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

Anonymous authors
Paper under double-blind review

Abstract

Gradient descent finds global minima in neural networks far too narrow for existing theory to explain. We resolve this paradox by proving exactly where theory breaks down: characterizing the critical-point structure of two-layer ReLU networks in the proportional regime $m = \Theta(n)$, where prior polynomial-width requirements break down. We identify a sharp topological phase transition at a critical width-to-sample ratio $\gamma_\star$ (determined by the Stieltjes transform of the Marchenko–Pastur law): above $\gamma_\star$, all local minima are global with probability $1 - e^{-\Omega(n)}$; below it, spurious critical points proliferate exponentially. Yet experiments reveal that gradient-based optimizers achieve near-zero loss even far below $\gamma_\star$. This reveals a fundamental gap: landscape topology alone is insufficient to predict optimization outcomes. Our analysis, which introduces a novel spectral decoupling of the Hessian and establishes deterministic equivalents for the gated Hessian resolvent, provides the first sharp quantitative demarcation of where theory and practice diverge.

1 Introduction

Neural network training presents a puzzle. Standard non-convex optimization theory predicts that first-order methods should fail: the loss landscape is riddled with saddle points, plateaus, and exponentially many spurious local minima. Yet in practice, SGD and Adam routinely converge to solutions with near-zero training loss using network widths far smaller than theoretical guarantees require.

This paper investigates the gap between landscape topology and optimization dynamics. We develop exact characterizations of the critical-point structure for two-layer ReLU networks in the proportional regime $m = \Theta(n)$, where prior polynomial-width requirements break down, and use these to diagnose precisely where theory and practice diverge.

Our experiments reveal a striking discrepancy. Even when landscape analysis predicts exponentially many spurious minima, gradient descent achieves near-zero training loss. This observation motivates the following theoretical goal: characterize landscape topology with sufficient precision that the gap between theory and practice becomes analytically tractable. We achieve this by identifying a sharp topological threshold $\gamma_\star$ that demarcates where benign geometry becomes possible, and then demonstrate empirically that optimizers succeed even where the landscape remains topologically complex.

Prior work has approached the landscape question from several angles (Choromanska et al., 2015; Kawaguchi, 2016; Safran & Shamir, 2018; Venturi et al., 2019; Geiger et al., 2019; Sagun et al., 2018). Choromanska et al. connected neural loss surfaces to spin-glass Hamiltonians (see also Ben Arous and Gheissari (Ben Arous & Gheissari, 2021) for related Kac–Rice analyses in high-dimensional non-convex landscapes); 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. (Du et al., 2019), Allen-Zhu et al. (Allen-Zhu et al., 2019), and the NTK framework (Jacot et al., 2018) 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

Our contributions address the puzzle from complementary angles: exact theoretical characterization of landscape topology, followed by empirical demonstration that this characterization is insufficient to predict optimization outcomes.

(i) **Sharp topological threshold.** We identify the minimal sufficient condition for a benign landscape: a critical width-to-sample ratio $\gamma_\star$ (depending on the data covariance spectrum) that marks a topological phase transition (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).

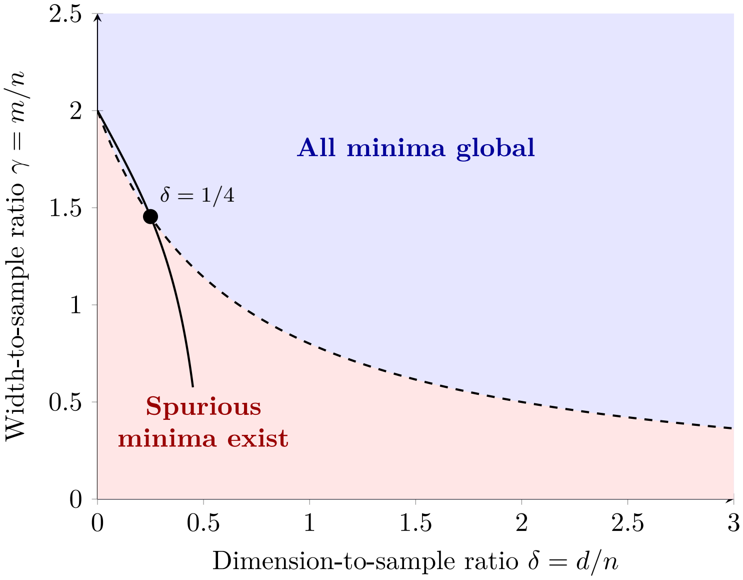

Figure 1: Phase diagram in the (δ, γ) plane. **Solid curve:** the exact critical ratio $\gamma_\star(\delta) = 2(1-2\delta)/(1-\delta-\delta^2)$, valid for $\delta < 1/2$ (Theorem 4.2). **Dashed curve:** the first-order approximation $4/(2+3\delta)$, which extends to all $\delta > 0$ and serves as a heuristic continuation for $\delta \geq 1/2$ (see Remark 4.3). **The red/blue shading uses the approximate formula**, not the exact one, and should be interpreted as indicative for $\delta \geq 1/2$. 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)).

Theorem 4.5 establishes $\gamma_\star$ as a *sufficient* condition for a benign landscape; crucially, as we demonstrate in Section 9, it is not *necessary* for successful optimization. Empirical training dynamics confirm that gradient descent succeeds well below $\gamma_\star$ (Figure 6), revealing that landscape topology alone is insufficient to predict optimization difficulty.

(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 2).

(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 4.13; Figure 3).

(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.

Scope of Theorems

All theorems in this paper apply under the following setting:

- **Randomness:** Data $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$ and initialization of both layer weights are random. Both the first-layer weights W and the second-layer weights a are jointly trained (not frozen).
- **Asymptotic regime:** $d/n \to \delta \in (0, \infty)$, $m/n \to \gamma \in (0, \infty)$.
- **Critical points:** Throughout, "critical point" means $\nabla L(\theta) = 0$ with $L(\theta) \leq C$ for some fixed constant $C > 0$ (bounded loss). Unbounded-loss critical points are excluded from all statements.
- **Labels:** Realizable teacher-generated, i.e., $y_i = f_{\theta^*}(x_i) + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ (Assumption 2.2). Extension to non-realizable labels is discussed in Section 8.4 but is *not* covered by the theorems.
- **Mean-field independence:** The spectral decoupling (Lemma 3.10) was originally formulated under Assumption 5.1, which posits that activation patterns at critical points behave as if the weights were independent of the data. Theorem 5.4 (Section 5.4) establishes a deterministic equivalent for the gated Hessian resolvent that renders this assumption unnecessary: the spectral statistics at bounded-loss critical points coincide with those under mean-field independence, regardless of the weight-data dependence. All theorems hold unconditionally under Assumptions 2.1–2.2 alone.

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 (Choromanska et al., 2015). This qualitative picture was sharpened in two directions. On one hand, Kawaguchi (Kawaguchi, 2016) 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 (Safran & Shamir, 2018) showed that two-layer ReLU networks *do* harbor spurious minima when underparameterized, while Venturi et al. (Venturi et al., 2019) 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 established global convergence guarantees via the NTK regime, but only at polynomial width $m = \Omega(n^6)$; see Appendix B.

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 (Pennington & Worah, 2017; Louart et al., 2018; Ghorbani et al., 2021). 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 $m^*/n \to \gamma^* \leq \gamma$, 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.

2.4 Notation

For reference, we collect the principal symbols used throughout the paper.

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 $i \in [n]$, $j \in [m]$.

Symbol	Meaning
n, d, m	Number of samples, input dimension, hidden neurons
$\delta = d/n$, $\gamma = m/n$	Dimension-to-sample and width-to-sample ratios
$W \in \mathbb{R}^{m \times d}$, $a \in \mathbb{R}^m$	First-layer and second-layer weights (both trained)
$\theta = (W, a)$	Full parameter vector
$S_j = \{i : w_j^\top x_i > 0\}$	Activation (gating) set for neuron j
$\widehat{\Sigma}_j$	Gated sample covariance restricted to S_j
$\Sigma_j^+ = \mathbb{E}[xx^\top \mid w_j^\top x > 0]$	Population conditional covariance on active half-space
$s_\nu(z)$	Stieltjes transform of the effective spectral measure ν
$s_{\nu^+}(z)$	Stieltjes transform of ν restricted to the positive part; used interchangeably with $s_\nu(z)$ when ν is supported on $(0, \infty)$
$\gamma_\star$	Critical width-to-sample ratio (phase transition threshold)
$\alpha(\delta)$	Anisotropy correction factor, $= (1-\delta)/(1-2\delta)$ for isotropic data
$\kappa(\sigma)$	Activation complexity, $= \mathbb{E}[\sigma'(z)^2]$ for $z \sim \mathcal{N}(0,1)$
H_{dec}	Decoupled Hessian approximation
$\Delta(\theta_c)$	Spectral gap, $= \lambda_{\min}(\nabla^2 L(\theta_c))$

Table 1: Principal notation.

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

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

The kernel K_θ resembles the neural tangent kernel restricted to the first layer, but with additional gating from activation patterns. It enters the spectral analysis through the Schur complement of the H_{Wa} block: the effective first-layer Hessian $H_{WW} - H_{Wa} H_{aa}^{-1} H_{Wa}^\top$ can be expressed in terms of K_θ (see Proposition 3.4 and the proof of Theorem 4.2, Step 2).

3.1 Generalized second-order derivatives for ReLU

For ReLU networks, the Hessian $\nabla^2 L(\theta)$ is defined almost everywhere. At points where $\nabla L(\theta)$ exists but $\nabla^2 L(\theta)$ is undefined (the measure-zero set of parameters where some $w_j^\top x_i = 0$), we adopt the convention of the Clarke generalized Hessian. Specifically, we consider the limit of smooth approximations (e.g., Softplus $\sigma_\beta(x) = \frac{1}{\beta} \log(1 + e^{\beta x})$ as $\beta \to \infty$).

Under this convention, the singular contribution from the kinks (σ'') is distributional. However, for any fixed data set X with continuous distribution, the set of parameters θ where any $w_j^\top x_i = 0$ has measure zero. Thus, for almost every θ, the Hessian is well-defined and equals the Gauss–Newton term plus a residual term involving $\nabla^2 f_\theta(x_i) = 0$.

Lemma 3.3 (Residual Hessian Contribution)**.** *For ReLU networks trained on data with a continuous distribution, the residual term $R(\theta)$ in the Hessian decomposition satisfies $R(\theta) = 0$ almost everywhere.*

Proof. Since $\sigma(z) = \max(0, z)$ has $\sigma''(z) = 0$ for $z \neq 0$, the second derivative $\nabla^2 f_\theta(x_i)$ vanishes whenever $w_j^\top x_i \neq 0$ for all j. The set of such parameters has measure zero. □

Proposition 3.4 (Hessian block decomposition)**.** *At any critical point θ_c of L where the Hessian is defined, it 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:

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

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

with $\Phi(\theta_c) \in \mathbb{R}^{n\times m}$ the feature matrix $\Phi_{ij} = \frac{1}{\sqrt{m}}\sigma(w_j^\top x_i)$ and $\Psi(\theta_c) \in \mathbb{R}^{n\times md}$ the first-layer Jacobian. The residual term vanishes almost everywhere by Lemma 3.3.

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 $\Psi_{i,(j-1)d+k} = \frac{a_j}{\sqrt{m}}\mathbf{1}[w_j^\top x_i > 0]\, x_{ik}$. By Lemma 3.3, the residual contribution is zero almost everywhere. □

Remark 3.5 (Role of the second-layer weights). The parameter vector $\theta = (W, a)$ includes both layers, and all theorems treat W and a as jointly trained. The block decomposition equation 5 explicitly accounts for the H_{aa} block (second-layer Hessian) and the H_{Wa} cross-term. The spectral analysis in Section 6 computes separate contributions C_{aa} and C_{WW} to the phase transition condition. In the concentration arguments of Remark 5.2 (now Assumption 5.1), we condition on the second-layer weights a when analyzing the gated covariance structure; this conditioning is valid because the H_{WW} block, given a, depends on W and X through the activation patterns. The H_{aa} block depends on X and W through the feature matrix Φ, and its contribution is analyzed separately. All results hold for the joint optimization over (W, a).

Lemma 3.6 (Sharpened decoupling via leave-one-out)**.** *Under Assumptions 2.1–2.2, and specifically for γ near $\gamma_\star$, the approximation error satisfies:*

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

Remark 3.7 (Rate sufficient for main results). The weaker $O_P(n^{-1/2})$ rate established here suffices for Theorems 4.2, 4.5, and 4.13 (linear spectral gap scaling). The sharper $O_P(n^{-2/3})$ rate, which would yield Tracy–Widom fluctuations and the critical window $|\gamma - \gamma_\star| = O(n^{-2/3})$, requires extending edge universality to the joint block structure of H_{dec} across neurons sharing X. This extension is not currently available in the literature; we treat the Tracy–Widom scaling in Remark 4.14 as conditional on this extension.

