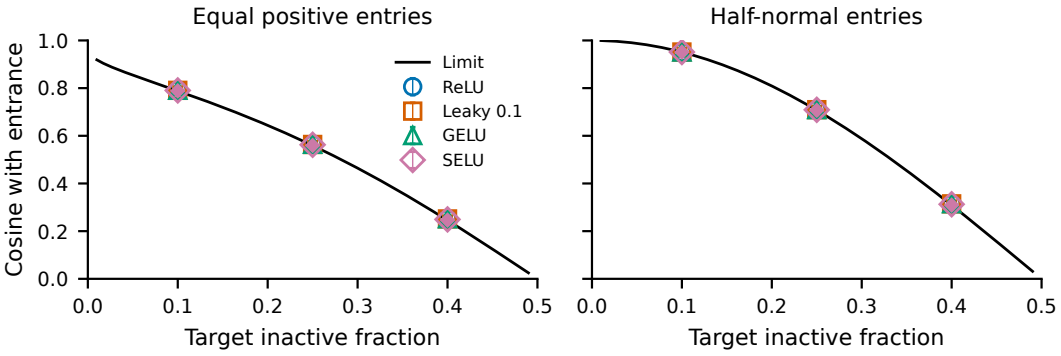


Figure 9: Input–terminal cosine against Corollary C.3.

correction uses $a_{\mathrm{ReLU}}/(1-\beta)$; the separate variance-matching control is $a_{\mathrm{ReLU}}/\sqrt{1+\beta^2}$. At $\beta = 0$ both controls use the same network rather than duplicate observations.

Observables and statistical reporting. The 144 matched-energy and 60 distinct inverse-transfer settings each use ten independent network seeds. Figures show means and standard errors, computed as sample SD divided by $\sqrt{10}$. Different settings share random streams within each seed; uncertainty is computed across seeds. The inactive-entry fraction and zero-column fraction are measured before the terminal gate; the latter equals the proportion of all-zero columns in the resulting $N \times n$ ReLU feature matrix $\mathcal{F}$. For the width-normalized Gram matrix $G = \mathcal{F}\mathcal{F}^{\mathsf{T}}/n$, feature correlation means the off-diagonal cosine entries $G_{ij}/\sqrt{G_{ii}G_{jj}}$. Figure 12 subtracts the ReLU target's predicted cosine entry from their empirical average. The cosine removes magnitude. Energy ratios q/q_{ODE}, with $q = \|X_L\|_{\mathrm{F}}^2/(nN)$ and $q_{\mathrm{ODE}} = \bar{q}(1)$, show substantial finite-depth bias at large coefficients. Across the matched-energy and transfer grids, absolute deviations of individual networks from their respective limiting predictions average 0.00641 and 0.00756.

For the transfer comparison in Section 6.2, we estimate uncertainty by resampling the ten paired seed blocks $20{,}000$ times, preserving every setting and both methods within a block. The joint

Dataset	Branch	At $a = 0$	Target 0.30	Target 0.40	Target 0.45
MNIST	ReLU	50.5 ± 0.5	30.1 ± 0.9	40.4 ± 2.9	45.3 ± 3.7
MNIST	Leaky 0.1	50.5 ± 0.5	30.0 ± 1.0	40.4 ± 3.0	45.2 ± 3.8
MNIST	GELU	50.5 ± 0.5	30.1 ± 1.1	40.4 ± 2.8	45.3 ± 3.6
MNIST	SELU	50.5 ± 0.5	30.3 ± 1.3	40.2 ± 1.5	45.3 ± 3.2
FashionMNIST	ReLU	50.5 ± 0.4	29.5 ± 1.1	39.9 ± 3.2	44.9 ± 3.7
FashionMNIST	Leaky 0.1	50.5 ± 0.4	29.5 ± 1.2	39.7 ± 3.4	44.7 ± 3.7
FashionMNIST	GELU	50.5 ± 0.4	29.6 ± 1.1	39.7 ± 3.0	44.9 ± 3.7
FashionMNIST	SELU	50.5 ± 0.4	29.9 ± 1.1	39.6 ± 1.9	44.6 ± 3.8
KMNIST	ReLU	50.8 ± 0.4	29.6 ± 0.8	40.1 ± 2.7	44.9 ± 3.5
KMNIST	Leaky 0.1	50.8 ± 0.4	29.6 ± 0.8	39.9 ± 2.8	44.8 ± 3.5
KMNIST	GELU	50.8 ± 0.4	29.6 ± 0.7	40.0 ± 2.6	45.0 ± 3.3
KMNIST	SELU	50.8 ± 0.4	29.6 ± 0.6	39.8 ± 1.5	45.1 ± 3.0
CIFAR10	ReLU	49.9 ± 0.3	30.8 ± 1.1	41.8 ± 2.1	47.0 ± 2.7
CIFAR10	Leaky 0.1	49.9 ± 0.3	30.8 ± 1.1	41.8 ± 2.2	47.0 ± 2.7
CIFAR10	GELU	49.9 ± 0.3	30.8 ± 1.1	41.8 ± 1.9	47.0 ± 2.4
CIFAR10	SELU	49.9 ± 0.3	30.6 ± 0.3	41.1 ± 1.9	46.4 ± 2.9

Table 9: Inactive coordinates (%) with a ReLU input. Means and standard deviations over three shared-weight networks.

correction reduces the average absolute target error of the network means by 1.350 percentage points for $\beta = 0.1$ and 4.335 points for $\beta = 0.5$. The corresponding exploratory 95% percentile intervals are $[1.133, 1.393]$ and $[3.888, 4.524]$ points. These intervals describe variation across initialization seeds for the tested settings.

Energy at finite depth. For a homogeneous branch, put $k_\beta = (1 + \beta^2)/2$. Taking conditional expectations in (5) gives

$$\mathbb{E}[Q_L \mid X_0] = Q_0(1 + k_\beta a^2/L)^L, \qquad q_{\mathrm{ODE}} = q_0 e^{k_\beta a^2}. \tag{66}$$

For the Gaussian input law, $q_0 = \mathbb{E}Q_0 = s^2$, so averaging over inputs gives $\mathbb{E}Q_L = q_0(1+k_\beta a^2/L)^L$. The energy deficit relative to the ODE can therefore be substantial when a is large, even if the zero-column fraction is close to its target. Table 10 separates these measurements. Writing $b = k_\beta a^2$, the inequalities $y - y^2/2 \le \log(1 + y) \le y$ for $y \ge 0$ give

$$\exp\left(-\frac{b^2}{2L}\right) \le \frac{\mathbb{E}Q_L}{q_{\mathrm{ODE}}} \le 1. \tag{67}$$

Thus L large relative to b^2 suffices for small relative error in the mean energy. Under correlation correction with fixed a_0, b^2 grows as $(1 - \beta)^{-4}$ as $\beta \uparrow 1$.

The most demanding case is $N = 32$, $d = 0.1$, $\beta = 0.5$, where joint calibration requires $a = 6.5810$. At depths 128 and 512, the observed energy ratios are 0.063 and 0.487, respectively. We repeat this case at depths 2048 and 8192 with ten fresh initializations per setting, including ReLU and variance-matching controls (Table 11). The joint correction gives energy ratios 0.894 and 1.030, and the mean off-diagonal ReLU Gram entry divided by the theoretical energy changes from 0.340 to 0.387, against its limit 0.376. The large finite-depth energy deficit diminishes, while sampling variation remains visible. The variance-matching control converges to a different zero-column probability and normalized Gram entry. Here $b = 27.0688$, and (67) guarantees less than 5% relative error in the mean energy for $L \ge 7143$.

The full normalized feature Gram. Table 12 compares G/h with the same limiting ReLU matrix, retaining amplitude as well as correlation. At $\beta = 0.5$ and $L = 128$, correlation correction has larger error than variance matching, despite its smaller inactivity error. The gap narrows at $L = 512$, consistently with the energy comparison above. Using the empirical ReLU mean as reference gives the same unfavorable aggregate comparison: over the twelve (N, L, d) settings at $\beta = 0.5$, the mean relative Frobenius discrepancy is 0.329 for correlation correction and 0.265 for variance matching. For each setting this discrepancy is $\|\overline{G/h} - \overline{G_0/h_0}\|_F / \|\overline{G_0/h_0}\|_F$, with bars denoting ten-network means.

