DATA-DEPENDENT COMPLEXITY BOUNDS FOR SATURATING NETWORKS VIA THE PUSHFORWARD DENSITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We derive a new Rademacher complexity bound for neural networks with saturating activations that requires neither input truncation nor activation surrogates. Our key insight is that the data distribution regularizes the inverse Jacobian's boundary singularity: under mild tail conditions, the pushforward density remains bounded despite divergent inverse derivatives. We characterize this boundedness via asymptotic analysis and show that depth increases the density supremum only polynomially without worsening the singularity, extending to high-dimensional and dimension-reducing maps. Experiments confirm that our density-based measure correlates more closely with generalization gaps and yields strictly tighter bounds than prior methods.

1 INTRODUCTION

Understanding the generalization capability of deep neural networks remains a fundamental challenge in machine learning. Generalization bounds have been derived from multiple perspectives, including VC dimension (Harvey et al., 2017; Anthony & Bartlett, 2009), weight norms (Neyshabur et al., 2015; Bartlett et al., 2017; Golowich et al., 2018; Neyshabur et al., 2018), PAC-Bayes analysis, and Rademacher complexity (Mohri et al., 2018). A recent line of work characterizes the complexity of hypothesis classes through the network Jacobian, particularly via Koopman-operator-based analyses (Hashimoto et al., 2024; 2026). Within this framework, a change of variables transfers the integral from the input space x to the output space $y = f(x)$, and the inverse Jacobian determinant appears in the bound:

$$\int |v(f(x))|^2 \, dx \leq \left(\sup_y |\det Df^{-1}(y)| \right) \|v\|_2^2 \leq \left(\prod_{l=1}^{L} \sup_{y \in \mathcal{Y}_l} \left\| D\sigma_l^{-1}(y) \right\| \right) \|v\|_2^2,$$

where σ_l denotes the activation function of the l-th layer and $\mathcal{Y}_l$ is its output range.

This approach requires invertibility and consequently restricts activation functions to smooth bijections. However, the smooth bijections most commonly employed in practice are typically saturating, meaning that the derivatives of their inverses diverge at the boundaries of the output range. For $\tanh$, one has $(\tanh^{-1})'(y) = 1/(1-y^2)$, so each of the suprema above is infinite. The sigmoid and softplus functions exhibit similar boundary behavior, leading to the same divergence. Existing works circumvent this divergence in two ways: either restricting the input domain so that activations never saturate (Hashimoto et al., 2026), or replacing the original activations with globally Lipschitz invertible surrogates (Hashimoto et al., 2024). Both strategies introduce additional approximation error. Domain restrictions implicitly discard the tails of the input distribution, while surrogate functions modify the network being analyzed.

Nevertheless, generalization is governed by the network behavior on the data distribution, rather than on all mathematically admissible outputs. If the input distribution rarely pushes the network into near-saturating regimes, the boundary singularities may have a negligible impact on the effective complexity, rendering such approximations unnecessary. This motivates the question: *Can the data distribution itself regularize the boundary singularities of the inverse Jacobian?*

We answer this question affirmatively. Classical worst-case bounds become vacuous because they examine the inverse Jacobian factor in isolation, which diverges at the boundaries. In reality, the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

relevant quantity in the change-of-variables formula is not the inverse Jacobian alone, but its product with the input density, which together form the pushforward density p_θ. Although the inverse Jacobian blows up at the boundary, the input density decays rapidly at the same locations. As long as the tails of the input distribution are sufficiently light, this decay is enough to suppress the divergence of the inverse Jacobian, thereby ensuring that the output density p_θ remains globally bounded. We characterize the conditions under which the pushforward density p_θ stays bounded and derive the corresponding convergence rates. By analyzing the asymptotic behavior of the density near the boundary of the output space, we provide criteria for the boundedness of p_θ in terms of the tail properties of the input distribution.Furthermore, we show that the essential supremum of p_θ grows only polynomially with network depth, while the order of the boundary singularities remains unchanged. We further extend this analytical framework to high-dimensional mixed mappings and dimensionality-reducing mappings.

Building on these insights, we derive a data-dependent Rademacher complexity bound:

$$\mathfrak{R}_n(\mathcal{F}) \;\le\; B\sqrt{\frac{Mr}{n}} \;+\; \frac{C\,B\,L_\Phi\,\sqrt{d}\,L\,\rho^{L-1}\,R_\Theta\,\sqrt{P}}{\sqrt{n}},$$

where $M := \sup_\theta \operatorname{ess\,sup}_y p_\theta(y)$ is the uniform density bound. This bound is obtained by direct computation of the data density and no longer requires any boundedness assumptions on the activation functions. We validate the theoretical predictions through extensive numerical simulations and apply the framework to assess the generalization capability of practical neural networks. Compared with existing baselines, our density-based metric correlates more closely with the observed generalization gap and yields tighter bounds than existing Rademacher-based counterparts.

2 Problem Setup

Invertible networks. We consider an L-layer invertible network of width d, $f_\theta = W_L \circ \sigma_{L-1} \circ W_{L-1} \circ \cdots \circ \sigma_1 \circ W_1$, where each $W_l \in \mathbb{R}^{d\times d}$ is invertible and each σ_l is a C^1 diffeomorphism. In the coordinatewise cases we analyze, each σ_l acts componentwise. The input is a random variable $X \sim \mu$, where μ is a probability measure on the input space with density p_X. We denote Lebesgue measure by λ, the essential supremum by $\operatorname{ess\,sup}$, and the pushforward of a measure μ under a map f by $f_\#\mu$. All notation is collected in Appendix A.

Factorization of the inverse Jacobian. The inverse map $\Psi_\theta := f_\theta^{-1}$ is again a C^1 diffeomorphism. Its absolute Jacobian determinant factorizes as $|\det D\Psi_\theta(y)| = C_W(\theta)\,\rho_\theta(y)$, where $C_W(\theta) := \prod_{l=1}^{L} |\det W_l|^{-1}$ collects the volume factors of the linear layers, and $\rho_\theta(y) := \prod_{l=1}^{L-1} \left|\det D\sigma_l^{-1}(u_l(y))\right|$ collects the corresponding factors from the inverse activations. Here $u_l(y)$ denotes the intermediate preactivation input to the inverse of the l-th activation. The chain-rule derivation is given in Appendix C.1.

