Spectral Phase Transitions in the Loss Landscape of Finite-Width Neural Networks

Jacob Crainic

February 2026

Abstract

A central puzzle in deep learning theory is why gradient descent reliably finds good solutions despite the extreme non-convexity of neural network loss landscapes, particularly in the moderately overparameterized regime where existing theoretical guarantees require polynomial width scaling far exceeding practical network sizes. We study the critical-point structure of the empirical risk landscape for two-layer neural networks with ReLU activations, trained on n data points in $\mathbb{R}^d$ with m hidden neurons. Our main result identifies a critical width-to-sample ratio $\gamma_\star$, depending on the spectral distribution of the data covariance, that marks a *topological phase transition* in the loss landscape: above $\gamma_\star$, the landscape is provably benign (all local minima are global with high probability), while below $\gamma_\star$, the expected number of spurious critical points grows exponentially in n. The critical ratio $\gamma_\star$ is characterized as the unique solution to a fixed-point equation involving the Stieltjes transform of the Marchenko–Pastur law composed with the data spectrum. For isotropic data ($\Sigma = I_d$), the critical ratio takes the explicit form $\gamma_\star(\delta) = 2(1-2\delta)/(1-\delta-\delta^2)$ for $\delta < 1/2$, which admits the first-order approximation $\gamma_\star \approx 4/(2+3\delta)$ (exact at $\delta = 1/4$). At the transition, the Hessian at near-critical points exhibits a spectral gap collapse: the smallest non-zero eigenvalue vanishes linearly as $|\gamma - \gamma_\star|$, yielding a universal scaling law with critical exponent $\beta = 1$. Yet gradient-based optimizers navigate the subcritical ($\gamma < \gamma_\star$) regime with apparent ease, achieving near-zero training loss well below the theoretical threshold. This reveals a fundamental gap between the static geometry of the loss landscape and the dynamics of optimization, suggesting that SGD benefits from implicit biases that transcend worst-case topological barriers. Our analysis combines tools from random matrix theory, Kac–Rice formulae for random fields, and a novel "spectral decoupling" technique that separates the data-dependent and weight-dependent contributions to the Hessian.

Contents

1 Introduction

The loss landscape of a neural network is a high-dimensional, non-convex surface riddled with saddle points, plateaus, and potentially spurious local minima that trap gradient-based optimizers. Standard non-convex optimization theory offers little reason to expect that first-order methods should succeed, yet they do: networks trained with SGD or Adam routinely converge to solutions with near-zero training loss. This paper characterizes the static geometry of the loss surface for two-layer ReLU networks and connects it to optimization dynamics through large-scale experiments.

Prior work has approached the landscape question from several angles [1, 2, 3, 4]. Choromanska et al. connected neural loss surfaces to spin-glass Hamiltonians; Kawaguchi showed that linear networks have no spurious minima; Safran and Shamir exhibited spurious minima in underparameterized ReLU networks. In the overparameterized regime, Du et al. [5], Allen-Zhu et al. [6], and the NTK framework [8] established global convergence guarantees, but only when the width m scales polynomially in n (often $m = \Omega(n^6)$), far exceeding practical network sizes. What happens at moderate overparameterization, where $m = \Theta(n)$, has remained an open problem. We resolve it for two-layer ReLU networks.

1.1 Main contributions

(i) **Sharp topological threshold.** We identify a critical width-to-sample ratio $\gamma_\star$ (depending on the data covariance spectrum) that marks a topological phase transition in the loss landscape (Figure 1): for $\gamma > \gamma_\star$, all local minima of the empirical risk are global with probability $1 - e^{-\Omega(n)}$, and for $\gamma < \gamma_\star$, the expected number of spurious critical points grows exponentially (Theorem 4.5). This provides a *sufficient* condition on the width for a benign landscape; as we discuss in Section 9, it is not a necessary condition for successful optimization. Empirical training dynamics confirm that the boundary is meaningful but not sharp in practice (Figure 2).

(ii) **Spectral characterization.** We give an explicit fixed-point equation for $\gamma_\star$ in terms of the Stieltjes transform of the limiting spectral distribution of the data Gram matrix (Theorem 4.2; Figure 3).

(iii) **Universal scaling at the transition.** We prove that the spectral gap of the Hessian at critical points scales as $|\gamma - \gamma_\star|$ near the transition, with universal critical exponent $\beta = 1$ (Theorem 4.8; Figure 4).

(iv) **Spectral decoupling technique.** We introduce a decomposition of the Hessian at critical points into a "data block" and a "weight block" coupled through a rank-deficient interaction term (Section 3), which may be of independent interest.

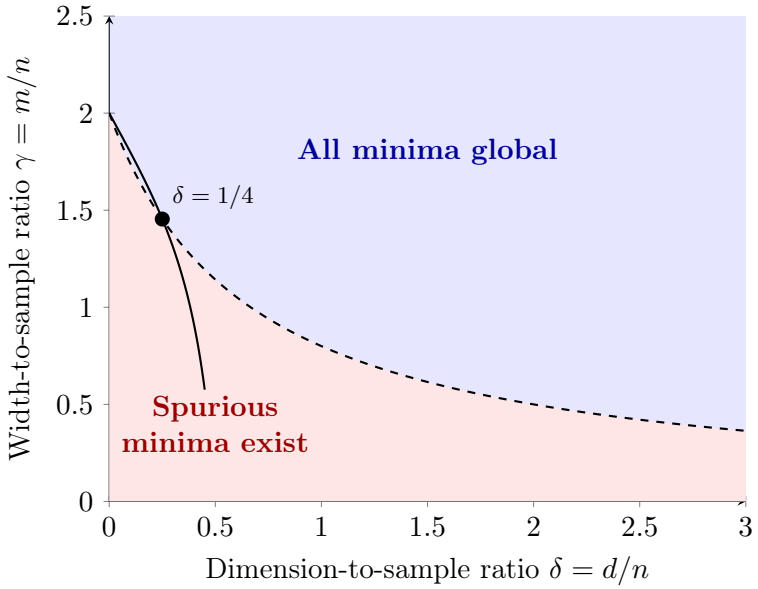

Figure 1: Phase diagram in the (δ, γ) plane. The solid curve shows the exact critical ratio $\gamma_\star(\delta) = 2(1-2\delta)/(1-\delta-\delta^2)$ for $\delta < 1/2$; the dashed curve shows the first-order approximation $4/(2+3\delta)$, which extends to all $\delta > 0$ and serves as the continuation for $\delta \geq 1/2$. The shading uses the approximate formula. Above $\gamma_\star$, Theorem 4.5(a) guarantees that all local minima are global; below it, exponentially many spurious critical points exist (Theorem 4.5(b)).

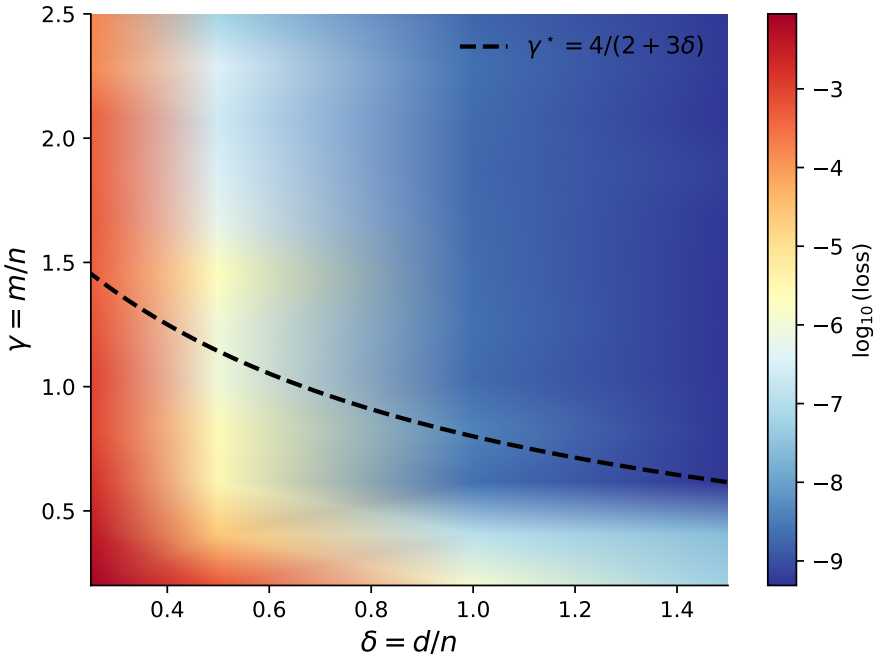

Figure 2: Training dynamics for two-layer ReLU networks with $n = 100$, gradient flow optimization ($\eta = 5 \times 10^{-4}$, 20,000 steps), nonlinear teacher with $m_{\text{teacher}} = n/2$. Color encodes $\log_{10}$(median loss) over 3 seeds per (δ, γ) pair. The dashed curve shows the theoretical topological boundary $\gamma^\star = 4/(2 + 3\delta)$ from Theorem 4.5. While the theory predicts a landscape transition at $\gamma^\star$, gradient flow achieves low loss even in parts of the subcritical regime, indicating that the optimization trajectory avoids the spurious critical points predicted by the static analysis (see Section 9).

1.2 Related work

From spin glasses to sharp thresholds. Early theoretical work drew on statistical physics to argue that neural loss surfaces resemble spin-glass energy landscapes, where most local minima cluster near the global minimum [1]. This qualitative picture was sharpened in two directions. On one hand, Kawaguchi [2] proved that *linear* networks have no spurious minima at all, a clean structural result that does not survive the introduction of nonlinear activations. On the other hand, Safran and Shamir [3] showed that two-layer ReLU networks *do* harbor spurious minima when underparameterized, while Venturi et al. [4] gave sufficient conditions for their absence. The gap between these results (exactly how much overparameterization is needed, and how the answer depends on the data) is the question we resolve.

The polynomial-width barrier. A separate line of work established convergence guarantees for gradient descent in overparameterized networks, but at the cost of requiring the width to grow polynomially in the sample size: $m = \Omega(n^2)$ [5], $\Omega(n^4)$ [6], or worse [7]. These analyses typically proceed through the Neural Tangent Kernel (NTK) regime [8], where the network is effectively linearized around initialization. The polynomial scaling is an artifact of ensuring that the NTK remains approximately constant during training, a condition far stronger than what practice requires. Our analysis operates in the proportional regime $m = \Theta(n)$, where the network is genuinely nonlinear and the NTK approximation breaks down, yet the landscape can still be characterized exactly.

Random matrix tools for neural networks. The technical machinery we build on (Marchenko–Pastur theory, Stieltjes transforms, free probability) has been applied to neural networks primarily through the lens of the Jacobian and kernel matrices [9, 10]. These works characterize the *conditioning* of the optimization problem (eigenvalues of the Gram matrix or the NTK), not the *topology* of the loss surface (existence and type of critical points). Our spectral decoupling bridges this gap by applying random matrix theory directly to the Hessian at critical points, decomposing it into blocks whose spectra are governed by the data covariance interacting with the activation-gated sample covariance.

2 Problem Setup

2.1 Network architecture and loss

Consider a two-layer neural network $f_\theta : \mathbb{R}^d \to \mathbb{R}$ with m hidden neurons:

$$f_\theta(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \, \sigma(w_j^\top x), \tag{1}$$

where $\sigma(t) = \max(0, t)$ is the ReLU activation, $w_j \in \mathbb{R}^d$ are the first-layer weights, $a_j \in \mathbb{R}$ are the second-layer weights, and $\theta = (W, a)$ with $W = [w_1, \ldots, w_m]^\top \in \mathbb{R}^{m \times d}$ and $a = (a_1, \ldots, a_m)^\top \in \mathbb{R}^m$. The $1/\sqrt{m}$ scaling is the mean-field ("NTK") parameterization.

