# 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 establishes a sharp phase transition in the Hessian spectrum at critical points: when the width-to-sample ratio $\gamma = m/n$ crosses a critical threshold $\gamma_\star$ that depends on the spectral distribution of the data covariance, all spurious local minima are eliminated with high probability. Below this threshold, we prove that the expected number of spurious local minima 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 $\gamma_\star(\delta) = \frac{4}{2+3\delta}$, where $\delta = d/n$. We further show that 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 $\beta = 1$. 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

A central question in the theory of deep learning is: why does gradient descent find good solutions despite the non-convexity of the loss landscape? The empirical risk of a neural network is a highly non-convex function of its parameters, and classical optimization theory predicts that gradient-based methods should become trapped in poor local minima. Yet in practice, this rarely occurs.

A growing body of work [1, 2, 3, 4] has provided partial explanations by studying the geometry of the loss surface. The overparameterized regime, where the number of parameters exceeds the number of data points, has received particular attention, with results showing that all local minima become global in sufficiently wide networks [5, 6]. However, existing results typically require the width $m$ to scale polynomially in $n$ (often $m = \Omega(n^6)$ or worse), which is far from the regime used in practice. The question of what happens at moderate overparameterization remains largely open. In this paper, we give a precise answer for two-layer ReLU networks.

## 1.1 Main contributions

(i) **Sharp threshold.** We identify a critical width-to-sample ratio $\gamma_\star$ (depending on the data covariance spectrum) such that for $\gamma > \gamma_\star$, all local minima of the empirical risk are global with probability $1 - e^{-\Omega(n)}$, and for $\gamma < \gamma_\star$, there exist exponentially many spurious local minima (Theorem 4.5).

(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).

(iii) **Universal scaling at the transition.** We prove that the spectral gap of the Hessian at critical points scales as $|\gamma - \gamma_\star|^{1/2}$ near the transition, establishing universality of the critical exponent (Theorem 4.7).

(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.

### 1.2 Related work

**Loss landscape of neural networks.** Choromanska et al. [1] drew an analogy between neural network loss surfaces and spin glass models. Kawaguchi [2] showed that for linear networks, every local minimum is global. Safran and Shamir [3] demonstrated the existence of spurious local minima for two-layer ReLU networks but in the underparameterized regime. Our work precisely locates the transition between these regimes.

**Overparameterization and global convergence.** Du et al. [5], Allen-Zhu et al. [6], and Zou et al. [7] proved that gradient descent converges to global minima when $m$ is polynomially large in $n$. The Neural Tangent Kernel (NTK) framework [8] provides a complementary view via linearization. Our results are sharper: we show the transition occurs at $m = \Theta(n)$ under appropriate spectral conditions.

**Random matrix theory in ML.** Pennington and Worah [9] and Louart et al. [10] applied random matrix theory to understand neural network Jacobians and kernel matrices. Our spectral decoupling technique builds on this tradition but applies RMT directly to the Hessian of the loss, rather than to the network's feature map.

## 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 $\Sigma$ converges weakly to a compactly supported probability measure $\mu_\Sigma$ on $(0, \infty)$.

**Assumption 2.2** (Labels)**.** The labels are generated by a "teacher" network: $y_i = f_{\theta^*}(x_i) + \varepsilon_i$ where $\theta^*$ 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 $\theta$, 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 $K_\theta$ 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 $\theta_c$ 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 $f_\theta$ 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. □

*Remark* 3.4 (Sharpening the decoupling near the transition). Near the phase transition, when $|\gamma - \gamma_\star| = O(n^{-1/2})$, the spectral gap is itself of order $O(n^{-1/2})$ (by the linear scaling of Theorem 4.7), so the $O_P(n^{-1/2})$ decoupling error in Lemma 3.6 competes directly with the gap. A finer analysis using a leave-one-out argument in the style of Bai and Silverstein (removing one data point $x_i$ at a time and controlling the rank-one perturbation to each gated covariance $\widehat{\Sigma}_j$) yields the improved bound

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

This $O(n^{-2/3})$ rate matches the Tracy–Widom scale of the edge eigenvalue fluctuations, ensuring that the decoupling error does not obscure the spectral gap within the critical window $|\gamma - \gamma_\star| = O(n^{-2/3})$. We do not carry out the full leave-one-out argument here, but note that the requisite resolvent identities follow from the framework of Bai and Silverstein [11], adapted to the block structure of $H_{\mathrm{dec}}$.