Pushforward density. By the change-of-variables formula, the pushforward measure $\nu_\theta := f_{\theta\#}\mu$ admits the density $p_\theta(y) = C_W(\theta)\,p_X(\Psi_\theta(y))\,\rho_\theta(y)\,\mathbf{1}_{Y_\theta}(y)$ supported on the image Y_θ. This data-induced density is the central object of the paper. For saturating activations, $\rho_\theta(y)$ may diverge as y approaches the boundary of the activation range, whereas the pulled-back input density $p_X(\Psi_\theta(y))$ typically decays there. The remainder of the paper investigates whether this product remains bounded, how its supremum grows with depth, and what implications such boundedness has for generalization.

3 Tail Competition and Depth Growth at the Boundary

In Section 4, we will provide a complexity bound that depends on the uniform upper bound of the pushforward density $M = \sup_\theta \operatorname{ess\,sup}_y p_\theta(y)$. The core task of this section is to answer two questions: When is M finite? Given that neural networks are multi-layer compositions, how does M grow with depth? The inverse Jacobian determinant diverges at the saturation boundary, while the input density decays when pulled back to this boundary. M is finite if and only if the decay rate of the input tail is sufficient to counteract the Jacobian singularity. We refer to this boundary asymptotic behavior as *tail competition*. In this section, we primarily consider the one-dimensional scenario, which is more intuitive and easier to understand; high-dimensional extensions are sketched

in Appendix G. Section 3.1 establishes the basic competition criterion via the scalar map $\tanh(wx)$; Section 3.2 extends this to deep networks, revealing that depth affects the density only through a slowly varying shift constant without altering the order of the singularity.

3.1 THE ONE-DIMENSIONAL TAIL COMPETITION CRITERION

Consider the scalar map $f_w(x) = \tanh(wx)$ $(w > 0)$. Its pushforward density is

$$p_w(y) = \underbrace{\frac{1}{w(1-y^2)}}_{\text{Jacobian singularity}} \cdot \underbrace{p_X\big(w^{-1}\operatorname{arctanh}(y)\big)}_{\text{input tail}}, \quad y \in (-1,1).$$

As $y \to \pm 1$, the first factor diverges at a rate of $(1-y^2)^{-1}$, while the second factor takes values in the tail of the input distribution. Boundedness depends entirely on the relative strength of these two factors. To precisely characterize this competition, let $y = 1 - t$ $(t \downarrow 0)$. Using $\operatorname{arctanh}(1-t) = \frac{1}{2}\log(2/t) + O(t)$, we obtain the asymptotic form near the boundary:

$$p_w(1-t) \asymp t^{-1}\, p_X\left(\frac{1}{2w}\log\frac{1}{t}\right), \qquad t \downarrow 0.$$

This expression clearly separates the competing parties: t^{-1} represents the boundary singularity, and p_X probes the input tail on a logarithmic scale. The following theorem provides criteria under standard tail families.

Theorem 3.1 (Boundedness criteria for the tanh pushforward density)**.** (i) Exponential tails. *If $p_X(x) \asymp e^{-cx^s}$ as $x \to +\infty$ with $c, s > 0$, then $p_w(1-t) \asymp t^{-1}\exp(-c(2w)^{-s}\log^s(1/t))$ as $t \downarrow 0$. The boundary integral $\int_0^\delta p_w(1-t)\,dt$ is finite for every $s > 0$. Moreover, p_w is bounded near $y = 1$ if and only if either $s > 1$, or $s = 1$ and $c \geq 2w$. For $0 < s < 1$, p_w is unbounded but integrable.* (ii) Polynomial tails. *If $p_X(x) \asymp x^{-r}$ as $x \to +\infty$ with $r > 0$, then $p_w(1-t) \asymp t^{-1}\left(\log\frac{1}{t}\right)^{-r}$. This density is integrable if and only if $r > 1$, and it is never bounded near $y = 1$. The corresponding statements at $y = -1$ are obtained by replacing the right tail of p_X with the left tail.*

When $s > 1$, the tail decay dominates the t^{-1} singularity and the density is bounded; at the critical exponent $s = 1$ the outcome depends on whether $c \geq 2w$; polynomial tails never yield a bounded density, although they are integrable when $r > 1$. For saturating activations whose inverses grow logarithmically at the boundary, such as sigmoid and softplus, the same criteria apply with a different constant preceding $\log(1/t)$ (Appendices D.4 and D.5). Since M in Section 4 takes the supremum over parameters, we require boundedness that holds uniformly with respect to weights. The following theorem guarantees this uniformity under Gaussian inputs.

Theorem 3.2 (Gaussian inputs regularize tanh uniformly)**.** *Let $X \sim \mathcal{N}(0,1)$ and $Y = \tanh(wX)$. Then for every compact interval $[a,b] \subset (0,\infty)$, $\sup_{w\in[a,b],\, y\in(-1,1)} p_w(y) < \infty$.*

Compactly supported inputs are similarly regularized, in which case the output support is a compact subinterval of $(-1,1)$ (Appendix D.3). The Gaussian assumption itself can be relaxed: any Gaussian mixture retains a super-exponential tail in the sense of Theorem 3.1(i) with $s = 2$, so boundedness survives; the point of the next result is that the uniform bound remains explicit and depends on the mixture only through its largest component scale.

Theorem 3.3 (Mixture inputs regularize tanh uniformly)**.** *Let $p_X = \sum_{j=1}^m \pi_j \mathcal{N}(\mu_j, \sigma_j^2)$ be a Gaussian mixture and $Y = \tanh(wX)$. Then for every compact interval $[a,b] \subset (0,\infty)$,*

$$\sup_{w\in[a,b],\, y\in(-1,1)} p_w(y) \;\leq\; \frac{1}{a\sqrt{2\pi}} \sum_{j=1}^m \frac{\pi_j}{\sigma_j} \max\left\{\tfrac{4}{3},\; e^{\,2b|\mu_j| + 2b^2\sigma_j^2}\right\} < \infty.$$

For a single standard Gaussian component this recovers the bound in the proof of Theorem 3.2 (Appendix D.3).

3.2 DEPTH DOES NOT WORSEN THE BOUNDARY SINGULARITY

Consider the deep composition $f_K = T^{\circ K}$. Successive layers progressively compress the output: from $(-1,1)$ to $(-\tanh(1), \tanh(1))$, and ultimately to an interval $(-b_K, b_K)$ that shrinks toward

zero as $K \to \infty$. Although one might conjecture that this intensifies the density singularity at the saturation boundary, we show that depth does not change its essential order. Regularization by the data distribution remains effective at all depths.

Theorem 3.4 (Asymptotic radius of the output interval)**.** *Define the sequence $\{a_m\}$ by $a_0 = 1$ and $a_{m+1} = T(a_m)$. Then f_K takes values in $(-b_K, b_K)$, where $b_K = a_{K-1}$. As $K \to \infty$, $b_K \sim \sqrt{\frac{3}{2(K-1)}}$.*

The boundary moves with K, but the key question is whether repeated composition makes the singularity sharper.