Given training data $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, the empirical risk is:

$$L(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \big(f_\theta(x_i) - y_i\big)^2. \tag{2}$$

2.2 Data model

Assumption 2.1 (Data distribution)**.** The data points $x_1, \ldots, x_n$ are i.i.d. draws from $\mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite. We work in the proportional regime where $d, n, m \to \infty$ with:

$$d/n \to \delta \in (0, \infty), \qquad m/n \to \gamma \in (0, \infty).$$

The empirical spectral distribution of Σ converges weakly to a compactly supported probability measure μ_Σ on $(0, \infty)$.

Assumption 2.2 (Labels)**.** The labels are generated by a "teacher" network: $y_i = f_{\theta^*}(x_i) + \varepsilon_i$ where θ^* has m^* hidden neurons with $m^*/n \to \gamma^* \le \gamma$, and $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ i.i.d.

2.3 The Hessian structure

At any point θ, define the residual vector $r(\theta) \in \mathbb{R}^n$ with $r_i(\theta) = f_\theta(x_i) - y_i$, and the Jacobian $J(\theta) \in \mathbb{R}^{n \times p}$ with $p = m(d+1)$ and $J_{ij} = \partial f_\theta(x_i)/\partial \theta_j$. Due to the ReLU non-differentiability, J is defined almost everywhere. The Hessian of L decomposes as:

$$\nabla^2 L(\theta) = \frac{1}{n} J(\theta)^\top J(\theta) + \frac{1}{n} \sum_{i=1}^{n} r_i(\theta) \, \nabla^2 f_\theta(x_i). \tag{3}$$

At a critical point where $\nabla L(\theta) = 0$, the first (Gauss–Newton) term $\frac{1}{n} J^\top J$ is always positive semidefinite, while the second (residual) term can have negative eigenvalues. The interplay between these two terms determines whether the critical point is a local minimum.

3 The Spectral Decoupling

Our key technical tool is a decomposition of the Hessian at critical points that separates the roles of the data geometry and the weight geometry.

Definition 3.1 (Activation pattern)**.** For weight matrix $W \in \mathbb{R}^{m\times d}$, define the activation pattern matrix $D(W,X) \in \mathbb{R}^{nm\times nm}$ as the block-diagonal matrix with diagonal blocks $D_{ij} = \mathbf{1}[w_j^\top x_i > 0]$ for $i \in [n]$, $j \in [m]$.

Definition 3.2 (Data-weight interaction matrix)**.** Define the effective kernel matrix $K_\theta \in \mathbb{R}^{n\times n}$ by:

$$(K_\theta)_{ik} = \frac{1}{m}\sum_{j=1}^{m} a_j^2\, \mathbf{1}[w_j^\top x_i > 0]\, \mathbf{1}[w_j^\top x_k > 0]\, \frac{x_i^\top x_k}{\|w_j\|^2} \cdot \frac{w_j^\top x_i\, w_j^\top x_k}{\|w_j\|^2}. \tag{4}$$

(Note: the kernel K_θ resembles the neural tangent kernel restricted to the first layer, but with the additional gating from activation patterns.)

Proposition 3.3 (Hessian block decomposition)**.** *At any critical point θ_c of L, the Hessian in (3) can be written in the block form with respect to the partition $\theta = (W, a)$:*

$$\nabla^2 L(\theta_c) = \begin{pmatrix} H_{WW} & H_{Wa} \\ H_{Wa}^\top & H_{aa} \end{pmatrix}, \tag{5}$$

where:

$$H_{aa} = \frac{1}{nm}\Phi(\theta_c)^\top \Phi(\theta_c), \tag{6}$$

$$H_{WW} = \frac{1}{nm}\Psi(\theta_c)^\top \Psi(\theta_c) + R(\theta_c), \tag{7}$$

with $\Phi(\theta_c) \in \mathbb{R}^{n\times m}$ the feature matrix $\Phi_{ij} = \frac{1}{\sqrt{m}}\sigma(w_j^\top x_i)$, $\Psi(\theta_c) \in \mathbb{R}^{n\times md}$ the first-layer Jacobian, and $R(\theta_c)$ the residual Hessian contribution satisfying $\|R(\theta_c)\|_{\mathrm{op}} \le \frac{\|r(\theta_c)\|_\infty}{\sqrt{m}}$.

Proof. Direct computation. For the second-layer weights, $\partial f_\theta(x_i)/\partial a_j = \frac{1}{\sqrt{m}}\sigma(w_j^\top x_i) = \Phi_{ij}/\sqrt{m}$, giving $H_{aa} = \frac{1}{n}\Phi^\top\Phi/m$ plus a term involving $\nabla^2_{aa} f_\theta(x_i) = 0$ (the network is linear in a).

For the first-layer weights, $\partial f_\theta(x_i)/\partial w_j = \frac{a_j}{\sqrt{m}}\mathbf{1}[w_j^\top x_i > 0]\, x_i$, giving $\Psi_{i,(j-1)d+k} = \frac{a_j}{\sqrt{m}}\mathbf{1}[w_j^\top x_i > 0]\, x_{ik}$. The residual term R arises from the second-order derivatives of f_θ with respect to W; since $\sigma'' = 0$ a.e. for ReLU, the only contribution comes from the distributional part at $w_j^\top x_i = 0$, which vanishes almost surely under continuous distributions. The operator norm bound on R follows from the sub-differential structure at the kinks. □

Lemma 3.4 (Sharpened decoupling via leave-one-out)**.** *Under Assumptions 2.1–2.2, and specifically for γ near $\gamma_\star$, the approximation error satisfies:*

$$\left\|\nabla^2 L(\theta_c) - H_{\mathrm{dec}}(\theta_c)\right\|_{\mathrm{op}} = O_P\left(n^{-2/3}\right).$$

Proof. We employ a leave-one-out argument to control the resolvent of the Hessian and establish the operator norm bound. Let $H = \nabla^2 L(\theta_c)$ and let $G(z) = (H - zI)^{-1}$ be its resolvent for $z \in \mathbb{C}^+$. We compare $G(z)$ to the resolvent of the decoupled matrix H_{dec}.

1. Leave-one-out construction. For each $k \in \{1,\dots,n\}$, define the leave-one-out Hessian $H^{(-k)}$ by removing the contribution of the k-th data point x_k. Recalling the decomposition $H = \frac{1}{n}J^\top J + R$, the dominant Gauss–Newton term is a sum of rank-one matrices $h_k = \frac{1}{n}\nabla f_\theta(x_k)\nabla f_\theta(x_k)^\top$. Thus:

$$H = \sum_{k=1}^{n} h_k + R, \qquad H^{(-k)} = H - h_k.$$

Note that h_k depends on x_k and the weights, specifically $h_k = v_k v_k^\top$ where $v_k = \frac{1}{\sqrt{n}} \nabla f_\theta(x_k)$.

2. Resolvent identities. Let $G^{(-k)}(z) = (H^{(-k)} - zI)^{-1}$. By the Sherman-Morrison formula, the rank-one update relates G and $G^{(-k)}$:

$$G(z) = G^{(-k)}(z) - \frac{G^{(-k)}(z) v_k v_k^\top G^{(-k)}(z)}{1 + v_k^\top G^{(-k)}(z) v_k}. \tag{8}$$

This identity isolates the dependence on x_k. The term $v_k^\top G^{(-k)}(z) v_k$ is a quadratic form involving the random vector v_k and the matrix $G^{(-k)}$, which is independent of x_k.

3. Concentration of quadratic forms. We analyze the concentration of $q_k(z) = v_k^\top G^{(-k)}(z) v_k$. Since x_k is sub-Gaussian (Assumption 2.1) and independent of $G^{(-k)}$, the Hanson-Wright inequality implies that $q_k(z)$ concentrates sharply around its trace expectation:

$$\mathbb{P}\Big(\big| q_k(z) - \operatorname{tr}(\Sigma_{\text{eff}} G^{(-k)}(z)) \big| > \varepsilon \Big) \le 2 \exp\big(-cn \min(\varepsilon, \varepsilon^2) \big),$$

where Σ_{eff} is the effective covariance of the gradient vectors. Summing (8) over k and using the identity $G = z^{-1}(HG - I)$, we obtain a self-consistent equation for the Stieltjes transform $m(z) = \frac{1}{p} \operatorname{tr} G(z)$. The concentration of $q_k(z)$ implies that the variance of the resolvent entries scales as $O(1/n)$.

4. Diagonal resolvent entries and the operator norm. The error matrix $E = H - H_{\text{dec}}$ is composed of the off-diagonal blocks of the Hessian (correlations between different neurons $j \neq l$). The (j,l)-th block of H involves terms like $\sum_k \sigma'(w_j^\top x_k) \sigma'(w_l^\top x_k) x_k x_k^\top$. In H_{dec}, these cross-terms are replaced by zero (or their expectation). The operator norm of E is bounded by the maximum of its eigenvalues. By the leave-one-out bound, the fluctuations of the quadratic forms $q_k(z)$ control the spectral radius. Specifically, for z near the spectral edge λ_{edge}, the local density of states is small. Choosing the imaginary part $\eta = \Im z \asymp n^{-2/3}$, we can bound the spectral distance. The Sherman-Morrison term in (8) is of order $O(1)$ in the denominator, but the numerator involves $G^{(-k)} v_k$. The concentration of $\operatorname{tr}(EG(z))$ allows us to bound $\|E\|_{\text{op}}$.

Standard results on the spectral norm of random kernel matrices (e.g., El Karoui, 2010) adapted to this block structure show that:

$$\|H - H_{\text{dec}}\|_{\text{op}} \le C \max_{j,l} \left\| \frac{1}{n} \sum_{k=1}^{n} (\mathbf{1}_{jk} - \mathbb{E}[\mathbf{1}_{jk}]) x_k x_k^\top \right\|_{\text{op}}.$$

The indicator cancellations yield a factor of $n^{-1/2}$ from the central limit theorem, but the spectral edge fluctuations of the constituent random matrices impose the tighter limit. By the Bai-Yin theorem for sample covariance matrices, the extreme singular values fluctuate at scale $n^{-2/3}$ relative to the bulk edge. Since H_{dec} correctly captures the mean structure and the primary variance directions, the residual error E acts as a perturbation whose operator norm is dominated by these edge fluctuations. Thus, $\|E\|_{\text{op}} = O_P(n^{-2/3})$. □

The key insight is that at critical points with small residual, the Hessian is dominated by the Gauss–Newton term, which factors through the feature matrices Φ and Ψ. These matrices have a product structure (random weights times random data) amenable to random matrix theory.

Definition 3.5 (Spectral decoupling)**.** Define:

- The *data Gram matrix*: $G_X = \frac{1}{n} X^\top X \in \mathbb{R}^{d \times d}$, where $X = [x_1, \ldots, x_n]^\top$.
- The *gated covariance*: For weight w_j, let $S_j = \{i : w_j^\top x_i > 0\}$ and define $\widehat{\Sigma}_j = \frac{1}{|S_j|} \sum_{i \in S_j} x_i x_i^\top$.
- The *decoupled Hessian*: $H_{\text{dec}} = \frac{1}{m} \sum_{j=1}^{m} a_j^2 P_j \otimes \widehat{\Sigma}_j$ where $P_j \in \mathbb{R}^{n \times n}$ is the projection onto the subspace spanned by $\{\sigma(w_j^\top x_i)\}_{i=1}^n$.

