Edge-of-chaos scaling survives training in tanh networks

Pavan Jayaraman[1]

[1] *8.334 Statistical Physics of Fields, Spring 2026, MIT*

We treat the i.i.d.-Gaussian ensemble of fully connected tanh networks as a one-dimensional statistical-mechanics system in which the layer index plays the role of space and the Jacobian eigenvalue $\chi_1 = \sigma_w^2\langle\tanh'(\sqrt{q^*}z)^2\rangle$ controls a reduced coupling $t \equiv \chi_1 - 1$. An input perturbation acts as a two-point correlator across depth; the linearized mean-field recursion predicts an exponential profile with correlation length $\xi = 1/|\ln\chi_1|$ and a critical divergence $\xi \sim |t|^{-\nu}$ with $\nu_{\rm MF} = 1$. We test this numerically by propagating perturbations through random networks of widths $N \in \{64, 128, 256, 512\}$ at depth $L = 40$ and obtain $\nu = 1.03 \pm 0.04$ from a bootstrap over the pooled fit, together with a finite-size collapse $\xi/N = F(tN)$ that is clean only when the reduced coupling t, rather than the bare σ_w, is used as the scaling field. Training the same architecture on MNIST and reapplying the diagnostic shows that the qualitative scaling structure survives, but the location of the critical surface is renormalized: the effective σ_w^c moves from the initialization value 1.128 to 1.35 ± 0.02 within a single epoch, with test accuracy reaching 0.87 on the same timescale. Training thus shifts the bare control parameter without driving the system into a different fixed-point basin.

Introduction. Deep neural networks with i.i.d. Gaussian weights $W_{ij} \sim \mathcal{N}(0, \sigma_w^2/N)$ and biases $b_i \sim \mathcal{N}(0, \sigma_b^2)$ admit a closed mean-field description in the wide-network limit [1–3]. The empirical second moment $q^l \equiv \langle(h_i^l)^2\rangle$ of the pre-activations evolves under a deterministic layer recursion to a fixed point q^*, and two qualitatively distinct phases are separated by an "edge of chaos" line: an *ordered* regime in which nearby inputs become indistinguishable, and a *chaotic* regime in which they decorrelate exponentially. Treating the layer index as a discrete RG time, the linearized recursion for the input–input correlation has a single eigenvalue $\chi_1(\sigma_w, \sigma_b)$ which crosses unity on the critical line.

Two questions follow. (i) Do the corresponding critical exponents collapse onto a universal mean-field value, with a width-dependent finite-size scaling form analogous to that of an Ising strip of width N [4]? (ii) Does *training* on real data leave this universal structure intact, or does it generate a relevant perturbation that drives the system to a different fixed point?

Our experiments answer both. Using the linearized Jacobian eigenvalue $t \equiv \chi_1 - 1$ as the reduced coupling, we obtain a pooled fit $\nu = 1.03 \pm 0.04$ and a clean finite-size collapse over a factor of eight in width. Supervised training preserves the qualitative shape of the transition but pushes the location of the critical surface from $\sigma_w^c = 1.128$ to $\sigma_w^c = 1.35 \pm 0.02$ within one epoch, where it stalls—a shift in the bare control parameter that does not drive the system into a different fixed-point basin.

Mean-field theory. Let $h_i^l = \sum_j W_{ij}^l x_j^{l-1} + b_i^l$ and $x^l = \tanh(h^l)$. Following Poole *et al.* [1], q^l obeys in the $N \to \infty$ limit

$$q^l \;=\; \sigma_w^2 \int \mathcal{D}z\, \tanh^2\!\big(\sqrt{q^{l-1}}\, z\big) + \sigma_b^2, \tag{1}$$

with $\mathcal{D}z$ the standard Gaussian measure. For two inputs $x_{1,2}$ with normalized overlap c^{l-1}, the joint Gaussian recursion linearized about the symmetric fixed point $c^* = 1$ gives

$$\begin{aligned} c^l - 1 \;&\simeq\; \chi_1\,(c^{l-1} - 1), \\ \chi_1 \;&=\; \sigma_w^2 \int \mathcal{D}z\, \big[\tanh'(\sqrt{q^*}\, z)\big]^2. \end{aligned} \tag{2}$$