Theorem 3.5 (Depth invariance of the boundary asymptotics)**.** *As $t \downarrow 0$,*

$$\Psi_K(b_K - t) = \tfrac{1}{2}\log\tfrac{1}{t} + c_K + o(1), \quad \Psi_K'(b_K - t) = \frac{1 + o(1)}{2t},$$

$$p_K(b_K - t) = \frac{1 + o(1)}{2t}\, p_X\left(\tfrac{1}{2}\log\tfrac{1}{t} + c_K + o(1)\right),$$

where c_K is a depth-dependent constant. Thus the singularity remains of order $1/t$ at every depth.

Regardless of network depth, the singularity at the boundary is always of order $1/t$, identical to the single-layer case. The entire effect of depth is absorbed into the shift constant c_K, which determines the growth of the density supremum: with $D_K := \prod_{i=0}^{K-2} \cosh^2(a_i)$, one has $c_K = \frac{1}{2}\log(2/D_K)$, and the tanh iterate asymptotics give $D_K = \Theta(K^{3/2})$, i.e. $c_K = -\frac{3}{4}\log K + O(1)$ (Theorem E.1).

Theorem 3.6 (Gaussian pushforward density grows polynomially)**.** *If $X \sim \mathcal{N}(0,1)$ and $Y = f_K(X)$, then $\|p_K\|_\infty = \Theta(K^{3/2})$.*

For Gaussian inputs, evaluating the density at the shift c_K introduces a factor of $\exp(c_K^2/2) = \Theta(K^{3/2})$. Layer by layer, $\|p_l\|_\infty / \|p_{l-1}\|_\infty = 1 + O(1/l)$, so $\log \|p_K\|_\infty = \frac{3}{2}\sum_{l \le K} 1/l + O(1)$, indicating harmonic series growth. This explains why the order of the boundary singularity remains unchanged, yet the overall upper bound still grows polynomially with depth (Appendices E.1 and E.2). The analysis for softplus is entirely parallel and, owing to its simpler structure, fully explicit: under the change of variables $r = e^x$, softplus iteration is conjugate to the translation $r \mapsto r + 1$, which yields the closed form $S^{\circ K}(x) = \log(e^x + K)$ taking values in $(\log K, \infty)$. The boundary singularity remains of order $1/t$, but under Gaussian inputs the density supremum grows only linearly, $\|p_K\|_\infty = \Theta(K)$, rather than as $K^{3/2}$ for $\tanh$ (Theorems F.1 and F.2 in Appendix F.1).

The one-dimensional analysis above can be directly extended. For product networks, $p_\theta(y) = \prod_j p_{\theta_j}(y_j)$, and Theorem 3.1 applies coordinate-wise. For Gaussian two-layer tanh networks with affine mixing, uniform density bounds can be obtained under constraints on the singular values and determinant of the mixing matrix. For deep chains with arbitrary invertible mixing, cut-norm recursion yields explicitly computable density bounds (Appendix G).

4 Data Dependent Complexity

This section translates the boundedness results of the pushforward density into generalization complexity guarantees for neural networks. We first establish the fundamental conversion from density boundedness to Rademacher bounds (Section 4.1) and then extend it to the full parameter learning setting using parameter space covering (Section 4.2). Next, we analyze how the density upper bound accumulates across network layers, revealing its essential distinction from existing Koopman methods (Section 4.3). Finally, we present a sample level density upper bound estimator that can be computed on trained networks (Section 4.4).

4.1 From Density Boundedness to Rademacher Bounds

Let $\Omega \subseteq \mathbb{R}^m$ be a measurable output domain equipped with the Lebesgue measure λ, let $f_\theta : X \to \Omega$ be a measurable map, and let p_θ be the pushforward density of $Y = f_\theta(X)$. Denote the empirical Rademacher complexity of a function class $\mathcal{F}$ on a sample $S = \{x_i\}_{i=1}^n$ by $\widehat{\mathfrak{R}}_S(\mathcal{F}) :=$

$\mathbb{E}_\epsilon \sup_{h\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i)$, where ϵ_i are independent Rademacher variables, and denote its expectation over samples by $\mathfrak{R}_n(\mathcal{F})$.

Proposition 4.1 (Complexity bounds from density boundedness). *Assume that $p_\theta(y) \le M_\theta$ almost everywhere and let $M := \sup_{\theta\in\Theta} M_\theta$. For the orthogonal feature class $\mathcal{F}_\theta = \{x \mapsto \langle a, \Phi(f_\theta(x))\rangle : \|a\|_2 \le B\}$, where $\Phi = (\psi_1, \dots, \psi_r)$ is orthonormal in $L^2(\Omega, \lambda)$, we have $\mathfrak{R}_n(\mathcal{F}_\theta) \le B\sqrt{M_\theta r/n}$. For the RKHS class $\mathcal{V}_B = \{v \in \mathcal{H}_k : \|v\|_{\mathcal{H}_k} \le B\}$, if $D_k := \int_\Omega k(y,y)\,dy < \infty$, then $\mathfrak{R}_n(\mathcal{F}_\theta) \le B\sqrt{MD_k/n}$. For any $v \in L^2(\Omega,\lambda)$, we have $\mathbb{E}_X|v(f_\theta(X))|^2 \le M_\theta \|v\|^2_{L^2(\Omega)}$. (Appendix H.1)*

4.2 Full Parameter Learning

Theorem 4.1 applies only when the network parameters θ are fixed and only the readout vector a is learnable. To cover full parameter learning, we perform a covering argument over the parameter space, adding the geometric complexity of the parameter space to the density bound.

Consider parameters $\theta = (W_1, \dots, W_L)$, where $W_l \in \mathbb{R}^{d\times d}$ and $\|W_l\|_{\mathrm{op}} \le \rho$. Assume the parameter space Θ is contained within a ball of radius R_Θ in $\mathbb{R}^P$ (with $P = Ld^2$), equipped with the metric $\|\theta - \theta'\| := \max_l \|W_l - W_l'\|_{\mathrm{op}}$. The network is defined as $f_\theta = T \circ W_L \circ \cdots \circ T \circ W_1$, where T is the coordinate wise $\tanh$ mapping and biases are absorbed into the weights. Assume the feature map Φ has Lipschitz constant L_Φ on $(-1,1)^d$, its components are orthonormal in $L^2(\Omega,\lambda)$, and $p_\theta \le M$ holds uniformly over Θ.