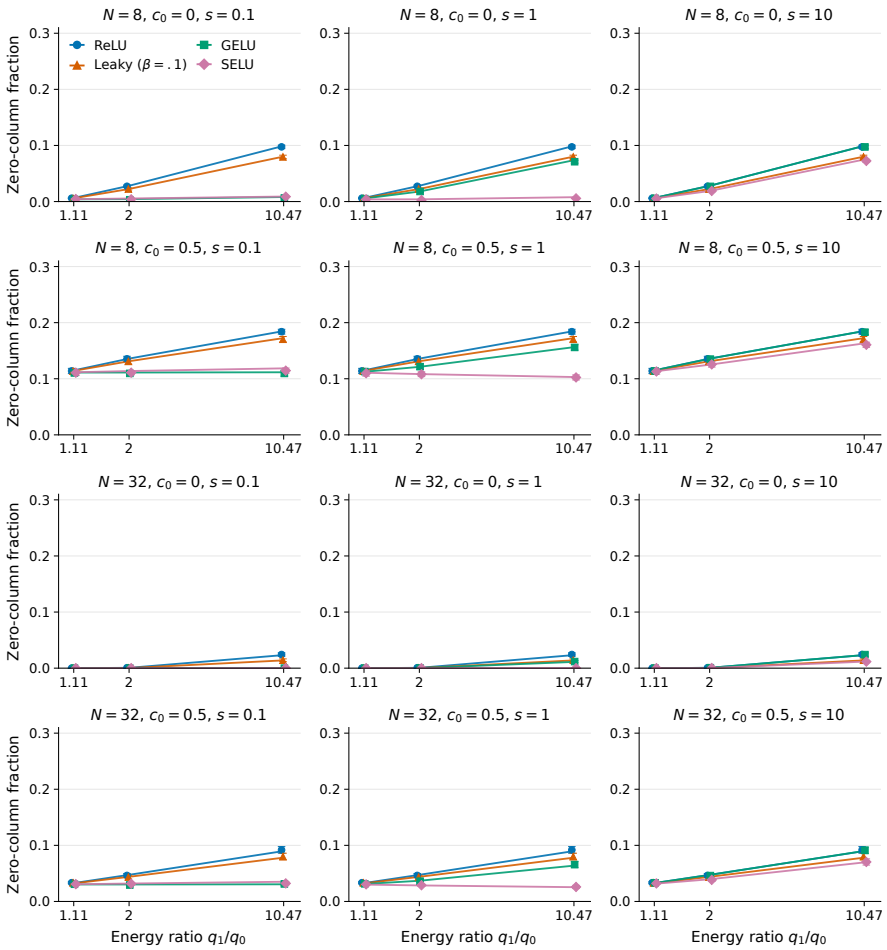


Figure 10: Matched-energy zero-column results. Panels cover every input count, initial correlation, and input scale; colors denote the four branch activations. Lines are covariance/orthant predictions and points are finite-network means with standard errors over ten runs. The horizontal axis is the limiting terminal/input energy ratio.

Positive empirical input law. A separate check uses $N = 3$ and the four positive coordinate atoms $(0.1, 0.7, 1.4)$, $(0.9, 0.2, 0.3)$, $(1.8, 1.1, 0.4)$, and $(0.4, 1.9, 0.6)$ with exact frequencies $(0.1, 0.2, 0.3, 0.4)$ at $n = 1000$. For each slope $\beta \in \{0, 0.1, 0.5\}$ and target $d \in \{0.05, 0.15, 0.3\}$, we independently invert the weighted Gaussian mixture orthant probability with added covariance $\bar{C}_a(1) - C_0$. The Gaussian CDF uses absolute and relative tolerances 10^{-8}. An independent evaluation by nested one-dimensional quadrature gives target residuals below 6.88×10^{-9} for all nine coefficients. The grid crosses $m \in \{1, 32\}$ and $L \in \{64, 512\}$, with ten networks per condition (360 networks). The largest absolute target error of a network mean is 0.0188, attained at $\beta = 0.5$, $d = 0.3$, $m = 1$, $L = 64$: the mean is 0.3188 and its sample standard deviation is 0.0600 (standard error 0.0190). This largest observed deviation is therefore comparable to the uncertainty of its mean.

G.1 Finite-depth correction and positive-law comparison

Finite-depth calibration. We test the correction in Appendix E.8 at $n = 1024$, $m = 32$, $N \in \{8, 32\}$, $L \in \{128, 512\}$, $d \in \{0.025, 0.05, 0.1\}$, and $\beta \in \{0.1, 0.5\}$. The signed Gaussian input law has $C_0 = I_N$. The original method uses the continuous-depth inverse and $a_\beta = a_0/(1 - \beta)$; the corrected method first inverts the finite ReLU recursion and then applies (62). All requested targets satisfy its strict feasibility condition. Including the corresponding two ReLU references gives 72

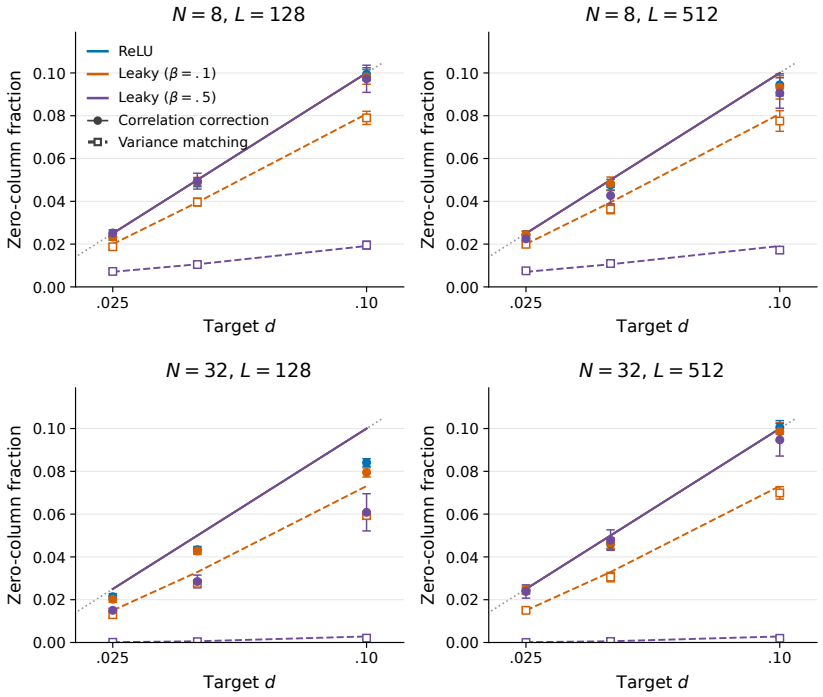


Figure 11: Gaussian inverse-transfer results. Both input counts and depths are shown, with ReLU as a reference. Solid lines and filled points use correlation correction; dashed lines and open points use variance matching. Lines are limiting predictions and points are means with standard errors over ten finite networks. The dotted diagonal is $D_N = d$.

L	β	Calibration	Target error	q/q_{ODE}	Predicted ratio
128	0	ReLU reference	0.0049	0.923	0.967
128	0.1	Correlation	0.0061	0.901	0.949
128	0.1	Variance	0.0188	0.927	0.967
128	0.5	Correlation	0.0124	0.533	0.574
128	0.5	Variance	0.0517	0.950	0.967
512	0	ReLU reference	0.0023	0.991	0.991
512	0.1	Correlation	0.0024	0.985	0.987
512	0.1	Variance	0.0168	0.992	0.991
512	0.5	Correlation	0.0046	0.811	0.826
512	0.5	Variance	0.0520	0.993	0.991

Table 10: Finite-depth energy and target error for the full inverse-transfer grid. Each row averages the six (N, d) settings. The error is the absolute difference between the ten-network mean zero-column fraction and its target, averaged across conditions. The energy column averages the observed ratio q/q_{ODE}, and the final column is the exact finite-depth expectation $(1 + k_\beta a^2/L)^L / \exp(k_\beta a^2)$ averaged over the same conditions, with $k_\beta = (1 + \beta^2)/2$. ReLU controls coincide and appear once.

conditions, each with ten fresh seeds (720 completed networks). The same seed couples Gaussian draws across methods.

Table 13 compares both methods with the same continuum ReLU target matrix. The full normalized Gram error decreases most at $\beta = 0.5$: corrected minus original is -0.06794 ± 0.00481 at $L = 128$ and -0.02589 ± 0.00455 at $L = 512$ (paired mean $\pm$ standard error). The corresponding changes in per-network silence-target error are much smaller and noisier: -0.00235 ± 0.00169 and -0.00029 ± 0.00074. Across all 24 paired leaky-branch conditions, mean silence-target error decreases in 16, is unchanged in 3, and increases in 5; the full Gram error decreases in 23 and

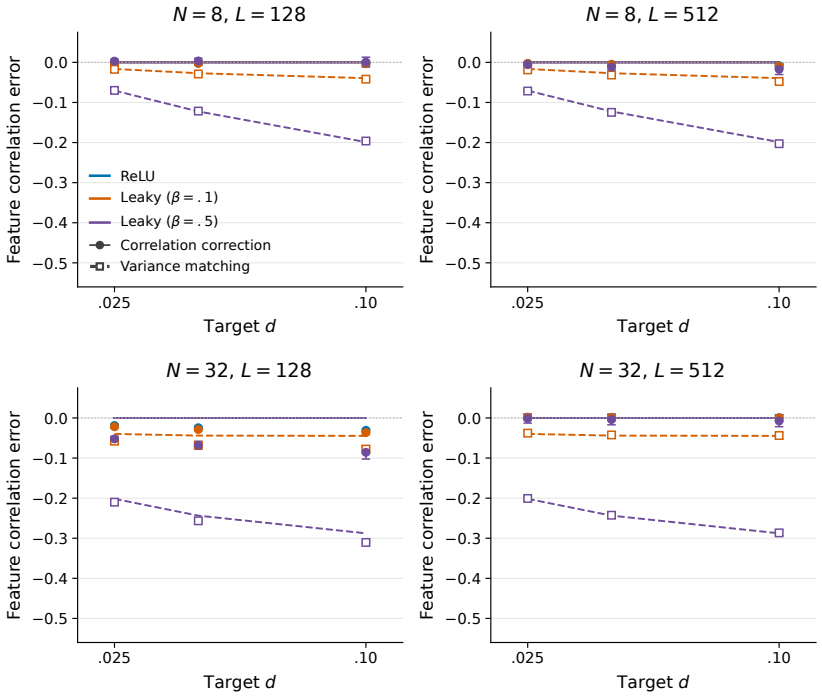


Figure 12: Feature-cosine error relative to the ReLU target. Solid lines and filled points use correlation correction; dashed lines and open points use variance matching. Lines are limiting predictions; points show means and standard errors across ten initializations.