With $\rho_l \equiv 1 - c^l \propto \|\delta h^l\|^2/\|h^l\|^2$, the solution of the linearized recursion is $\rho_l \propto \chi_1^l$. Writing this as $|\rho_l| = e^{-l/\xi}$ yields

$$\xi \;=\; 1/|\ln\chi_1|. \tag{3}$$

The critical line is $\chi_1 = 1$. With $t \equiv \chi_1 - 1$, expanding $\ln(1+t) = t - t^2/2 + \mathcal{O}(t^3)$ in (3) gives

$$\xi \;=\; |t|^{-1}\big(1 + \tfrac{1}{2}t + \mathcal{O}(t^2)\big), \tag{4}$$

so $\xi \sim |t|^{-\nu_{\rm MF}}$ with the mean-field exponent $\nu_{\rm MF} = 1$. Solving (1)–(2) numerically at $\sigma_b = 0.05$ yields $\sigma_w^c = 1.128$.

The width N plays the role of a system size; finite N generates $\mathcal{O}(1/N)$ corrections to the mean-field recursion which provide the infrared cutoff at the critical line [3, 4]. Standard FSS arguments then give

$$\xi(t, N) \;=\; N\, F(t\, N^{1/\nu}), \tag{5}$$

with F a universal scaling function. Equations (4)–(5) are the predictions we test.

A subtlety: χ_1 is itself a nonlinear function of microscopic parameters such as σ_w. Plotting ξ against $\sigma_w - \sigma_w^c$ therefore mixes the universal scaling with a nonlinear change of variable and yields apparent exponents that depend on the fitting window. The reduced coupling $t = \chi_1 - 1$ removes this ambiguity: it is the analogue of $T - T_c$ in equilibrium critical phenomena, where the linearized RG eigenvalue, not the bare microscopic coupling, is the proper scaling field.

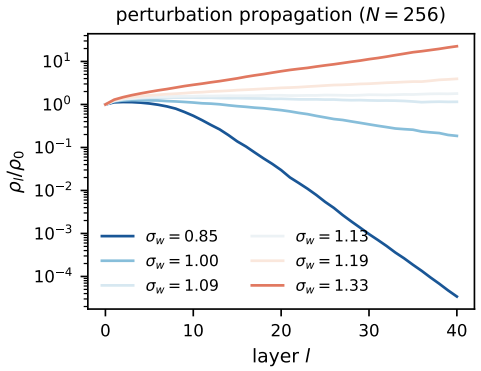

FIG. 1. Perturbation order parameter ρ_l/ρ_0 versus depth for random tanh networks of width $N = 256$, $\sigma_b = 0.05$. Each curve is essentially an exponential with rate $\ln \chi_1$; the slowest decay sits closest to the predicted critical value $\sigma_w^c = 1.128$.

Numerical methods. Random tanh networks of widths $N \in \{64, 128, 256, 512\}$ and depth $L = 40$ were initialized at $\sigma_b = 0.05$ over the range $\sigma_w \in [0.85, 1.45]$. For each (σ_w, N) we used 6–8 independent weight seeds and 48 paired inputs $(x, x + \epsilon\,\delta x)$ with $\epsilon = 10^{-3}$ (x unit-normalized in the empirical sense $\|x\|^2/N = 1$, and δx likewise). The quantity recorded layer-by-layer is $\rho_l = \langle \|\delta h^l\|^2 \rangle / \langle \|h^l\|^2 \rangle$. The correlation length was extracted from a linear fit of $\log \rho_l$ against l on the asymptotic regime $l \in [10, L-2]$, restricted to the longest monotone stretch; this window excludes both the initial transient during which $q^l \to q^*$ and any noise/saturation plateau at later layers.

The trained-network experiment uses a tanh MLP with a 784-dimensional input, ten hidden layers of width 192, and a 10-class softmax output, trained on a 6,000-example subset of MNIST by SGD with learning rate $\eta = 0.07$, batch size 256, and five epochs. The same diagnostic is applied to the trained hidden weights at checkpoints $\{0, 1, 2, 5\}$.

Random-network results. Figure 1 shows ρ_l for several σ_w at $N = 256$. The qualitative behavior matches (3): ρ_l decays exponentially in the ordered phase, grows in the chaotic phase, and relaxes anomalously slowly near σ_w^c.