Theorem 4.2 (Full parameter data dependent Rademacher bound). *Under the above setup, there exists a universal constant C such that*

$$\mathfrak{R}_n(\mathcal{F}) \le B\sqrt{\frac{Mr}{n}} + \frac{C\,B\,L_\Phi\,\sqrt{d}\,L\,\rho^{L-1}\,R_\Theta\,\sqrt{P}}{\sqrt{n}}.$$

This bound consists of two terms: the first corresponds to the intrinsic complexity of the fixed network, controlled by the density upper bound M; the second arises from the covering number of the parameter space and is independent of M. Since $\tanh$ is 1 Lipschitz and bounded, intermediate layer activations satisfy $\|h_l\|_2 \le \sqrt{d}$. By induction, we obtain a global Lipschitz estimate of the network output with respect to parameters: $\|f_\theta(x) - f_{\theta'}(x)\|_2 \le \sqrt{d}\,L\,\rho^{L-1}\|\theta - \theta'\|$, which renders the induced Rademacher process sub Gaussian in θ. Applying Dudley's entropy integral to the P dimensional ball yields the second term (Appendix I.1).

4.3 The Layerwise Profile

The above bound compresses the entire network into a single scalar M. A natural question is whether a finer layer-wise decomposition can be obtained, or in other words, whether it can be expressed as a product over individual layers rather than only considering the final outcome. The answer is negative.

Proposition 4.3 (Layerwise isometry). *Let $f_\theta = g_L \circ \cdots \circ g_1$, $\mu_0 = \mu$, and $\mu_l := (g_l)_{\#}\mu_{l-1}$. Define the operator $\mathcal{K}_l : L^2(\mu_l) \to L^2(\mu_{l-1})$ by $\mathcal{K}_l v = v \circ g_l$. Each $\mathcal{K}_l$ is an isometry, and thus*

$$\|v \circ f_\theta\|_{L^2(\mu_0)} = \|v\|_{L^2(\mu_L)} \le \sqrt{M(\theta)}\,\|v\|_{L^2(\lambda)}.$$

This implies that the comparison with the Lebesgue measure occurs solely between the final layer output μ_L and λ. Any layer wise decomposition must originate from the decomposition of $M(\theta)$ itself; forcibly inserting a Lebesgue reference measure at intermediate layers would introduce a factor of $\sup_z |\det Dg_l^{-1}(z)|$, which is infinite for saturating activations and degenerates to worst case bounds. (Appendix I.2)

Although independent layer wise bounding is impossible, $M(\theta)$ itself possesses a clear layer wise accumulation structure. Let p_l denote the density of the l th layer output, satisfying the recurrence $p_l(y) = p_{l-1}(g_l^{-1}(y))\,|\det Dg_l^{-1}(y)|$. Defining the layer wise amplification factor $\Lambda_l := \|p_l\|_\infty / \|p_{l-1}\|_\infty$, we have $M(\theta) = M_0 \prod_{l=1}^L \Lambda_l$, where $M_0 = \|p_X\|_\infty$. When $g_l = \sigma_l \circ W_l$, the contributions of the linear transformation and nonlinear saturation can be separated as $\Lambda_l = |\det W_l|^{-1} S_l$, where

$$S_l := \frac{\sup_z p_{l-1}(W_l^{-1}z)\,|\det D\sigma_l^{-1}(z)|}{\|p_{l-1}\|_\infty}.$$

In the worst case, S_l is relaxed to $\sup_z |\det D\sigma_l^{-1}(z)|$, which is infinite for $\tanh$; however, from a data dependent perspective, S_l is finite, with its value determined by the asymptotic analysis in Section 3. For the $\tanh$ chain, the amplification factors of layers $l \geq 2$ also admit an explicit, distribution free upper bound: after the first layer the activations lie in the bounded cube $(-1,1)^d$, so only deterministic Jacobian accumulation remains. Define the layerwise saturation potential

$$\Phi_{\mathrm{sat}}(W) := \max_{z\in[-1,1]^d} \sum_{j=1}^{d} \log\cosh^2\big((Wz)_j\big) = \max_{s\in\{\pm1\}^d} \sum_{j=1}^{d} \log\cosh^2\big((Ws)_j\big),$$

where the maximum is attained at a vertex since $\log\cosh^2$ is convex.

Proposition 4.4 (Cut norm recursion for deep mixed chains)**.** *For the chain $U_l = W_l Z_{l-1}$, $Z_l = T(U_l)$ with invertible $W_l \in \mathbb{R}^{d\times d}$, and every $l \geq 2$, $\log\|p_{Z_l}\|_\infty \leq \log\|p_{Z_{l-1}}\|_\infty + \Phi_{\mathrm{sat}}(W_l) - \log|\det W_l|$.*

$$\log\|p_{Z_L}\|_\infty \leq \log\|p_{Z_1}\|_\infty + \sum_{l=2}^{L}\big[\Phi_{\mathrm{sat}}(W_l) - \log|\det W_l|\big].$$

Moreover $2\|W\|_{\infty\to1} - 2d\log 2 \leq \Phi_{\mathrm{sat}}(W) \leq 2\|W\|_{\infty\to1}$, so Φ_{sat} agrees with the cut norm up to a factor of 4 and an additive constant, and is computable to within Grothendieck's constant $K_G < 1.783$ by SDP relaxation. A corner path lower bound matches the first layer term, so the recursion is nearly tight; the residual gap is the combinatorial question of whether the vertices maximizing each layer's potential can be realized along a single orbit of $z \mapsto T(Wz)$. (Appendix G)

This accumulation structure admits explicit asymptotic characterization in the deep limit. For deep $\tanh$ chains, Theorem 3.6 gives $\|p_K\|_\infty = \Theta(K^{3/2})$, corresponding to a layer wise factor $\Lambda_K = 1 + 3/(2K) + O(K^{-2})$, hence $\log M_K \approx \log M_0 + \frac{3}{2}\sum_{l\leq K} 1/l$. For deep softplus, Theorem F.2 gives $\|p_K\|_\infty = \Theta(K)$, corresponding to $\Lambda_l \approx 1 + 1/l$. The growth is distributed uniformly across layers in the form of a harmonic series: each layer contributes only a multiplicative factor of $1 + O(1/l)$. This explains why the order of boundary singularities does not change with depth (Theorem 3.5), while the total density upper bound grows only polynomially. This layer wise structure stands in term by term alignment yet essential contrast to Koopman methods (Hashimoto et al., 2026). Both approaches share identical factors for the linear layers, $|\det W_l|^{-1/2}$, and for the readout and sampling terms; the critical difference lies in the activation layer: Koopman methods use the worst case factor $\sup_z |\det D\sigma_l^{-1}(z)|^{1/2}$, which diverges for $\tanh$, whereas we use the data dependent finite quantity $\sqrt{S_l}$. This is the reason our method avoids infinities. Using the coordinate decomposition (Appendix G), the above results extend directly to high dimensions: for d dimensional $\tanh$ networks, $\|p_\theta\|_\infty = \Theta(K^{3d/2})$, and for softplus, $\Theta(K^d)$. Accordingly, the Rademacher bounds are $CBK^{3d/4}\sqrt{r/n}$ and $CBK^{d/2}\sqrt{r/n}$, respectively, with constants independent of K, n, and the assumed parameters.