L	Calibration	$\widehat{d}_N$	Q_L/q_{ODE}	Predicted ratio	G_{off}/h
2048	ReLU	0.1002 ± 0.0027	1.0364 ± 0.0272	0.9929	0.3954 ± 0.0137
2048	Correlation	0.1028 ± 0.0063	0.8939 ± 0.0545	0.8375	0.3404 ± 0.0278
2048	Variance	0.0026 ± 0.0005	1.0142 ± 0.0108	0.9929	0.2374 ± 0.0046
8192	ReLU	0.0996 ± 0.0031	1.0296 ± 0.0270	0.9982	0.3889 ± 0.0147
8192	Correlation	0.1078 ± 0.0055	1.0297 ± 0.0410	0.9564	0.3870 ± 0.0206
8192	Variance	0.0038 ± 0.0006	1.0078 ± 0.0115	0.9982	0.2357 ± 0.0054

Table 11: Deeper networks at $N = 32$, $d = 0.1$, $n = 1024$, $m = 32$. Means $\pm$ standard errors over ten networks; leaky branches use $\beta = 0.5$. The predicted energy ratio is given by (66). G_{off} averages off-diagonal terminal ReLU Gram entries and $h = q_{\mathrm{ODE}}$ because population input energy is $q_0 = 1$. Its normalized limits are 0.37641 for ReLU and correlation correction, and 0.23257 for variance matching. The latter has limiting zero-column fraction 0.002826.

increases in 1. Both the coefficient and the deterministic normalization change, so this comparison does not isolate their contributions. A control using the same normalization rule for both methods does not show a Gram-error benefit from the coefficient change (Table 14). With each network divided by its own finite-depth h, the corrected coefficient has a slightly larger mean error in all four groups; using the continuous-depth h for both methods gives the same direction. The Gram improvement is therefore attributable to finite-energy normalization. The coefficient correction gives small, condition-dependent changes in the silent-fraction error.

Positive laws with identical pair marginals. We also test Proposition E.4 with $\delta = 0.99$, $N = 3$, $n = 1024$, $m = 32$, ReLU branches, $a \in \{0.2, 0.5, 1\}$, and $L \in \{128, 512\}$. Each of the four atoms of a parity law occurs exactly 256 times in the input array. Both laws use ten fresh paired seeds per setting, for 120 networks. Figure 14 compares the three coefficients at both depths. At $L = 512$, the paired differences $\widehat{d}_{N,-} - \widehat{d}_{N,+}$ are 0.06709 ± 0.00169, 0.07041 ± 0.00171, and 0.05186 ± 0.00224

β	L	Correlation	Variance	Paired difference
0.1	128	0.0569 ± 0.0033	0.0567 ± 0.0036	0.0003 ± 0.0009
0.1	512	0.0381 ± 0.0022	0.0382 ± 0.0020	-0.0001 ± 0.0019
0.5	128	0.1697 ± 0.0035	0.1055 ± 0.0031	0.0642 ± 0.0022
0.5	512	0.0903 ± 0.0041	0.0934 ± 0.0023	-0.0031 ± 0.0038

Table 12: Mean entrywise absolute error $N^{-2} \sum_{i,j} |(G/h)_{ij} - T_{ij}|$ against the limiting ReLU target T, whose diagonal is $1/2$ and off-diagonal entries are $\kappa(c_{\mathrm{ReLU}})/2$. Each seed averages the six (N, d) settings; entries show mean $\pm$ standard error across ten paired seeds. The last column is correlation minus variance.

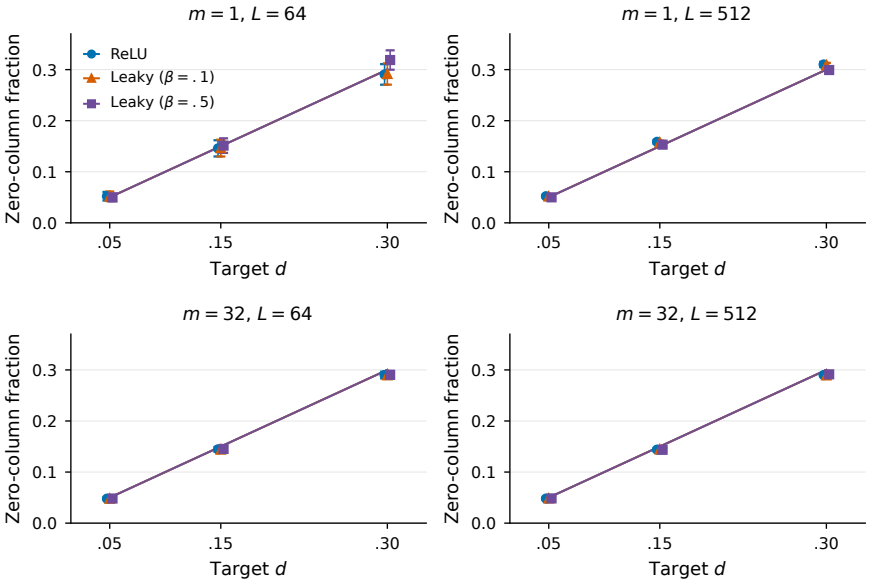


Figure 13: Calibration on positive inputs. The panels cover three leaky slopes, three targets, two branch widths, and two depths, with $N = 3$ and $n = 1000$. Coefficients are inverted separately for each slope and target from the weighted mixture-Gaussian orthant prediction. Lines are predictions; points are means with standard errors over ten finite networks. Small horizontal offsets separate coincident targets.

for the three coefficients, respectively. The limiting differences are 0.06604, 0.06908, and 0.05158. The finite differences in marginal inactivity and pairwise Gram entries are sampling fluctuations: their distributions agree between the two balanced input laws even at finite depth, as shown in Appendix E.10. This equality coexists with the predicted separation in joint silence.

G.2 Readout with a restricted feature budget

An exploratory subsampling analysis of those 120 networks suggested the finite-feature comparison in Proposition 4.3. We then fixed a confirmation protocol and generated 120 new networks, using the same twelve parity-law, coefficient and depth settings and ten new network seeds. For each network, 5000 uniformly sampled ordered eight-tuples of distinct coordinates give nested feature budgets $M \in \{3, 4, 8\}$. The sampling bank is independent for each network seed and shared across the two laws and conditions within that seed. There is no intercept. We average $1 - \mathrm{rank}(F^{(M)})/3$ over the subsets, which integrates the relative least-squares residual over independent isotropic sphere labels. No labels or readout weights need to be sampled or trained. Numerical rank uses singular values above 10^{-10} times the largest; replacing this by 10^{-8} changes none of the 360 network–budget results.

β	L	Original silence error	Finite silence error	Original Gram error	Finite Gram error
0.1	128	0.0107 ± 0.0013	0.0098 ± 0.0011	0.0507 ± 0.0036	0.0439 ± 0.0029
0.1	512	0.0077 ± 0.0011	0.0076 ± 0.0011	0.0414 ± 0.0018	0.0399 ± 0.0017
0.5	128	0.0193 ± 0.0017	0.0170 ± 0.0016	0.1688 ± 0.0032	0.1008 ± 0.0044
0.5	512	0.0149 ± 0.0021	0.0146 ± 0.0024	0.0973 ± 0.0051	0.0714 ± 0.0029

Table 13: Fresh finite-depth correction, mean $\pm$ standard error over ten paired seeds; each seed averages the six (N, d) settings. Silence error averages $|\widehat{d}_N - d|$ computed separately for each network. Gram error averages all entries of $|G/h - T|$, with the same target T for both methods: diagonal $1/2$, off-diagonal $\kappa(c_d)/2$. The original uses $h = \exp(k_\beta a^2)$; the correction uses $h = (1 + k_\beta a^2/L)^L$, where $k_\beta = (1 + \beta^2)/2$. Thus the comparison changes both coefficient and normalizer.

β	L	Both finite-depth h	Both continuous-depth h
0.1	128	0.000412 ± 0.000256	0.000764 ± 0.000076
0.1	512	0.000008 ± 0.000071	0.000008 ± 0.000051
0.5	128	0.006566 ± 0.001186	0.011461 ± 0.000593
0.5	512	0.000925 ± 0.000310	0.002339 ± 0.000073

Table 14: Corrected minus original full Gram error when both methods use the same normalization rule, evaluated at their respective coefficients. The two rules are $(1 + k_\beta a^2/L)^L$ and $\exp(k_\beta a^2)$; the target matrix and paired aggregation are unchanged from Table 13. Values are means $\pm$ standard errors over ten paired seeds. Positive means favor the original coefficient; the smallest differences are within their sampling uncertainty.

The primary setting, chosen before simulation, is $a = 0.2$, $L = 512$, $M = 3$. Mean residuals are 0.025747 and 0.070653 for parity plus and minus, giving a paired difference of 0.044907 ± 0.001391 (mean $\pm$ standard error), compared with the limiting prediction 0.045390. Table 16 includes every coefficient, depth and budget. Differences at $M = 4$ are smaller; at $M = 8$ they are near zero and some estimates have the opposite sign. The theoretical differences at $M = 8$ are below 1.6×10^{-6}. The fitting difference is concentrated at small feature budgets. Table 15 pairs the change in zero columns with this fitting difference at the primary setting.