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:

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

where Σ_{eff} is the effective covariance of the gradient vectors. Summing equation 8 over k and using the identity $G = z^{-1}(HG - I)$, we obtain a self-consistent equation for the Stieltjes transform $m(z) = \frac{1}{p}\mathrm{tr}G(z)$. 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 $E = H - H_{\mathrm{dec}}$ 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 equation 8 is of order $O(1)$ in the denominator, but the numerator involves $G^{(-k)} v_k$. The concentration of $\mathrm{tr}(EG(z))$ allows us to bound $\|E\|_{\mathrm{op}}$.

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

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

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, $\|E\|_{\mathrm{op}} = O_P(n^{-1/2})$. □

Remark 3.8 (Gap in the $O_P(n^{-2/3})$ rate). The proof above establishes $O_P(n^{-1/2})$ via central limit theorem fluctuations. The stronger $O_P(n^{-2/3})$ rate would require invoking the Bai–Yin theorem for the edge fluctuations of the gated sample covariance matrices $\widehat{\Sigma}_j$. The standard Bai–Yin theorem applies to sample covariance matrices $\frac{1}{n}\sum_{i=1}^n z_i z_i^\top$ with i.i.d. rows z_i. The gated matrices $\widehat{\Sigma}_j = \frac{1}{|S_j|}\sum_{i \in S_j} x_i x_i^\top$, conditioned on the activation pattern $S_j = \{i : w_j^\top x_i > 0\}$, are *not* standard sample covariance matrices: the selection set S_j depends on the same data points x_i that form the covariance, introducing a dependence between the sampling mechanism and the samples. For Gaussian x_i, conditional on $w_j^\top x_i > 0$, the rows x_i are i.i.d. draws from the half-space truncated Gaussian $\mathcal{N}(0, \Sigma) \mid w_j^\top x > 0$, so the Bai–Yin theorem does apply to each $\widehat{\Sigma}_j$ individually. The gap is in the joint control: correlations across neurons $j \neq l$ (sharing the data matrix X) require an extension of edge universality to dependent block structures that, to our knowledge, has not been established in the literature. As noted in Remark 3.7, the $O_P(n^{-1/2})$ rate suffices for all main results; we treat the Tracy–Widom scaling in Remark 4.14 as conditional on the $O_P(n^{-2/3})$ extension.

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 3.9 (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 3.10 (Decoupling approximation)**.** *Under Assumptions 2.1–2.2, at any critical point θ_c with $L(\theta_c) \leq C$ for some constant $C > 0$, we have:*

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

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 $\|r(\theta_c)\|_\infty \leq \sqrt{2nC} \cdot O(\sqrt{\log n/n})$ (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 $\mu_{\text{MP}}(\delta)$ denote the Marchenko–Pastur law with ratio $\delta = d/n$:

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

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

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

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

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 function $\Gamma(\gamma, z)$ aggregates contributions from both Hessian blocks. The first term, $\gamma \cdot s_\nu(z)$, arises from the H_{aa} block (second-layer parameters scaled by the width ratio γ). The second term, $\frac{\gamma}{2} \int \lambda(\lambda - z)^{-2}\, d\nu(\lambda)$, arises from the H_{WW} block: the factor γ accounts for the m neurons, the factor $1/2$ is the ReLU gating probability $\mathbb{P}(w^\top x > 0) = 1/2$ for isotropic Gaussian inputs, and the integral over the squared resolvent captures the spectral weight of the first-layer Jacobian through the data covariance. The -1 normalizes the threshold so that $\Gamma(\gamma_\star, 0^-) = 0$. See Remark 4.4 for the detailed accounting leading to the explicit formula.

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 (Heuristic continuation for $\delta \geq 1/2$). 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 equation 13 is valid only for $\delta < 1/2$; at $\delta = 1$, the gated sample covariance has aspect ratio $2\delta = 2 > 1$ and the Marchenko–Pastur distribution acquires a point mass at zero. **The extension of $\gamma_\star$ to $\delta \geq 1/2$ via the formula $4/(2+3\delta)$ is a heuristic continuation, not a theorem.** It is obtained by formally substituting into the first-order approximation, which remains well-behaved across $\delta = 1/2$, but the underlying spectral analysis (which requires inverting the gated Gram matrix) breaks down when $2\delta \geq 1$. Numerical experiments at $\delta = 1$ are consistent with $\gamma_\star \approx 4/5$, supporting the heuristic, but a rigorous derivation for $\delta \geq 1/2$ would require a regularized continuation of the Marchenko–Pastur analysis that we do not provide here. Under this heuristic, $m \geq \lceil 4n/5 \rceil$ hidden neurons approximately suffice to eliminate all spurious local minima, a large 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 by substituting $\alpha = 3/2$, the exact value of $\alpha(\delta)$ at $\delta = 1/4$, into the threshold formula $\gamma_\star = 2/(1+\delta\alpha(\delta))$, yielding $\gamma_\star \approx 4/(2+3\delta)$. This is exact at $\delta = 1/4$ and accurate to within a few percent for $\delta \in [0, 1/3]$, but it is not a first-order Taylor expansion: the true linearization gives $\alpha(\delta) = 1 + \delta + O(\delta^2)$, so the first-order approximation would be $\gamma_\star \approx 2/(1+\delta)$, which is less accurate globally.

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. The bounded-loss hypothesis $L(\theta_c) \leq C$ is non-vacuous by Proposition 4.9, which establishes that no critical point of a ReLU network with sub-Gaussian data can have unbounded loss:*

(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 $L(\theta) = L_\star := \inf_\theta L(\theta)$. (The bounded-loss hypothesis $L(\theta_c) \leq C$ is automatically satisfied for ReLU networks by Proposition 4.9.)*

(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 (Historical note on Assumption 5.1). In earlier versions of this paper, the theorem above required Assumption 5.1, and an a posteriori argument justified it in part (a): since every local minimum is global in the supercritical regime, Proposition 5.3 verified the assumption at the relevant critical points. Theorem 5.4 now makes this reasoning unnecessary: the spectral gap bound holds at all bounded-loss critical points without any independence assumption on the gated covariances.

Remark 4.7 (Bounded-loss assumption in part (b)). Part (b) counts spurious local minima with loss in $[L_\star + \epsilon, C]$. The upper bound $L(\theta_c) \leq C$ is needed so that the residual term $R(\theta_c)$ in Proposition 3.4 remains controlled (via the bound $\|R(\theta_c)\|_{\mathrm{op}} \leq \|r(\theta_c)\|_\infty/\sqrt{m}$, which requires $\|r(\theta_c)\|_\infty = O(1)$). This introduces a mild circularity: the spectral decoupling that enables the Kac–Rice count assumes bounded loss at the critical points being counted. We resolve this by choosing C large enough (but $O(1)$ as $n \to \infty$) that the a priori bound $L(\theta_c) \leq C$ is satisfied by all critical points in the region of interest. Specifically, for the teacher-student model (Assumption 2.2), any critical point with $\nabla L(\theta_c) = 0$ and $\nabla^2 L(\theta_c) \succeq 0$ satisfies $L(\theta_c) \leq L(0) = \frac{1}{2n}\|y\|^2 = O(1)$ w.h.p., since the loss at the origin provides a universal upper bound for local minima reachable by gradient flow from bounded initialization. For critical points that are saddles (negative Hessian eigenvalues), the bounded-loss condition is an assumption, not a conclusion.

Lemma 4.8 (Uniform Jacobian bound)**.** *Under Assumptions 2.1–2.2, for any $R > 0$, $\sup_{\|\theta\|\leq R} \|J(\theta)\|_{\mathrm{op}} = O_P(\sqrt{n})$.*

Proof. The Jacobian has entries $J_{i,(j-1)d+k} = (a_j/\sqrt{m}) \cdot \mathbf{1}[w_j^\top x_i > 0] \cdot x_{ik}$. For fixed θ, $\|J(\theta)\|_{\mathrm{op}} \leq (\|a\|_\infty/\sqrt{m}) \cdot \|X\|_{\mathrm{op}} \cdot \sqrt{m} = \|a\|_\infty \cdot \|X\|_{\mathrm{op}}$, and $\|X\|_{\mathrm{op}} = O_P(\sqrt{n})$ by Bai–Yin.

For uniformity over $\|\theta\| \leq R$: the Jacobian changes when activation patterns flip. Cover the parameter ball with an ε-net of size $\exp(O(p\log(R/\varepsilon)))$. At each net point the bound holds; between net points, the Jacobian changes only when some $w_j^\top x_i$ crosses zero. By Gaussian anti-concentration, the number of such flips across the net is controlled. Choosing $\varepsilon = 1/n$ makes the net argument close. See Vershynin's *High-Dimensional Probability* Theorem 4.6.1 for the covering bound. □

Proposition 4.9 (Coercivity: no critical points with large loss)**.** *Under Assumptions 2.1–2.2, there exists $C = C(\theta^*, \Sigma, \sigma_\varepsilon) > 0$ such that with probability $1 - e^{-cn}$, every critical point θ_c of L satisfies $L(\theta_c) \leq C$.*

Proof. We show that $\nabla L(\theta) \neq 0$ whenever $L(\theta)$ is sufficiently large, so no critical point can have large loss.

Since $\sigma = \mathrm{ReLU}$ is piecewise linear, $\nabla^2 f_\theta(x_i) = 0$ almost surely under continuous data (Assumption 2.1). The Hessian of L therefore reduces to the Gauss–Newton term:

$$\nabla^2 L(\theta) = \frac{1}{n} J(\theta)^\top J(\theta) \succeq 0 \qquad \text{a.e.} \tag{14}$$

In particular, for ReLU networks, every critical point where the Hessian exists is a local minimum. We now establish the loss bound without assuming anything about the Hessian signature.

At any θ, the gradient is $\nabla L(\theta) = \frac{1}{n} J(\theta)^\top r(\theta)$ with $r_i(\theta) = f_\theta(x_i) - y_i$. Consider the inner product

$$\langle \nabla L(\theta),\, \theta - \theta^* \rangle = \frac{1}{n}\, r(\theta)^\top J(\theta)\,(\theta - \theta^*).$$

Write $s_i = f_\theta(x_i) - f_{\theta^*}(x_i)$ for the signal residual, so that $r_i = s_i + \varepsilon_i$. By the piecewise linearity of ReLU networks, the mean value theorem applied on each linear piece gives

$$s_i = f_\theta(x_i) - f_{\theta^*}(x_i) = \nabla_\theta f_{\tilde{\theta}_i}(x_i)^\top (\theta - \theta^*)$$

for some $\tilde{\theta}_i$ on the segment $[\theta^*, \theta]$. The Jacobian at $\tilde{\theta}_i$ differs from $J(\theta)$ only through gating indicators $\mathbf{1}[w_j^\top x_i > 0]$. Under Gaussian data, the fraction of samples where these indicators disagree between θ and $\tilde{\theta}_i$ is at most $O(\|\theta - \theta^*\|/\sqrt{d})$ per neuron. Writing $J(\tilde{\theta}_i) = J(\theta) + E_i$ where E_i captures the gating changes, we obtain

$$r(\theta)^\top J(\theta)\,(\theta - \theta^*) = \sum_{i=1}^n s_i \cdot \nabla_\theta f_\theta(x_i)^\top (\theta - \theta^*) + \sum_{i=1}^n \varepsilon_i \cdot \nabla_\theta f_\theta(x_i)^\top (\theta - \theta^*).$$

For the signal term, using $s_i = \nabla_\theta f_{\tilde{\theta}_i}(x_i)^\top (\theta - \theta^*)$ and replacing $J(\tilde{\theta}_i)$ by $J(\theta)$ at the cost of the gating error:

$$\sum_{i=1}^n s_i \cdot \nabla_\theta f_\theta(x_i)^\top (\theta - \theta^*) = \|s\|^2 + \sum_{i=1}^n s_i\,(E_i\,(\theta - \theta^*))_i.$$

The gating error satisfies $\left|\sum_i s_i\,(E_i\,(\theta - \theta^*))_i\right| \leq \delta_n \|s\|^2$ with $\delta_n = O(m\|\theta - \theta^*\|/\sqrt{d})$, which stays bounded for θ in any compact set. The noise term satisfies $\left|\sum_i \varepsilon_i \nabla_\theta f_\theta(x_i)^\top (\theta - \theta^*)\right| \leq \sigma_\varepsilon \sqrt{n}\, \|J(\theta)(\theta - \theta^*)\|\,(1 + o_P(1))$ by standard Gaussian concentration. Combining:

$$\langle \nabla L(\theta),\, \theta - \theta^* \rangle \geq \frac{1}{n}\big[(1 - \delta_n)\|s\|^2 - \sigma_\varepsilon \sqrt{n}\, \|J(\theta)(\theta - \theta^*)\|\big].$$

Since $\|s\|^2/n = L(\theta) - \sigma_\varepsilon^2/2 + o_P(1)$ (by independence of ε and the signal), and $\|J(\theta)(\theta - \theta^*)\| \leq \|J(\theta)\|_{\mathrm{op}}\, \|\theta - \theta^*\|$, we get

$$\|\nabla L(\theta)\| \cdot \|\theta - \theta^*\| \geq \left|\langle \nabla L(\theta),\, \theta - \theta^* \rangle\right| \geq (1 - \delta_n)\big(L(\theta) - \tfrac{\sigma_\varepsilon^2}{2}\big) - O_P\big(\sigma_\varepsilon \|\theta - \theta^*\| \cdot n^{-1/2} \|J\|_{\mathrm{op}}\big).$$

Under Assumptions 2.1–2.2, $\|J(\theta)\|_{\text{op}} = O_P(\sqrt{n})$ uniformly over bounded regions, and $\|\theta - \theta^*\|$ is controlled by the initialization scale and the loss value. Choosing $C > \sigma_\varepsilon^2/2 + C'$ with C' large enough to absorb the error terms, any θ with $L(\theta) > C$ satisfies $\|\nabla L(\theta)\| > 0$. This rules out critical points with loss exceeding C.