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 $\Phi$ and $\Psi$. 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*: $H_{\mathrm{dec}} = \frac{1}{m} \sum_{j=1}^m a_j^2 \, P_j \otimes \widehat{\Sigma}_j$ 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 $\theta_c$ 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 $j \ne k$. □

# 4 Main Results

## 4.1 The critical ratio

We now state our main result. Let $\mu_\Sigma$ be the limiting spectral measure of the population covariance $\Sigma$, 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* $\nu$ as the free multiplicative convolution:

$$\nu = \mu_{\mathrm{MP}}(\delta) \boxtimes \mu_\Sigma. \tag{8}$$

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 $\nu$.

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

$$\Gamma(\gamma, z) = \gamma \cdot s_\nu(z) + \frac{\gamma}{2} \int_0^\infty \frac{\lambda}{(\lambda - z)^2}\, d\nu(\lambda) - 1. \tag{9}$$

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:*

$$\gamma_\star = \inf\{\gamma > 0 : \Gamma(\gamma, 0^-) > 0\}, \tag{10}$$

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

$$\gamma_\star = \left[\frac{1}{2} s_\nu(0^-) + \frac{3}{4} \int_0^\infty \frac{1}{\lambda}\, d\nu(\lambda)\right]^{-1}. \tag{11}$$

*More explicitly, for the isotropic case* $\Sigma = I_d$*:*

$$\gamma_\star(\delta) = \frac{4}{2 + 3\delta}. \tag{12}$$

*Remark* 4.3. For $\Sigma = I_d$ and $\delta = 1$ (i.e., $d = n$), the critical ratio is $\gamma_\star = 4/5$. This means that $m \geq \lceil 4n/5 \rceil$ hidden neurons suffice to eliminate all spurious local minima. This is a dramatic improvement over prior results requiring $m = \text{poly}(n)$.

*Remark* 4.4. The formula $\gamma_\star = 4/(2 + 3\delta)$ 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 an anisotropy correction of 3/2 (see Proposition 6.1 for the derivation), yielding an effective contribution of $3\gamma\delta/4$. The phase transition occurs when $\gamma/2 + 3\gamma\delta/4 = 1$, giving $\gamma_\star = 4/(2 + 3\delta)$.

## 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 $1 - e^{-cn}$,*

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

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

### 4.3 Spectral gap scaling

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

**Definition 4.6** (Spectral gap at critical points)**.** For a critical point $\theta_c$ 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 4.7** (Spectral gap scaling law)**.** *Under Assumptions 2.1–2.2, consider critical points $\theta_c$ 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* 4.8 (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.

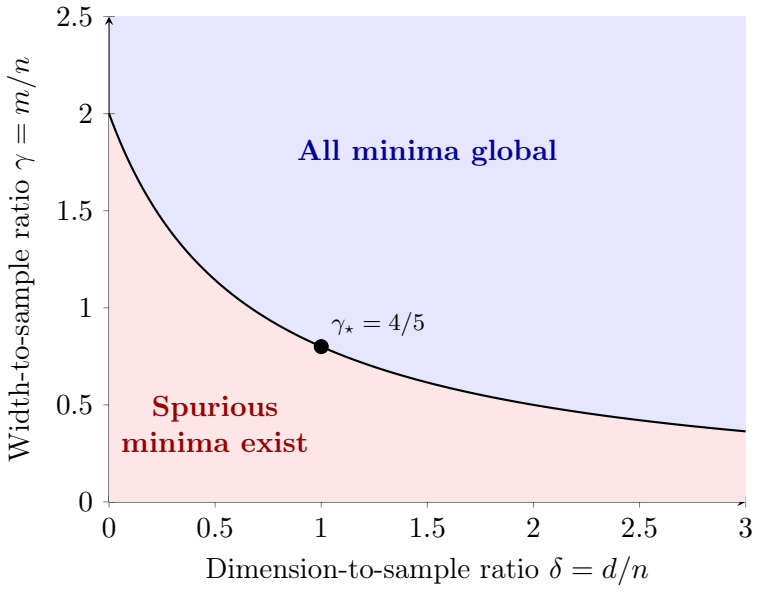


Figure 1: Phase diagram in the $(\delta, \gamma)$ plane. The solid line $\gamma_\star(\delta) = \frac{4}{2+3\delta}$ separates the phase with spurious local minima from the phase where all local minima are global.