H FINITE-WIDTH FEATURE CAPACITY

Design. We test whether silence on the input set reveals a finite-width capacity constraint that marginal inactivity does not describe. We use the explicit shared Gaussian A_k, B_k matrices of (1), with $n = N \in \{64, 128\}$, $m = 32$, $L = 64$, and $a \in \{0, 1, 2, 3, 4\}$. The branches are ReLU, leaky ReLU with slope 0.1, exact GELU, SELU, or linear; every network has the same terminal ReLU. Inputs are signed Gaussian vectors $u^{(s)} = \sqrt{c_0}\, g + \sqrt{1 - c_0}\, g_s$, where $g, g_1, \ldots, g_N$ are independent $N(0, I_n)$ and $c_0 \in \{0, 0.5, 0.9\}$. Each of the 150 settings uses ten independent initializations. Within each repetition and width, coefficients and activations share the same matrices, Gaussian input draws and labels. Computation uses float64 on an NVIDIA A10, with CPU singular-value calculations. These finite networks have $N = n$; the joint-limit theorem instead keeps N fixed.

Readout and diagnostics. Let $F_{si} = \mathrm{ReLU}(x_{L,i}^{(s)})$, $P = I_N - \mathbf{1}\mathbf{1}^\mathsf{T}/N$, and $F_c = PF$. The primary labels are an independent, uniformly permuted vector with $N/2$ entries each of $+1$ and -1. A secondary target is the centered linear score $PU_{\mathrm{in}}v$, rescaled to squared norm N, where the fixed teacher has independent coordinates $v_i \in \{-1/\sqrt{n}, 1/\sqrt{n}\}$ and $U_{\mathrm{in}} \in \mathbb{R}^{N \times n}$ has the inputs as rows. We report the relative least-squares residual

$$R(y; F) = \min_w \frac{\|F_c w - y\|^2}{\|y\|^2}.$$

Centering corresponds to eliminating a free intercept in least squares. Numerical rank and projection residuals use singular values exceeding 10^{-10} times the largest; cutoffs $10^{-12}, 10^{-8}, 10^{-6}$ are retained as sensitivity checks. We also record marginal inactivity, exact zero columns, centered kernel spectra, and frozen-head GD from zero, using loss $\|F_c w - y\|^2/(2N)$ and the per-condition step size

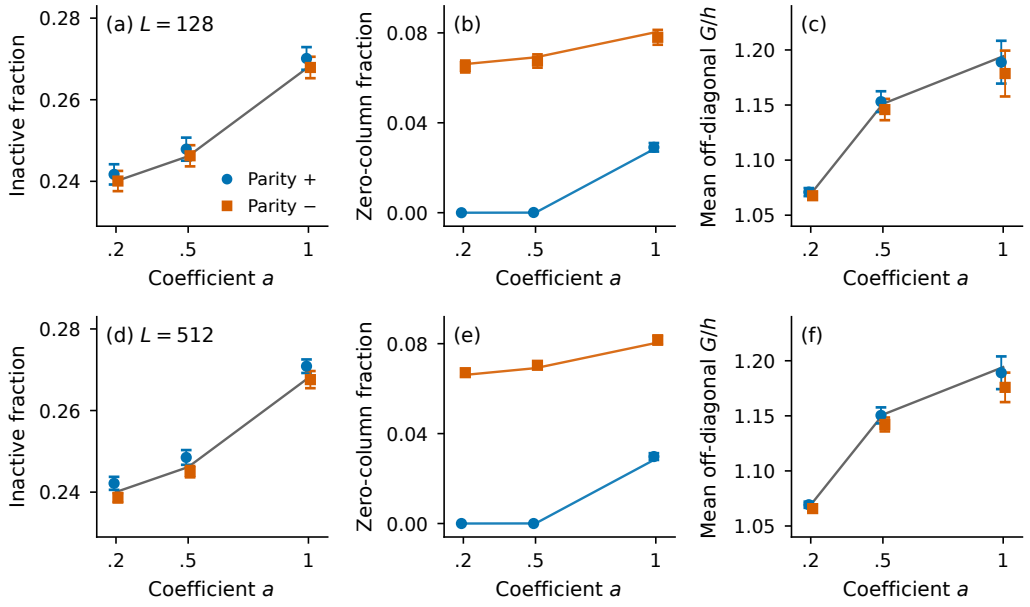


Figure 14: Positive parity input laws at both tested depths. Points are means with standard errors over ten networks; lines are limiting predictions. The laws have the same limiting marginal inactivity and pairwise feature Gram, but different silence probabilities on all three inputs. The Gram panel averages off-diagonal entries of G/h, with $h = \exp(a^2/2)$; total predicted energy is $(1+\delta^2)h$. Small horizontal offsets separate the two laws.

Minus–plus difference (percentage points)	Limiting prediction	Measured mean ± SE
Zero feature fraction	6.60	6.46 ± 0.15
Relative least-squares residual, $M = 3$	4.54	4.49 ± 0.14

Table 15: Primary comparison at $a = 0.2$, $L = 512$, from ten paired network initializations. The residual uses no intercept and 5000 coordinate subsets per network.

$N/s_{\max}(F_c)^2$ at $0, 1, 10, 100, 1000$ steps. This step normalization removes a common feature-scale effect.

Capacity constraint. If $Z = n\widehat{d}_N$ columns of F vanish, they also vanish in F_c. Consequently

$$r := \operatorname{rank}(F_c) \le \min\{N-1, n-Z\}. \tag{68}$$

For an independent uniform balanced label vector, $\mathbb{E}_y[yy^{\mathsf{T}}] = NP/(N-1)$ and $\|y\|^2 = N$. Projecting onto the column space of F_c gives

$$\mathbb{E}_y[R(y;F) \mid F] = 1 - \frac{r}{N-1} \ge 1 - \frac{\min\{N-1, n-Z\}}{N-1}. \tag{69}$$

The last bound averages over the random label vector. The rank inequality holds algebraically for every feature matrix.

Results. Table 17 compares $a = 0$ and $a = 4$. The zero-column fraction increased and centered rank decreased in all ten repetitions. The centered states before the terminal ReLU retained numerical rank 127 at both endpoints in every repetition; all these ranks and the primary terminal ranks were stable across the four cutoffs. The zero-column ceiling was attained in nine of the ten $a = 4$ terminal matrices. The lost columns constrain which labels the readout can fit. The linear control illustrates a separate limitation: at $a = 4, c_0 = 0$ it had no zero columns and centered terminal rank 127 in every repetition, yet its 1000-step relative GD loss was 0.372 ± 0.083, compared with 0.328 ± 0.028

a	L	$M = 3$	$M = 4$	$M = 8$
0.2	128	0.041633 ± 0.002024	0.005327 ± 0.000539	-0.000007 ± 0.000007
0.2	512	0.044907 ± 0.001391	0.006087 ± 0.000472	0.000000 ± 0.000010
0.5	128	0.042720 ± 0.002059	0.005753 ± 0.000614	-0.000007 ± 0.000007
0.5	512	0.045680 ± 0.001649	0.006467 ± 0.000569	0.000007 ± 0.000012
1.0	128	0.028787 ± 0.002303	0.006040 ± 0.000743	0.000000 ± 0.000010
1.0	512	0.031760 ± 0.001847	0.006547 ± 0.000596	-0.000007 ± 0.000012

Table 16: Difference $R_{-}(M) - R_{+}(M)$ in expected relative least-squares residual for the two parity laws, without an intercept. Each network contributes its average over 5000 subsets; entries are means and standard errors of paired differences across ten independent repetitions, each with its own network seed and sampling bank. Subsets are Monte Carlo draws within each network.

Table 17: Readout capacity: ReLU, $n = N = 128$, $c_0 = 0$. Entries are mean $\pm$ sample SD over ten independent repetitions; the last column summarizes within-seed differences. Residuals below 10^{-25} are displayed as $\simeq 0$.

Diagnostic	$a = 0$	$a = 4$	Paired difference
Inactive-entry fraction	0.501 ± 0.003	0.490 ± 0.043	-0.011 ± 0.042
Zero-column fraction	0	0.085 ± 0.041	$+0.085 \pm 0.041$
Zero-column count	0	10.9 ± 5.2	$+10.9 \pm 5.2$
Centered rank	127.0 ± 0.0	117.0 ± 5.1	-10.0 ± 5.1
Random-label residual	$\simeq 0$	0.080 ± 0.052	$+0.080 \pm 0.052$
Linear-teacher residual	$\simeq 0$	0.068 ± 0.039	$+0.068 \pm 0.039$

for ReLU. Zero columns therefore do not provide a sufficient diagnostic of finite-time learning; conditioning also matters. Figure 15 gives the broader descriptive grid.

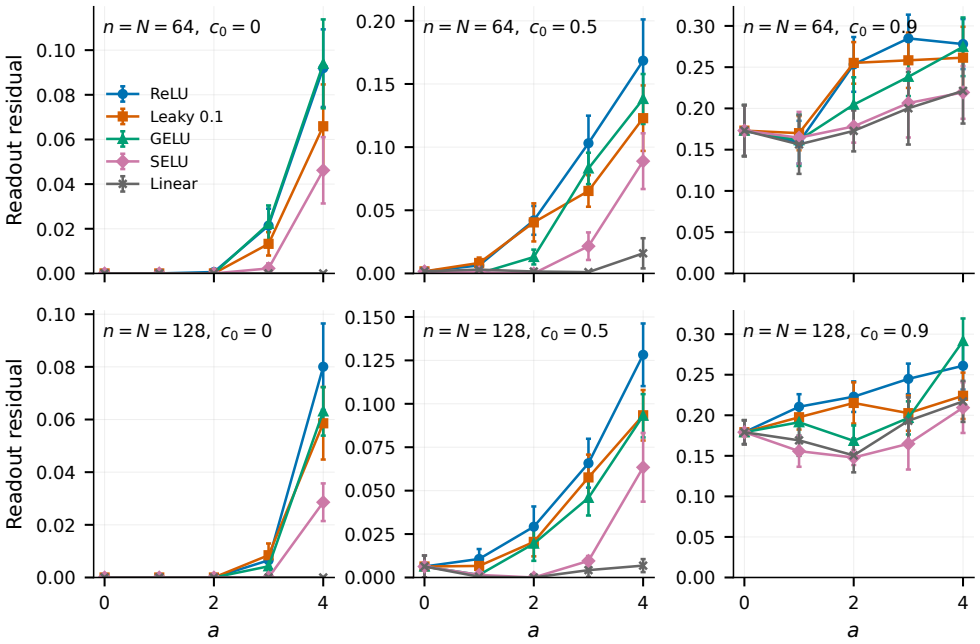


Figure 15: In-sample random-label least-squares residuals of centered, frozen terminal features across both widths and all three input correlations. Curves show mean $\pm$ standard error across ten independent repetitions for each condition, with shared draws within a repetition. All networks use $m = 32$, $L = 64$, and a common terminal ReLU.