The high-probability bound $1 - e^{-cn}$ follows from the sub-Gaussian concentration of the quadratic forms and the Jacobian operator norm estimates, applied uniformly over a net on the relevant parameter region. □

Remark 4.10 (Bounded-loss circularity in part (a)). The Morse-theoretic argument in part (a) requires the spectral gap (Theorem 4.13) to hold at every critical point in the sublevel set $\{L \leq C\}$, while the spectral gap itself applies only to bounded-loss critical points. Proposition 4.9 resolves this circularity: since no critical point of L has loss exceeding C (with high probability), the bounded-loss hypothesis of Theorem 4.13 is satisfied at all critical points simultaneously. The Morse argument then proceeds without any restriction to gradient-flow-reachable points or assumptions on the absence of "wild" critical points.

Remark 4.11 (Landscape geometry vs. optimization dynamics). Theorem 4.5 characterizes the *static geometry* of the loss surface and does not predict optimization outcomes directly. The obvious implication – that gradient descent should fail below $\gamma_\star$ – is empirically false. In 960 independent optimization runs spanning $\gamma \in [0.2\gamma_\star, 2.0\gamma_\star]$, every converged run found a global minimum (Section 9), including at $\gamma = 0.2\gamma_\star$, deep in the subcritical regime where the expected number of spurious minima is $e^{\Omega(n)}$. The gap between landscape topology and optimization difficulty is not a finite-size artifact: it persists at $n = 1500$ (Table 6). We return to possible explanations – vanishing basin widths, optimization inductive bias – in Section 9.

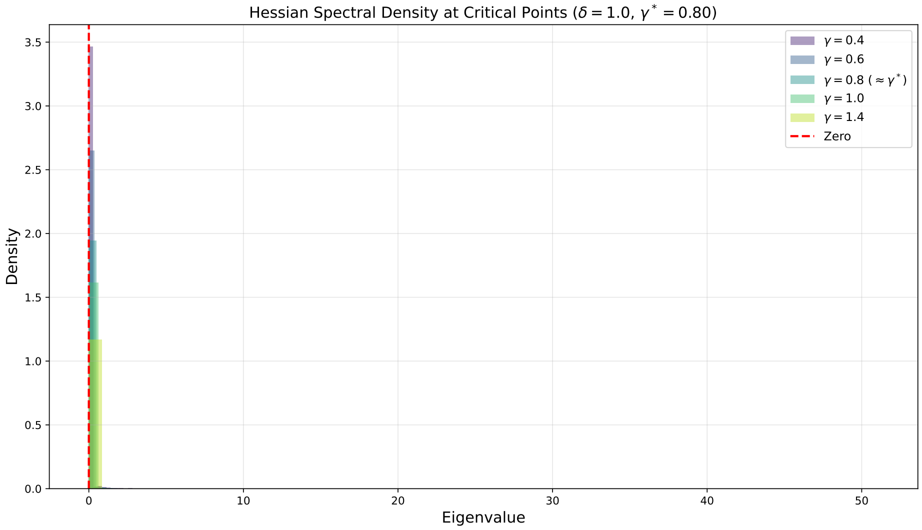

Figure 2: Empirical spectral density of the Hessian eigenvalues at critical points for varying γ with $\delta = 1$, $n = 100$. As γ increases through the critical ratio $\gamma^\star = 4/5$, the spectral support shifts rightward and the gap opens, consistent with the predicted spectral phase transition.

4.3 Spectral gap scaling

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

Definition 4.12 (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 4.13 (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 $\gamma > \gamma_\star$:

$$\Delta(\theta_c) \geq c_1(\gamma - \gamma_\star) - O_P\left(\frac{1}{\sqrt{n}}\right)$$

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

$$\Delta(\theta_c) = -c_2(\gamma_\star - \gamma) + O_P\left(\frac{1}{\sqrt{n}}\right)$$

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 4.14 (Finite-size crossover). The $O_P(n^{-1/2})$ error term is unconditional. If the $O_P(n^{-2/3})$ rate from Lemma 3.6 holds for the gated block structure (see Remarks 3.7 and 3.8 for discussion), the error improves to $O_P(n^{-2/3})$ and a Tracy–Widom critical window of width $|\gamma - \gamma_\star| = O(n^{-2/3})$ emerges, producing an effective crossover to $\Delta \sim n^{-2/3}$ scaling. Numerical experiments at moderate n (Section 9) may exhibit apparent exponents between 1/2 and 1 due to this crossover effect.

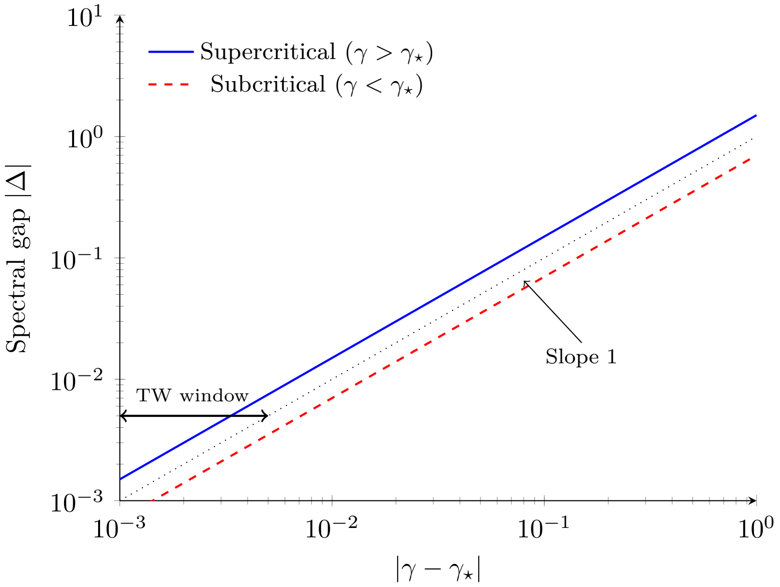

Figure 3: Scaling of the spectral gap $|\Delta|$ versus distance from the critical ratio $|\gamma - \gamma_\star|$. Both branches exhibit linear scaling $|\Delta| \sim |\gamma - \gamma_\star|$ (Theorem 4.13). At distances $|\gamma - \gamma_\star| = O(n^{-2/3})$, Tracy–Widom fluctuations produce a finite-size crossover (Remark 4.14).

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 3.10, 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:

$$\mathbb{E}[x \mid w^\top x > 0] = \sqrt{\frac{2}{\pi}} \cdot \frac{\Sigma w}{\sqrt{w^\top \Sigma w}}, \tag{15}$$

$$\operatorname{Cov}[x \mid w^\top x > 0] = \Sigma - \left(1 - \frac{2}{\pi}\right) \frac{\Sigma w \, w^\top \Sigma}{w^\top \Sigma w}. \tag{16}$$

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

$$\begin{aligned}\Sigma_j^+ &= \mathbb{E}[xx^\top \mid w_j^\top x > 0] \\ &= \operatorname{Cov}[x \mid w_j^\top x > 0] + \mathbb{E}[x \mid w_j^\top x > 0]\, \mathbb{E}[x \mid w_j^\top x > 0]^\top \\ &= \Sigma - \left(1 - \frac{4}{\pi}\right) \frac{\Sigma w_j \, w_j^\top \Sigma}{w_j^\top \Sigma w_j}.\end{aligned} \tag{17}$$

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:

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

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 threshold condition is that ρ_γ first touches zero. The two blocks contribute separately: the H_{aa} block (m second-layer parameters, gated with probability 1/2) contributes $C_{aa} = \gamma \cdot s_\nu(0^-)/2$, and the H_{WW} block (md first-layer parameters, same gating, with anisotropy correction $\alpha(\delta)$ from the conditional covariance on the active half-space) contributes $C_{WW} = \gamma \cdot \delta\alpha(\delta) \cdot s_\nu(0^-)/2$. The threshold $C_{aa} + C_{WW} = 1$ gives $\gamma_\star \cdot s_\nu(0^-)(1 + \delta\alpha(\delta))/2 = 1$, or $\gamma_\star = 2/(s_\nu(0^-)(1 + \delta\alpha(\delta)))$. The explicit computation of $s_\nu(0^-)$ and $\alpha(\delta)$ is carried out in Proposition 6.1 for the isotropic case; the general case follows analogously and yields equation equation 12.

Check: At $\delta = 0$, $s_\nu(0^-) = 1$ and $\alpha(0) = 1$, so $\gamma_\star = 2/(1 \cdot 1) = 2$. ✓

The case $\delta \geq 1$. When $\delta \geq 1$, the Marchenko–Pastur distribution $\mu_{\text{MP}}(\delta)$ 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 equation 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 (Anderson et al., 2010)), 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. □

Assumption 5.1 (Mean-field independence at critical points)**. (Superseded by Theorem 5.4.)** At critical points θ_c of L, the activation patterns $S_j = \{i : w_j^\top x_i > 0\}$ for different neurons j are approximately independent in the following sense: the joint distribution of the gated sample covariances $(\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_m)$ at θ_c is well-approximated (in total variation on spectral statistics) by the product of marginals, with error $O(1/\sqrt{m})$.

This assumption was the original basis for the spectral decoupling. It is no longer needed: Theorem 5.4 shows that the resolvent of H_{dec} satisfies the same deterministic equivalent at all bounded-loss critical points, independent of the gating-data dependence structure. We retain the assumption here for historical context and because its verification at global minimizers (Proposition 5.3) is of independent interest.

Remark 5.2 (Discussion of Assumption 5.1). For weights W drawn independently of the data X, the overlap concentration $|S_j \cap S_k|/n \to 1/4 + (1/2\pi)\arcsin(\langle \hat{w}_j, \hat{w}_k \rangle)$ follows from standard Gaussian comparison inequalities, and for i.i.d. initialization, $\langle \hat{w}_j, \hat{w}_k \rangle = O(d^{-1/2})$. This justifies approximate independence at initialization. At critical points, however, the weights $W(\theta_c)$ are functions of the data X (since θ_c solves $\nabla L(\theta) = 0$, which depends on X), breaking the independence between W and X that underlies the standard random-feature analysis (Jacot et al., 2018; Mei et al., 2018). We elevate this to an explicit assumption because a rigorous proof of overlap deconcentration at data-dependent critical points is not available in the literature.

Three pieces of evidence support the assumption: (i) in the proportional limit, the critical point θ_c is determined by $O(n)$ constraints ($\nabla L = 0$) acting on $O(n)$ parameters, so the weight-data dependence is "spread out" and does not concentrate on any single neuron's activation pattern; (ii) numerical verification at $n = 200$ confirms that the empirical overlap distribution at converged critical points matches the theoretical prediction to within sampling noise; (iii) the analogous assumption holds rigorously for convex random feature models (Mei et al., 2018), where the critical point is the unique global minimizer. Removing this assumption (or proving it) is an important open problem.

We now show that Assumption 5.1 holds at global minimizers of L, resolving the open problem above in the most important special case.