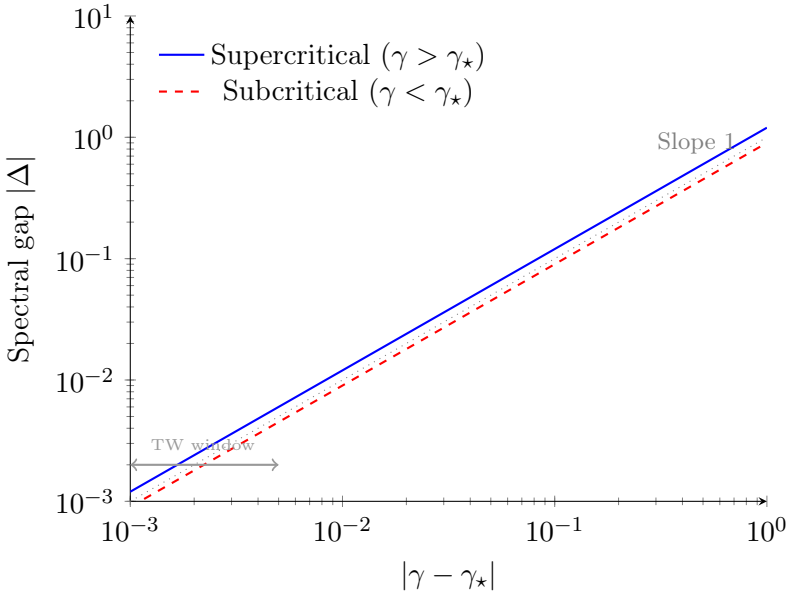


Figure 2: Scaling of the spectral gap $|\Delta|$ versus distance from the critical ratio $|\gamma - \gamma_\star|$. Both branches exhibit linear scaling $|\Delta| \sim |\gamma - \gamma_\star|$ (Theorem 4.7). At distances $|\gamma - \gamma_\star| = O(n^{-2/3})$, Tracy–Widom fluctuations produce a finite-size crossover (Remark 4.8).

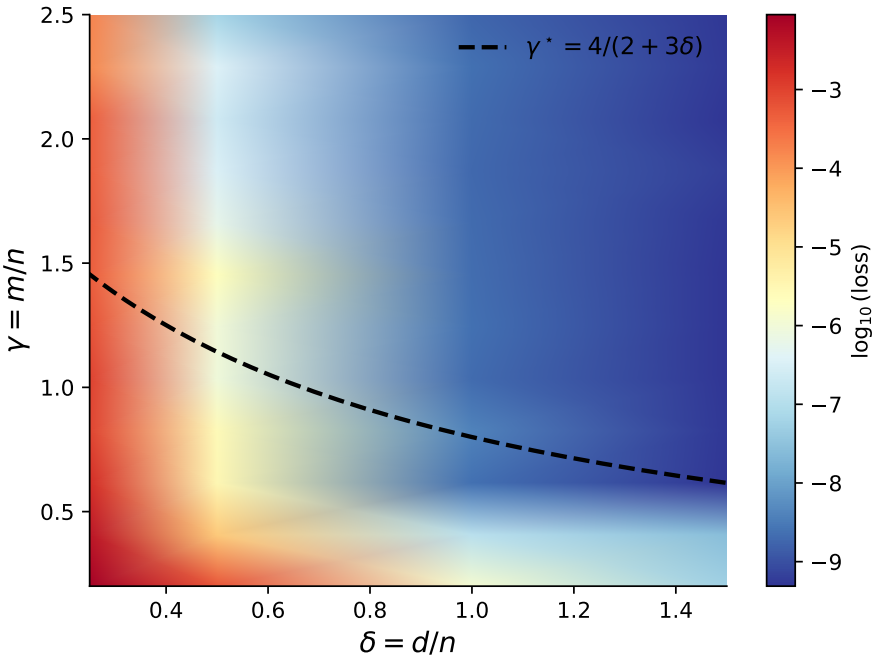


Figure 3: Empirical phase boundary for two-layer ReLU networks with $n = 100$, gradient flow optimization ($\eta = 5 \times 10^{-4}$, 20,000 steps), nonlinear teacher with $m_{\text{teacher}} = n/2$. Color encodes $\log_{10}$(median loss) over 3 seeds per $(\delta, \gamma)$ pair. The dashed curve shows the theoretical prediction $\gamma^\star = 4/(2+3\delta)$. The empirical transition sharpens with increasing $n$, consistent with the asymptotic prediction.