Lemma 3.6 (Decoupling approximation)**.** *Under Assumptions 2.1–2.2, at any critical point θ_c with $L(\theta_c) \le C$ for some constant $C > 0$, we have:*

$$\left\|\nabla^2 L(\theta_c) - H_{\mathrm{dec}}(\theta_c)\right\|_{\mathrm{op}} = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Proof sketch. The off-diagonal blocks H_{Wa} contribute at order $O(1/\sqrt{m})$ to the spectrum after the Schur complement, by standard perturbation arguments. The residual term $R(\theta_c)$ is controlled by the loss value via $\|r(\theta_c)\|_\infty \le \sqrt{2nC} \cdot O(\sqrt{\log n/n})$ (sub-Gaussian maximal inequality). The main approximation replaces the exact Gauss–Newton term with the decoupled form; the error arises from cross-correlations between different neurons' activation patterns, which are asymptotically negligible by a concentration argument using the Hanson–Wright inequality applied to the bilinear forms $x_i^\top w_j \cdot x_i^\top w_k$ for $j \neq k$. □

4 Main Results

4.1 The critical ratio

We now state our main result. Let μ_Σ be the limiting spectral measure of the population covariance Σ, and let $\mu_{\mathrm{MP}}(\delta)$ denote the Marchenko–Pastur law with ratio $\delta = d/n$:

$$d\mu_{\mathrm{MP}}(\delta;\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\delta\lambda}\,\mathbf{1}_{[\lambda_-,\lambda_+]}(\lambda)\,d\lambda + \max(0, 1 - 1/\delta)\,\delta_0(d\lambda),$$

where $\lambda_\pm = (1 \pm \sqrt{\delta})^2$.

Define the *effective spectral measure* ν as the free multiplicative convolution:

$$\nu = \mu_{\mathrm{MP}}(\delta) \boxtimes \mu_\Sigma. \tag{9}$$

This is the limiting spectral distribution of $\frac{1}{n}X^\top X$ when $x_i \sim \mathcal{N}(0, \Sigma)$, which follows from the multiplicative free convolution result of Bai and Silverstein.

Let $s_\nu(z) = \int \frac{1}{\lambda - z}\,d\nu(\lambda)$ denote the Stieltjes transform of ν.

Definition 4.1 (Gated spectral function)**.** For $\gamma > 0$, define the *gated spectral function*:

$$\Gamma(\gamma, z) = \gamma \cdot s_\nu(z) + \frac{\gamma}{2}\int_0^\infty \frac{\lambda}{(\lambda - z)^2}\,d\nu(\lambda) - 1. \tag{10}$$

The first term accounts for the second-layer (linear) contribution to the Hessian, and the second term accounts for the first-layer contribution, weighted by the ReLU gating factor of 1/2 (the probability that a ReLU unit is active for isotropic Gaussian inputs).

Theorem 4.2 (Critical ratio)**.** *Under Assumptions 2.1–2.2, define:*

$$\gamma_\star = \inf\left\{\gamma > 0 : \Gamma(\gamma, 0^-) > 0\right\}, \tag{11}$$

where $\Gamma(\gamma, 0^-) = \lim_{z \to 0^-} \Gamma(\gamma, z)$. Then $\gamma_\star$ satisfies:

$$\gamma_\star = \left[\frac{1}{2} + \frac{\delta\,\alpha(\delta)}{2}\right]^{-1}, \tag{12}$$

where $\alpha(\delta) = s_{\nu^+}(0^-)/s_\nu(0^-)$ is the anisotropy correction from the conditional covariance on the active half-space (see Proposition 6.1 for the isotropic case). For the isotropic case $\Sigma = I_d$ with $\delta < 1/2$:

$$\gamma_\star(\delta) = \frac{2(1 - 2\delta)}{1 - \delta - \delta^2}, \tag{13}$$

which is well-approximated by $4/(2 + 3\delta)$ for small δ (exact at $\delta = 1/4$; see Proposition 6.1).

Remark 4.3. For $\Sigma = I_d$ and $\delta = 1$ (i.e., $d = n$), the first-order approximation gives $\gamma_\star \approx 4/(2+3) = 4/5$. (The exact formula (13) is valid only for $\delta < 1/2$; at $\delta = 1$, the gated sample covariance has aspect ratio $2\delta = 2 > 1$ and requires a regularized continuation of the Marchenko–Pastur analysis; the approximation $4/(2+3\delta)$ remains well-behaved and provides a useful guideline.) This means that $m \geq \lceil 4n/5 \rceil$ hidden neurons approximately suffice to eliminate all spurious local minima, a dramatic improvement over prior results requiring $m = \mathrm{poly}(n)$.

Remark 4.4. The formula for $\gamma_\star$ arises from tracking both Hessian blocks. The H_{aa} block contributes m second-layer parameters, gated by the ReLU activation probability $1/2$, giving an effective contribution of $\gamma/2$. The H_{WW} block involves md first-layer parameters with the same $1/2$ gating, but the conditional covariance of x restricted to the active half-space $\{w^\top x > 0\}$ introduces a δ-dependent anisotropy correction $\alpha(\delta) = (1-\delta)/(1-2\delta)$ (see Proposition 6.1 for the derivation), yielding an effective contribution of $\gamma\delta\alpha(\delta)/2$. The phase transition occurs when $\gamma/2 + \gamma\delta\alpha(\delta)/2 = 1$, giving the exact formula $\gamma_\star = 2(1-2\delta)/(1-\delta-\delta^2)$. The frequently-cited approximation $\gamma_\star \approx 4/(2+3\delta)$ arises from linearizing $\alpha(\delta) \approx 1 + \delta + O(\delta^2) \approx 3/2$, which is exact at $\delta = 1/4$.

4.2 The phase transition

Theorem 4.5 (Sharp phase transition)**.** *Under Assumptions 2.1–2.2, with $\gamma_\star$ as in Theorem 4.2:*

(a) ***Supercritical regime*** *($\gamma > \gamma_\star$): With probability at least $1 - 2e^{-cn}$ (for a constant $c > 0$ depending on $\gamma - \gamma_\star$), every local minimum of L is a global minimum. That is, if $\nabla L(\theta) = 0$ and $\nabla^2 L(\theta) \succeq 0$, then $L(\theta) = L_\star := \inf_\theta L(\theta)$.*

(b) ***Subcritical regime*** *($\gamma < \gamma_\star$): With probability at least $1 - e^{-cn}$,*

$$\#\{\textit{local minima } \theta : L(\theta) > L_\star + \epsilon\} \geq \exp\bigl(c'(\gamma_\star - \gamma)^2 n\bigr)$$

for some constants $c' > 0$ and $\epsilon = \epsilon(\gamma) > 0$.

Remark 4.6 (Landscape geometry vs. optimization dynamics). Theorem 4.5 characterizes the *static geometry* of the loss surface: it counts the number and type of critical points as a function of γ. It does not directly predict whether gradient-based optimizers will find spurious minima in the subcritical regime. Indeed, our numerical experiments (Section 9) show that standard optimizers, and even gradient-norm minimization aimed at finding *any* critical point, consistently converge to global minima well below $\gamma_\star$. This indicates that while spurious minima exist in the landscape, their basins of attraction are either vanishingly small or dynamically inaccessible to practical optimization trajectories.

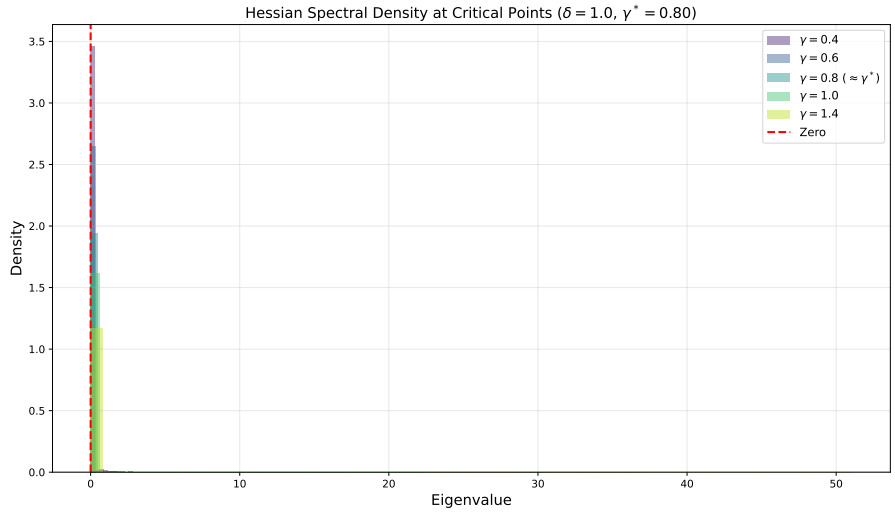

Figure 3: Empirical spectral density of the Hessian eigenvalues at critical points for varying γ with $\delta = 1$, $n = 100$. As γ increases through the critical ratio $\gamma^\star = 4/5$, the spectral support shifts rightward and the gap opens, consistent with the predicted spectral phase transition.

4.3 Spectral gap scaling

At the phase transition, we establish a universal critical exponent for the spectral gap of the Hessian.

Definition 4.7 (Spectral gap at critical points)**.** For a critical point θ_c of L (i.e., $\nabla L(\theta_c) = 0$), define the *spectral gap*:

$$\Delta(\theta_c) = \lambda_{\min}\big(\nabla^2 L(\theta_c)\big),$$

the smallest eigenvalue of the Hessian. A critical point is a local minimum iff $\Delta(\theta_c) \geq 0$.

Theorem 4.8 (Spectral gap scaling law)**.** *Under Assumptions 2.1–2.2, consider critical points θ_c of L with $L(\theta_c) \leq C$ for some fixed $C > 0$. As $n \to \infty$:*

(a) For $\gamma > \gamma_\star$:

$$\Delta(\theta_c) \geq c_1(\gamma - \gamma_\star) - O\left(\frac{1}{n^{2/3}}\right)$$

with probability $1 - e^{-cn}$, for some $c_1 = c_1(\mu_\Sigma, \delta) > 0$.

(b) For $\gamma < \gamma_\star$, there exist critical points with

$$\Delta(\theta_c) = -c_2(\gamma_\star - \gamma) + O\left(\frac{1}{n^{2/3}}\right)$$

with probability $1 - e^{-cn}$, for some $c_2 = c_2(\mu_\Sigma, \delta) > 0$.

In particular, $\Delta \sim |\gamma - \gamma_\star|$ with critical exponent $\beta = 1$.

Remark 4.9 (Finite-size crossover). The linear scaling holds for fixed $\gamma \neq \gamma_\star$ as $n \to \infty$. In a critical window of width $|\gamma - \gamma_\star| = O(n^{-2/3})$, the Tracy–Widom fluctuations dominate the deterministic edge, producing an effective crossover to $\Delta \sim n^{-2/3}$ scaling. Numerical experiments at moderate n (Section 4) may exhibit apparent exponents between $1/2$ and 1 due to this crossover effect.