4.4 Estimating the Density Bound from the Sample

Theoretically, $M = \sup_\theta \operatorname{ess\,sup}_y p_\theta(y)$ involves a global supremum and is not directly computable in practice. Leveraging the recursive structure of the pushforward density, we design a sample level estimator that requires only a single forward pass. For the $\tanh$ chain in Section 2, the change of variables recursion scales to a pointwise log density identity along each sample path (Appendix H.2). To avoid tracking its terms layer by layer, we fit a Gaussian distribution $\widehat{\Sigma}_L$ to the pre activations U_L of the final layer. Its sample covariance automatically encodes the mixing, saturation, and volume change effects of the preceding $L-1$ layers. The final estimator splits into two parts at the last nonlinearity:

$$\log\widehat{M} = \max_{i\leq n}\Big[\log\varphi_{\widehat{\Sigma}_L}\big(u_L(x_i)\big) + \sum_{j=1}^{d}\log\cosh^2 u_{Lj}(x_i)\Big].$$

The first term summarizes the cumulative density contribution of the first $L-1$ layers via a Gaussian approximation; the second term is the inverse Jacobian of the final $\tanh$. The second term cannot be absorbed into the Gaussian approximation because the readout layer attaches to Z_L, and direct density estimation near the boundary of $(-1,1)^d$ is numerically unstable. This splitting is anchored at the readout interface, so increasing depth introduces no additional terms; $\log p_X$ is independent of θ and naturally cancels out when comparing models.

5 Dimension Reduction via the Coarea Formula

The preceding theory requires the network to be square and invertible. However, in practical applications, the encoder and bottleneck layers are typically dimensionality-reducing, rendering the Jacobian matrix non-invertible and causing the Change of Variables formula to fail directly. This section addresses how to establish density bounds and complexity bounds for such non-invertible and dimensionality-reducing networks.

5.1 Coarea Density for Dimension-Reducing Maps

Let $E : \mathbb{R}^{d_0} \to \mathbb{R}^{d_1}$ (with $d_0 \geq d_1$) be a Lipschitz map with full row rank almost everywhere. Its d_1-dimensional Jacobian is defined as $J_E(x) := \sqrt{\det(DE(x)\, DE(x)^\top)}$. The coarea formula states that for any non-negative measurable function u,

$$\int_{\mathbb{R}^{d_0}} u(x)\, J_E(x)\, dx = \int_{\mathbb{R}^{d_1}} \left(\int_{E^{-1}(z)} u(x)\, d\mathcal{H}^{d_0-d_1}(x) \right) dz,$$

where $\mathcal{H}^{d_0-d_1}$ denotes the Hausdorff measure on the fiber $E^{-1}(z)$. By choosing a test function ϕ and setting $u(x) = \phi(E(x))\, p_X(x)/J_E(x)$, we can identify the density of $Z = E(X)$.

Proposition 5.1 (Pushforward Density for Dimension-Reducing Maps (Appendix J.1)). *If X has density p_X and $J_E > 0$ almost everywhere, then the density of $Z = E(X)$ is given by*

$$p_Z(z) = \int_{E^{-1}(z)} \frac{p_X(x)}{\sqrt{\det(DE(x)\, DE(x)^\top)}}\, d\mathcal{H}^{d_0-d_1}(x).$$

Although E is non-invertible, we can view $p_Z(z)$ as the average or accumulation of the input density p_X over the entire preimage set $E^{-1}(z)$ divided by a generalized Jacobian factor $\sqrt{\det(DEDE^\top)}$. This step transforms the issue of non-invertibility into a problem of computing fiber integrals.

5.2 Density Bounds for Rectangular Affine-Tanh Networks

The bound follows by the same tail-competition argument as Section 3, replacing $|\det W|^{-1}$ with $\det(WW^\top)^{-1/2}$ and reducing the Gaussian pre-activation to the d_1-dimensional variable $U = WX \sim \mathcal{N}(0, WW^\top)$. The Gaussian tails suppress the $(1-s^2)^{-1}$ singularity coordinate-wise; only the prefactor on the logarithmic scale changes from 1 to d_1.

Theorem 5.2 (Uniform Density Bound for Rectangular Affine-Tanh Encoders). *Let $X \sim \mathcal{N}(0, I_{d_0})$ and $E(x) = T(Wx)$, where $W \in \mathbb{R}^{d_1 \times d_0}$ ($d_1 \leq d_0$) has full row rank with singular values in $[\alpha, \beta] \subset (0, \infty)$, and T is the coordinate-wise $\tanh$ function. Then the density of $Z = E(X)$ on $(-1,1)^{d_1}$ satisfies*

$$\|p_Z\|_\infty \leq \frac{M_\beta^{d_1}}{(2\pi)^{d_1/2}\, \alpha^{d_1}}, \qquad M_\beta := \sup_{s \in (-1,1)} \frac{\exp(-\operatorname{arctanh}(s)^2/(2\beta^2))}{1-s^2} < \infty.$$

This bound depends only on α, β, and d_1; consequently, $\sup_W \|p_Z\|_\infty < \infty$ over the constrained parameter class.(Appendix J.2)

5.3 Bottleneck Autoencoders

The complexity argument in Section 4 invokes the pushforward density only once at the readout layer: $\mathbb{E}_X\, U(f_\theta(X)) = \int U(y)\, p_\theta(y)\, dy \leq M \int U(y)\, dy$. Subsequent layers do not enter the bound. Therefore, the complexity bound itself never requires the network to be invertible; invertibility is merely one route to computing the density bound M, while the coarea formula provides an alternative route applicable to dimension-reducing maps.