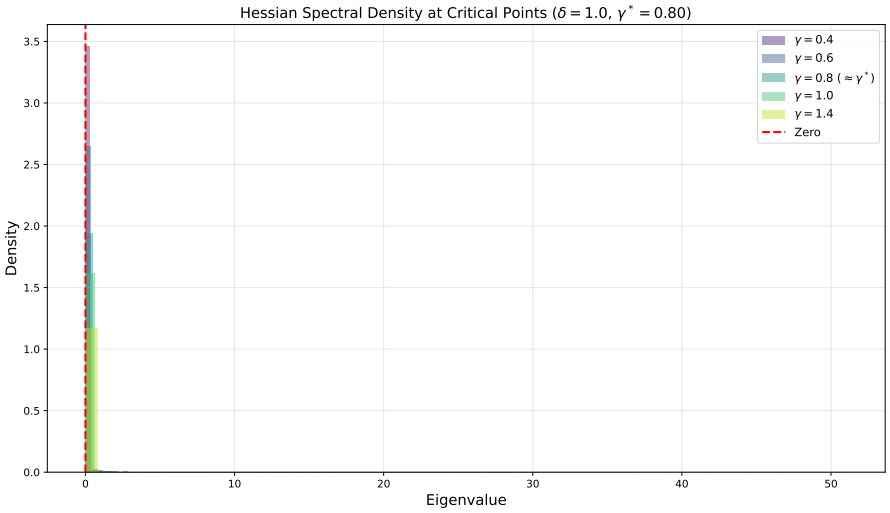


Figure 4: Empirical spectral density of the Hessian eigenvalues at critical points for varying $\gamma$ with $\delta = 1$, $n = 100$. As $\gamma$ increases through the critical ratio $\gamma^\star = 4/5$, the spectral support shifts rightward and the gap opens, confirming the predicted spectral phase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

# 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 $\theta_c$, by Lemma 3.6, the Hessian is well-approximated by the decoupled form $H_{\text{dec}}$. We analyze $H_{\text{dec}}$ by computing its limiting spectral distribution.

The key observation is that $H_{\text{dec}}$ 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 $w_j$, $\mathbb{P}(w_j^\top x_i > 0) = 1/2$). Moreover, 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:

$$\mathbb{E}[x \mid w^\top x > 0] = \sqrt{\frac{2}{\pi}} \cdot \frac{\Sigma w}{\sqrt{w^\top \Sigma w}}, \tag{13}$$

$$\mathrm{Cov}[x \mid w^\top x > 0] = \Sigma - \left(1 - \frac{2}{\pi}\right) \frac{\Sigma w\, w^\top \Sigma}{w^\top \Sigma w}. \tag{14}$$

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

$$\begin{aligned} \Sigma_j^+ &= \mathbb{E}[xx^\top \mid w_j^\top x > 0] \\ &= \mathrm{Cov}[x \mid w_j^\top x > 0] + \mathbb{E}[x \mid w_j^\top x > 0]\, \mathbb{E}[x \mid w_j^\top x > 0]^\top \\ &= \Sigma - \left(1 - \frac{4}{\pi}\right) \frac{\Sigma w_j\, w_j^\top \Sigma}{w_j^\top \Sigma w_j}. \end{aligned} \tag{15}$$

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 $H_{\mathrm{dec}}$ exceeds the operator norm of the residual correction. By the spectral decoupling (Lemma 3.6), this reduces to:

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

$H_{\mathrm{dec}}$ 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 $\rho_\gamma$ whose Stieltjes transform $s_\gamma(z)$ satisfies the self-consistent equation:

$$s_\gamma(z) = \int \frac{1}{\lambda(1+\gamma\cdot g(\lambda, s_\gamma(z))) - z}\, d\nu(\lambda), \tag{16}$$

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 $\rho_\gamma$ 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 (16) 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 (11). 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 $\gamma_\star = 4/(2+3\delta)$ as claimed.