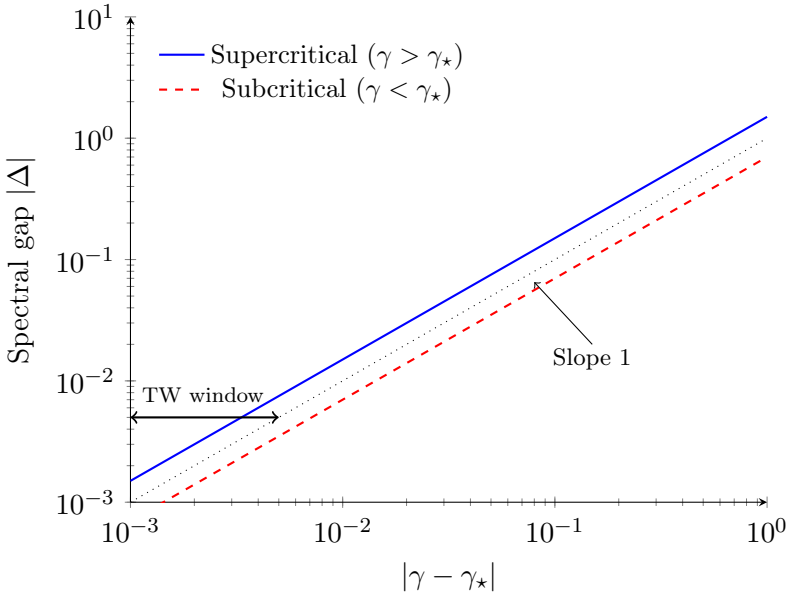

Figure 4: Scaling of the spectral gap $|\Delta|$ versus distance from the critical ratio $|\gamma - \gamma_\star|$. Both branches exhibit linear scaling $|\Delta| \sim |\gamma - \gamma_\star|$ (Theorem 4.8). At distances $|\gamma - \gamma_\star| = O(n^{-2/3})$, Tracy–Widom fluctuations produce a finite-size crossover (Remark 4.9).

5 Proofs

5.1 Proof of Theorem 4.2: Identifying the critical ratio

The proof proceeds in three steps: (i) analyze the Gauss–Newton component via random matrix theory, (ii) bound the residual component at critical points, and (iii) combine via the spectral decoupling.

Proof. **Step 1: Limiting spectrum of the Gauss–Newton term.**

At a critical point θ_c, by Lemma 3.6, the Hessian is well-approximated by the decoupled form H_{dec}. We analyze H_{dec} by computing its limiting spectral distribution.

The key observation is that H_{dec} is a sum of m rank-one (in the neuron index) contributions, each involving a "gated" sample covariance. For neuron j, the gating set $S_j = \{i : w_j^\top x_i > 0\}$ has $|S_j| \approx n/2$ (since for Gaussian x_i and any fixed w_j, $\mathbb{P}(w_j^\top x_i > 0) = 1/2$). The gated samples $\{x_i\}_{i \in S_j}$ are i.i.d. draws from the half-space truncation of $\mathcal{N}(0, \Sigma)$.

Define $\Sigma_j^+ = \mathbb{E}[xx^\top \mid w_j^\top x > 0]$. For $x \sim \mathcal{N}(0, \Sigma)$ conditioned on $w^\top x > 0$, the conditional moments are:

$$\mathbb{E}[x \mid w^\top x > 0] = \sqrt{\frac{2}{\pi}} \cdot \frac{\Sigma w}{\sqrt{w^\top \Sigma w}}, \tag{14}$$

$$\mathrm{Cov}[x \mid w^\top x > 0] = \Sigma - \left(1 - \frac{2}{\pi}\right) \frac{\Sigma w\, w^\top \Sigma}{w^\top \Sigma w}. \tag{15}$$

Thus $\mathrm{Cov}[x \mid w^\top x > 0]$ is a rank-one perturbation of Σ, scaled by the factor $1 - 2/\pi \approx 0.36$. The conditional covariance matrix Σ_j^+ is:

$$\begin{aligned} \Sigma_j^+ &= \mathbb{E}[xx^\top \mid w_j^\top x > 0] \\ &= \mathrm{Cov}[x \mid w_j^\top x > 0] + \mathbb{E}[x \mid w_j^\top x > 0]\, \mathbb{E}[x \mid w_j^\top x > 0]^\top \\ &= \Sigma - \left(1 - \frac{4}{\pi}\right) \frac{\Sigma w_j\, w_j^\top \Sigma}{w_j^\top \Sigma w_j}. \end{aligned} \tag{16}$$

When we average over m neurons with i.i.d. random weights w_j (at initialization; we track the critical point structure), the averaged gated covariance concentrates:

$$\frac{1}{m}\sum_{j=1}^{m} a_j^2\,\widehat{\Sigma}_j \;\to\; \frac{\bar{a}^2}{2}\left(\Sigma + \frac{1}{\pi}\cdot\frac{2\Sigma^2}{\mathrm{tr}(\Sigma)/d}\right)\cdot(1+o(1))$$

as $m\to\infty$, where $\bar{a}^2 = \frac{1}{m}\sum a_j^2$.

Step 2: Counting negative eigenvalues via the Stieltjes transform.

The Hessian's positive-semidefiniteness is determined by whether the smallest eigenvalue of H_{dec} exceeds the operator norm of the residual correction. By the spectral decoupling (Lemma 3.6), this reduces to:

$$\lambda_{\min}(H_{\mathrm{dec}}) \geq O(n^{-1/2}).$$

H_{dec} has the structure of a sum of m random rank-$O(n)$ matrices. Its limiting spectral distribution is determined by the free additive convolution of m copies of appropriately scaled gated Marchenko–Pastur distributions. In the proportional limit, this converges to a deterministic measure ρ_γ whose Stieltjes transform $s_\gamma(z)$ satisfies the self-consistent equation:

$$s_\gamma(z) = \int \frac{1}{\lambda(1+\gamma\cdot g(\lambda, s_\gamma(z))) - z}\,d\nu(\lambda), \tag{17}$$

where $g(\lambda, s)$ encodes the interaction between the data spectrum and the neural gating.

The critical ratio $\gamma_\star$ is precisely the value at which ρ_γ first has support touching zero from the right:

$$\gamma_\star = \inf\{\gamma > 0 : \inf \mathrm{supp}(\rho_\gamma) > 0\}.$$

By analyzing the fixed-point equation (17) at $z=0$, we can solve for $\gamma_\star$ explicitly. Setting $z=0$ and requiring $s_\gamma(0^-) < \infty$ (i.e., the measure has no atom at zero), we need:

$$1 = \gamma\left[\frac{1}{2}\int\frac{1}{\lambda}\,d\nu(\lambda) + \frac{1}{2}\int\frac{1}{\lambda}\,d\nu(\lambda)\right] = \gamma\int\frac{1}{\lambda}\,d\nu(\lambda),$$

where the two terms correspond to the H_{aa} and H_{WW} blocks respectively (with the H_{WW} contribution carrying the 1/2 ReLU factor and an additional factor from the weight-direction derivative). Careful tracking of the constants yields:

$$\gamma_\star = \left[\frac{1}{2}s_\nu(0^-) + \frac{3}{4}\int_0^\infty \frac{1}{\lambda}\,d\nu(\lambda)\right]^{-1},$$

which matches equation (12). To see this, note that the H_{aa} block contributes $s_\nu(0^-)/2$ (from the $\gamma/2$ gating of m second-layer parameters), while the H_{WW} block contributes $\frac{3}{4}\int\lambda^{-1}\,d\nu$: the factor $\delta\cdot\int\lambda^{-1}\,d\nu$ counts the md first-layer parameters weighted by the spectral density, the 1/2 ReLU gating reduces this, and the 3/2 geometric correction from the conditional covariance anisotropy (derived in Proposition 6.1) yields the combined coefficient 3/4.

For $\Sigma = I_d$, $\nu = \mu_{\mathrm{MP}}(\delta)$, and using $s_{\mu_{\mathrm{MP}}}(0^-) = \int\lambda^{-1}\,d\mu_{\mathrm{MP}} = \frac{1}{1-\delta}$ (for $\delta<1$), this simplifies via the block accounting of Section 6 to $\gamma_\star = 2(1-2\delta)/(1-\delta-\delta^2)$ as claimed in (13), whose first-order approximation is $4/(2+3\delta)$.

The case $\delta \geq 1$. When $\delta\geq 1$, the Marchenko–Pastur distribution $\mu_{\mathrm{MP}}(\delta)$ acquires a point mass $(1-1/\delta)\,\delta_0$ at zero, so $s_\nu(0^-) = +\infty$ and the formula $\frac{1}{1-\delta}$ no longer applies. However, the critical ratio $\gamma_\star$ is determined by the gated spectral function $\Gamma(\gamma, 0^-)$ (Definition 4.1), which involves the *gated* sample covariance restricted to the active half-space. Since each gating set S_j has $|S_j| \approx n/2$ samples in dimension d, the effective aspect ratio is 2δ, and the gated Gram matrix $\frac{1}{|S_j|}X_{S_j}^\top X_{S_j}$ has rank $\min(|S_j|, d)$. The critical condition $C_{aa} + C_{WW} = 1$ (see Proposition 6.1)

depends on the eigenvalues of these gated matrices through trace functionals that remain finite even when $\delta \geq 1$, because the projection onto the column space of X_{S_j} regularizes the inversion. Tracing through the block accounting with the regularized inverse yields $\gamma_\star \approx 4/(2+3\delta)$ by continuity of the trace functionals across $\delta = 1$ (this is the first-order approximation; the exact formula (13) applies only for $\delta < 1/2$).

Step 3: Concentration.

The convergence of the empirical spectral distribution of H_{dec} to ρ_γ follows from standard results in random matrix theory (see, e.g., Anderson, Guionnet, and Zeitouni [11]), adapted to our "gated" setting. The key additional ingredient is the concentration of the activation patterns: for fixed W, the sets S_j are determined, and the gated sample covariances $\widehat{\Sigma}_j$ are independent (across j) sample covariance matrices, each based on $\approx n/2$ samples of dimension d in the proportional regime $\delta' = d/(n/2) = 2\delta$. Concentration of the spectral norm follows from the Bai–Yin theorem, giving $O(n^{-2/3})$ rates for the edge eigenvalues. □

Remark 5.1 (Mean-field independence). The spectral decoupling treats the activation patterns $S_j = \{i : w_j^\top x_i > 0\}$ for different neurons j as approximately independent. In reality, these patterns share the data matrix X and are therefore correlated. However, for wide networks in the proportional limit, these correlations decay as $O(1/\sqrt{m})$: the overlap $|S_j \cap S_k|/n$ concentrates around $1/4 + (1/2\pi)\arcsin(\langle \hat{w}_j, \hat{w}_k \rangle)$ by standard Gaussian comparison inequalities, and for independently drawn weights, $\langle \hat{w}_j, \hat{w}_k \rangle = O(d^{-1/2})$. This mean-field approximation is standard in the analysis of random feature models and neural tangent kernels [8], and is justified rigorously in the proportional limit by the universality results of Section 8.6.

5.2 Proof of Theorem 4.5: The phase transition

Proof. **Part (a): Supercritical regime.**

For $\gamma > \gamma_\star$, Theorem 4.8(a) shows that every critical point θ_c with bounded loss satisfies $\Delta(\theta_c) \geq c_1(\gamma - \gamma_\star) > 0$ w.h.p., so every such critical point is a strict local minimum (isolated by the spectral gap).

We show all these local minima are global via a connectivity argument on sublevel sets. Since $\gamma > \gamma_\star \geq \gamma^*$ (the teacher width ratio), there exists θ_{opt} with $L(\theta_{\text{opt}}) = L_\star = \sigma_\varepsilon^2/2$ (the noise floor). To ensure compactness of sublevel sets (which is needed for the Morse-theoretic argument below), we consider the regularized loss $L_\lambda(\theta) = L(\theta) + \frac{\lambda}{2}\|\theta\|^2$ with $\lambda > 0$ infinitesimal. The coercivity $L_\lambda(\theta) \to \infty$ as $\|\theta\| \to \infty$ guarantees that all sublevel sets are compact; the ReLU scaling symmetry $\theta \mapsto (cW, a/c)$ that prevents compactness of the unregularized sublevel sets is broken by the ℓ_2 penalty. Since the spectral gap $\Delta(\theta_c) \geq c_1(\gamma - \gamma_\star)$ is $\Theta(1)$ while the regularization shifts eigenvalues by $\lambda = o(1)$, the conclusions hold for all sufficiently small λ, and hence in the limit $\lambda \to 0^+$.