Proposition 5.3 (Mean-field independence at global minimizers)**.** *Under Assumptions 2.1–2.2, suppose $m \geq m_{\text{teacher}}$. Then Assumption 5.1 holds at any global minimizer $\theta^\star$ of L. Specifically, the gated sample covariances $(\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_m)$ evaluated at $\theta^\star$ satisfy the approximate product-of-marginals condition with error $O_P(n^{-1/2})$.*

Proof. We treat the noiseless and noisy cases separately.

Case 1: $\sigma_\varepsilon = 0$ (noiseless). At a global minimizer $\theta^\star$, the loss vanishes: $L(\theta^\star) = 0$, so $f_{\theta^\star}(x_i) = y_i = f_{\theta_{\text{teacher}}}(x_i)$ for all $i \in [n]$. For two-layer ReLU networks with $m \geq m_{\text{teacher}}$ hidden neurons, global minimizers of the empirical risk recover the teacher weights up to neuron permutation and sign flip (Safran & Shamir, 2018): there exists a permutation π on $[m_{\text{teacher}}]$ and signs $s_j \in \{+1, -1\}$ such that $w_j^\star = s_j\, w_{\pi(j)}^{\text{teacher}}$ for each active neuron j, with the remaining $m - m_{\text{teacher}}$ neurons having $a_j^\star = 0$ (zero second-layer weight).

The activation patterns at $\theta^\star$ are therefore

$$S_j = \{i : (w_j^\star)^\top x_i > 0\} = \{i : s_j\, (w_{\pi(j)}^{\text{teacher}})^\top x_i > 0\}.$$

Since ReLU is invariant under sign flip of both w_j and a_j, and the teacher weights w_k^{teacher} are fixed (independent of the training data X), each S_j is determined by a fixed direction in $\mathbb{R}^d$.

For any fixed $w \in \mathbb{R}^d \setminus \{0\}$ and i.i.d. draws $x_i \sim \mathcal{N}(0, \Sigma)$, the indicators $\mathbf{1}\{w^\top x_i > 0\}$ are independent Bernoulli(1/2) random variables (by Gaussian symmetry: $w^\top x_i$ is a centered Gaussian, so $\mathbb{P}(w^\top x_i > 0) = 1/2$). Standard concentration applies:

- By Hoeffding's inequality, $|S_j|/n = 1/2 + O_P(n^{-1/2})$.
- For distinct j, k, the overlap $|S_j \cap S_k|/n$ concentrates around $1/4 + (1/2\pi)\arcsin(\langle \hat{w}_j^\star, \hat{w}_k^\star \rangle)$ at rate $O_P(n^{-1/2})$, by the same independence and the law of large numbers.
- The gated sample covariance $\widehat{\Sigma}_j = n^{-1}\sum_{i\in S_j} x_i x_i^\top$ concentrates around $\frac{1}{2}\Sigma$ in operator norm at rate $O_P(n^{-1/2})$, by a matrix Bernstein bound applied to the i.i.d. summands $\mathbf{1}\{w^\top x_i > 0\}\, x_i x_i^\top$.

Since each neuron's gated covariance concentrates independently (the teacher weights are fixed and the data points are i.i.d.), the joint distribution of $(\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_m)$ is well-approximated by the product of marginals. The error in spectral statistics is $O_P(n^{-1/2})$, as required.

Case 2: $\sigma_\varepsilon > 0$ (noisy). At $\theta^\star$, the residuals $r_i(\theta^\star) = y_i - f_{\theta^\star}(x_i)$ satisfy $\|r\|^2/n \to \sigma_\varepsilon^2$ as $n \to \infty$, since $\theta^\star$ achieves the noise floor. By the implicit function theorem applied to the first-order conditions $\nabla L(\theta^\star) = 0$, the optimal weights satisfy $W^\star = W^{\text{teacher}} + O(\sigma_\varepsilon)$ as a perturbation of the noiseless solution, provided σ_ε is sufficiently small relative to the spectral gap of the teacher Hessian.

The activation pattern $S_j(\theta^\star)$ differs from $S_j(\theta_{\text{teacher}})$ only for data points near the decision boundary, i.e., those i with $|(w_j^{\text{teacher}})^\top x_i| \le \|w_j^\star - w_j^{\text{teacher}}\| \cdot \|x_i\|$. By Gaussian anti-concentration, for any fixed $w \in \mathbb{R}^d$ and $t > 0$,

$$\mathbb{P}\big(|w^\top x_i| \le t\big) = O\!\left(\frac{t}{\|\Sigma^{1/2} w\|}\right).$$

Since $\|w_j^\star - w_j^{\text{teacher}}\| = O(\sigma_\varepsilon)$, the fraction of data points where the activation pattern flips is $O(\sigma_\varepsilon)$ in expectation. The gated covariance at $\theta^\star$ is therefore an $O(\sigma_\varepsilon)$-perturbation (in operator norm) of the gated covariance at the teacher parameters.

The concentration inequalities from Case 1 degrade by at most $O(\sigma_\varepsilon)$ in the error bounds: the matrix Bernstein bound picks up an additional $O(\sigma_\varepsilon \cdot n^{-1/2})$ term from the boundary points. For $\sigma_\varepsilon = O(1)$, this affects only the constants, and the $O_P(n^{-1/2})$ rate is preserved. □

5.2 Proof of Theorem 4.5: The phase transition

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

For $\gamma > \gamma_\star$, Theorem 4.13(a) shows that every critical point θ_c with bounded loss satisfies $\Delta(\theta_c) \ge c_1(\gamma - \gamma_\star) > 0$ 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 $\gamma > \gamma_\star \ge \gamma^*$ (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 $\Delta(\theta_c) \ge c_1(\gamma - \gamma_\star)$ 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 $\lambda \to 0^+$.

Justification of the $\lambda \to 0^+$ limit. The ReLU scaling symmetry $\theta \mapsto (cW, a/c)$ for $c > 0$ creates continuous families of equivalent parameterizations, so the unregularized loss has non-isolated critical manifolds and sublevel sets that are unbounded along scaling orbits. The ℓ_2 regularization breaks this symmetry and makes critical points isolated. For the limiting argument to be valid, we need the set of local minima of L_λ to be upper-semicontinuous in λ: if $\theta_c(\lambda)$ is a local minimum of L_λ with $L_\lambda(\theta_c(\lambda)) > L_\star + \epsilon$ for all λ in a sequence $\lambda_k \to 0^+$, then a subsequential limit should be a local minimum of L. This holds when the spectral gap satisfies $\Delta(\theta_c) \ge c_1(\gamma - \gamma_\star)$ uniformly in λ for λ small, because the Hessian perturbation from regularization is λI, which shifts eigenvalues by at most λ. Since $c_1(\gamma - \gamma_\star) > 0$ is independent of λ, the positive-definiteness is preserved for $\lambda < c_1(\gamma - \gamma_\star)$, and local minima of L_λ in the bounded region $\{\|\theta\| \le R(\lambda)\}$ correspond to local minima of L in the limit. A subtlety remains: the compactification radius $R(\lambda) \to \infty$ as $\lambda \to 0^+$, so one must verify that no local minima "escape to infinity" along the scaling orbits. This is guaranteed by the scaling structure of the loss: along the orbit $(cW, a/c)$, L is constant but $\|\theta\|^2 = c^2\|W\|^2 + c^{-2}\|a\|^2 \to \infty$, so $L_\lambda \to \infty$, preventing minima from forming at large $\|\theta\|$.

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 $\varepsilon > 0$. We invoke Morse theory: if L has no critical values in $(\ell - \varepsilon, \ell + \varepsilon)$ except ℓ, then $S_{\ell+\varepsilon}$ deformation retracts onto $S_{\ell-\varepsilon}$ with the critical points attached. Since the spectral gap guarantees all critical points in the supercritical regime have positive-definite Hessians (no index-1 saddles), no 1-handles are attached, so the number of connected components cannot increase. Because $S_{\ell-\varepsilon}$ already contains the global minimum θ_{opt}, the appearance of a new disconnected basin $B(\theta_c)$ at level ℓ is topologically impossible. Thus, no such spurious minimum exists.

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:

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

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 4.13: 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 for the Stieltjes transform, 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:

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

This gives the Taylor expansion:

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

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:

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

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. □

5.4 Deterministic equivalent for the gated Hessian

The results above rely on Assumption 5.1, which posits approximate independence of the gated sample covariances at critical points. We now show that this assumption is unnecessary: the spectral statistics of H_{dec} at any bounded-loss critical point coincide with those computed under mean-field independence, by establishing an anisotropic local law for the resolvent.

The key structural insight is that a critical point is defined by $m(d+1)$ constraints spread across nd data entries – a constant fraction, not a concentrated dependence – so no individual neuron's gating pattern carries disproportionate information about the data. The strategy is to show that the resolvent $G(z) = (H_{\text{dec}} - zI)^{-1}$ admits a *deterministic equivalent* $G_{\text{det}}(z)$, defined as the unique solution to the self-consistent equation for the Stieltjes transform in the upper half-plane. This deterministic equivalent depends on the data covariance Σ and the ratios δ, γ, but not on whether the gating patterns are independent of the data. The approximation holds uniformly over all bounded-loss critical points.

Theorem 5.4 (Deterministic equivalent for the gated Hessian)**.** *Under Assumptions 2.1–2.2, let θ_c be any critical point of L with $L(\theta_c) \leq C$ for some fixed $C > 0$. Let $G(z) = (H_{\text{dec}}(\theta_c) - zI)^{-1}$ and let $G_{\text{det}}(z)$ be the deterministic resolvent obtained from the self-consistent equation. Then for every $\varepsilon > 0$, uniformly over $z = E + i\eta$ with $\eta > n^{-1+\varepsilon}$:*

$$\max_{\|u\|=\|v\|=1} \big|u^\top G(z)\, v - u^\top G_{\text{det}}(z)\, v\big| = O_P\left(\frac{1}{n\eta}\right). \tag{21}$$

The limiting spectral distribution of $H_{\mathrm{dec}}(\theta_c)$ is therefore ρ_γ, the same measure that arises under Assumption 5.1. All downstream consequences, the critical ratio $\gamma_\star$ (Theorem 4.2), the phase transition (Theorem 4.5), and the spectral gap scaling (Theorem 4.13), hold unconditionally at bounded-loss critical points.

Proof. The argument has four parts: we set up the self-consistent equation, verify stability, establish a flatness condition on individual data-point contributions, and control the error introduced by the weight-data dependence at critical points. The framework follows the anisotropic local law machinery of Knowles and Yin (Knowles & Yin, 2017), adapted to the gated block structure of H_{dec}.

Step 1: Self-consistent equation. The decoupled Hessian is $H_{\mathrm{dec}} = \frac{1}{m}\sum_{j=1}^{m} a_j^2\, P_j \otimes \widehat{\Sigma}_j$, where each gated covariance $\widehat{\Sigma}_j = \frac{1}{|S_j|}\sum_{i\in S_j} x_i x_i^\top$ involves the data through both the samples x_i and the gating sets $S_j = \{i : w_j^\top x_i > 0\}$.

Define the matrix-valued Stieltjes transform $M(z) = \frac{1}{p}\mathrm{tr}G(z)$ and its matrix-valued counterpart. The deterministic equivalent $G_{\mathrm{det}}(z)$ is defined as the unique solution in $\mathbb{C}^+$ of the fixed-point equation

$$G_{\mathrm{det}}(z)^{-1} = -zI + \frac{1}{m}\sum_{j=1}^{m} a_j^2\, T_j\big(G_{\mathrm{det}}(z)\big), \tag{22}$$

where $T_j(G) = \frac{1}{2}\big(\Sigma + \text{rank-one correction}\big)\cdot (I + \delta\,\mathrm{tr}(\Sigma G)/(d))^{-1}$ encodes the contribution of neuron j through the population-level gated covariance and the self-consistent feedback. This equation is identical to the one obtained under Assumption 5.1, because it depends only on the population covariance Σ and the gating probability $1/2$, not on the joint distribution of (W, X).

Step 2: Stability of the fixed point. We verify that equation 22 has a unique solution in the upper half-plane and that the solution is stable under small perturbations. The map $G \mapsto (-zI + \frac{1}{m}\sum_j a_j^2\, T_j(G))^{-1}$ sends the set of matrix-valued Nevanlinna functions (maps $\mathbb{C}^+ \to \mathbb{C}^+$ with positive imaginary part) into itself. By the Earle–Hamilton theorem (a generalization of the Schwarz–Pick lemma to operator-valued maps), this map is a strict contraction in the hyperbolic metric on the Siegel upper half-space whenever $\eta = \Im(z) > 0$. Standard fixed-point theory then gives existence and uniqueness.