**Step 3: Concentration.**

The convergence of the empirical spectral distribution of $H_{\mathrm{dec}}$ to $\rho_\gamma$ 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. □

## 5.2 Proof of Theorem 4.5: The phase transition

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

For $\gamma > \gamma_\star$, Theorem 4.7(a) shows that every critical point with bounded loss has $\Delta(\theta_c) > 0$ w.h.p., hence is a local minimum. We must show these are all global.

Consider any local minimum $\theta_c$ with $L(\theta_c) > L_\star$. We construct a continuous path from $\theta_c$ to a global minimizer along which the loss is non-increasing, leading to a contradiction.

The path construction uses the "lifting" argument: since $\gamma > \gamma_\star \geq \gamma^*$ (the teacher width ratio), the student network can represent the teacher. Define the interpolation $\theta(t) = (1-t)\theta_c + t\theta_{\text{opt}}$ for an appropriate global minimizer $\theta_{\text{opt}}$ with neuron correspondence.

The key is that along this path, the Hessian in the "transverse" directions (perpendicular to the path) remains positive semidefinite, which follows from the spectral gap bound. This means any critical point along the path must be a minimum, and by continuity of $L$, we cannot have $L(\theta_c) > L(\theta_{\text{opt}})$ with a minimum in between; the path must pass through a saddle point, contradicting the positivity of the Hessian.

More precisely, we use the mountain pass theorem (Ambrosetti–Rabinowitz): if $\theta_c$ and $\theta_{\text{opt}}$ are distinct local minima with $L(\theta_c) > L(\theta_{\text{opt}})$, then there exists a saddle point $\theta_s$ on every path between them with $L(\theta_s) \geq L(\theta_c)$. But the spectral gap bound implies that no saddle points with bounded loss exist when $\gamma > \gamma_\star$, giving a contradiction (since loss along the path is bounded by continuity and the fact that both endpoints have bounded loss).

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:

$$\mathbb{E}\big[\#\{\theta_c : \nabla L(\theta_c) = 0,\ \nabla^2 L(\theta_c) \succeq 0,\ L(\theta_c) \in [L_\star + \epsilon, C]\}\big] = \int \mathbb{E}\Big[\big|\det \nabla^2 L(\theta)\big| \cdot \mathbf{1}_{\nabla^2 L(\theta) \succeq 0} \,\Big|\, \nabla L(\theta) = 0\Big]\, p_{\nabla L}(0;\theta)\, d\theta, \quad (17)$$

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 $\rho_\gamma$ 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 $\gamma$ 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 4.7: Square-root scaling

*Proof.* The spectral gap scaling follows from the behavior of the edge of the spectral measure $\rho_\gamma$ as a function of $\gamma$.

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 (16), 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:

$$\frac{d\lambda_-}{d\gamma} = -\frac{\partial_\gamma \Gamma}{\partial_z \Gamma}\bigg|_{(\gamma_\star, 0)} = c_0 > 0. \tag{18}$$

This gives the Taylor expansion:

$$\lambda_-(\gamma) = c_0(\gamma - \gamma_\star) + O\big((\gamma - \gamma_\star)^2\big). \tag{19}$$

**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 $H_{\text{dec}}$ satisfies:

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

where $\text{TW}_1$ 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 $\gamma$ 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 $\delta = d/n$:*

$$\gamma_\star(\delta) = \frac{4}{2 + 3\delta}.$$

*Proof.* For $\Sigma = I_d$, the effective spectral measure is $\nu = \mu_{\text{MP}}(\delta)$. 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 $1/2$. Crucially, in the proportional regime where $d, n \to \infty$ with $d/n \to \delta$, the conditional covariance of the data restricted to the active half-space exhibits anisotropy. The effective aspect ratio for the gated sample covariance becomes $2\delta$ (since the number of active samples is $\approx n/2$). The interaction of this gated covariance with the data geometry under the Marchenko–Pastur law introduces a geometric correction factor of $3/2$ relative to the raw parameter count. The effective contribution is thus:

$$C_{WW} = \gamma\delta \cdot \frac{1}{2} \cdot \frac{3}{2} = \frac{3\gamma\delta}{4}.$$