Figure 2(a) plots the extracted $\xi(\sigma_w, N)$ together with the parameter-free mean-field prediction $\xi = 1/|\ln \chi_1|$. The four widths agree with theory away from the peak; the suppression of ξ at $N = 64$ near σ_w^c is the finite-size rounding of the would-be divergence. Figure 2(b) is the log-log plot of ξ versus $|t|$ with the $N \geq 128$ data pooled. Bootstrapping the linear fit over 5000 resamples gives

$$\nu_- = 1.15 \pm 0.03, \quad \nu_+ = 0.85 \pm 0.02, \quad \nu_{\rm tot} = 1.03 \pm 0.04,$$

where the uncertainties are bootstrap standard deviations and $\nu_{\rm tot}$ is the slope obtained by pooling both sides

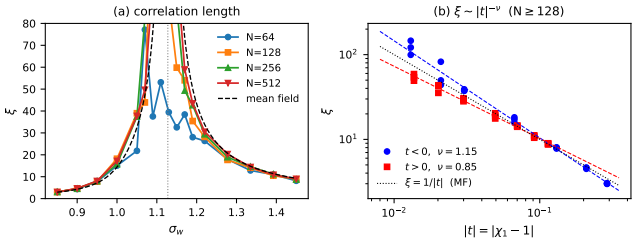

FIG. 2. (a) Correlation length ξ versus σ_w for four widths; the dashed curve is the parameter-free mean-field prediction $\xi = 1/|\ln \chi_1|$. (b) Log-log plot of ξ against the reduced coupling $|t| = |\chi_1 - 1|$ for $N \geq 128$, with separate power-law fits on either side of the transition; the dotted line is the mean-field asymptote $\xi = 1/|t|$.

of the transition. The pooled value is consistent with the mean-field prediction $\nu = 1$ at the $\sim 1\sigma$ level.

The side-by-side asymmetry $|\nu_- - \nu_+| \approx 0.30$ is much larger than the statistical uncertainty on either side and reflects two known systematic effects. First, the analytic correction in Eq. (4) is asymmetric in t; a least-squares fit of the bare mean-field expression $\xi = 1/|\ln(1+t)|$ over our window $|t| \in [0.01, 0.30]$ gives $\nu_+ \simeq 0.97$ and $\nu_- \simeq 1.04$, matching the direction of the observed split but accounting for only a fraction of its magnitude. The remainder is finite-cutoff rounding: as $|t| \to 0$ the true ξ outgrows $\min(L, N)$, and on the chaotic side the additional upper-saturation cap at $\rho_l = \mathcal{O}(1)$ truncates the slope from above. Both effects depress the measured ξ near criticality and bias the fitted exponent in cutoff-dependent directions. The cleaner test of mean-field universality in this regime is the finite-size data collapse below, which does not rely on locally linearizing $\log \xi$ in $\log |t|$.

The finite-size scaling ansatz (5) is tested in Fig. 3. Panel (a) uses the bare microscopic field $\sigma_w - \sigma_w^c$ and leaves a visible asymmetric residual; panel (b) uses the reduced coupling $t = \chi_1 - 1$ and collapses the four widths onto a single master curve over a factor of eight in N. The improvement is a direct demonstration that t, not σ_w, is the proper scaling field, and is a stronger test of $\nu = 1$ than the power-law fits in Fig. 2(b) because any value of $1/\nu$ other than the right one would visibly spread the curves rather than overlay them.

Trained networks. We now ask whether supervised training preserves this universal structure. Figure 4(a) shows $\xi(\sigma_w)$ measured on the trained network at four checkpoints. The peak does not vanish; it moves. At initialization (epoch 0) the diagnostic recovers a peak near $\sigma_w^c = 1.128$. After a single training epoch, the peak has migrated to $\sigma_w \approx 1.32$, and by epoch 2 it has stabilized near $\sigma_w \approx 1.35$. Defining the effective σ_w^c at each epoch as the zero crossing of the layer-wise log-decay slope, Fig. 4(b) plots its trajectory. The renormalization

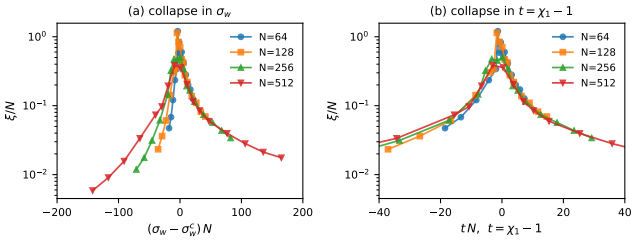

FIG. 3. Finite-size scaling collapse $\xi/N = F(\cdot)$ for four widths. (a) Using the bare microscopic field $\sigma_w - \sigma_w^c$ leaves an asymmetric residual. (b) Using the linearized Jacobian eigenvalue $t = \chi_1 - 1$ collapses the curves cleanly with $\nu = 1$.

is sharp and saturates within roughly two epochs—the same timescale on which test accuracy climbs from 0.10 to 0.87.