I ADDITIONAL TRAINING OBSERVATIONS

These experiments compare short-run learning after prescribing the marginal initial fraction. The seven methods give no consistent accuracy ranking. Each network uses explicit matrices shared by all images.

Data. We use MNIST (Le Cun et al., 1998), Fashion-MNIST (Xiao et al., 2017), KMNIST (Clanuwat et al., 2018) and CIFAR-10 (Krizhevsky, 2009). From each official training split we draw 10,000 balanced training examples (1000 per class) and 2000 disjoint balanced validation examples (200 per class); the official 10,000-image test split is used unchanged. Images are flattened and standardized with one scalar mean and one scalar standard deviation computed over all training pixels, with no augmentation.

Network. A trainable affine embedding and a softplus map the standardized image ξ to the input $u = s\,\mathrm{softplus}(W_0\xi + b_0)$, so every $u_i > 0$; the input scale is $s = 1$ in the primary grid. The residual stack is (1) with $n = m = 128$, $L \in \{32, 128\}$ and one of the four branch activations, with explicit matrices A_k, B_k shared by all inputs and no bias or normalization in the hidden layers. The terminal ReLU features are divided by one scalar, the root-mean-square of the initial terminal features over a fixed calibration subset of 512 training examples, which is held fixed during training; a linear readout with bias gives the logits. Embedding and readout use Xavier normal initialization, A_k has $N(0, 1/n)$ entries and B_k has $N(0, \sigma^2/m)$ entries, so σ enters only through the initial B_k. All weights and the embedding and readout biases train.

Initialization methods. Each condition compares seven initial scales, all of the form $\sigma = a/\sqrt{L}$: $a = 0$, which starts every branch output at zero; $a = 1$; the full rule (9) at targets $p \in \{0.1, 0.25, 0.4\}$; the ReLU coefficient for $p = 0.25$ transferred unchanged; and that coefficient multiplied by $\sqrt{(1/2)/k_\phi(\bar{q}_0)}$, where $\bar{q}_0$ is the mean input energy over the calibration subset. For ReLU the last two coincide with the $p = 0.25$ rule and give identical runs under paired random draws. For leaky ReLU, the input-variance and full rules coincide. The full rule follows Appendix C.5: each of the 512 calibration inputs, taken at initialization, has its own energy curve started from its own Q_0 and its own coordinate histogram, so the rule averages 512 predicted fractions and solves for the a that brings the average to p. The input energies are never pooled into one curve.

Grid. The primary grid crosses the four datasets, four activations and two depths at $s = 1$, giving 32 conditions. A secondary grid runs GELU and SELU at $L = 128$ with $s \in \{0.25, 4\}$ on all four datasets, 16 conditions. This grid focuses on the two activations whose calibration changes with input scale.

Optimization and selection. Training runs 600 updates of batch 128 over successive random permutations of the training set, using Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) without weight decay or schedule and with the total gradient norm clipped at 10. For each condition and method the learning rate is chosen from $\{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}\}$ by the mean final validation cross-entropy over two tuning seeds, with nonfinite runs ranked worst; only the chosen rate is run on three independent evaluation seeds, whose test values never enter selection. Checkpoints at 0, 100, 300 and 600 updates record the terminal fraction and the validation and test metrics, all at the rate selected at 600. We also record the fraction of terminal coordinates that are inactive on every one of the 512 fixed calibration inputs. This counts coordinates unused by that subset, whereas the terminal inactive fraction averages inactivity over inputs and coordinates.

Reporting. Results are means and sample standard deviations over three evaluation seeds. For an average across conditions, we first average the conditions within each seed and then compute the mean and standard deviation of the three seed averages. Random initial directions, data splits and minibatch order are paired across methods.

I.1 RESULTS

Table 18 summarizes the 32 primary conditions, using the reporting convention above. The mean initial inactive fractions are 0.102, 0.255 and 0.410 at targets 0.1, 0.25 and 0.4. Training does not

Method	Accuracy (%)	Difference from $a = 0$	Accuracy $\leq 11\%$
$a = 0$	75.25 ± 0.43	0.00 ± 0.00	1/96
$a = 1$	74.97 ± 0.76	-0.28 ± 0.97	1/96
$p = 0.1$	75.60 ± 0.13	0.35 ± 0.32	0/96
$p = 0.25$	74.44 ± 0.76	-0.81 ± 0.79	3/96
$p = 0.4$	74.10 ± 0.65	-1.14 ± 0.50	2/96
input variance	74.38 ± 0.85	-0.86 ± 0.86	3/96
ReLU rule	74.14 ± 1.07	-1.11 ± 0.84	4/96

Table 18: All primary-study methods, averaged over 32 conditions within each seed. Accuracy differences are in percentage points; uncertainties are sample standard deviations of three seed averages. The final column counts low-accuracy runs among 32×3 evaluations; this descriptive threshold excludes no runs. Target 0.25 and input variance coincide for ReLU and leaky ReLU; the ReLU rule also coincides for ReLU. These evaluations and counts overlap.

preserve these values: after 600 updates, the mean fractions lie between 0.56 and 0.64. The paired accuracy difference between $p = 0.1$ and $a = 0$ is 0.35 ± 0.32 percentage points, but no initialization is consistently best across the dataset–activation–depth conditions.

Some depth-128 CIFAR-10 runs reach chance accuracy. In every run with test accuracy at most 11%, all terminal features are zero on the 512 calibration inputs at updates 300 and 600. For GELU and SELU over all 24 depth-128 conditions with $s \in \{0.25, 1, 4\}$, separately selecting each method's learning rate gives a paired integral-minus-ReLU-rule accuracy difference of -0.49 ± 0.54 percentage points. Thus better matching the initial fraction does not imply better test accuracy in these experiments.

I.2 Comparison at a shared learning rate

To separate calibration from method-specific learning-rate selection, we also use one common rate for the three calibration rules in each of the 24 depth-128 conditions: four datasets, GELU and SELU, and $s \in \{0.25, 1, 4\}$. For each candidate rate, we average final validation cross-entropy over the original two tuning seeds and all three methods. The minimizing rate is used by all three methods, with ties resolved in favor of the smaller rate. Three fresh evaluation seeds run the same 600-update protocol with paired random draws.

The ReLU rule, input-variance correction and integral have mean test accuracies $74.50 \pm 0.32\%$, $74.34 \pm 0.10\%$ and $74.13 \pm 0.54\%$, respectively, after averaging conditions within each seed. The paired integral-minus-ReLU difference is -0.38 ± 0.77 percentage points. The largest condition-level difference is -9.25 points for CIFAR-10 with GELU at $s = 0.25$; the other five activation–scale groups range from -0.19 to 0.07 points after averaging over datasets. The corresponding mean absolute initial-fraction errors are 5.10, 2.27 and 2.24 points. The two rate-selection procedures therefore agree qualitatively: correcting initial activity does not produce an accuracy advantage here. The three-seed calibration errors also fluctuate across initializations, as the next diagnostic shows.

I.3 Variation across initializations

The learning study uses three evaluation initializations. To assess how much their initial fractions fluctuate, we repeat the ReLU setup at target 0.25, $n = m = 128$ and depths 32 and 128 on all four datasets, without training. One hundred fresh seeds repeat the data-selection, pixel standardization, 512 calibration inputs, positive embedding and weight initialization. Thus each seed repeats the input-sampling and initialization procedure, rather than holding one embedded input set fixed.

For the same embedded inputs and coefficient, we also generate independent one-input transitions using (17). This second calculation preserves each input's marginal law but removes the dependence created by sharing matrices across images. The two calculations have the same conditional expectation by (17); their comparison checks the implementation of this marginal law. Table 19 reports signed errors relative to 25%. Uncertainty is the sample standard deviation of the 100 per-seed means divided by $\sqrt{100}$; datasets and depths are paired within seeds.

Dataset	Depth	Learning seeds	Fresh matrices	Independent inputs
MNIST	32	+2.59	-0.03 ± 0.20	-0.20 ± 0.01
MNIST	128	-2.05	$+0.46 \pm 0.21$	-0.10 ± 0.02
Fashion-MNIST	32	+3.13	$+0.02 \pm 0.20$	-0.18 ± 0.02
Fashion-MNIST	128	-1.82	$+0.47 \pm 0.21$	-0.10 ± 0.02
KMNIST	32	+2.88	-0.02 ± 0.19	-0.18 ± 0.01
KMNIST	128	-1.89	$+0.27 \pm 0.20$	-0.09 ± 0.01
CIFAR-10	32	+1.97	-0.35 ± 0.24	-0.18 ± 0.01
CIFAR-10	128	-0.51	-0.27 ± 0.20	-0.08 ± 0.02

Table 19: ReLU initialization at target 25%: signed errors in percentage points. The last two columns use 100 fresh repetitions; $\pm$ denotes the standard error of the mean. Each repetition uses the same embedded inputs for its shared-matrix and independent-input calculations.