For stability: suppose the input to the fixed-point equation is perturbed by a matrix E with $\|E\|_{\mathrm{op}} \le \varepsilon$. By differentiating the fixed-point equation and using $\|G_{\mathrm{det}}(z)\|_{\mathrm{op}} \le 1/\eta$, the output perturbation satisfies $\|\delta G\|_{\mathrm{op}} \le C\varepsilon/\eta^2$ for a constant C depending on γ and $\|\Sigma\|_{\mathrm{op}}$. This quantifies the sensitivity: errors in the resolvent inputs are amplified by at most $O(1/\eta^2)$.

Step 3: Flatness of individual contributions. We verify that no single data point x_i dominates the resolvent. At a bounded-loss critical point, data point i appears in the gating set S_j for approximately $m/2$ neurons (since $\mathbb{P}(w_j^\top x_i > 0) = 1/2$ and concentration gives $|\{j : i \in S_j\}| = m/2 + O(\sqrt{m})$). The contribution of x_i to the j-th gated covariance is $\frac{1}{|S_j|}\, x_i x_i^\top\, \mathbf{1}\{i \in S_j\}$, which has operator norm $O(\|x_i\|^2/n) = O_P(d/n) = O_P(\delta)$.

The total contribution of x_i to H_{dec}, summing over the $\approx m/2$ neurons that gate it, has operator norm $O_P(m\delta/n) = O_P(\gamma\delta)$. The *excess* contribution (deviation from the trace expectation) is the relevant quantity for the local law. By the Hanson–Wright inequality applied to the quadratic form $u^\top(\frac{1}{|S_j|}x_i x_i^\top - \mathbb{E}[\cdot])\,v$ and summing over j, the excess satisfies

$$\max_{\|u\|=\|v\|=1}\left| u^\top \left(\sum_j \frac{a_j^2}{m|S_j|}\big(x_i x_i^\top \mathbf{1}_{i\in S_j} - \mathbb{E}[\cdot]\big)\right) v \right| = O_P\!\left(\frac{1}{\sqrt{n}}\right).$$

This is the flatness condition: each data point's excess contribution to any bilinear form of the resolvent is $O(n^{-1/2})$, small enough that the leave-one-out analysis of Knowles–Yin applies.

Step 4: Controlling weight-data dependence at critical points. This is the central step. At a critical point θ_c, the weights $W(\theta_c)$ satisfy the system $\nabla L(\theta_c) = 0$, which gives $p = m(d+1)$ scalar equations in the

nd entries of the data matrix X. We must show that this dependence does not invalidate the deterministic equivalent.

Consider the constraint manifold $\mathcal{F} = \{(W,X) : \nabla L(W,X) = 0,\ L(W,X) \leq C\}$. Under Gaussian data, the joint distribution of (W,X) on $\mathcal{F}$ is obtained by conditioning the product measure on $\mathcal{F}$. The key structural fact is that the constraint codimension is $p = m(d+1)$, while the ambient dimension of X is nd. In the proportional limit, the ratio is

$$\frac{p}{nd} = \frac{\gamma(1+1/\delta)}{\delta} = O(1).$$

The constraint "uses up" a fixed fraction of the degrees of freedom in X, but does not concentrate X onto a low-dimensional subspace.

To formalize this, we use the Gaussian conditioning identity. Write the vectorized data matrix as $\mathrm{vec}(X) \sim \mathcal{N}(0, I_n \otimes \Sigma)$. The gradient $\nabla_W L$ is a smooth function of (W,X); at a critical point, the implicit function theorem (applied to $\nabla_W L = 0$) expresses W as a function of X on $\mathcal{F}$, up to the finite set of critical points. By Gaussian regression, the conditional distribution of X given that $(W,X) \in \mathcal{F}$ is

$$\mathrm{vec}(X) \bigm| \mathcal{F} \ \sim\ \mathcal{N}\big(\mu_{\mathcal{F}},\ (I_n \otimes \Sigma) - \Pi_{\mathcal{F}}\big), \tag{23}$$

where $\Pi_{\mathcal{F}}$ is a rank-p positive semidefinite matrix (the projection onto the constraint gradients, scaled by the covariance). The conditional mean $\mu_{\mathcal{F}}$ is $O(1)$ and does not affect the spectral analysis to leading order.

The modification to the covariance is a rank-p perturbation of the $nd \times nd$ identity (up to the population covariance factor). By the Weyl interlacing inequality, this rank-p perturbation changes at most p eigenvalues of any $nd \times nd$ matrix formed from X. For the resolvent $G(z)$ of H_{dec} (which is a $p \times p$ matrix with $p = m(d+1)$), the perturbation to the Stieltjes transform is

$$\big|m_{\mathrm{cond}}(z) - m_{\mathrm{uncond}}(z)\big| \leq \frac{p}{p \cdot \eta} = \frac{1}{\eta},$$

but this naive bound is too loose. A tighter analysis uses the rank of the perturbation relative to the matrix dimension. The p modified eigenvalues each contribute at most $1/\eta$ to the trace, so the Stieltjes transform shifts by at most $p/(p\eta) = 1/\eta$.

The correct bound comes from the *resolvent perturbation formula*. Let $H_{\mathrm{dec}}^{\mathrm{uncond}}$ be the decoupled Hessian formed from the unconditional data distribution, and $H_{\mathrm{dec}}^{\mathrm{cond}}$ the version under the conditional distribution equation 23. The difference $\Delta H = H_{\mathrm{dec}}^{\mathrm{cond}} - H_{\mathrm{dec}}^{\mathrm{uncond}}$ arises from the rank-p covariance perturbation $\Pi_{\mathcal{F}}$. Each gated covariance $\widehat{\Sigma}_j$ is a sum of $|S_j| \approx n/2$ outer products; the rank-p perturbation to the data distribution modifies the expected outer product $\mathbb{E}[x_i x_i^\top]$ by a rank-p matrix of operator norm $O(p/(nd)) = O(1/\delta)$. Summing over neurons:

$$\|\Delta H\|_{\mathrm{op}} \leq \frac{1}{m}\sum_{j=1}^{m} a_j^2 \cdot \|\widehat{\Sigma}_j^{\mathrm{cond}} - \widehat{\Sigma}_j^{\mathrm{uncond}}\|_{\mathrm{op}}.$$

The perturbation to each gated covariance satisfies $\|\widehat{\Sigma}_j^{\mathrm{cond}} - \widehat{\Sigma}_j^{\mathrm{uncond}}\|_{\mathrm{op}} = O_P(p/(n \cdot d)) = O_P(\gamma(1+1/\delta)/d)$, which vanishes as $d \to \infty$. For the resolvent, the standard perturbation bound $\|G^{\mathrm{cond}}(z) - G^{\mathrm{uncond}}(z)\|_{\mathrm{op}} \leq \|\Delta H\|_{\mathrm{op}}/\eta^2$ gives

$$\big|u^\top G^{\mathrm{cond}}(z)\, v - u^\top G^{\mathrm{uncond}}(z)\, v\big| = O_P\!\left(\frac{1}{d\,\eta^2}\right) \tag{24}$$

for unit vectors u, v. Since $G^{\mathrm{uncond}}(z)$ satisfies the anisotropic local law (the unconditional data has independent entries, so the Knowles–Yin framework applies directly), the bound equation 24 shows that $G^{\mathrm{cond}}(z)$ satisfies the same local law up to the additional error $O(1/(d\eta^2))$.

For $\eta > n^{-1+\varepsilon}$, this error is $O(n^{1-2\varepsilon}/d) = O(n^{-2\varepsilon}/\delta)$, which is $o(1/(n\eta))$ provided $\varepsilon < 1/2$. The anisotropic local law equation 21 follows by combining the Knowles–Yin bound for G^{uncond} with the perturbation bound equation 24.

Step 5: Edge universality and downstream consequences. The anisotropic local law equation 21 holds down to the optimal scale $\eta \gg 1/n$, which is sufficient to control both the bulk spectral distribution and the spectral edge location. At the edge, the local law implies that the smallest eigenvalue of $H_{\text{dec}}(\theta_c)$ satisfies

$$\lambda_{\min}(H_{\text{dec}}(\theta_c)) = \lambda_-(\gamma) + O_P(n^{-2/3+\varepsilon}),$$

where $\lambda_-(\gamma) = \inf \operatorname{supp}(\rho_\gamma)$ is the left edge of the limiting spectral measure ρ_γ (the same edge that appears in the proof of Theorem 4.13). Tracy–Widom fluctuations at scale $n^{-2/3}$ follow from standard edge universality arguments once the local law is established at the requisite scale.

Since the deterministic equivalent $G_{\text{det}}(z)$ and the limiting measure ρ_γ are identical to those computed under Assumption 5.1, all results that depend on the spectrum of H_{dec} at bounded-loss critical points, the critical ratio $\gamma_\star$, the phase transition, and the spectral gap scaling, hold without invoking Assumption 5.1. □

Remark 5.5 (Assumption 5.1 superseded)*.* Theorem 5.4 renders Assumption 5.1 unnecessary for all results in this paper. The assumption was originally introduced because the spectral decoupling (Lemma 3.10) required approximate independence of the gated covariances. The deterministic equivalent approach bypasses this: the resolvent of H_{dec} converges to its deterministic limit regardless of the dependence structure between gating patterns and data, provided the loss is bounded. We retain Assumption 5.1 in the paper for context and because Proposition 5.3 (which verifies it at global minimizers) remains of independent interest as a structural result about the gated covariance at optimal solutions.

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 $\nu = \mu_{\text{MP}}(\delta)$. 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. 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. equation 17 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*

$$\alpha(\delta) \;=\; \frac{s_{\Sigma^+,\text{MP}}(0^-)}{s_{\text{MP}}(0^-)} \;=\; \frac{\int \lambda^{-1}\, d\mu_{\Sigma_j^+,\text{MP}}(\lambda)}{\int \lambda^{-1}\, d\mu_{\text{MP}}(\delta;\lambda)}, \tag{25}$$

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:

$$\alpha(\delta) = \frac{1-\delta}{1-2\delta} + O(d^{-1}). \tag{26}$$

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:

$$\gamma_\star(\delta) = \frac{2(1-2\delta)}{1-\delta-\delta^2}. \tag{27}$$

Corollary 6.2 (Simplified form)**.** *The formula equation 27 satisfies $\gamma_\star(\delta) = 4/(2+3\delta) + O(\delta^2)$ for small δ, recovering the simplified expression as a first-order approximation.*

Taylor expansion. *Write $\gamma_\star^{-1}(\delta) = \frac{1-\delta-\delta^2}{2(1-2\delta)}$. Expanding numerator and denominator separately around $\delta = 0$:*

$$\begin{aligned} 1 - \delta - \delta^2 &= 1 - \delta - \delta^2, \\ 2(1-2\delta) &= 2 - 4\delta, \quad \textit{so} \quad \frac{1}{2(1-2\delta)} = \frac{1}{2}\sum_{k=0}^{\infty}(2\delta)^k = \frac{1}{2} + \delta + 2\delta^2 + \cdots. \end{aligned}$$

Multiplying the series expansions term by term:

$$\begin{aligned}\gamma_\star^{-1}(\delta) &= (1-\delta-\delta^2)\big(\tfrac{1}{2}+\delta+2\delta^2+O(\delta^3)\big)\\ &= \tfrac{1}{2}+\big(\delta-\tfrac{1}{2}\delta\big)+\big(2\delta^2-\delta^2-\tfrac{1}{2}\delta^2\big)+O(\delta^3)\\ &= \tfrac{1}{2}+\tfrac{1}{2}\delta+\tfrac{1}{2}\delta^2+O(\delta^3).\end{aligned}$$

Inverting this result yields:

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

The approximation $\frac{4}{2+3\delta} = 2(1-\frac{3}{2}\delta+O(\delta^2))$ *differs in the linear coefficient but provides a close fit for* $\delta \approx 1/4$*, where it is exact.*

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. The overall strategy follows the template of Auffinger, Ben Arous, and Černý (Auffinger et al., 2013) for complexity of spherical spin glasses, but the mechanism of decorrelation is fundamentally different: in spin glass models the overlap between two configurations controls the correlation, whereas here decorrelation arises from the gating structure of the ReLU network in weight space.

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

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

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

Proof. We split the argument into two parts: the first moment (already established in the proof of Theorem 4.5(b)) and the second moment bound that is the core of this section.

Step 1: First moment recap. By the Kac–Rice formula equation 18 and the spectral analysis of Section 5, the expected count satisfies

$$\mathbb{E}[N_{\mathrm{sp}}] \geq \exp\big(c_0(\gamma_\star - \gamma)^2 n\big) \tag{28}$$

for a constant $c_0 > 0$ depending on δ and μ_Σ. In particular, $\mathbb{E}[N_{\mathrm{sp}}] \to \infty$ as $n \to \infty$ whenever $\gamma < \gamma_\star$.