Corollary 5.3 (Rademacher Bound for Bottleneck Encoders). *Let $E : \mathbb{R}^{d_0} \to \mathbb{R}^{d_1}$ ($d_1 < d_0$) have full row rank almost everywhere, and let $M_E := \|p_Z\|_\infty < \infty$ (computed via Theorem 5.1; explicitly controlled via Theorem 5.2 in the rectangular affine-tanh case). For the readout class on the latent space $\mathcal{V}_B = \{v \in \mathcal{H}_k : \|v\|_{\mathcal{H}_k} \leq B\}$, where $D_k := \int_{\mathbb{R}^{d_1}} k(z,z)\, dz < \infty$, the function class $\mathcal{F}_E = \{x \mapsto v(E(x)) : v \in \mathcal{V}_B\}$ satisfies $\mathfrak{R}_n(\mathcal{F}_E) \leq B\sqrt{\frac{M_E\, D_k}{n}}$.*

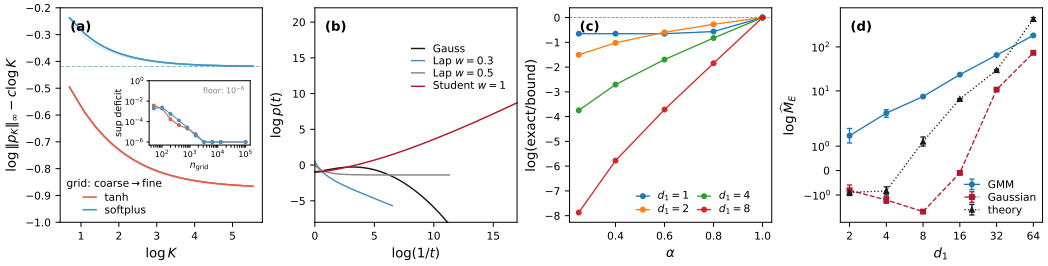


Figure 1: **Numerical verification.** (a) Density supremum versus depth under Gaussian input. (b) Pushforward density boundary profiles for four input tails. (c) Ratio of exact to predicted values for the rectangular affine-$\tanh$ bound versus the lower singular value α with $\beta = 1$. (d) Estimated density bound $\log \widehat{M}_E$ for trained bottleneck encoders versus latent space width.

This bound follows directly from applying Theorem 4.1 (RKHS case) to $Z = E(X)$, requiring only that the density be bounded by M_E. Crucially, the decoder Jacobian does not enter the bound—only the encoder's density M_E matters. Appendix J.3 develops this for reconstruction losses, lifting the previous restriction to square autoencoders.

6 Experiments

The experiments comprise two parts: numerical verification of theoretical predictions in Sections 3 and 5, and validation of generalization bounds on real datasets. All experiments use MNIST with settings consistent with Hashimoto et al. (2026) and employ MLPs with $\tanh$ activations; full configurations are given in Appendix K.

6.1 Numerical Verification

Depth scaling law (Figure 1a). For deep composite functions $f_K = \tanh^{\circ K}$ and $S_K = \mathrm{sp}^{\circ K}$ under Gaussian input, Theorems 3.6 and F.2 predict density supremum growth rates of $\Theta(K^{3/2})$ and $\Theta(K)$ respectively, with softplus yielding leading constant $e^{1/2}/\sqrt{2\pi} \approx 0.6577$. Plotting $\log \|p_K\|_\infty - \frac{3}{2}\log K$ and $\log \|p_K\|_\infty - \log K$ against depth shows both curves flattening, with softplus stabilizing near the predicted constant. Fitted slopes for $K \in [8, 256]$ are 1.476 and 0.993, aligning closely with theory.

Tail saturation phase diagram (Figure 1b). Theorem 3.1 states that bounded pushforward density requires input tail decay sufficient to suppress the $1/t$ singularity. Plotting $\log p(t)$ against $\log(1/t)$ for four input distributions validates this: Gaussian bends downward (bounded); Laplace ($c = 1$) shows slope $1 - \frac{1}{2w} \approx -0.67$ at $w = 0.3$ (bounded) and zero slope at $w = 0.5$ (critical case $c = 2w$); Student-t ($\nu = 3$) rises with local slope approaching 1 (unbounded). Results fully match the theoretical phase diagram.

Tightness of rectangular network bounds (Figure 1c). For affine-$\tanh$ encoder $E(x) = T(Wx)$ with singular values in $[\alpha, \beta]$, Theorem 5.2 gives upper bound $M_\beta^{d_1}/((2\pi)^{d_1/2}\alpha^{d_1})$ for $\|p_Z\|_\infty$. Since WX is Gaussian, the exact supremum is computable. The ratio $\log(\text{exact}/\text{bound})$ remains non-positive across all α when $\beta = 1$, confirming the bound holds everywhere. The gap is zero at $\alpha = \beta$ (isotropic) and widens with increasing anisotropy and output dimension.

Density bounds for trained bottleneck encoders (Figure 1d). Theorem 5.2 predicts $\log M_E$ increases monotonically with bottleneck dimension d_1. Comparing three estimators—single Gaussian, BIC-selected GMM, and the theoretical upper bound—on trained bottleneck autoencoders shows both the theoretical bound and GMM increase monotonically with d_1, matching theory. The single Gaussian estimate decreases non-monotonically for $d_1 \in [4, 8]$ before rising at $d_1 = 16$, indicating estimator choice affects bound accuracy. Nevertheless, we retain Gaussian estimation for subsequent baseline comparisons due to its strong empirical performance.

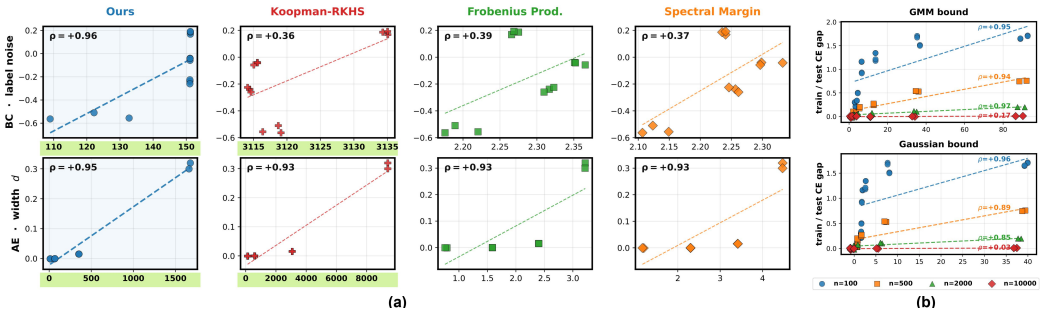


Figure 2: **Generalization performance evaluation.** Scatter plots of bounds on a logarithmic scale versus empirical generalization gaps. (a) Top row shows label noise and bottom row shows width variation where ρ denotes the Spearman rank correlation coefficient. The green bars below highlight the numerical difference between our method and the Koopman bound. (b) Downstream validity check for Gaussian versus GMM density estimators where n encodes downstream sample size.