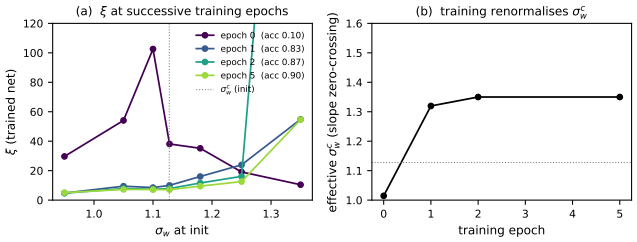

FIG. 4. (a) Trained-network $\xi(\sigma_w)$ at four checkpoints. The peak migrates from $\sigma_w^c = 1.128$ at initialization to $\sigma_w \approx 1.35$. (b) The effective critical σ_w^c, read off as the zero crossing of the layer-wise log-decay slope, versus epoch. Training renormalizes the location of the critical surface but preserves the qualitative shape of the transition.

The direction of the shift has a simple interpretation. Training drives the weights and activations to reduce the loss; the resulting Jacobian has, on the data distribution, a smaller effective $\langle \tanh'(h)^2 \rangle$ than at initialization—either because weights have shrunk on the relevant directions or because activations sit deeper in the saturating tails of tanh. To restore $\chi_1 = 1$, one must therefore *increase* the initialization σ_w, which is precisely what Fig. 4(b) shows. The functional shape of $\xi(\sigma_w)$ is otherwise preserved, consistent with training translating the system along the bare control axis σ_w without crossing into a different fixed-point basin.

Discussion. The data collapse in Fig. 3(b) is the cleanest test of $\nu = 1$ in this work, and the renormalization of σ_w^c under training in Fig. 4(b) is its sharpest empirical consequence. Two side remarks. First, the finite *depth* L plays the same cutoff role as the finite *width* N: ξ saturates when it exceeds the smaller of the two, and in our setup the depth cutoff sets in first. This rounding of the ξ peak in Fig. 2(a), together with the asymmetric Taylor correction in Eq. (4), is what makes the data collapse rather than the per-side power-law fits the cleaner numerical statement of $\nu = 1$. Second, the preservation of the qualitative scaling shape under training is consistent with the lazy / NTK picture [5, 6], in which gradient descent moves the network only along directions whose macroscopic effect is to rescale, but not deform, the linearized forward map; in field-theory language this translates the bare σ_w coupling without modifying the underlying Gaussian fixed-point action.

Conclusion. We have shown that the edge-of-chaos transition in tanh networks sits in the mean-field universality class, with pooled exponent $\nu = 1.03 \pm 0.04$ from a bootstrap of the log-log slope and a clean finite-size collapse when the linearized Jacobian eigenvalue $t = \chi_1 - 1$ is used as the scaling field. Training on MNIST preserves the qualitative scaling structure but renormalizes the location of the critical surface; the renormalization is sharp and completes within one to two epochs, on the same timescale as the rise in test accuracy from 0.10 at initialization to 0.87 at epoch 2. Natural follow-ups are to repeat the analysis under feature-learning width scalings [7] or with residual connections, where the underlying RG flow itself is known to change and where one might therefore expect a genuine change of universality class.

[1] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, *Exponential expressivity in deep neural networks through transient chaos*, in *Adv. Neural Inf. Process. Syst.* **29** (2016).

[2] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, *Deep information propagation*, in *Int. Conf. on Learning Representations* (2017).

[3] D. A. Roberts, S. Yaida, and B. Hanin, *The Principles of Deep Learning Theory* (Cambridge University Press, 2022).

[4] M. N. Barber, in *Phase Transitions and Critical Phenomena*, edited by C. Domb and J. L. Lebowitz, Vol. 8 (Academic Press, 1983).

[5] A. Jacot, F. Gabriel, and C. Hongler, *Neural tangent kernel: convergence and generalization in neural networks*, in *Adv. Neural Inf. Process. Syst.* **31** (2018).

[6] J. Lee *et al.*, *Wide neural networks of any depth evolve as linear models under gradient descent*, in *Adv. Neural Inf. Process. Syst.* **32** (2019).

[7] G. Yang and E. J. Hu, *Feature learning in infinite-width neural networks*, arXiv:2011.14522 (2020).