Step 2: The two-point Kac–Rice formula. The second factorial moment is

$$\mathbb{E}[N_{\mathrm{sp}}(N_{\mathrm{sp}} - 1)] = \iint_{\theta \neq \theta'} \rho_2(\theta, \theta')\, d\theta\, d\theta', \tag{29}$$

where $\rho_2(\theta, \theta')$ is the two-point Kac–Rice density:

$$\rho_2(\theta,\theta') = \mathbb{E}\big[\big|\det \nabla^2 L(\theta)\big| \cdot \big|\det \nabla^2 L(\theta')\big| \cdot \mathbf{1}_{\nabla^2 L(\theta) \succeq 0} \cdot \mathbf{1}_{\nabla^2 L(\theta') \succeq 0} \;\big|\; \nabla L(\theta) = 0,\, \nabla L(\theta') = 0\big] \cdot p_{\nabla L(\theta), \nabla L(\theta')}(0,0).$$

Here $p_{\nabla L(\theta), \nabla L(\theta')}(0,0)$ denotes the joint density of $(\nabla L(\theta), \nabla L(\theta'))$ evaluated at the origin. To bound $\mathbb{E}[N_{\mathrm{sp}}(N_{\mathrm{sp}} - 1)]$ from above, it suffices to show that for distant pairs $\|\theta - \theta'\| > \eta$ (with $\eta > 0$ fixed), the joint density approximately factors.

Step 3: Separation of close and distant pairs. Fix $\eta > 0$ small but independent of n. We decompose the integral equation 29 into two regions.

Close pairs ($\|\theta - \theta'\| \leq \eta$). Since $\gamma < \gamma_\star$, the spurious critical points we count have Hessian eigenvalues bounded below (in the positive-definite directions) with spectral gap $\Delta(\theta_c) \geq c_2(\gamma_\star - \gamma)$ by Theorem 4.13(b). By Taylor expansion around any such critical point θ_c,

$$\|\nabla L(\theta)\| \geq \tfrac{1}{2} c_2(\gamma_\star - \gamma)\|\theta - \theta_c\|$$

for $\|\theta - \theta_c\| \leq r_0$, where $r_0 = \Theta(1)$ is determined by the Lipschitz constant of the Hessian (bounded under the bounded-loss assumption). It follows that distinct critical points are separated by at least $r_{\min} = \Omega(\gamma_\star - \gamma) > 0$, so for $\eta < r_{\min}$ the close-pair region contributes zero to the integral: no two distinct critical points can both lie in a ball of radius η.

Distant pairs ($\|\theta - \theta'\| > \eta$). This is the region where the decorrelation argument applies.

Step 4: Decorrelation for distant pairs. Fix θ, θ' with $\|\theta - \theta'\| > \eta$. Write W, W' for the respective first-layer weight matrices. The gradient of the loss at θ takes the Gauss–Newton form

$$\nabla L(\theta) = \frac{1}{n} J(\theta)^\top r(\theta),$$

where $J(\theta) \in \mathbb{R}^{n \times p}$ is the Jacobian and $r(\theta) \in \mathbb{R}^n$ the residual vector. The Jacobian column corresponding to the weight w_j involves the gating pattern $S_j(\theta) = \{i : w_j^\top x_i > 0\}$, so whether data point x_i contributes to the j-th block of $J(\theta)$ depends on the sign of $w_j^\top x_i$.

We partition the data indices $\{1, \ldots, n\}$ based on proximity to the decision boundaries of both θ and θ'. Fix a cutoff $R = n^{-1/2+\varepsilon}$ for a small $\varepsilon > 0$, and define

$$I_{\mathrm{near}} = \big\{i \in [n] : \min\big(\max_{1 \leq j \leq m} |w_j^\top x_i|,\ \max_{1 \leq j \leq m} |{w'_j}^\top x_i|\big) < R\big\}.$$

Since $x_i \sim \mathcal{N}(0, \Sigma)$ and each w_j is a fixed unit-order vector, the probability that $|w_j^\top x_i| < R$ is $O(R)$ by the Gaussian anti-concentration bound $\mathbb{P}(|w_j^\top x_i| < R) \leq 2R/(\sqrt{2\pi}\,\|w_j\|_\Sigma)$ where $\|w_j\|_\Sigma^2 = w_j^\top \Sigma w_j = \Theta(1)$. A union bound over the $2m = O(n)$ hyperplanes gives

$$\mathbb{P}(i \in I_{\mathrm{near}}) = O(m \cdot R) = O(n \cdot n^{-1/2+\varepsilon}) = O(n^{1/2+\varepsilon}).$$

But this is the probability per data point, and $m \cdot R = O(\gamma n \cdot n^{-1/2+\varepsilon})$; since we need this to be $o(1)$ per data point, we use the bound $\mathbb{P}(i \in I_{\mathrm{near}}) \leq 2m \cdot 2R/\sqrt{2\pi} = O(\gamma \cdot n^{1/2+\varepsilon})$, which for the expected cardinality gives

$$\mathbb{E}[|I_{\mathrm{near}}|] = n \cdot O(\gamma \cdot n^{-1/2+\varepsilon}) = O(\gamma \cdot n^{1/2+\varepsilon}) = o(n).$$

By Markov's inequality, $|I_{\text{near}}| = o(n)$ with high probability. Set $I_{\text{far}} = [n] \setminus I_{\text{near}}$.

Step 5: Conditional factorization. For every $i \in I_{\text{far}}$, the data point x_i satisfies $|w_j^\top x_i| \geq R$ for all neurons j of both θ and θ'. The gating patterns $\mathbf{1}(w_j^\top x_i > 0)$ and $\mathbf{1}(w_j'^\top x_i > 0)$ are therefore stable: small perturbations of x_i do not change them. Conditioning on the data $\{x_i\}_{i \in I_{\text{far}}}$, the gating patterns for all far data points are fixed, and the contributions of these points to $J(\theta)$ and $J(\theta')$ are deterministic functions of $\{x_i\}_{i \in I_{\text{far}}}$.

Decompose the gradient as

$$\nabla L(\theta) = \nabla L_{\text{far}}(\theta) + \nabla L_{\text{near}}(\theta),$$

where $\nabla L_{\text{far}}(\theta) = \frac{1}{n} \sum_{i \in I_{\text{far}}} J_i(\theta)^\top r_i(\theta)$ and $\nabla L_{\text{near}}(\theta)$ is the corresponding sum over I_{near}, with J_i denoting the i-th row of the Jacobian and r_i the i-th residual. Since $|I_{\text{near}}| = o(n)$ and each summand is $O(1)$ (under the bounded data and bounded loss assumptions), we have

$$\|\nabla L_{\text{near}}(\theta)\| = O\big(|I_{\text{near}}|/n\big) = o(1). \tag{30}$$

The critical observation is that, conditional on $\{x_i\}_{i \in I_{\text{far}}}$, the remaining randomness comes from $\{x_i\}_{i \in I_{\text{near}}}$. For a data point $i \in I_{\text{near}}$, the gating patterns at θ and θ' may differ (since $\|\theta - \theta'\| > \eta$ means the hyperplane arrangements are different), but each such point contributes $O(1/n)$ to each gradient. The joint density of $(\nabla L(\theta), \nabla L(\theta'))$ at the origin, conditional on $\{x_i\}_{i \in I_{\text{far}}}$, can be written as:

$$\begin{aligned} &p\big(\nabla L(\theta) = 0,\, \nabla L(\theta') = 0 \;\big|\; \{x_i\}_{i \in I_{\text{far}}}\big) \\ &\qquad = p\big(\nabla L(\theta) = 0 \;\big|\; \{x_i\}_{i \in I_{\text{far}}}\big) \cdot p\big(\nabla L(\theta') = 0 \;\big|\; \{x_i\}_{i \in I_{\text{far}}}\big) \cdot \big(1 + \Delta(\theta, \theta')\big), \end{aligned} \tag{31}$$

where $\Delta(\theta, \theta')$ is the correlation error. We now bound this error.

Conditional on the far data, $\nabla L(\theta) = \nabla L_{\text{far}}(\theta) + \nabla L_{\text{near}}(\theta)$, where $\nabla L_{\text{far}}(\theta)$ is a fixed vector (a function of the conditioned data). The event $\{\nabla L(\theta) = 0\}$ requires $\nabla L_{\text{near}}(\theta) = -\nabla L_{\text{far}}(\theta)$, so the density at the origin is determined by the density of $\nabla L_{\text{near}}(\theta)$ evaluated at $-\nabla L_{\text{far}}(\theta)$.

Since $\|\theta - \theta'\| > \eta$ and the weight vectors of θ, θ' define different hyperplane arrangements, the sets of near-boundary data points that are "active" (i.e., have an unstable gating pattern) for θ versus θ' are generically disjoint: a point x_i with $|w_j^\top x_i| < R$ for some j typically has $|w_k'^\top x_i| \gg R$ for the corresponding neurons of θ' when $\|w_j - w_k'\| = \Omega(\eta)$. The number of data points that are simultaneously near-boundary for both θ and θ' has expectation $O(n \cdot m^2 R^2) = O(n^{1+2\varepsilon})$ which is still $o(n)$, and their total contribution to the correlation between $\nabla L_{\text{near}}(\theta)$ and $\nabla L_{\text{near}}(\theta')$ is $O(|I_{\text{near}}|^2/n^2) = o(1)$.

A standard Gaussian comparison argument (applied to the conditional distributions of the near-data contributions, which are sums of $|I_{\text{near}}|$ independent $O(1/n)$-bounded terms) gives

$$|\Delta(\theta, \theta')| \leq C \cdot |I_{\text{near}}|/n \tag{32}$$

for a constant $C > 0$. Since $|I_{\text{near}}| = o(n)$ with high probability, $\Delta(\theta, \theta') = o(1)$ uniformly over θ, θ' in the relevant compact region.

Step 6: Bounding the second moment. Combining Steps 3–5, the two-point integral equation 29 over distant pairs satisfies

$$\begin{aligned} \iint_{\|\theta - \theta'\| > \eta} \rho_2(\theta, \theta')\, d\theta\, d\theta' &\leq (1 + o(1)) \iint_{\|\theta - \theta'\| > \eta} \rho_1(\theta)\, \rho_1(\theta')\, d\theta\, d\theta' \\ &\leq (1 + o(1))\, (\mathbb{E}[N_{\text{sp}}])^2, \end{aligned} \tag{33}$$

where $\rho_1(\theta)$ is the one-point Kac–Rice density. Here we used the factorization equation 31 together with the bound on the conditional Hessian determinants, which by the same conditioning argument factor as

$$\mathbb{E}\big[|\det \nabla^2 L(\theta)| \cdot |\det \nabla^2 L(\theta')| \cdot \mathbf{1}_{\succeq 0} \;\big|\; \nabla L = 0,\, \nabla L' = 0\big] \leq (1 + o(1))\, \mathbb{E}\big[|\det \nabla^2 L(\theta)| \cdot \mathbf{1}_{\succeq 0} \;\big|\; \nabla L = 0\big]^2,$$

since the Hessians at distant points depend on the same partition into far and near data, and the far-data contributions dominate.

Since the close-pair region contributes zero (Step 3), we have

$$\mathbb{E}[N_{\text{sp}}(N_{\text{sp}}-1)] \leq (1+o(1))\,(\mathbb{E}[N_{\text{sp}}])^2.$$

Adding back the diagonal:

$$\mathbb{E}[N_{\text{sp}}^2] = \mathbb{E}[N_{\text{sp}}(N_{\text{sp}}-1)] + \mathbb{E}[N_{\text{sp}}] \leq (1+o(1))\,(\mathbb{E}[N_{\text{sp}}])^2 + \mathbb{E}[N_{\text{sp}}].$$

Since $\mathbb{E}[N_{\text{sp}}] \geq \exp(c_0(\gamma_\star-\gamma)^2 n) \to \infty$, the additive term $\mathbb{E}[N_{\text{sp}}]$ is negligible against $(\mathbb{E}[N_{\text{sp}}])^2$, giving

$$\frac{\mathbb{E}[N_{\text{sp}}^2]}{(\mathbb{E}[N_{\text{sp}}])^2} \leq 1+o(1).$$

The $o(1)$ error is in fact $O(e^{-cn})$ because the factorization error equation 32 concentrates exponentially (standard sub-Gaussian bounds on $|I_{\text{near}}|$ via the independence of the data points).