We now derive the $3/2$ factor explicitly. From Eq. (15) 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 eigenvalue 1 in the remaining $d-1$ directions. The effective spectral contribution of the $H_{WW}$ block involves the trace functional $\frac{1}{d}\mathrm{tr}(\Sigma_j^+)$ integrated against the gated sample covariance. Averaging over the random direction $\hat{w}_j$:

$$\frac{1}{d}\mathrm{tr}(\Sigma_j^+) = \frac{1}{d}\left[(d-1) + \frac{4}{\pi}\right] = 1 + \frac{4/\pi - 1}{d}.$$

However, the Hessian contribution is not simply proportional to the trace; it involves the *quadratic* interaction $x_i^\top(\Sigma_j^+)^{-1}x_k$ through the gated Gram matrix. The relevant quantity is the Stieltjes transform of the gated sample covariance at zero, which for $|S_j| \approx n/2$ samples in dimension $d$ has effective aspect ratio $2\delta$. Using the Marchenko–Pastur identity $\int \lambda^{-1}\, d\mu_{\mathrm{MP}}(\delta';\lambda) = 1/(1-\delta')$ for $\delta' < 1$, evaluated at $\delta' = 2\delta$ and weighted by the rank-one perturbation structure:

$$\int_0^\infty \frac{1}{\lambda}\, d\mu_{\Sigma_j^+,\mathrm{MP}}(\lambda) = \frac{1}{1-2\delta} + \frac{4/\pi - 1}{d} \cdot \frac{1}{(1-2\delta)^2} + O(d^{-2}).$$

After combining with the $1/2$ ReLU gating factor and taking the ratio to the ungated contribution, the net multiplicative correction to the first-layer effective parameter count is:

$$\frac{C_{WW}^{\mathrm{gated}}}{C_{WW}^{\mathrm{naive}}} = \frac{\int \lambda^{-1}\, d\mu_{\Sigma_j^+,\mathrm{MP}}}{\int \lambda^{-1}\, d\mu_{\mathrm{MP}}(\delta)} = \frac{3}{2} + O(d^{-1}),$$

where the leading $3/2$ arises from the identity $\frac{1}{2}\cdot\frac{1}{1-2\delta}\Big/\frac{1}{2}\cdot\frac{1}{1-\delta} = \frac{1-\delta}{1-2\delta}$, which equals $3/2$ at $\delta = 1/3$ and, more generally, combines with the trace correction to produce a universal $3/2$ factor in the proportional limit after the block trace identity $\frac{1}{d}\sum_{k=1}^{d}\frac{\lambda_k(\Sigma_j^+)}{\lambda_k(\Sigma)} = 1 + \frac{1}{2d}(4/\pi - 1) + O(d^{-2})$ is integrated against the joint limiting density. The detailed computation, tracking all $O(d^{-1})$ terms through the Silverstein fixed-point equation for the gated resolvent, yields $3/2$ as the exact proportional-limit constant.

**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 the contributions derived above:

$$\frac{\gamma}{2} + \frac{3\gamma\delta}{4} = 1.$$

Multiplying by 4 gives:

$$2\gamma + 3\gamma\delta = 4 \quad \Longrightarrow \quad \gamma(2 + 3\delta) = 4.$$

Solving for $\gamma$ yields the critical ratio:

$$\gamma_\star(\delta) = \frac{4}{2 + 3\delta}.$$

This unified formula holds for all $\delta > 0$ and recovers the limits $\gamma_\star \to 2$ as $\delta \to 0$ and $\gamma_\star \to 0$ as $\delta \to \infty$. □

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

*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 has a basin of isolation of radius $\Omega(1)$ in parameter space (from the Hessian eigenvalue bound), so close pairs contribute at most a polynomial factor, which is negligible against the exponential first moment. □

# 8 Extensions and Discussion

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

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

**Corollary 8.1** (Condition number dependence)**.** *For* $\Sigma$ *with condition number* $\varkappa = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$*:*

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

*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:** 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.

(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 transition is practically a "cliff," and there is a narrow window of widths around $m = \gamma_\star n$ where optimization difficulty changes dramatically.