The three seeds used in the learning study have errors from -2.05 to $+3.13$ percentage points. With 100 fresh repetitions, the matrix means range from -0.35 to $+0.47$ points, with standard errors 0.19–0.24 points. The independent-input estimates are 0.08–0.20 points below target, with standard errors 0.01–0.02 points. The large depth-dependent shift in those three draws is not reproduced in the larger sample. Variation across shared-matrix initializations remains about two percentage points, much larger than the mean bias estimated from independent inputs. At depth 128, the paired matrix-minus-sampler differences for MNIST and Fashion-MNIST are 0.56 ± 0.21 and 0.57 ± 0.21 points, respectively (standard errors). These two contrasts share seed randomness across datasets and are about 2.6 standard errors from zero.

J A NORMALIZED BRANCH INPUT

The input-amplitude laws in Section 3.4 depend on the state energy seen by the branch. To isolate this dependence, we normalize the branch input by its root mean square, leaving the identity path unchanged:

$$x_{k+1} = x_k + \frac{a}{\sqrt{L}} B_k \phi\big(A_k \mathcal{R}(x_k)\big), \qquad \mathcal{R}(x) = \frac{\sqrt{n}\, x}{\|x\|} \quad (x \neq 0), \qquad \mathcal{R}(0) = 0. \tag{70}$$

This is an idealized RMS normalization (Zhang & Sennrich, 2019), with zero stabilization constant and unit gain, applied before A_k. It removes the dependence of the branch variance on the current state energy. Normalization placement also changes gradient scales at initialization (Xiong et al., 2020); here it changes how input amplitude enters the terminal-activity rule. The following result is finite-dimensional: neither width nor depth needs to tend to infinity. Throughout this appendix the input is positive and independent of the residual weights.

J.1 AN EXACT DISTRIBUTION AND A FINITE-SIZE BOUND

Let G be standard normal. We only require that ϕ is measurable, $v_\phi(1) = \mathbb{E}\phi(G)^2 > 0$, and $\mathbb{E}\phi(G)^4 < \infty$. Define

$$\kappa_\phi = \frac{\mathrm{Var}(\phi(G)^2)}{v_\phi(1)^2}, \qquad F_n(w) = \frac{1}{n}\sum_{i=1}^{n} \Phi(-u_i/\sqrt{w}) \quad (w > 0), \qquad F_n(0) = 0.$$

For positive u, F_n increases continuously from zero to one half. Write $w_{p,n}$ for its unique root at a prescribed $p \in (0, 1/2)$. Choosing $a^2 = w_{p,n}/v_\phi(1)$ matches the mean accumulated variance; the theorem below bounds the remaining error due to its random variation. This is the constant-profile case of (9). For ReLU, $\kappa_\phi = 5$; for leaky ReLU it is $6(1+\beta^4)/(1+\beta^2)^2 - 1$.

Theorem J.1. *Fix $n, m, L \geq 1$, $u \in (0,\infty)^n$, and $p \in (0, 1/2)$. In (70), take $a = a_{p,n} = \sqrt{w_{p,n}/v_\phi(1)}$. Then*

(i) The terminal state has the exact distribution

$$x_L \overset{d}{=} u + \sqrt{\tau}\, g, \qquad \tau = \frac{a^2}{mL}\sum_{j=1}^{mL} \phi(Z_j)^2, \qquad g \sim N(0, I_n),$$

where $Z_1, \ldots, Z_{mL}$ are independent standard normals and g is independent of them. In particular, $\mathbb{E}\tau = w_{p,n}$ and $\mathrm{Var}\,\tau = \kappa_\phi w_{p,n}^2/(mL)$.

(ii) The mean fraction and its mean-square error satisfy

$$\big|\mathbb{E}\widehat{p}_{n,L} - p\big| \leq \min\left\{\frac{1}{2}, \frac{2.28\kappa_\phi}{mL}\right\}, \tag{71}$$

$$\mathbb{E}(\widehat{p}_{n,L} - p)^2 \leq \frac{1}{4n} + \frac{1.06\kappa_\phi}{mL}. \tag{72}$$

(iii) For every $s > 0$, the calibrated coefficient obeys $a_{p,n}(su) = s a_{p,n}(u)$. More strongly, with paired weights the states satisfy $x_k(su; s\sigma) = s x_k(u; \sigma)$ at every layer, so the terminal fractions are identical.

Unlike the unnormalized rule, this coefficient changes linearly with input amplitude for every activation. Also, the mean bias has no state-width term, although a single network's fraction still fluctuates with n.

Proof. We first establish the mixture, then control its deviation from its mean variance. A nonzero state stays nonzero almost surely. Indeed, conditionally on $x_k \neq 0$ and A_k, a zero branch leaves it unchanged, whereas a nonzero branch followed by B_k adds a nondegenerate Gaussian vector. Induction therefore permits using $\|\mathcal{R}(x_k)\|^2 = n$ at every layer.

Given the past, $z_k = A_k \mathcal{R}(x_k)$ is standard Gaussian in $\mathbb{R}^m$, with a law independent of that past. Given z_k, the increment is centered Gaussian with covariance $a^2 H_k I_n / L$, where $H_k = m^{-1} \sum_j \phi(z_{kj})^2$. The conditional *joint* law of $(z_k, x_{k+1} - x_k)$ is thus constant given the past. These pairs are independent across layers and identically distributed. Summing their increments conditional on (H_k) proves (i), including the moments of τ.

For the error bounds, let φ denote the standard normal density. Direct differentiation, with $t = u/\sqrt{w}$, gives

$$\frac{d}{dw}\Phi(-u/\sqrt{w}) = \frac{t\varphi(t)}{2w}, \qquad \frac{d^2}{dw^2}\Phi(-u/\sqrt{w}) = \frac{t(t^2-3)\varphi(t)}{4w^2}.$$

Consequently $0 < F_n'(w) \le c_1/w$ and $|F_n''(w)| \le c_2/w^2$, where

$$c_1 = \frac{\varphi(1)}{2} < 0.121, \qquad c_2 = \frac{\sqrt{6(3-\sqrt{6})}}{4}\varphi(\sqrt{3-\sqrt{6}}) < 0.1377.$$

For the second constant, maximizing $t|t^2-3|\varphi(t)/4$ on either side of $\sqrt{3}$ gives the stationary points $t^2 = 3 \mp \sqrt{6}$; the first has the larger value.

Put $w = w_{p,n}$ and $\eta = \kappa_\phi/(mL)$. By (i), conditional on τ the terminal inactivity indicators are independent Bernoulli variables. At $\tau = 0$ they are all zero. Thus

$$\mathbb{E}[\widehat{p}_{n,L} \mid \tau] = F_n(\tau), \qquad \mathrm{Var}(\widehat{p}_{n,L} \mid \tau) \le \frac{1}{4n}.$$

On the event $\mathcal{A} = \{\tau \ge w/2\}$, Taylor's theorem bounds the first-order remainder by

$$|F_n(\tau) - F_n(w) - F_n'(w)(\tau - w)| \le 2c_2(\tau - w)^2/w^2.$$

On $\mathcal{A}^c$, its two terms have opposite signs: $F_n(\tau) - F_n(w) \in [-p, 0]$ and $F_n'(w)(w - \tau) \in [0, c_1]$. Its absolute value is therefore below $1/2$. Since $\mathbb{E}\tau = w$ and $\mathbb{P}(\mathcal{A}^c) \le 4\,\mathrm{Var}(\tau)/w^2 = 4\eta$,

$$|\mathbb{E}F_n(\tau) - F_n(w)| \le (2 + 2c_2)\eta < 2.28\eta.$$

The additional bound $1/2$ follows from $\mathbb{E}\widehat{p}_{n,L}, p \in [0, 1/2]$; it is not a bound on each realized fraction.

For the mean-square error, the conditional variance formula gives

$$\mathbb{E}(\widehat{p}_{n,L} - p)^2 \le \frac{1}{4n} + \mathbb{E}\big(F_n(\tau) - F_n(w)\big)^2.$$

The last squared difference is at most $4c_1^2(\tau - w)^2/w^2$ on $\mathcal{A}$, and at most $1/4$ on its complement. Its expectation is bounded by $(1 + 4c_1^2)\eta < 1.06\eta$, proving (ii).

Finally, $F_{n,su}(w) = F_{n,u}(w/s^2)$ gives $w_{p,n}(su) = s^2 w_{p,n}(u)$ and $a_{p,n}(su) = s a_{p,n}(u)$. Because $\mathcal{R}(sx) = \mathcal{R}(x)$, substitution in each block proves the paired-state identity in (iii) by induction. □

Several inputs sharing weights. For fixed positive inputs $u^{(1)}, \dots, u^{(N_{\mathrm{pr}})}$, solve $N_{\mathrm{pr}}^{-1} \sum_b F_n^{(b)}(w) = p$ and use $a^2 = w/v_\phi(1)$. Each marginal has the same mean clock w. Write $p_b = F_n^{(b)}(w)$, so $N_{\mathrm{pr}}^{-1} \sum_b p_b = p$. The bias bound follows by averaging the single-input bounds around their respective p_b. For the mean-square error, Jensen's inequality yields

$$\mathbb{E}\left(\frac{1}{N_{\mathrm{pr}}}\sum_b \widehat{p}^{(b)} - p\right)^2 \le \frac{1}{N_{\mathrm{pr}}}\sum_b \mathbb{E}(\widehat{p}^{(b)} - p_b)^2 \le \frac{1}{4n} + \frac{1.06\kappa_\phi}{mL}.$$

This uses no independence between the inputs' outputs. Random input embeddings independent of the residual weights can be conditioned on before applying the result.