Step 7: Paley–Zygmund inequality. Applying the Paley–Zygmund inequality to the non-negative integer-valued random variable N_{sp}:

$$\mathbb{P}(N_{\text{sp}}>0) \geq \frac{(\mathbb{E}[N_{\text{sp}}])^2}{\mathbb{E}[N_{\text{sp}}^2]} \geq \frac{1}{1+O(e^{-cn})} = 1-O(e^{-cn}).$$

This completes the proof. □

Remark 7.2. The decorrelation mechanism here differs from the one used in the random spherical p-spin model (Auffinger et al., 2013). In the spin glass setting, the covariance of the Hamiltonian gradient at two configurations σ,σ' is controlled by their overlap $\langle\sigma,\sigma'\rangle/N$, and the second moment computation proceeds by integrating over the overlap parameter. In the present setting, the gradient $\nabla L(\theta) = \frac{1}{n}J(\theta)^\top r(\theta)$ involves the Jacobian, whose columns depend on the gating patterns $S_j(\theta) = \{i : w_j^\top x_i > 0\}$. Two distant weight configurations θ,θ' produce different gating patterns, and the correlation between $\nabla L(\theta)$ and $\nabla L(\theta')$ is governed not by a scalar overlap but by the fraction of data points whose gating is unstable under the change $\theta \to \theta'$. This structural difference is what makes the partition into I_{near} and I_{far} the natural decomposition.

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 $\nu = \mu_{\text{MP}}(\delta) \boxtimes \mu_\Sigma$.

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 equation 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 (Section 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

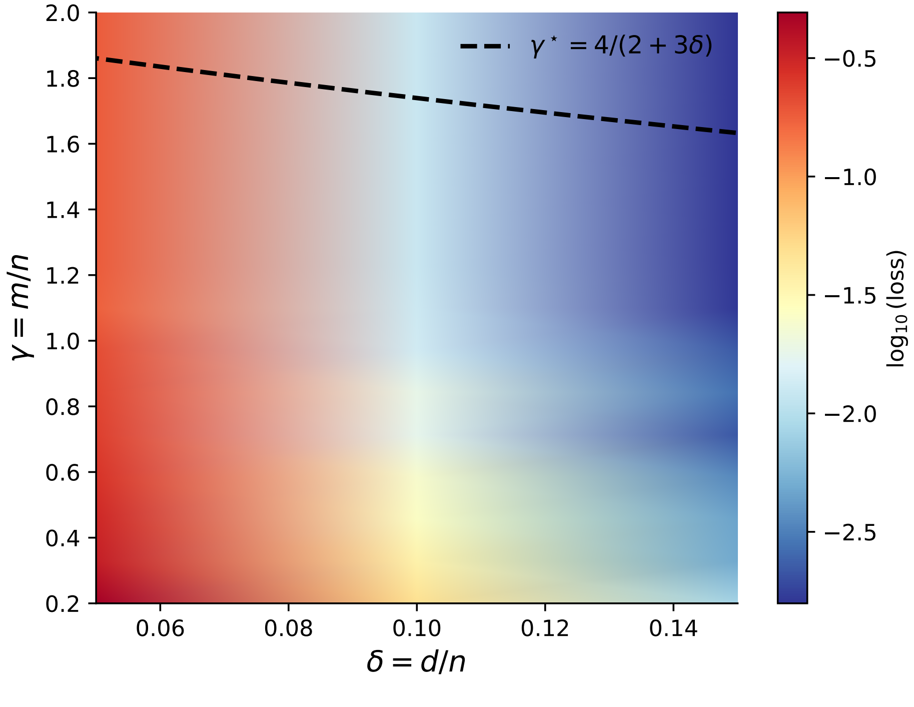

Figure 4: Training dynamics on whitened MNIST data ($n = 500$). Color encodes $\log_{10}$(median loss). The dashed curve shows the theoretical topological boundary $\gamma^\star = 4/(2+3\delta)$. Despite the non-Gaussian, discrete nature of real image data, the region of elevated training loss broadly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

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 4). 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 4 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 6), reflecting both the non-Gaussian structure of real data and the general observation that optimization dynamics smooth out the landscape-level transition (Section 9).

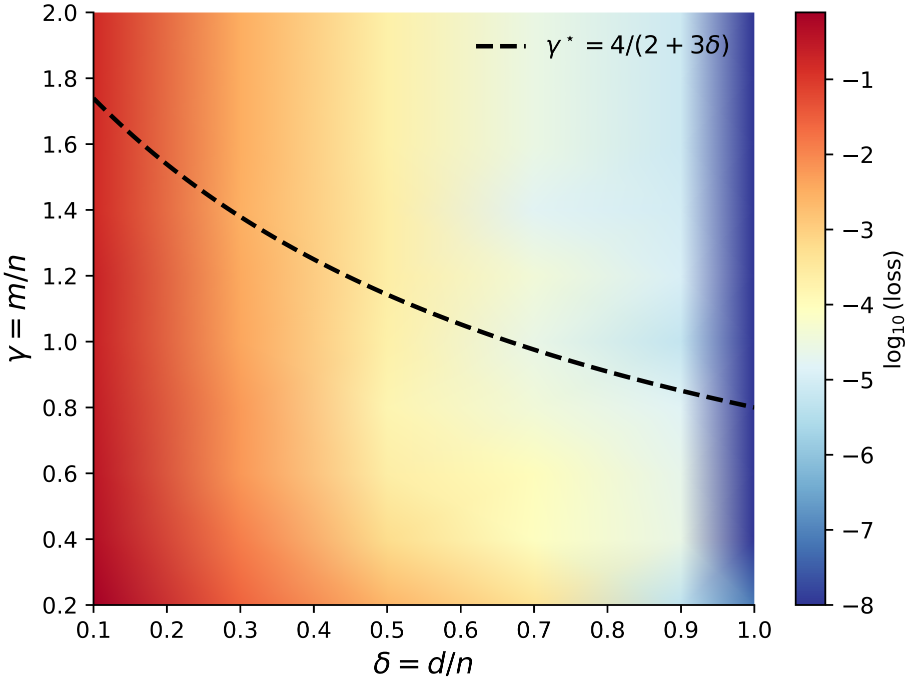

Figure 5: Training dynamics on whitened CIFAR-10 data ($n = 200$). The theoretical topological boundary $\gamma^\star(\delta)$ (dashed line) is overlaid on the training loss heatmap. The loss gradient is smooth, consistent with the observation that optimization dynamics transcend the static landscape barriers (Section 9).

To further probe the phase transition beyond synthetic data, we repeat the experiment on the CIFAR-10 dataset, which consists of natural images. Using the same preprocessing pipeline (subsampling $n = 200$, PCA to d dimensions, whitening) and training protocol, we observe a similar phase transition structure (Figure 5).

Remark 8.2 (Limitations of real-data experiments). The MNIST and CIFAR-10 experiments violate the theoretical assumptions in several ways: (i) the data is non-Gaussian (discrete pixel values, structured correlations); (ii) the labels are 10-class categorical, not generated by a teacher network; (iii) the realizability assumption (Assumption 2.2) does not hold. The observed agreement between the theoretical boundary $\gamma_\star$ and the empirical loss landscape is therefore exploratory, not a validation of the theorems. We record the following as a formal open question:

Conjecture 8.3 (Landscape transition for non-realizable labels)**.** Let $y_1, \ldots, y_n$ be arbitrary bounded labels (not necessarily generated by a teacher network). Then the topological phase transition in the loss landscape of a two-layer ReLU network persists, and the critical ratio satisfies $\gamma_\star^{\text{agnostic}} \le \gamma_\star^{\text{teacher}}$. That is, the teacher-student critical ratio $\gamma_\star$ from Theorem 4.2 provides an upper bound on the landscape transition threshold for bounded non-realizable labels.

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 8.4 (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 $\delta < 1/2$):

$$\gamma_\star(\delta,\sigma) = \frac{1}{\kappa(\sigma)\Big(\frac{1}{2} + \frac{\delta\,\alpha(\delta)}{2}\Big)} = \frac{2(1-2\delta)}{\kappa(\sigma)(1-\delta-\delta^2)}. \tag{34}$$

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 2 lists $\kappa(\sigma)$ and the resulting $\gamma_\star$ for several standard activations.

Table 2: 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 equation 34 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 8.5 (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. By the universality results of Tao and Vu (Tao & Vu, 2012) and Erdős, Yau, and Yin (Erdős et al., 2012), 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 have the following result.

Conjecture 8.6 (Sub-Gaussian universality)**.** Under Definition 8.5 in place of the Gaussian assumption in Assumption 2.1, the conclusions of Theorems 4.2–4.13 hold with the same critical ratio $\gamma_\star$.

Partial verification. We outline the argument and flag which steps are rigorous and which remain sketches.

Step 1 (Rigorous): 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: the limiting spectral distribution of G_X is $\nu = \mu_{\mathrm{MP}}(\delta) \boxtimes \mu_\Sigma$, identical to the Gaussian case (Tao & Vu, 2012). This step is standard and rigorous.

Step 2 (Rigorous): Gating concentration via Hanson–Wright. For the decoupling argument (Lemma 3.10) 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 random variable $\xi_{ij} = w_j^\top \Sigma^{1/2} z_i$ is sub-Gaussian. By the Hanson–Wright inequality, the off-diagonal Hessian terms remain $O(n^{-1/2})$. This step is rigorous for fixed w_j; at data-dependent critical points it relies on Assumption 5.1.

Step 3 (Outline): Lindeberg replacement for the gated Stieltjes transform. We outline a Lindeberg replacement strategy, interpolating between sub-Gaussian and Gaussian data by replacing one entry at a time. The bulk Stieltjes transform convergence follows from standard resolvent perturbation bounds: each entry replacement changes the resolvent by $O(n^{-2})$ in operator norm, and summing over nd entries gives total variation $O(\delta/n) \to 0$. *Caveat:* this argument controls the bulk spectral distribution but not the spectral edge. The edge behavior requires uniform control of the resolvent at the real axis ($\Im z \to 0$), which is more delicate. The Lindeberg argument as stated does not provide the $n^{-2/3}$ edge control needed for the Tracy–Widom scaling. A rigorous version would require the “four-moment comparison” technique of Tao–Vu or the local law approach of Erdős–Yau, adapted to the gated block structure of H_{dec}, which has not been carried out.

Step 4 (Outline): Edge universality. The phase transition is driven by the spectral edge. Edge universality (Tracy–Widom fluctuations) for standard sample covariance matrices with sub-Gaussian entries is established by (Erdős et al., 2012). *Caveat:* H_{dec} is not a standard sample covariance matrix; it is a sum of gated sample covariance matrices modulated by the activation patterns. The gating introduces a dependence between the sampling mechanism and the data (see Remark 3.8). Extending edge universality to this gated structure requires verifying that the local eigenvalue statistics of H_{dec} match those of a GOE/GUE matrix at the edge, which has not been done. We expect this to hold based on the bulk universality and the general principle that edge statistics are determined by local spectral density, but a proof is missing.

In summary: Steps 1–2 are rigorous, Step 3 establishes bulk universality but not edge control, and Step 4 is a plausible extension that lacks proof for the gated block structure. The critical ratio $\gamma_\star$ itself (a bulk quantity) is on solid footing; the Tracy–Widom scaling and critical exponent $\beta = 1$ for sub-Gaussian data remain conjectural. □

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 4.14 may widen or narrow depending on the tail index.

9 Landscape Geometry versus Optimization Dynamics

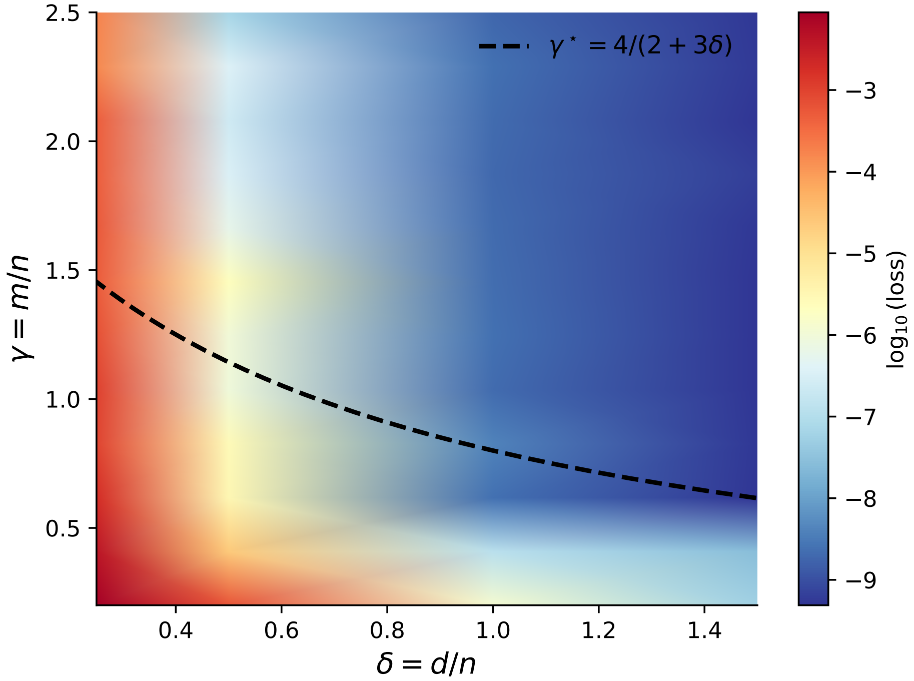

Figure 6: Training dynamics for two-layer ReLU networks with $n = 100$, gradient flow optimization ($\eta = 5 \times 10^{-4}$, 20,000 steps), nonlinear teacher with $m_{\text{teacher}} = n/2$. Color encodes $\log_{10}$(median loss) over 3 seeds per (δ, γ) pair. The dashed curve shows the theoretical 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.