6.2 Generalization Performance Validation

Relationship between generalization bounds and generalization gaps (Fig. 2a). We design two experimental axes including a label noise axis with $\tanh$ classifiers trained under different noise rates and a width axis with three-layer $\tanh$ autoencoders of varying hidden layer widths trained at fixed depth and sample size. We compare our proposed bound against three existing baselines including Koopman-RKHS (Hashimoto et al., 2026), Frobenius product norm (Neyshabur et al., 2015), and spectral normalization margin bound (Bartlett et al., 2017). On the label noise axis, the Spearman rank correlation coefficient between our bound and the empirical generalization gap is significantly higher than all baseline methods. On the width axis, our bound performs on par with the best baseline and is numerically tighter than Koopman-RKHS.

Downstream linear probe validation (Figure 2b). To validate the theoretical bound for bottleneck autoencoders derived in Theorem 5.3 for handling dimensionality reduction operations, we conduct linear probe experiments on MNIST. We train a set of independent bottleneck encoders with symmetric decoders and freeze them. Subsequently, for different downstream sample sizes n, we train linear classifiers on the bottleneck representation $z = \tanh(u)$ for 10-class classification. For each configuration, we compute two density-based bounds with detailed formulas provided in Appendix J.3. The density upper bound is estimated using either a single Gaussian distribution or a GMM. The Spearman correlation coefficient ρ is strongest at small n and decreases somewhat as n increases. Most importantly, the correlation for GMM is slightly higher than that for the single Gaussian distribution which confirms that more precise density estimation improves the predictive power of the bound.

7 Conclusion and Limitations

This paper demonstrates that the data distribution itself can regularize boundary singularities in saturated networks, thereby eliminating the need to truncate inputs or replace activation functions. Our analysis reveals that although increasing network depth causes the supremum of the pushforward density to grow polynomially, the order of the boundary singularity never deteriorates. This insight leads to a finite and data-dependent Rademacher complexity bound. However, many limitations remain. For heavy-tailed distributions, the tail decay rate is insufficient to suppress the boundary singularity, which may cause our bound to diverge. Computing the density upper bound M still relies on invertible structures or explicit density modeling. How to efficiently and accurately estimate M for general redundant networks lacking such structures remains an open computational problem.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, figure and table generation, code implementation, and literature searches for related work. We did not use generative AI tools to write the manuscript or perform mathematical derivations, and data synthesis is not applicable to this work. All AI-assisted outputs were carefully reviewed by the authors: polished text was manually verified to ensure fidelity to our original ideas; AI-generated code was rigorously tested to confirm its correctness; retrieved references were checked against primary sources; and generated figures and tables were validated for accuracy. We take full responsibility for the final content of this work, including the text, claims, code, and artifacts produced with the assistance of generative AI.

REFERENCES

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 2009.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 254–263, 2018.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pp. 6240–6249, 2017.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

Micah Goldblum, Jonas Geiping, Avi Schwarzschild, Michael Moeller, and Tom Goldstein. Truth or backpropaganda? an empirical investigation of deep learning theory. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, volume 75 of *Proceedings of Machine Learning Research*, pp. 297–299, 2018.

Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1064–1068, 2017.

Yuka Hashimoto, Sho Sonoda, Isao Ishikawa, Atsushi Nitanda, and Taiji Suzuki. Koopman-based generalization bound: New aspect for full-rank weights. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Yuka Hashimoto, Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Why high-rank neural networks generalize?: An algebraic framework with RKHSs. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026.

Haotian Ju, Dongyue Li, and Hongyang R. Zhang. Robust fine-tuning of deep neural networks with Hessian-based generalization guarantees. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10431–10461, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 10236–10245, 2018.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Shuai Li, Kui Jia, Yuxin Wen, Tongliang Liu, and Dacheng Tao. Orthogonal deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1352–1368, 2021.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2 edition, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1376–1401, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: Unified generalization error analysis of large compressible deep neural network. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via Lipschitz augmentation. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pp. 9722–9733, 2019.

Colin Wei and Tengyu Ma. Improved sample complexities for deep networks and robust classification via an all-layer margin. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Weinan E, Chao Ma, and Lei Wu. The Barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022.

A NOTATION

Symbol	Description
Measure and probability	
λ	Lebesgue measure
$\operatorname{ess\,sup}$	essential supremum
$f_{\#}\mu$	pushforward of measure μ under map f
μ	input probability measure on $X_0 \subseteq \mathbb{R}^d$
p_X	density of μ with respect to λ
P_X	$\|p_X\|_\infty$, possibly $+\infty$
ν_θ	pushforward $f_{\theta\#}\mu$, the data-induced measure on Y_θ
Network architecture	
f_θ	L-layer invertible network $W_L \circ \sigma_{L-1} \circ \cdots \circ \sigma_1 \circ W_1$
W_l	invertible $d \times d$ linear layer
σ_l	elementwise C^1 diffeomorphism activation
Ψ_θ	inverse f_θ^{-1}
$u_l(y)$	intermediate inverse preactivation
Y_θ	activation image, output domain of f_θ
X_0	input domain
Pushforward density	
$C_W(\theta)$	$\prod_{l=1}^{L} \lvert\det W_l\rvert^{-1}$, linear volume factor
$\rho_\theta(y)$	$\prod_{l=1}^{L-1} \lvert\det D\sigma_l^{-1}(u_l(y))\rvert$, activation inverse-Jacobian factor
$p_\theta(y)$	$C_W(\theta)\, p_X(\Psi_\theta(y))\, \rho_\theta(y)\, \mathbf{1}_{Y_\theta}(y)$, data-induced pushforward density
Relaxation hierarchy	
B_{path}	$\|v \circ f_\theta\|^2_{L^2(\mu)}$, tightest data-dependent quantity
$B_{\mathrm{sup,data}}$	$\|v\|^2_{L^2} \cdot \operatorname{ess\,sup}_y p_\theta(y)$, data-dependent bound (this paper)
$B_{\mathrm{wc}}^{\mathrm{glob}}$	$C_W P_X \|v\|^2_{L^2} \cdot \operatorname{ess\,sup}_y \rho_\theta(y)$, global worst-case
$B_{\mathrm{wc}}^{\mathrm{layer}}$	$C_W P_X \|v\|^2_{L^2} \cdot \prod_l \operatorname{ess\,sup}_y \lvert\det D\sigma_l^{-1}\rvert$, layerwise worst-case
Hypothesis classes and complexity	
v	output function $Y_\theta \to \mathbb{R}$ (or $\mathbb{C}$)
$h_{\theta,v}$	composed predictor $x \mapsto v(f_\theta(x))$
$\mathcal{H}_k$	RKHS with kernel k
$\mathcal{V}_B$	$\{v \in \mathcal{H}_k : \|v\|_{\mathcal{H}_k} \le B\}$
$\mathcal{F}_\theta$	$\{x \mapsto \langle a, \Phi(f_\theta(x))\rangle : \|a\|_2 \le B\}$
$\widehat{\mathfrak{R}}_S(\mathcal{F})$	empirical Rademacher complexity on sample S
Shallow network objects	
$f_w(x)$	$\tanh(wx)$, single-layer scalar network
ϕ	elementwise activation ($\tanh$, sigmoid, softplus)
$L(y)$	$\log(y/(1-y))$, logit (inverse of sigmoid)
$\mathrm{sp}^{-1}(y)$	$\log(e^y - 1)$, inverse softplus
Deep network objects	
f_K	$\tanh^{\circ K}$, K-fold composition of $\tanh$
a_m	m-fold tanh iterate, $a_0 = 1$, $a_{m+1} = \tanh(a_m)$
D_K	$\prod_{i=0}^{K-2} \cosh^2(a_i)$, interior inverse-Jacobian product
b_K	output boundary of f_K, $b_K = a_{K-1} \sim \sqrt{3/(2(K-1))} \to 0$
Ψ_K	inverse of f_K
p_K	pushforward density of f_K under Gaussian input