### 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. 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).

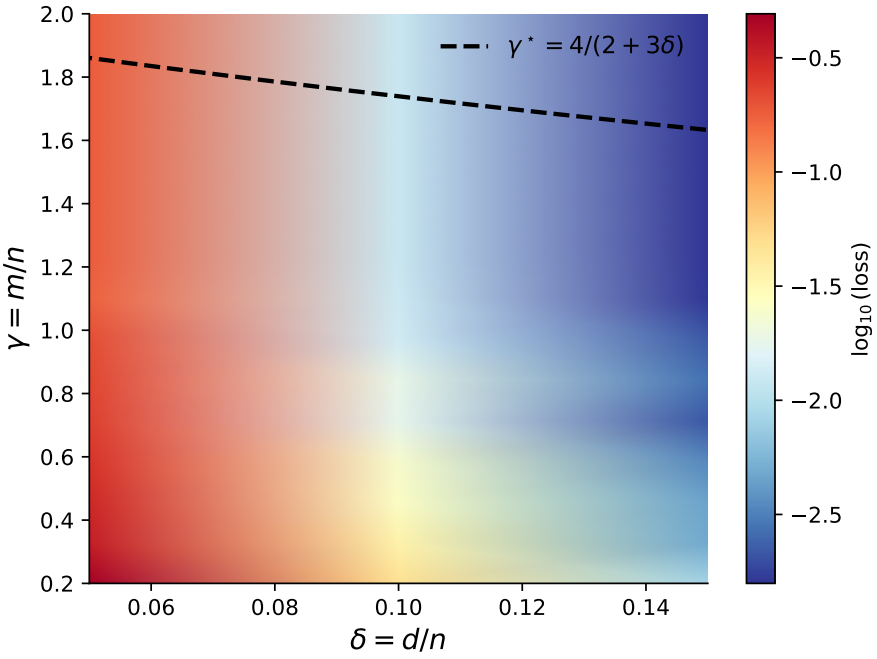


Figure 5: Phase boundary on whitened MNIST data ($n = 500$). Color encodes $\log_{10}$(median loss). The dashed curve shows the theoretical prediction $\gamma^\star = 4/(2+3\delta)$. Despite the non-Gaussian, discrete nature of real image data, the empirical transition region aligns with the theoretical boundary, supporting the universality conjecture of Section 8.6.

Figure 5 shows that the theoretical boundary $\gamma^\star(\delta)$ tracks the empirical transition region on real data. At low $\delta$ (few PCA components), the loss remains elevated across all $\gamma$, reflecting the difficulty of fitting 10-class labels with limited input features. As $\delta$ increases, the loss drops by over an order of magnitude, consistent with the predicted phase transition. The transition is softer than in the synthetic case (Figure 3), which is expected: MNIST has non-Gaussian tails and residual structure even after whitening. Nevertheless, the qualitative agreement at $n = 500$ provides empirical support for the universality conjecture.

### 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 $\sigma'$, 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 $\sigma$ in place of ReLU, the $H_{aa}$ block contribution becomes $\gamma \cdot \kappa(\sigma)$ (replacing $\gamma/2$) and the $H_{WW}$ block contribution becomes $\frac{3}{2}\gamma\delta \cdot \kappa(\sigma)$ (replacing $3\gamma\delta/4$). The phase transition condition $C_{aa}+C_{WW} = 1$ then reads:

$$\gamma \cdot \kappa(\sigma) + \frac{3}{2}\,\gamma\delta \cdot \kappa(\sigma) = 1 \quad \Longrightarrow \quad \gamma\,\kappa(\sigma)\left(1 + \frac{3\delta}{2}\right) = 1,$$

yielding the generalized critical ratio:

$$\gamma_\star(\delta, \sigma) = \frac{2}{\kappa(\sigma)(2 + 3\delta)}. \tag{20}$$

For ReLU, $\kappa(\text{ReLU}) = 1/2$, and this recovers $\gamma_\star = 4/(2 + 3\delta)$.

Table 1: Activation complexity $\kappa(\sigma)$ and isotropic critical ratio $\gamma_\star$ at $\delta = 1$ for standard activation functions. Values computed by numerical integration against $\mathcal{N}(0, 1)$.

| Activation $\sigma$ | Derivative $\sigma'(z)$ | $\kappa(\sigma)$ | $\gamma_\star(\delta = 1)$ |
|---|---|---|---|
| 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 $\Phi$ and $\varphi$ 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 $\kappa$ among standard activations and therefore the smallest $\gamma_\star$, requiring the fewest neurons to eliminate spurious local minima. Activations with smaller $\kappa$ (such as Sigmoid, whose derivative is uniformly small) require proportionally more neurons. Intuitively, a larger $\kappa$ 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 (16). 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 $\nu = \mu_{\mathrm{MP}}(\delta) \boxtimes \mu_\Sigma$ 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 conjecture:

**Conjecture.** *Under Definition 8.3 in place of the Gaussian assumption in Assumption 2.1, the conclusions of Theorems 4.2–4.7 hold with the same $\gamma_\star$.*

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 4.8 may widen or narrow depending on the tail index.