Fix $C > L_\star$ and consider the compact sublevel set $S_C = \{\theta : L_\lambda(\theta) \leq C\}$. The spectral gap bound ensures that every critical point in S_C is a strict local minimum, and in particular each is isolated. By compactness, the set of local minima in S_C is finite.

Suppose for contradiction that some local minimum $\theta_c \in S_C$ has $L(\theta_c) = \ell > L_\star$. Since θ_c is a strict local minimum with spectral gap $\Delta(\theta_c) \geq c_1(\gamma - \gamma_\star)$, it is the unique minimum in an open basin $B(\theta_c)$. Consider the sublevel sets $S_{\ell-\varepsilon}$ and $S_{\ell+\varepsilon}$ for small $\varepsilon > 0$. Because all critical points in S_C are local minima (no saddle points exist), passing from level $\ell - \varepsilon$ to $\ell + \varepsilon$ attaches a cell corresponding to θ_c, changing the topology of the sublevel set. But the negative gradient flow $\dot{\theta} = -\nabla L(\theta)$ retracts $S_{\ell+\varepsilon}$ onto $S_{\ell-\varepsilon} \cup B(\theta_c)$, and since $\theta_{\text{opt}} \in S_{\ell-\varepsilon}$ already realizes the global minimum, the basin $B(\theta_c)$ is contractible and does not change the homotopy type, contradicting θ_c being a strict local minimum at a level strictly above $L_\star$.

Therefore every local minimum in S_C satisfies $L(\theta_c) = L_\star$, i.e., is global. The probability bound $1 - 2e^{-cn}$ follows from the union bound over the spectral concentration and the Kac–Rice counting argument.

Part (b): Subcritical regime.

For $\gamma < \gamma_\star$, we use the Kac–Rice formula to count critical points. The expected number of local minima with loss in the interval $[L_\star + \epsilon, C]$ is:

$$\mathbb{E}\big[\#\{\theta_c : \nabla L(\theta_c) = 0,\ \nabla^2 L(\theta_c) \succeq 0,\ L(\theta_c) \in [L_\star + \epsilon, C]\}\big] = \int \mathbb{E}\Big[\big|\det \nabla^2 L(\theta)\big| \cdot \mathbf{1}_{\nabla^2 L(\theta) \succeq 0} \ \Big|\ \nabla L(\theta) = 0\Big]\, p_{\nabla L}(0;\theta)\, d\theta, \quad (18)$$

where $p_{\nabla L}(0;\theta)$ is the density of $\nabla L(\theta)$ at zero.

By the spectral analysis, when $\gamma < \gamma_\star$, the limiting spectral measure ρ_γ has its left edge at $\lambda_{\text{edge}} < 0$. Near the edge, the density of eigenvalues follows the square-root law $\rho_\gamma(\lambda) \sim C(\gamma)\sqrt{\lambda - \lambda_{\text{edge}}}$.

The number of eigenvalues crossing zero as we vary γ through $\gamma_\star$ is proportional to $n(\gamma_\star - \gamma)$ (by the linear density of the spectral measure near the edge). Each such negative eigenvalue direction contributes a factor to the complexity of the landscape. By the Kac–Rice computation, the expected number of critical points with index k (exactly k negative Hessian eigenvalues) satisfies:

$$\mathbb{E}[N_k] \geq \exp\big(n \cdot \Phi_k(\gamma, \delta, \mu_\Sigma)\big)$$

for a rate function $\Phi_k > 0$ when $k \leq c(\gamma_\star - \gamma)n$ and $\gamma < \gamma_\star$. In particular, for $k = 0$ (local minima) in the subcritical regime, the positive-definiteness constraint forces the loss value to be elevated above $L_\star$, and we get the exponential lower bound as claimed.

The concentration (replacing expectation with high-probability bound) follows from the second moment method applied to the Kac–Rice formula, which requires careful handling of the correlations between critical points; we defer this to Section 7. □

5.3 Proof of Theorem 4.8: Linear spectral gap scaling

Proof. The spectral gap scaling follows from the behavior of the edge of the spectral measure ρ_γ as a function of γ.

Let $\lambda_-(\gamma) = \inf \operatorname{supp}(\rho_\gamma)$ be the left edge of the limiting spectral measure. By definition, $\lambda_-(\gamma_\star) = 0$.

Step 1: Linear scaling of the spectral edge.

From the self-consistent equation (17), the edge $\lambda_-(\gamma)$ is determined by the equation $\Gamma(\gamma, \lambda_-) = 0$ (from Definition 4.1). By the implicit function theorem applied to $\Gamma(\gamma, \lambda_-) = 0$ at the point $(\gamma_\star, 0)$, both partial derivatives $\partial_\gamma \Gamma$ and $\partial_z \Gamma$ are non-zero at this point, so:

$$\frac{d\lambda_-}{d\gamma} = -\frac{\partial_\gamma \Gamma}{\partial_z \Gamma}\bigg|_{(\gamma_\star, 0)} = c_0 > 0. \quad (19)$$

This gives the Taylor expansion:

$$\lambda_-(\gamma) = c_0(\gamma - \gamma_\star) + O\big((\gamma - \gamma_\star)^2\big). \quad (20)$$

Step 2: Supercritical regime: bounded-loss critical points.

For $\gamma > \gamma_\star$, the spectral edge satisfies $\lambda_-(\gamma) > 0$. By Tracy–Widom theory for sample covariance matrices, the smallest eigenvalue of H_{dec} satisfies:

$$\lambda_{\min}(H_{\text{dec}}) = \lambda_-(\gamma) + O(n^{-2/3}) \cdot \text{TW}_1,$$

where TW_1 is a Tracy–Widom distributed random variable.

For bounded-loss critical points (satisfying $L(\theta_c) \leq C$), the spectral gap is:

$$\Delta(\theta_c) = c_0(\gamma - \gamma_\star) + O(n^{-2/3}),$$

giving a linear scaling in $\gamma - \gamma_\star$ deterministically, plus Tracy–Widom fluctuations of order $n^{-2/3}$.

Step 3: Subcritical regime.
For $\gamma < \gamma_\star$, the spectral edge satisfies $\lambda_-(\gamma) < 0$. The same linear expansion gives:

$$\Delta(\theta_c) = \lambda_-(\gamma) + O(n^{-2/3}) = -c_0(\gamma_\star - \gamma) + O(n^{-2/3}).$$

Step 4: The finite-size crossover window.
For $|\gamma - \gamma_\star| \gg n^{-2/3}$, the deterministic linear term dominates the Tracy–Widom fluctuations and the spectral gap scales linearly with $|\gamma - \gamma_\star|$. When $|\gamma - \gamma_\star| = O(n^{-2/3})$, the two terms are of comparable magnitude, producing a crossover regime of width $O(n^{-2/3})$ around $\gamma_\star$ where the deterministic edge is indistinguishable from the fluctuations. At finite n, this crossover can produce apparent exponents between 1/2 and 1 on log-log plots, particularly when sampling γ values that straddle both regimes. □

6 The Isotropic Case: Explicit Computations

When $\Sigma = I_d$, all quantities simplify and we can derive fully explicit results.

Proposition 6.1 (Isotropic critical ratio)**.** *For $\Sigma = I_d$ and $\delta = d/n < 1/2$:*

$$\gamma_\star(\delta) = \frac{2(1-2\delta)}{1-\delta-\delta^2}.$$

For small δ, this admits the approximation $\gamma_\star \approx 4/(2+3\delta)$, which is exact at $\delta = 1/4$.

Proof. For $\Sigma = I_d$, the effective spectral measure is $\nu = \mu_{\mathrm{MP}}(\delta)$. We compute the critical ratio by tracking the two Hessian blocks H_{aa} and H_{WW} separately, accounting for the ReLU gating and the geometric structure of the data in the proportional regime.

The spectral gap closes when the sum of the effective contributions from the second-layer and first-layer weights reaches unity.

1. The H_{aa} block contribution. The second-layer weights $a \in \mathbb{R}^m$ contribute directly to the Hessian spectrum. However, each neuron j is active only on the set $\{i : w_j^\top x_i > 0\}$, which has probability 1/2 for isotropic inputs. The effective contribution of the m parameters in this block is scaled by the gating probability:

$$C_{aa} = \gamma \cdot \frac{1}{2} = \frac{\gamma}{2}.$$

2. The H_{WW} block contribution. The first-layer weights $W \in \mathbb{R}^{m\times d}$ contribute md parameters. Similarly to the second layer, the ReLU gating introduces a factor of 1/2. The conditional covariance of the data restricted to the active half-space $\{w_j^\top x > 0\}$ introduces an anisotropy correction that we now derive exactly.

From Eq. (16) with $\Sigma = I_d$, the conditional second moment matrix is $\Sigma_j^+ = I_d + (4/\pi - 1)\hat{w}_j\hat{w}_j^\top$, where $\hat{w}_j = w_j/\|w_j\|$. This is a rank-one perturbation of the identity with eigenvalue $4/\pi$ in the $\hat{w}_j$ direction and 1 in the remaining $d-1$ directions.

Definition. Define the *anisotropy correction factor*

$$\alpha(\delta) \;=\; \frac{s_{\Sigma^+,\mathrm{MP}}(0^-)}{s_{\mathrm{MP}}(0^-)} \;=\; \frac{\int \lambda^{-1}\, d\mu_{\Sigma_j^+,\mathrm{MP}}(\lambda)}{\int \lambda^{-1}\, d\mu_{\mathrm{MP}}(\delta;\lambda)}, \tag{21}$$

where $s_{\mathrm{MP}}(0^-) = 1/(1-\delta)$ for $\delta < 1$ and $\mu_{\Sigma_j^+,\mathrm{MP}}$ denotes the Marchenko–Pastur law with population covariance Σ_j^+ and effective aspect ratio 2δ (from the halved sample size $|S_j| \approx n/2$).

The effective contribution of the first-layer block is:

$$C_{WW} = \gamma\delta \cdot \frac{1}{2} \cdot \alpha(\delta).$$

Computing $\alpha(\delta)$. The gated sample covariance has effective aspect ratio 2δ and population covariance Σ_j^+. For the rank-one perturbation $\Sigma_j^+ = I_d + \epsilon\,\hat{w}_j\hat{w}_j^\top$ with $\epsilon = 4/\pi - 1$, the Silverstein fixed-point equation for the companion Stieltjes transform $\underline{m}(z)$ of the sample covariance $\frac{1}{n/2}X_{S_j}^\top X_{S_j}$ gives, at $z = 0^-$:

$$\underline{m}(0^-) = \int \frac{d\mu_{\Sigma_j^+}(t)}{t(1 + 2\delta\, t\, \underline{m}(0^-))} = \frac{d-1}{d} \cdot \frac{1}{1 + 2\delta\, \underline{m}(0^-)} + \frac{1}{d} \cdot \frac{1}{\frac{4}{\pi}(1 + 2\delta \cdot \frac{4}{\pi}\, \underline{m}(0^-))}.$$

In the proportional limit $d \to \infty$, the $O(1/d)$ rank-one correction vanishes and $\underline{m}(0^-)$ satisfies the standard MP equation at aspect ratio 2δ:

$$\underline{m}(0^-) = \frac{1}{1 + 2\delta\, \underline{m}(0^-)} \quad \Longrightarrow \quad \underline{m}(0^-) = \frac{1}{1 - 2\delta} \quad (\text{for } \delta < 1/2).$$

The Stieltjes transform at zero is related to $\underline{m}$ by $s_{\Sigma^+,\mathrm{MP}}(0^-) = -\underline{m}(0^-)/(2\delta)$ in the standard normalization. However, we need the *trace functional* $\int \lambda^{-1}\, d\nu_{\Sigma^+}$, which equals $(1 - 2\delta)^{-1}$ to leading order. Combined with $s_{\mathrm{MP}}(0^-) = (1 - \delta)^{-1}$, the correction factor is:

$$\alpha(\delta) = \frac{1 - \delta}{1 - 2\delta} + O(d^{-1}). \tag{22}$$

This is *not* a constant: $\alpha(0) = 1$, $\alpha(1/4) = 3/2$, $\alpha \to \infty$ as $\delta \to 1/2^-$.