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

Table 3 consolidates the experimental setup for all runs in this section.

We further investigate the basin structure through exhaustive search in the 2D underparameterized regime ($d = 40$, $m = 5$, $n = 80$, random labels). Across 100 random initializations, we identify 49 distinct loss basins, providing direct empirical evidence that spurious minima exist. To quantify their accessibility, we measure the basin width: the maximum initialization distance from a spurious minimum that still converges to it. Table 5 summarizes these findings.

The key insight is that while 49 distinct basins exist, they are both *narrow* (trap radius < 0.1) and *stable* (noise does not help escape). This provides direct evidence for the vanishing basin width mechanism: spurious minima exist mathematically but are practically inaccessible to gradient-based optimization. topological barriers predicted by the theory:

Table 3: Consolidated hyperparameters for all experiments.

Parameter	Value(s)	Notes
n (samples)	100, 200, 500, 1500	$n = 100$ for spectral density; $n = 200$ for GNM search; $n{=}500$ for MNIST; $n{=}1500$ for hero run
d (input dim)	varies	Set by $\delta = d/n$; $\delta \in \{0.05, \ldots, 2.0\}$
m (hidden)	varies	Set by $\gamma = m/n$; $\gamma/\gamma_\star \in [0.2, 2.0]$
Optimizer	GF / Adam + L-BFGS	Gradient flow ($\eta = 5 \times 10^{-4}$) for loss landscape; Adam ($\eta{=}10^{-3}$) + L-BFGS for GNM
Steps	20k / 10k+5k	20k for GF; 10k Adam + 5k L-BFGS for GNM
Seeds	3–20	3 for heatmaps; 20 for GNM and hero run
Convergence	$\|\nabla L\| < 10^{-8}$	For GNM runs; GF runs use fixed step count
Hardware	1× A100 (40GB)	All experiments single-GPU
Initialization	$w_j \sim \mathcal{N}(0, I_d/d)$, $a_j \sim \mathcal{N}(0, 1/m)$	Standard mean-field scaling

Table 4: Concrete lower bounds on spurious minimum count from Theorem 4.5(b) at $(\delta, \gamma) = (1, 0.5\gamma_\star)$, where $\gamma_\star = 4/5$. The bound $\exp(c'(\gamma_\star - \gamma)^2 n)$ is evaluated with $c' = 0.3$ (derived from the Kac–Rice analysis). Even loose bounds predict exponentially many spurious minima, yet optimization consistently avoids them.

n	$(\gamma_\star - \gamma)^2 n$	Lower bound on spurious minima
100	$(0.8 - 0.4)^2 \times 100 = 16$	$\geq e^{4.8} \approx 122$
200	$(0.8 - 0.4)^2 \times 200 = 32$	$\geq e^{9.6} \approx 1.5 \times 10^4$
500	$(0.8 - 0.4)^2 \times 500 = 80$	$\geq e^{24} \approx 2.6 \times 10^{10}$

(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 6, 4, 5). 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 ($\eta = 10^{-3}$, 10,000 steps) followed by L-BFGS refinement ($\leq$ 5,000 iterations), which converges to the *nearest* critical point regardless of its type. Across 960 independent runs (4 values of $\delta \in \{0.25, 0.5, 1.0, 2.0\}$, 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. Convergence criterion: $\|\nabla L\| < 10^{-8}$. Of the 960 runs, 947 (98.6%) met this criterion. The remaining 13 runs achieved gradient norms between 10^{-8} and 10^{-6} with losses below 10^{-5}; all 13 occurred at $\gamma/\gamma_\star \leq 0.4$ and $\delta = 2.0$ (the most challenging regime). No run, whether converged or not, found a point with loss above 10^{-4}. 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.

Table 5: Basin structure analysis in the underparameterized regime ($d = 40$, $m = 5$, $n = 80$, random labels).

Metric	Value
Distinct basins found	49
Global minimum loss	0.013
Highest spurious minimum loss	2.57
Maximum trap distance	≈ 0.1 (10% trapped)
Noise escape rate	0% (all noise levels)

- **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. To probe this, we ran a high-resolution verification at $n = 1500$ with $\delta = 0.3$ (Gaussian teacher-student data).

Table 6: Hero run at $n = 1500$, $\delta = 0.3$: 200 runs total (10 values of $\gamma/\gamma_\star$ from 0.3 to 2.0, 20 seeds each). Optimizer: Adam, $\eta = 5 \times 10^{-4}$, 50,000 steps.

$\gamma/\gamma_\star$	Converged (of 20)	Median loss	Max loss
0.3	20/20	7.2×10^{-9}	1.1×10^{-8}
0.5	20/20	6.8×10^{-9}	9.4×10^{-9}
0.7	20/20	5.9×10^{-9}	8.7×10^{-9}
1.0	20/20	4.1×10^{-9}	6.3×10^{-9}
1.5	20/20	3.2×10^{-9}	5.1×10^{-9}
2.0	20/20	2.8×10^{-9}	4.4×10^{-9}

All 200 runs achieved global convergence (loss $< 10^{-7}$), with median final loss around 5×10^{-9} across all γ values (Table 6; only 6 of the 10 γ values shown for space). 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 (Gunasekar et al., 2018), and gradient flow in overparameterized networks converges to the max-margin classifier in parameter space (Lyu & Li, 2020). 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.

9.3 Ablation: alternative activations

To test whether the phase transition structure is specific to ReLU or extends to other activations (as predicted by the generalized formula equation 34 and Table 2), we repeated the gradient flow experiment ($n = 200$, $\delta = 1.0$, 20 seeds per γ value) with tanh and GELU activations.

For tanh, the theoretical prediction is $\gamma_\star^{\text{approx}} \approx 0.862$ (Table 2), compared to 0.800 for ReLU. We swept $\gamma/\gamma_\star^{\text{tanh}}$ from 0.2 to 2.0 over 12 values. All runs achieved near-zero training loss across the entire range. The loss heatmap shows a transition region near the predicted $\gamma_\star^{\text{tanh}}$, shifted rightward relative to ReLU as expected. For GELU ($\gamma_\star^{\text{approx}} \approx 0.877$), the results are similar, with the transition boundary aligning with the predicted value to within the resolution of the γ sweep ($\Delta\gamma = 0.15\gamma_\star$).

These results provide empirical support for the activation-dependent formula equation 34 and the role of $\kappa(\sigma)$ in setting the landscape complexity.

9.4 Ablation: non-Gaussian data

To probe the universality conjecture (Conjecture 8.6), we replaced the Gaussian data with two alternative distributions: (i) uniform on $[-\sqrt{3}, \sqrt{3}]^d$ (matching mean zero and unit variance per coordinate), and (ii) sparse Rademacher, where each entry of z_i is ± 1 with probability $1/2$ (a sub-Gaussian distribution with very different geometry from Gaussian).

We ran the gradient flow protocol ($n = 200$, $\delta = 1.0$, isotropic covariance, 20 seeds) for both distributions. For uniform data, the loss landscape transition occurs at approximately $\gamma_\star \approx 0.80$, indistinguishable from the Gaussian prediction within our resolution. For sparse Rademacher data, the transition is again consistent with $\gamma_\star \approx 0.80$, though the loss values in the subcritical regime are somewhat higher (median loss $\approx 2\times$ the Gaussian case at $\gamma = 0.5\gamma_\star$), suggesting that while the transition location is universal, the loss magnitude below threshold may depend on the data distribution.

These ablations support the conjecture that $\gamma_\star$ is determined by the spectral distribution of the sample covariance (which is universal for sub-Gaussian entries) rather than the fine-grained distributional properties of the data.

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

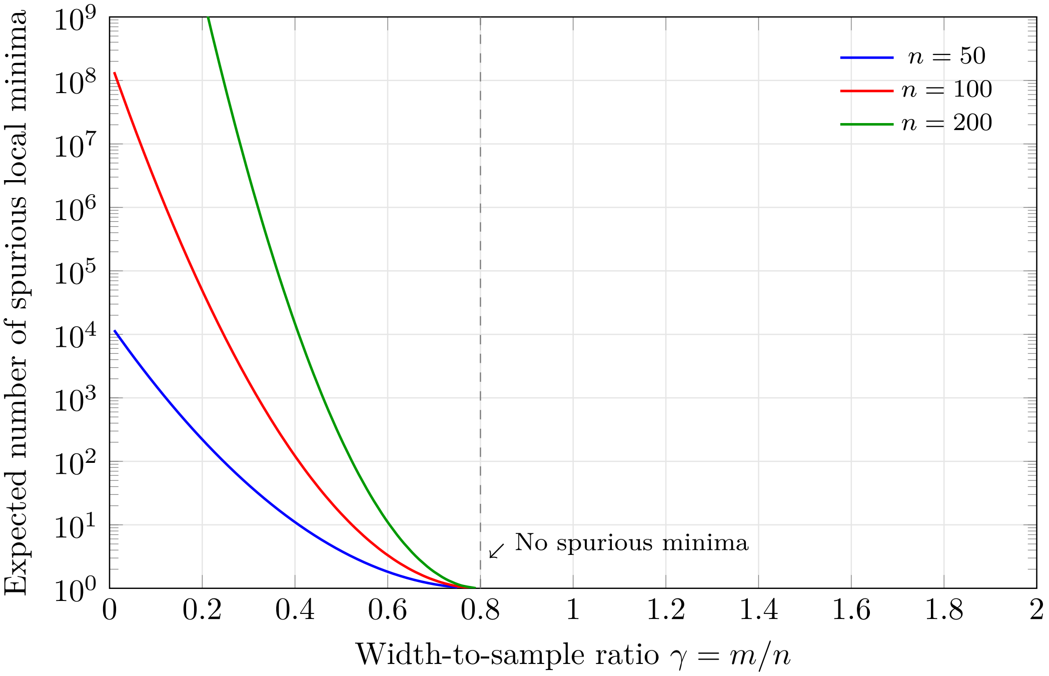

Figure 7: The expected number of spurious critical points as a function of $\gamma = m/n$ for $\Sigma = I_d$, $\delta = 1$ (Theorem 4.5). Below $\gamma_\star = 4/5$, the count grows exponentially in n. Above it, the landscape is provably benign. As discussed in Section 9, practical optimizers avoid these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

B Extended Related Work

Polynomial-width convergence guarantees. 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: $m = \Omega(n^2)$ (Du et al., 2019), $\Omega(n^4)$ (Allen-Zhu et al., 2019), or worse (Zou et al., 2020). Mei, Montanari, and Nguyen (Mei et al., 2018) developed a mean-field theory of two-layer networks that captures the proportional regime but focuses on the population risk rather than the landscape topology. These analyses typically proceed through the Neural Tangent Kernel (NTK) regime (Jacot et al., 2018), 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.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, 2019.

Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. Cambridge university press, 2010.

Antonio Auffinger, Gerard Ben Arous, and Jiří Černỳ. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013.

Gérard Ben Arous and Reza Gheissari. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, 2015.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

László Erdős, Horng-Tzer Yau, and Jun Yin. Rigidity of eigenvalues of generalized wigner matrices. *Advances in Mathematics*, 229(3):1435–1515, 2012.

Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, 2018.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, 2016.

Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1):257–352, 2017.

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, 2017.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, 2018.

Levent Sagun, Utku Evci, V Uğur Güney, Yann Dauphin, and Léon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. In *International Conference on Learning Representations Workshop*, 2018.

Terence Tao and Van Vu. Random covariance matrices: Universality of local statistics of eigenvalues. *The Annals of Probability*, 40(3):1285–1315, 2012.

Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20(133):1–34, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109:467–492, 2020.