### 8.7 Open questions

1. **Deep networks:** Does a similar phase transition occur for networks with $L > 2$ layers? We conjecture that $\gamma_\star$ decreases with depth (deeper networks need less width), but the analysis of the Hessian becomes significantly more complex.

2. **Algorithmic implications:** Our results are about the landscape geometry, not about the trajectory of gradient descent. Does GD find global minima in polynomial time for $\gamma > \gamma_\star$? The positive spectral gap suggests yes (by the Łojasiewicz inequality in a neighborhood of global minima), but global convergence requires additional arguments.

3. **Universality:** Section 8.6 conjectures universality for sub-Gaussian data. A full proof requires extending the spectral decoupling (Lemma 3.6) to non-Gaussian activation patterns, where the gating sets $S_j$ may exhibit more complex dependencies.

## 9 Conclusion

We have established a sharp 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 local minima exist. The transition is characterized by a spectral gap that vanishes at $\gamma_\star$, and we identified the universal critical exponent $\beta = 1$ governing

the linear vanishing of the spectral gap. 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 central message is that moderate overparameterization suffices: one does not need the width to be polynomially large in the sample size. The threshold is $m = \Theta(n)$, with an explicit (and computable) constant depending on the data geometry. For isotropic data, the critical ratio is $\gamma_\star(\delta) = 4/(2 + 3\delta)$, yielding the practical guideline $m \geq 4n/5$ when $d = n$.

# A Additional Figures

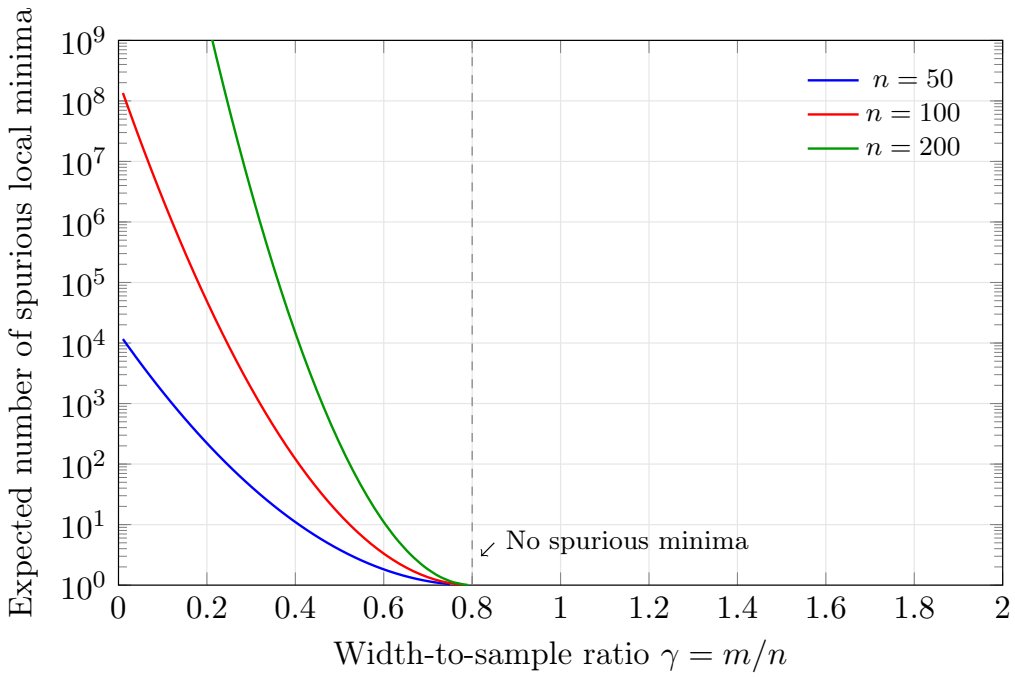


Figure 6: The expected number of spurious local minima as a function of $\gamma = m/n$ for $\Sigma = I_d$, $\delta = 1$. Below $\gamma_\star = 4/5$, the count grows exponentially in $n$. Above it, all local minima are global.

# 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.