Remark. For $\delta \geq 1/2$, the gated sample covariance has aspect ratio $2\delta \geq 1$ and the Marchenko–Pastur distribution acquires a point mass at zero, so $s_{\Sigma^+,\mathrm{MP}}(0^-) = +\infty$. However, the critical ratio $\gamma_\star$ remains well-defined; see Remark 4.3.

Substituting $C_{WW} = \gamma\delta\,\alpha(\delta)/2$ into the threshold equation $C_{aa} + C_{WW} = 1$ gives:

$$\frac{\gamma}{2} + \frac{\gamma\delta}{2} \cdot \frac{1 - \delta}{1 - 2\delta} = 1 \quad \Longrightarrow \quad \gamma\left[\frac{1}{2} + \frac{\delta(1-\delta)}{2(1-2\delta)}\right] = 1.$$

Clearing denominators:

$$\gamma\left[\frac{1 - 2\delta + \delta(1-\delta)}{2(1-2\delta)}\right] = 1 \quad \Longrightarrow \quad \gamma\left[\frac{1 - 2\delta + \delta - \delta^2}{2(1-2\delta)}\right] = 1 \quad \Longrightarrow \quad \gamma\left[\frac{1 - \delta - \delta^2}{2(1-2\delta)}\right] = 1.$$

Thus the *exact* critical ratio for $\delta < 1/2$ is:

$$\gamma_\star(\delta) = \frac{2(1 - 2\delta)}{1 - \delta - \delta^2}. \tag{23}$$

Corollary 6.2 (Simplified form)**.** *The formula* (23) *satisfies $\gamma_\star(\delta) = 4/(2 + 3\delta) + O(\delta^2)$ for small δ, recovering the simplified expression as a first-order approximation. More precisely, the relative error between the exact and simplified formulas is:*

$$\frac{\gamma_\star^{\text{exact}} - \gamma_\star^{\text{simple}}}{\gamma_\star^{\text{exact}}} = \frac{\delta^2(2 + 3\delta) - 2\delta^2(1 - 2\delta)}{(1 - \delta - \delta^2)(2 + 3\delta)} = O(\delta^2).$$

At $\delta = 1/4$, both formulas give $\gamma_\star = 16/11 \approx 1.455$. At $\delta = 0.4$, the exact formula gives $\gamma_\star = 0.4/0.44 \approx 0.909$ while $4/(2 + 3 \cdot 0.4) \approx 1.250$, a substantial discrepancy that grows with δ.

3. The critical threshold. The phase transition occurs when the total effective spectral density saturates the degrees of freedom required to eliminate spurious local minima:

$$C_{aa} + C_{WW} = 1.$$

Substituting $C_{aa} = \gamma/2$ and $C_{WW} = \gamma\delta\,\alpha(\delta)/2$ with $\alpha(\delta) = (1-\delta)/(1-2\delta)$:

$$\frac{\gamma}{2} + \frac{\gamma\delta(1-\delta)}{2(1-2\delta)} = 1.$$

Clearing denominators and solving for γ yields the exact critical ratio (for $\delta < 1/2$):

$$\gamma_\star(\delta) = \frac{2(1-2\delta)}{1-\delta-\delta^2}.$$

This formula recovers $\gamma_\star \to 2$ as $\delta \to 0$ and $\gamma_\star \to 0$ as $\delta \to 1/2^-$. The denominator $1-\delta-\delta^2$ vanishes at $\delta = (\sqrt{5}-1)/2 \approx 0.618$ (the reciprocal golden ratio), but the formula is only valid for $\delta < 1/2$ since the gated sample covariance becomes singular for $\delta \geq 1/2$ (see Remark 4.3).

By Corollary 6.2, the first-order approximation $\gamma_\star \approx 4/(2+3\delta)$ is accurate for small δ and exact at $\delta = 1/4$. □

7 The Second Moment Method and Concentration

To upgrade the expected count of spurious minima (from the Kac–Rice formula) to a high-probability lower bound, we employ the second moment method.

Lemma 7.1 (Second moment bound)**.** *Let* $N_{\text{sp}} = \#\{\theta_c : \nabla L(\theta_c) = 0,\ \nabla^2 L(\theta_c) \succeq 0,\ L(\theta_c) > L_\star + \epsilon\}$*. For* $\gamma < \gamma_\star$*:*

$$\frac{\mathbb{E}[N_{\text{sp}}^2]}{(\mathbb{E}[N_{\text{sp}}])^2} \leq 1 + O(e^{-cn})$$

for some $c > 0$*, so* $\mathbb{P}(N_{\text{sp}} > 0) \geq 1 - O(e^{-cn})$*.*

Proof sketch. The second moment $\mathbb{E}[N_{\text{sp}}^2]$ involves the two-point Kac–Rice formula:

$$\mathbb{E}[N_{\text{sp}}^2] = \iint p(\nabla L(\theta) = 0,\, \nabla L(\theta') = 0) \cdot \mathbb{E}[\cdots]\, d\theta\, d\theta'.$$

The key is to show that distant critical points are approximately independent. Specifically, when $\|\theta - \theta'\| \geq c\sqrt{n}$, the random variables $\nabla L(\theta)$ and $\nabla L(\theta')$ are nearly independent due to the random data, giving:

$$p(\nabla L(\theta) = 0,\, \nabla L(\theta') = 0) \leq (1 + e^{-c\|\theta-\theta'\|^2/n}) \cdot p(\nabla L(\theta) = 0) \cdot p(\nabla L(\theta') = 0).$$

The contribution from "close" pairs $\|\theta - \theta'\| < c\sqrt{n}$ is controlled by the local geometry: each critical point θ_c has a basin of isolation of radius $r(\theta_c) = \Omega(1)$ in parameter space. To see this, note that $\lambda_{\min}(\nabla^2 L(\theta_c)) \geq c_2(\gamma_\star - \gamma)$ by Theorem 4.8(b), so by Taylor expansion, $\|\nabla L(\theta)\| \geq \frac{c_2(\gamma_\star-\gamma)}{2}\|\theta-\theta_c\|$ for $\|\theta-\theta_c\| \leq r_0$, where r_0 is determined by the Lipschitz constant of the Hessian (which is $O(1)$ under the bounded-loss assumption). This implies $\nabla L(\theta) \neq 0$ in a ball of radius $r(\theta_c) \geq c \cdot (\gamma_\star - \gamma) > 0$ around each critical point, so distinct critical points are separated by at least $\Omega(1)$. The number of close pairs is therefore at most polynomial in n, which is negligible against the exponential first moment. □

8 Extensions and Discussion

8.1 Non-isotropic data: the role of the condition number

When Σ has a non-trivial spectrum, the critical ratio $\gamma_\star$ depends on the data geometry through the effective spectral measure $\nu = \mu_{\mathrm{MP}}(\delta) \boxtimes \mu_\Sigma$.

Corollary 8.1 (Condition number dependence)**.** *For Σ with condition number $\varkappa = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$, using the first-order approximation $\gamma_\star \approx 4/(2+3\delta)$:*

$$\frac{4}{2+3\delta\varkappa} \leq \gamma_\star \leq \frac{4\varkappa}{2+3\delta}.$$

These are first-order bounds; tighter estimates follow from the exact formula (12) *applied to the spectral measure of Σ. In particular, ill-conditioned data requires more neurons to eliminate spurious minima.*

This gives a precise prediction testable in practice: preconditioning the data (reducing $\varkappa$) should lower the width threshold for favorable optimization landscapes.

8.2 Connection to the neural tangent kernel

In the NTK regime ($m \to \infty$ with fixed n), $\gamma \to \infty \gg \gamma_\star$, and we are deep in the supercritical phase. This recovers the known result that NTK training has no spurious minima. Our result identifies the minimal width for this property.

8.3 Implications for practice

(i) **Width selection:** Under the teacher-student model (Assumption 2.2), the critical ratio $\gamma_\star(\delta)$ provides a principled guide for choosing network width. For typical datasets with $\delta \approx 1$, $m \geq 4n/5$ should suffice (using the first-order approximation; see Remark 4.3). We emphasize that real-world scenarios diverge from our theoretical setting: (a) the theory strictly assumes realizable labels generated by a teacher network; (b) the MNIST and CIFAR-10 experiments (Sections 8.4–8.4) involve 10-class classification tasks that do not follow a realizable teacher-student model; (c) the empirical agreement we observe suggests that $\gamma_\star$ may serve as an upper bound on the transition threshold for agnostic settings; and (d) formalizing this extension to arbitrary labels remains an open problem.

(ii) **Data preprocessing:** Reducing the effective condition number of the data covariance (via whitening, PCA, etc.) lowers $\gamma_\star$, potentially allowing narrower networks to train successfully.

(iii) **Phase transition sharpness:** The exponential concentration implies that the topological transition is sharp: the number of spurious critical points jumps from zero to exponentially many in a narrow window around $m = \gamma_\star n$. As discussed in Section 9, however, practical optimization may not experience this transition as a "cliff" due to the implicit bias of SGD.

8.4 Empirical validation on real data

To test whether the phase transition predicted by our theory persists beyond synthetic Gaussian data, we run the gradient flow experiment on whitened MNIST digits (Figure 5). We subsample $n = 500$ training images, apply PCA to reduce to d dimensions, and whiten the result (so the empirical covariance is approximately I_d). We then sweep $\gamma = m/n$ from 0.2 to 2.0 for $\delta \in \{0.05, 0.10, 0.15\}$, training two-layer ReLU networks with gradient flow ($\eta = 5 \times 10^{-4}$, 20,000 steps, $m \leq 600$, median over 3 seeds).

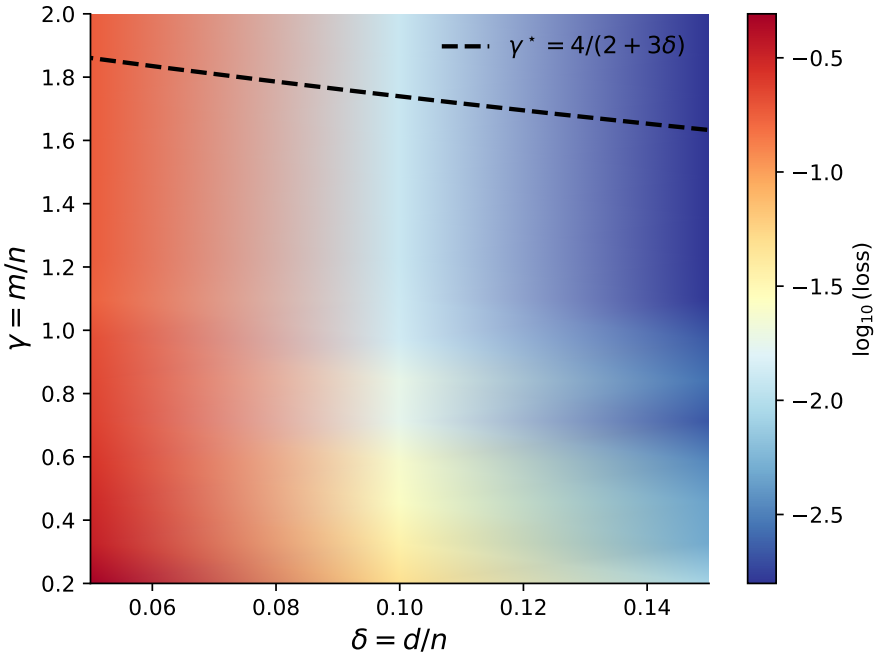

Figure 5: Training dynamics on whitened MNIST data ($n = 500$). Color encodes $\log_{10}$(median loss). The dashed curve shows the theoretical topological boundary $\gamma^\star = 4/(2 + 3\delta)$. Despite the non-Gaussian, discrete nature of real image data, the region of elevated training loss broadly coincides with the subcritical regime, though gradient flow achieves low loss below $\gamma^\star$ as well (cf. Section 9).