The leading bias. If $\mathbb{E}|\phi(G)|^6 < \infty$, the bias has the expansion

$$\mathbb{E}\widehat{p}_{n,L} - p = \frac{\kappa_\phi}{8mL}\frac{1}{n}\sum_i t_i(t_i^2 - 3)\varphi(t_i) + O((mL)^{-3/2}), \qquad t_i = \frac{u_i}{\sqrt{w_{p,n}}}. \tag{73}$$

Activation	κ_ϕ	$mL = 16$	$mL = 64$	$mL = 256$	$mL = 1024$
ReLU	5.000	-2.204 ± 0.011	-0.535 ± 0.005	-0.134 ± 0.002	-0.032 ± 0.001
Leaky ReLU	4.882	-2.143 ± 0.011	-0.521 ± 0.005	-0.131 ± 0.002	-0.031 ± 0.001
GELU	6.216	-2.689 ± 0.012	-0.667 ± 0.006	-0.166 ± 0.003	-0.040 ± 0.001
SELU	1.503	-0.624 ± 0.005	-0.156 ± 0.003	-0.040 ± 0.001	-0.009 ± 0.001

Table 20: Normalized branches: mean bias in percentage points and Monte Carlo standard error at target 0.25.

To see this, put $T = \tau/w$ and expand $F_n(wT)$ to second order at $T = 1$. The normalized third derivative is bounded on $T \geq 1/2$, since it is an average of $t(t^4 - 10t^2 + 15)\varphi(t)/(8T^3)$, with $t = t_i/\sqrt{T}$. For the average of the mL centered independent variables defining $T-1$, symmetrization and the third-moment bound for Rademacher sums give $\mathbb{E}|T-1|^3 \leq C(mL)^{-3/2}$. On $T < 1/2$, the Taylor terms are bounded and $\mathbb{P}(T < 1/2) \leq 8\mathbb{E}|T-1|^3$. Both remainders therefore have the stated order. Substituting $\mathrm{Var}(T) = \kappa_\phi/(mL)$ gives (73). For equal inputs at $p = 0.25$, the ReLU leading bias is $-0.34094/(mL)$. The $1/(mL)$ bias order is attained in this case; the leading coefficient can vanish at other targets.

J.2 Numerical checks

We check the bias and the input-scale identity separately. For the bias, we use equal positive inputs, $n = 128$, $m = 1$, target 0.25 and $L \in \{16, 64, 256, 1024\}$. For each activation, three seeds each generate 131072 independent sequences of branch energies. We evaluate the conditional Gaussian probabilities exactly given each sequence, rather than sampling the terminal Gaussian vector. Table 20 reports the estimated mean bias and its Monte Carlo standard error; the dashed curves in Figure 16(a) are the leading term in (73). At $mL = 16$, the ReLU bias is -2.204 ± 0.011 percentage points, compared with the leading prediction -2.131; higher-order terms are still visible at this depth. Normalization makes the input-scale rule common to the activations, but their finite-depth biases differ through κ_ϕ. This relative variance is 6.22 for GELU and 1.50 for SELU, compared with 5.00 for ReLU. The smaller relative variation of SELU branch energies gives a smaller leading bias in this comparison.

The image comparison uses MNIST, Fashion-MNIST, KMNIST and CIFAR-10, with the same scalar pixel standardization and positive softplus entrance as Appendix I. Each initialization uses 512 nearly class-balanced images, a fixed Xavier-normal embedding, $n = 128$, $m = 32$ and depths 64 and 256. For each activation and each of three initializations, we calibrate target 0.25 on the unit-scale inputs, then multiply those inputs by $s \in \{0.25, 1, 4\}$. We compare reusing the coefficient unchanged with multiplying it by s. All images and both rules share the actual matrices, and the two depths use a common matrix prefix. No weights are trained in this comparison. Figure 16(b) shows the depth-256 result, and Table 21 gives both depths. With the paired power-of-two scales, corrected states divided by s match the unit-scale states exactly in the computation. Reusing the coefficient instead gives depth-256 fractions, averaged over datasets and initializations, of 0.41–0.43 at $s = 0.25$ and about 0.07 at $s = 4$, depending on the activation. The correction preserves the original fraction, including its initialization-to-initialization variation around the target.

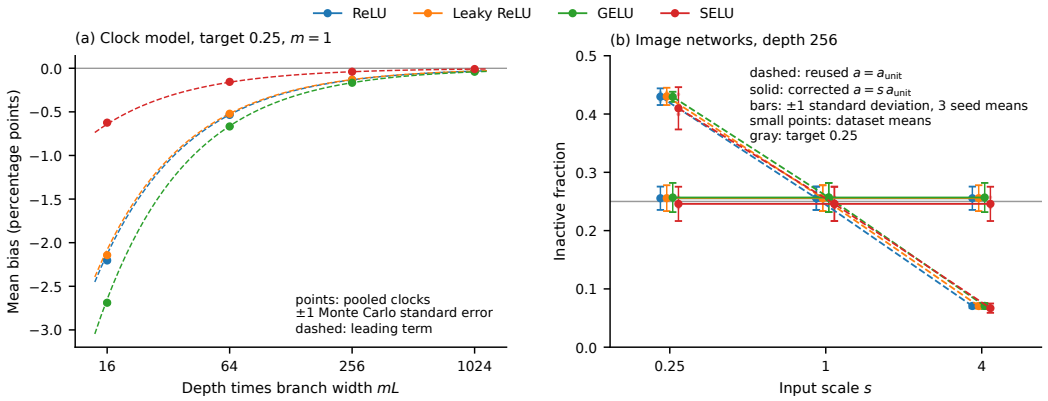


Figure 16: Normalized branches. (a) Mean bias and its leading prediction. (b) Input-scale transfer at depth 256: reused coefficient (dashed) and coefficient multiplied by input scale (solid). Bars show sample standard deviations of three means over datasets; faint dots are dataset means. Markers are offset horizontally for visibility.

		$L = 64$			$L = 256$		
Dataset	Activation	$s = 0.25$	$s = 1$	$s = 4$	$s = 0.25$	$s = 1$	$s = 4$
MNIST	ReLU	42.3 / 24.7	24.7 / 24.7	7.3 / 24.7	42.5 / 25.6	25.6 / 25.6	7.2 / 25.6
MNIST	Leaky ReLU	41.6 / 24.8	24.8 / 24.8	7.4 / 24.8	42.6 / 25.6	25.6 / 25.6	7.2 / 25.6
MNIST	GELU	40.5 / 24.4	24.4 / 24.4	7.2 / 24.4	42.5 / 25.8	25.8 / 25.8	7.2 / 25.8
MNIST	SELU	37.8 / 24.5	24.5 / 24.5	7.2 / 24.5	39.9 / 24.4	24.4 / 24.4	6.7 / 24.4
Fashion-MNIST	ReLU	42.2 / 25.0	25.0 / 25.0	6.8 / 25.0	43.7 / 25.5	25.5 / 25.5	6.9 / 25.5
Fashion-MNIST	Leaky ReLU	41.6 / 25.1	25.1 / 25.1	6.9 / 25.1	43.8 / 25.6	25.6 / 25.6	6.9 / 25.6
Fashion-MNIST	GELU	40.8 / 24.7	24.7 / 24.7	6.7 / 24.7	43.6 / 25.8	25.8 / 25.8	7.0 / 25.8
Fashion-MNIST	SELU	37.8 / 24.5	24.5 / 24.5	6.9 / 24.5	40.0 / 24.3	24.3 / 24.3	6.6 / 24.3
KMNIST	ReLU	42.6 / 24.7	24.7 / 24.7	6.9 / 24.7	43.5 / 25.7	25.7 / 25.7	7.1 / 25.7
KMNIST	Leaky ReLU	41.9 / 24.8	24.8 / 24.8	6.9 / 24.8	43.6 / 25.7	25.7 / 25.7	7.0 / 25.7
KMNIST	GELU	41.1 / 24.4	24.4 / 24.4	6.8 / 24.4	43.5 / 25.8	25.8 / 25.8	7.1 / 25.8
KMNIST	SELU	38.7 / 23.9	23.9 / 23.9	6.6 / 23.9	40.6 / 23.9	23.9 / 23.9	6.5 / 23.9
CIFAR-10	ReLU	41.2 / 24.6	24.6 / 24.6	7.0 / 24.6	42.2 / 25.4	25.4 / 25.4	7.1 / 25.4
CIFAR-10	Leaky ReLU	41.0 / 24.7	24.7 / 24.7	6.9 / 24.7	42.2 / 25.4	25.4 / 25.4	7.1 / 25.4
CIFAR-10	GELU	41.4 / 24.8	24.8 / 24.8	6.9 / 24.8	42.2 / 25.3	25.3 / 25.3	7.1 / 25.3
CIFAR-10	SELU	41.5 / 24.4	24.4 / 24.4	6.7 / 24.4	43.4 / 25.7	25.7 / 25.7	7.0 / 25.7

Table 21: Image inputs with normalized branches: mean inactive percentages over three initializations, reused / corrected coefficient. The target is 25%.

J.3 A POSITIVE STABILIZATION CONSTANT

Practical RMS normalization divides by the square root of the mean square plus a positive constant. This constant matters when the input energy is comparable to it. Let $\varepsilon > 0$ be fixed and let $\zeta > 0$ be a fixed scalar gain. Replace $\mathcal{R}$ in (70) by

$$\mathcal{R}_{\varepsilon,\zeta}(x) = \frac{\zeta x}{\sqrt{\|x\|^2/n + \varepsilon}}, \qquad h(q) = \frac{\zeta^2 q}{q + \varepsilon}.$$

The branch variance now depends on the current energy again, through $\widetilde{v}(q) = v_\phi(h(q))$. We return to the standing Lipschitz activation assumption of Section 3.