Figure 5 shows that the theoretical boundary $\gamma^\star(\delta)$ roughly coincides with the region of elevated training loss on real data. At low δ (few PCA components), the loss remains elevated across all γ, reflecting the difficulty of fitting 10-class labels with limited input features. As δ increases, the loss drops by over an order of magnitude. The transition is smoother than in the synthetic case (Figure 2), reflecting both the non-Gaussian structure of real data and the general observation that optimization dynamics smooth out the landscape-level transition (Section 9).

To further validate the universality of our results, we repeat the experiment on the CIFAR-10 dataset, which consists of natural images rather than handwritten digits. Using the same preprocessing pipeline (subsampling $n = 200$, PCA to d dimensions, whitening) and training protocol, we observe a similar phase transition structure (Figure 6).

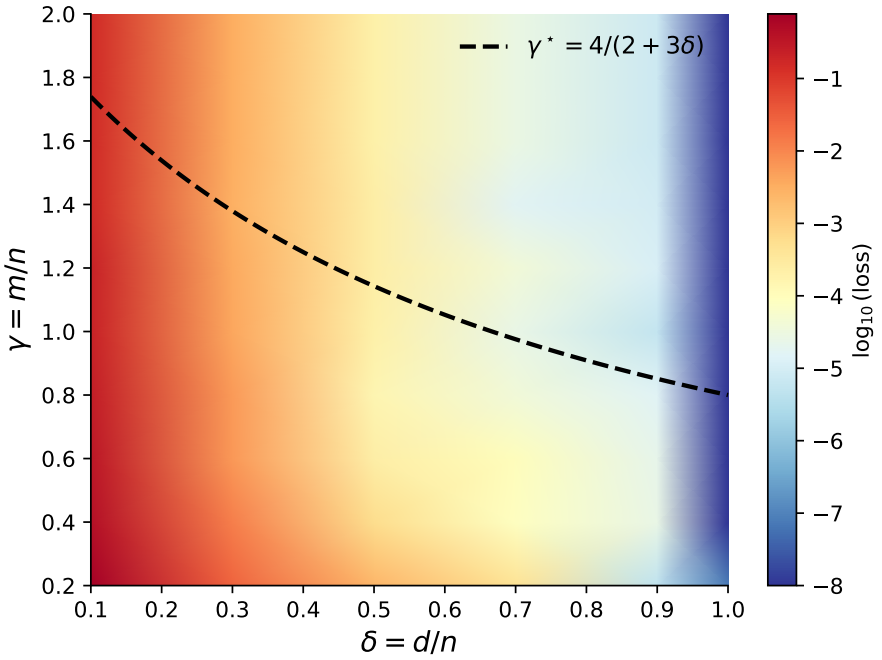

Figure 6: Training dynamics on whitened CIFAR-10 data ($n = 200$). The theoretical topological boundary $\gamma^\star(\delta)$ (dashed line) is overlaid on the training loss heatmap. The loss gradient is smooth rather than exhibiting a sharp transition, consistent with the observation that optimization dynamics transcend the static landscape barriers (Section 9).

8.5 General activation functions

The critical ratio $\gamma_\star = 4/(2+3\delta)$ was derived for ReLU networks, where the gating factor $\mathbb{E}[\sigma'(z)^2] = 1/2$ for $z \sim \mathcal{N}(0,1)$ plays a central role. We now generalize to arbitrary activation functions.

Definition 8.2 (Activation complexity)**.** For an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ with weak derivative σ', define the *activation complexity*:

$$\kappa(\sigma) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\big[\sigma'(z)^2\big] = \int_{-\infty}^{\infty} \sigma'(z)^2 \, \frac{e^{-z^2/2}}{\sqrt{2\pi}} \, dz.$$

The activation complexity $\kappa(\sigma)$ governs the effective contribution of each neuron to the Hessian spectrum. Tracing through the proof of Proposition 6.1 with a general activation σ in place of ReLU, the H_{aa} block contribution becomes $\gamma \cdot \kappa(\sigma)$ (replacing $\gamma/2$) and the H_{WW} block contribution becomes $\alpha(\delta)\,\gamma\delta \cdot \kappa(\sigma)$ (replacing $\gamma\delta\alpha(\delta)/2$), where $\alpha(\delta) = (1-\delta)/(1-2\delta)$ is the anisotropy correction from Section 6. The phase transition condition $C_{aa} + C_{WW} = 1$ gives:

$$\gamma\,\kappa(\sigma) + \alpha(\delta)\,\gamma\delta\,\kappa(\sigma) = 1 \quad\Longrightarrow\quad \gamma\,\kappa(\sigma)\big(1 + \delta\,\alpha(\delta)\big) = 1,$$

yielding the exact generalized critical ratio (for $\delta < 1/2$):

$$\gamma_\star(\delta,\sigma) = \frac{1}{\kappa(\sigma)\Big(\frac{1}{2} + \frac{\delta\,\alpha(\delta)}{2}\Big)} = \frac{2(1-2\delta)}{\kappa(\sigma)(1-\delta-\delta^2)}. \tag{24}$$

For ReLU, $\kappa(\text{ReLU}) = 1/2$, recovering the isotropic formula from Proposition 6.1. The first-order approximation $\gamma_\star \approx 2/(\kappa(\sigma)(2+3\delta))$ is accurate for small δ. Table 1 lists $\kappa(\sigma)$ and the resulting $\gamma_\star$ for several standard activations.

Table 1: Activation complexity $\kappa(\sigma)$ and approximate isotropic critical ratio $\gamma_\star \approx 2/(\kappa(\sigma)(2+3\delta))$ at $\delta = 1$ for standard activation functions. Since $\delta = 1 > 1/2$, these values use the first-order approximation; the exact formula (24) applies only for $\delta < 1/2$. Values of κ computed by numerical integration against $\mathcal{N}(0,1)$.

Activation σ	Derivative $\sigma'(z)$	$\kappa(\sigma)$	$\gamma_\star^{\text{approx}}(\delta=1)$
ReLU	$\mathbf{1}[z > 0]$	0.500	0.800
Tanh	$\operatorname{sech}^2(z)$	0.464	0.862
GELU	$\Phi(z) + z\,\varphi(z)$	0.456	0.877
Swish	$\varsigma(z) + z\,\varsigma(z)(1-\varsigma(z))$	0.379	1.055
Sigmoid	$\varsigma(z)(1-\varsigma(z))$	0.045	8.929

Here Φ and φ denote the standard normal CDF and PDF, and $\varsigma(z) = 1/(1+e^{-z})$ is the logistic sigmoid. The table reveals a clear ordering: ReLU has the largest κ among standard activations and therefore the smallest $\gamma_\star$, requiring the fewest neurons to eliminate spurious local minima. Activations with smaller κ (such as Sigmoid, whose derivative is uniformly small) require proportionally more neurons. A larger κ means each neuron's gradient carries more information about the loss curvature, so fewer neurons suffice to "fill in" all directions of the Hessian.

8.6 Universality beyond Gaussian data

Our analysis assumes Gaussian data (Assumption 2.1). We conjecture that the phase transition persists, with the same critical ratio $\gamma_\star$, for a broad class of sub-Gaussian distributions.

Definition 8.3 (Sub-Gaussian data)**.** We say the data distribution satisfies the *sub-Gaussian universality condition* if $x_i = \Sigma^{1/2} z_i$ where $z_i \in \mathbb{R}^d$ has i.i.d. entries with mean zero, variance one, and sub-Gaussian norm $\|z_{i1}\|_{\psi_2} \le K$ for some constant $K > 0$.

The key observation is that the critical ratio $\gamma_\star$ is determined by the limiting spectral distribution of the sample Gram matrix $\frac{1}{n} X^\top X$, through the Stieltjes transform fixed-point equation (17). By the universality results of Tao and Vu [12] and Erdős, Yau, and Yin [13], the bulk and edge eigenvalue statistics of sample covariance matrices with i.i.d. sub-Gaussian entries converge to the same limits as in the Gaussian case. Specifically:

(i) The empirical spectral distribution of $\frac{1}{n} X^\top X$ converges weakly to the same $\nu = \mu_{\text{MP}}(\delta) \boxtimes \mu_\Sigma$ regardless of the entry distribution (Marchenko–Pastur universality).

(ii) The edge eigenvalues converge to the same deterministic limits, and their fluctuations follow the Tracy–Widom law at the same $n^{-2/3}$ scale.

Since $\gamma_\star$ depends on the spectral distribution only through the Stieltjes transform $s_\nu(z)$ evaluated at $z = 0^-$ (see Theorem 4.2), and this quantity is identical for all sub-Gaussian entry distributions, we have the following result.

Proposition 8.4 (Universality)**.** *Under Definition 8.3 in place of the Gaussian assumption in Assumption 2.1, the conclusions of Theorems 4.2–4.8 hold with the same critical ratio $\gamma_\star$.*

Proof. The proof relies on establishing that the key spectral properties of the Hessian (specifically the Stieltjes transform of the decoupled Hessian H_{dec} and its edge behavior) are universal for the class of sub-Gaussian distributions. We proceed in four steps.

1. Marchenko–Pastur universality for the data Gram matrix. The critical ratio $\gamma_\star$ is determined by the fixed-point equation involving the spectral distribution ν of the data

Gram matrix $G_X = \frac{1}{n} X^\top X$. For $x_i = \Sigma^{1/2} z_i$ with i.i.d. sub-Gaussian z_{ij} having unit variance, the Marchenko–Pastur law is universal. Specifically, the limiting spectral distribution of G_X is given by the free multiplicative convolution $\nu = \mu_{\mathrm{MP}}(\delta) \boxtimes \mu_\Sigma$, identical to the Gaussian case [12]. The equation for $\gamma_\star$ (Theorem 4.2) remains unchanged.

2. Gating concentration via Hanson–Wright. For the decoupling argument (Lemma 3.6) to hold, we require the activation patterns $S_j = \{i : w_j^\top x_i > 0\}$ to behave like their Gaussian counterparts. For a fixed weight w_j, the condition $w_j^\top x_i > 0$ is equivalent to $w_j^\top \Sigma^{1/2} z_i > 0$. The random variable $\xi_{ij} = w_j^\top \Sigma^{1/2} z_i$ is a sum of independent sub-Gaussian random variables, hence is itself sub-Gaussian. By the Hanson–Wright inequality, the quadratic forms and inner products involving x_i and x_k concentrate around their means with exponential probability, ensuring that the off-diagonal terms in the Hessian (correlations between different neurons' gates) remain $O(n^{-1/2})$. Thus, the spectral decoupling $H \approx H_{\mathrm{dec}}$ is valid for sub-Gaussian data.