Proposition J.2. *Under the input and dimension assumptions of Theorem 3.2, its conclusion holds with v_ϕ replaced by $\widetilde{v}$ in the radial ODE. For positive inputs as in Theorem 3.3, the unique limiting coefficient is*

$$a_p^2 = \int_{q_0}^{q_0+w_p} \frac{dq}{v_\phi(\zeta^2 q/(q+\varepsilon))}. \tag{74}$$

Replacing q_0, w_p by their finite-input counterparts gives a consistent calibration rule. In particular, if $v_\phi(q) = c_\phi q$, then

$$a_p(su)^2 = \frac{s^2 w_p + \varepsilon \log(1 + w_p/q_0)}{c_\phi \zeta^2}, \qquad s > 0, \tag{75}$$

where q_0, w_p refer to the unscaled input law.

Proof. Conditionally on the state, the coordinates of $A_k \mathcal{R}_{\varepsilon,\zeta}(x_k)$ are independent $N(0, h(Q_k))$ variables. The representation (17) therefore holds with $H_k = m^{-1} \sum_j \phi(\sqrt{h(Q_k)} z_{kj})^2$. Set $K_* = K\zeta/\sqrt{\varepsilon}$. The inequalities $|v_\phi(q) - v_\phi(r)| \le K^2 |q - r|$ and $0 \le h'(q) \le \zeta^2/\varepsilon$ give

$$|\widetilde{v}(q) - \widetilde{v}(r)| \le K_*^2 |q - r|, \qquad \widetilde{v}(q) \le K_*^2 q, \qquad \mathbb{E}\phi(\sqrt{h(q)}G)^4 \le 3K_*^4 q^2.$$

These are the bounds used in Appendix C to control radial moments, compare the radial chain with its ODE, and replace conditional characteristic functions by their Gaussian limits. Thus the same argument proves convergence of the empirical fraction, with radial mean-square error of order $1/n + 1/(mL) + 1/L^2$. The constants may now depend on ε and ζ.

The ODE remains global because $\widetilde{v}(q) \le K^2 \zeta^2$. Along its increasing path from $q_0 > 0$,

$$\widetilde{v}(q) \ge c_- \zeta^2 q_0/(q_0 + \varepsilon) > 0.$$

Its accumulated variance consequently increases continuously from zero to infinity as a increases. The Gaussian mixture root w_p is unchanged, and separation of variables gives (74). The integrand is continuous and positive on a neighborhood of the limiting integration interval. The finite-input roots and energies converge as in Theorem 3.3, proving consistency. For a linear profile, the integrand is $(1 + \varepsilon/q)/(c_\phi \zeta^2)$. Integrating and using $q_0(su) = s^2 q_0$ and $w_p(su) = s^2 w_p$ proves (75). □

For ReLU and leaky ReLU, the two terms in (75) separate the regimes. The logarithmic term dominates as $s \downarrow 0$, so the coefficient approaches a positive constant. For large s, the variance term dominates and the coefficient is asymptotically linear in input amplitude. Setting $\varepsilon = 0$ in the formula recovers the ideal RMS rule. At positive ε, the branch energies depend on the state, so the exact mixture and explicit finite-size constants of Theorem J.1 are replaced by the joint-limit result above.

More generally, all four activations have a finite positive $k_\phi(0+)$, and (74) gives

$$\lim_{s \downarrow 0} a_p(su)^2 = \frac{\varepsilon \log(1 + w_p/q_0)}{\zeta^2 k_\phi(0+)}, \qquad \lim_{s \to \infty} \frac{a_p(su)^2}{s^2} = \frac{w_p}{v_\phi(\zeta^2)}.$$

Change variables $q = s^2 t$ in the integral and use the positive linear lower bound on v_ϕ to apply dominated convergence on $[q_0, q_0 + w_p]$. Thus the same two regimes persist beyond homogeneous activations, with activation-dependent constants. For a homogeneous profile, the squared coefficient is exactly the sum of its ideal-RMS value and ε/ζ^2 times the unnormalized value in Theorem 3.4(i).

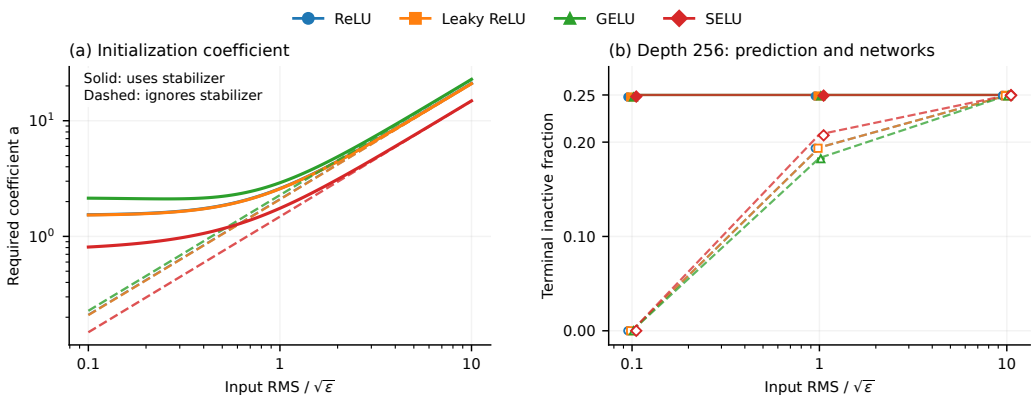


Figure 17: Positive stabilization constant. (a) Required coefficients. (b) Limiting predictions and depth-256 networks; bars are Monte Carlo standard errors. Solid/filled includes the constant; dashed/open ignores it. Markers are offset horizontally for visibility.

The intermediate scales distinguish GELU from SELU. Put $\xi = s^2$ and $h_\xi(t) = \zeta^2 \xi t/(\xi t + \varepsilon)$. Differentiating the integrand after the same change of variables gives

$$\frac{\partial}{\partial \xi} \frac{\xi}{v_\phi(h_\xi(t))} = \frac{k_\phi(h_\xi(t)) - (\zeta^2 - h_\xi(t))k'_\phi(h_\xi(t))}{\zeta^2 k_\phi(h_\xi(t))^2}.$$

SELU has $k'_\phi < 0$, so its coefficient increases strictly with input amplitude. GELU has $k_\phi(0+) = 1/4$ and $k'_\phi(0+) = 3/(2\pi)$. If $\zeta^2 > \pi/6$, including unit gain, its coefficient initially decreases as the input grows, before eventually increasing linearly. Thus a positive stabilization constant does not entirely remove the activation-dependent behavior of small inputs.

J.4 Checking the transition between input scales

We set $\varepsilon = \zeta = 1$ and use constant positive inputs with RMS 0.1, 1 and 10, spanning energies below and above the stabilization constant. The networks have $n = 128$, $m = 32$ and depth 64 or 256. For each activation and depth, three seeds each generate 1536 independent networks through the exact one-input Gaussian transition. Each rule evaluates the branch variance at its own current state. Gaussian draws are paired across rules and input scales within each depth.

We compare (74) with the ideal-RMS rule that ignores the stabilization constant, both at target 0.25. Figure 17 shows the coefficient and depth-256 fractions; Table 22 includes both depths. The corrected mean fractions are within 0.37 percentage points of the target over the 24 activation–depth–scale settings. Ignoring the constant produces fractions near zero at input RMS 0.1 and between 0.18 and 0.21 at RMS 1; the two rules nearly coincide at RMS 10. The relevant comparison is the input RMS divided by $\sqrt{\varepsilon}$: for a smaller stabilization constant, the same transition occurs at correspondingly smaller inputs.

Activation	Input RMS	$L = 64$		$L = 256$	
		Includes ε	Ignores ε	Includes ε	Ignores ε
ReLU	0.1	24.64 ± 0.05	0.00 ± 0.00	24.79 ± 0.05	0.00 ± 0.00
ReLU	1	24.88 ± 0.06	19.30 ± 0.05	24.89 ± 0.05	19.35 ± 0.05
ReLU	10	24.97 ± 0.06	24.92 ± 0.06	24.96 ± 0.06	24.91 ± 0.06
Leaky ReLU	0.1	24.64 ± 0.05	0.00 ± 0.00	24.79 ± 0.05	0.00 ± 0.00
Leaky ReLU	1	24.88 ± 0.06	19.30 ± 0.05	24.89 ± 0.05	19.35 ± 0.05
Leaky ReLU	10	24.97 ± 0.06	24.92 ± 0.06	24.96 ± 0.06	24.91 ± 0.06
GELU	0.1	24.64 ± 0.05	0.00 ± 0.00	24.78 ± 0.05	0.00 ± 0.00
GELU	1	24.85 ± 0.06	18.15 ± 0.05	24.90 ± 0.05	18.24 ± 0.05
GELU	10	24.98 ± 0.06	24.91 ± 0.06	24.96 ± 0.06	24.89 ± 0.06
SELU	0.1	24.71 ± 0.05	0.00 ± 0.00	24.83 ± 0.05	0.00 ± 0.00
SELU	1	24.90 ± 0.06	20.77 ± 0.05	24.93 ± 0.06	20.72 ± 0.05
SELU	10	24.97 ± 0.06	24.93 ± 0.06	24.99 ± 0.06	24.95 ± 0.06

Table 22: Positive stabilization constant: mean inactive percentages and Monte Carlo standard errors over 4608 independent networks per cell. The prescribed target is 25%.