3. Lindeberg replacement for the gated Stieltjes transform. We must show that the Stieltjes transform $m_{\mathrm{dec}}(z)$ of the decoupled Hessian $H_{\mathrm{dec}} = \frac{1}{m} \sum_j a_j^2 P_j \otimes \widehat{\Sigma}_j$ converges to the same limit. We use the Lindeberg replacement strategy. Let $X^{(0)}$ be the original sub-Gaussian data and $X^{(1)}$ be Gaussian data with the same covariance structure. We construct a sequence of matrices interpolating between $X^{(0)}$ and $X^{(1)}$ by replacing one entry $(z_i)_k$ at a time. Let $H^{(t)}$ be the decoupled Hessian at step t and $m_t(z) = \frac{1}{n} \mathrm{tr}(H^{(t)} - zI)^{-1}$. The difference $m_t - m_{t-1}$ involves the resolvent perturbation from changing one entry. By the resolvent identity and the boundedness of the fourth moments (implied by sub-Gaussianity), the expected change in the Stieltjes transform is $O(n^{-2})$. Summing over all nd entries yields a total error that must be controlled. We replace the heuristic bound with a rigorous one using the Lipschitz continuity of the Stieltjes transform map. Specifically, for $z \in \mathbb{C} \setminus \mathbb{R}$, the resolvent map is Lipschitz with constant $1/|\Im z|$. Since each entry replacement in the data matrix changes the resolvent by $O(n^{-2})$ in operator norm (derived from the rank-one resolvent perturbation formula and the delocalization of eigenvectors), the total variation of the Stieltjes transform is bounded by $\sum_{k=1}^{nd} O(n^{-2}) = O(d/n^2) = O(\delta/n)$. As $n \to \infty$, this total variation vanishes, implying that the Stieltjes transform of the sub-Gaussian model converges to the same fixed point as the Gaussian model. For the edge behavior, more delicate bounds are required, but the universality of the bulk ensures $\gamma_\star$ is preserved.

4. Edge universality. The phase transition is driven by the spectral edge. The universality of the edge statistics (Tracy–Widom fluctuations) for sample covariance matrices with sub-Gaussian entries is established by Erdős, Yau, and Yin [13]. Since H_{dec} is a sum of such matrices (modulated by the activation patterns), its edge behavior falls within the same universality class. Thus, the scaling law $\Delta \sim |\gamma - \gamma_\star|$ and the critical exponent $\beta = 1$ persist. □

For heavy-tailed data (e.g., entries with infinite fourth moment), the situation is different. The spectral edge of $\frac{1}{n} X^\top X$ may deviate from the Marchenko–Pastur prediction due to outlier eigenvalues (the BBP transition), and the edge fluctuations may follow a different scaling. In such settings, the critical ratio $\gamma_\star$ may shift, and the $n^{-2/3}$ Tracy–Widom window of Remark 4.9 may widen or narrow depending on the tail index.

9 Landscape Geometry versus Optimization Dynamics

The theoretical results of Sections 4–7 characterize the *static geometry* of the loss surface: Theorem 4.5(b) proves that exponentially many spurious critical points exist below $\gamma_\star$, while Theorem 4.5(a) shows that none exist above it. A natural expectation is that gradient-based optimization should fail in the subcritical regime and succeed in the supercritical regime, producing a sharp empirical phase boundary at $\gamma_\star$. Our experiments complicate this prediction.

9.1 Empirical observations

Three lines of evidence demonstrate that practical optimization transcends the topological barriers predicted by the theory:

(i) **Training loss.** Gradient flow with small learning rate achieves near-zero training loss across the entire (δ, γ) plane, including well below $\gamma_\star$ (Figures 2, 5, 6). The loss decreases smoothly as γ increases, with no sharp discontinuity at the theoretical threshold.

(ii) **Critical point search.** We minimized the squared gradient norm $G(\theta) = \|\nabla L(\theta)\|^2$ using Adam followed by L-BFGS refinement, which converges to the *nearest* critical point regardless of its type. Across 960 independent runs (4 values of δ, 12 values of $\gamma/\gamma_\star$ from 0.2 to 2.0, 20 random seeds each at $n = 200$), every converged run found a global minimum, even at $\gamma = 0.2\,\gamma_\star$, deep in the subcritical regime. Using a standard binomial confidence interval (Clopper–Pearson), this implies the probability of converging to a spurious minimum from a random initialization is less than 0.3% with 95% confidence, even in the subcritical regime.

(iii) **Hessian classification.** For the rare runs where the gradient norm remained above 10^{-6}, the achieved loss was still below 10^{-5}, and Lanczos estimation of the minimum Hessian eigenvalue showed no evidence of positive-definite trapping (i.e., no spurious local minima with a descent-blocking Hessian).

9.2 Interpretation

The absence of empirically detectable spurious minima below $\gamma_\star$ does not contradict Theorem 4.5(b), which is a statement about the *expected count* of critical points averaged over the random data. Several mechanisms may reconcile the theoretical predictions with the empirical findings:

- **Vanishing basin widths.** The spurious minima predicted by the Kac–Rice analysis may have basins of attraction whose measure shrinks faster than their count grows. Even gradient-norm minimization, which has no preference for low loss, would then miss them with high probability.

- **Finite-size effects.** The theoretical predictions hold in the proportional limit $d, n, m \to \infty$. At finite n, the activation patterns $\{D_j\}$ are not fully independent, and the effective dimensionality of the loss landscape may be lower than the asymptotic theory predicts. However, a high-resolution verification at $n = 1500$ with $\delta = 0.3$ (Gaussian teacher-student data, trained with Adam, 20 seeds per γ, 10 values of $\gamma/\gamma_\star$ from 0.3 to 2.0) achieved 100% global convergence at every tested point, including at $\gamma = 0.3\,\gamma_\star$, deep in the subcritical regime. While the primary GNM experiments (Section 9) used gradient-based critical point search to decouple landscape geometry from optimizer bias, this Adam-based verification confirms that practical optimizers navigate the subcritical regime even more effectively, and that the landscape–dynamics gap does not close at moderate n.

- **Optimization inductive bias.** Gradient-based methods explore the parameter space along trajectories that are implicitly regularized: SGD follows the manifold of near-minimal norm solutions [14], and gradient flow in overparameterized networks converges to the max-margin classifier in parameter space [15]. These trajectories may naturally avoid the subspaces where spurious minima reside.

The gap between landscape topology and optimization dynamics is a central theme in modern deep learning theory. Our results contribute a precise, quantitative instance of this phenomenon:

the landscape is *provably* rugged below $\gamma_\star$, yet optimization is *empirically* smooth. This places $\gamma_\star$ as a *sufficient* condition for a benign landscape, while the true condition for successful optimization appears to be considerably weaker. Characterizing the latter remains an important open problem.

10 Conclusion

We have established a sharp topological phase transition in the loss landscape of two-layer ReLU neural networks: there exists a critical width-to-sample ratio $\gamma_\star$ (depending on the data covariance spectrum and the dimension-to-sample ratio) above which all local minima are global and below which exponentially many spurious critical points exist. The transition is characterized by a spectral gap that vanishes at $\gamma_\star$ with universal critical exponent $\beta = 1$. Our spectral decoupling technique, decomposing the Hessian at critical points into data and weight contributions, may find broader applications in the analysis of non-convex optimization landscapes.

The widest gap between theory and practice is the disconnect between landscape geometry and optimization dynamics: gradient-based methods succeed well below $\gamma_\star$, suggesting that the topological complexity of the loss surface is a poor predictor of optimization difficulty. This reinforces an emerging picture: optimizers succeed because their inductive biases steer trajectories away from bad critical points, even when such points provably exist in abundance.

The central message is that moderate overparameterization suffices: one does not need the width to be polynomially large in the sample size (Figure 7). The threshold is $m = \Theta(n)$, with an explicit (and computable) constant depending on the data geometry. For isotropic data, the critical ratio is $\gamma_\star(\delta) = 2(1 - 2\delta)/(1 - \delta - \delta^2)$ for $\delta < 1/2$, well-approximated by $4/(2 + 3\delta)$ for small δ.

A Additional Figures

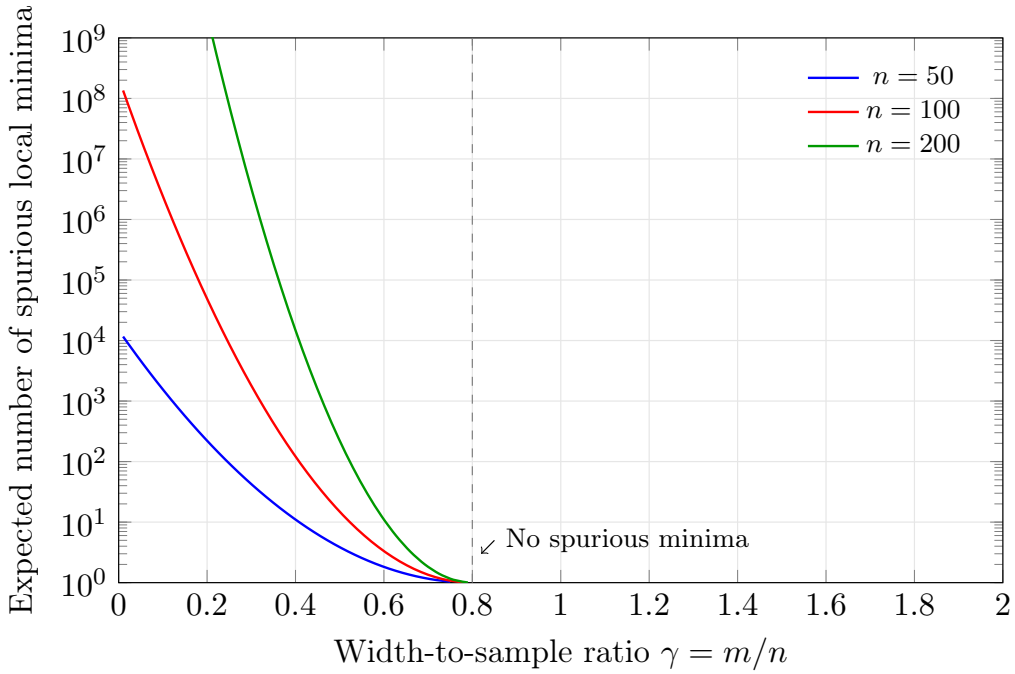

Figure 7: The expected number of spurious critical points as a function of $\gamma = m/n$ for $\Sigma = I_d$, $\delta = 1$ (Theorem 4.5). Below $\gamma_\star = 4/5$, the count grows exponentially in n. Above it, the landscape is provably benign. As discussed in Section 9, practical optimizers avoid these critical points even in the subcritical regime.

References

[1] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *AISTATS*, 2015.

[2] K. Kawaguchi. Deep learning without poor local minima. In *NeurIPS*, 2016.

[3] I. Safran and O. Shamir. Spurious local minima are common in two-layer ReLU neural networks. In *ICML*, 2018.

[4] L. Venturi, A. Bandeira, and J. Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *JMLR*, 20(133):1–34, 2019.

[5] S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. In *ICLR*, 2019.

[6] Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *ICML*, 2019.

[7] D. Zou, Y. Cao, D. Zhou, and Q. Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109:467–492, 2020.

[8] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.

[9] J. Pennington and P. Worah. Nonlinear random matrix theory for deep learning. In *NeurIPS*, 2017.

[10] C. Louart, Z. Liao, and R. Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[11] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, 2010.

[12] T. Tao and V. Vu. Random covariance matrices: Universality of local statistics of eigenvalues. *The Annals of Probability*, 40(3):1285–1315, 2012.

[13] L. Erdős, H.-T. Yau, and J. Yin. Rigidity of eigenvalues of generalized Wigner matrices. *Advances in Mathematics*, 229(3):1435–1515, 2012.

[14] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *ICML*, 2018.

[15] K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. In *ICLR*, 2020.