Table 1: Notation used in the paper.

B RELATED WORK AND COMPARISON WITH EXISTING BOUNDS

We compare our proposed data-dependent bound with representative norm-based, compression-based, and Koopman-based generalization bounds. The last column indicates the main derivation route underlying each bound, including Rademacher complexity, margin analysis, and compression. Here, S denotes the sample size, while notation specific to each bound follows the corresponding reference.

Table 2: Comparison of the proposed bound with representative existing generalization bounds.

Authors	Rate	Type
Neyshabur et al. (2015)	$\frac{2^L \prod_{l=1}^L \|W_l\|_{2,2}}{\sqrt{S}}$	Norm-based
Neyshabur et al. (2018)	$\frac{L \max_l d_l \prod_{l=1}^L \|W_l\|}{\sqrt{S}} \left(\sum_{l=1}^L \frac{\|W_l\|_{2,2}^2}{\|W_l\|^2} \right)^{1/2}$	
Golowich et al. (2018)	$\left(\prod_{l=1}^L \|W_l\|_{2,2} \right) \min \left\{ \frac{1}{S^{1/4}}, \sqrt{\frac{L}{S}} \right\}$	
Bartlett et al. (2017)	$\frac{\prod_{l=1}^L \|W_l\|}{\sqrt{S}} \left(\sum_{l=1}^L \frac{\|W_l^T - A_l^T\|_{2,1}^{2/3}}{\|W_l\|^{2/3}} \right)^{3/2}$	
Wei & Ma (2020)	$\frac{(\sum_{l=1}^L \kappa_l^{2/3} \min\{L^{1/2}\|W_l - A_l\|_{2,2}, \|W_l - B_l\|_{1,1}\}^{2/3})^{3/2}}{\sqrt{S}}$	
Ju et al. (2022)	$\frac{\sum_{l=1}^L \theta_l \|W_l - A_l\|_{2,2}}{\sqrt{S}}$	
Li et al. (2021)	$\|\mathbf{x}\| \, \lvert \prod_{l=1}^L \|W_l\| - 1 \rvert + \gamma_{\mathbf{x}} + \sqrt{\frac{c_{\mathcal{X}}}{S}}$	
Arora et al. (2018)	$\hat{r} + \frac{L \max_i \|f(x_i)\|}{\hat{r}\sqrt{S}} \left(\sum_{l=1}^L \frac{1}{\mu_l^2 \mu_{l\to}^2} \right)^{1/2}$	Compression
Suzuki et al. (2020)	$\frac{\hat{r}}{\sqrt{S}} + \sqrt{\frac{L}{S}} \left(\sum_{l=1}^L \tilde{r}_l (\tilde{d}_{l-1} + \tilde{d}_l) \right)^{1/2}$	
Hashimoto et al. (2024)	$\frac{\|v\|_{H_L}}{\sqrt{S}} \prod_{l=1}^L \frac{G_l \|K_{\sigma_l}\|_{H_l} \|W_l\|^{s_l - 1}}{\det(W_l^* W_l)^{1/4}}$	Koopman-based
Hashimoto et al. (2026)	$\frac{\|v\|_{\mathcal{L}_L}}{\sqrt{S}} \prod_{l=1}^L \frac{G_l \|K_{\sigma_l}\|_{\mathcal{L}_l}}{\det(W_l^* W_l)^{1/4}}$	
This paper	$B\sqrt{\frac{Mr}{n}} + \frac{CBL_\Phi \sqrt{d}\, L \rho^{L-1} R_\Theta \sqrt{P}}{\sqrt{n}}$	Data-dependent

The first few classes of methods in the table rely only on the weight matrices W_l and are therefore referred to as norm-based bounds. Norm products penalize large weights and implicitly favor low-rank solutions, yet full-rank networks can exhibit equally good generalization performance in practice. This raises a natural question: how can high-rank weight matrices nevertheless support good generalization? Hashimoto et al. (2024) address this question by lifting each layer to a Koopman operator acting between function spaces and measuring complexity through operator norms. By a change of variables, the integration can be transferred to the output space, introducing an inverse Jacobian determinant that admits a layerwise decomposition. For a linear layer, the resulting contribution is the factor $\det(W_l^* W_l)^{-1/4}$, which decreases as the rank of the weight matrix increases. The nonlinear layers, in turn, contribute activation factors $\|K_{\sigma_l}\|$. For saturating activation functions, the inverse Jacobian diverges at the boundary of the output range, causing the corresponding operator norm to become infinite. To avoid this divergence, Hashimoto et al. (2024) replace the activation functions with globally Lipschitz and invertible surrogates.

Table 3: Comparison of the layerwise constants in the Koopman-based bounds and our data-dependent bound. The two bounds have the same Koopman structure, with different reference measures at the activation interfaces. For saturating activations, the Lebesgue-based factor is infinite, whereas the data-dependent factor $\sqrt{S_l}$ is finite by Theorem 3.1.

Factor	Koopman bound	This paper
Linear layer l	$\lvert \det W_l \rvert^{-1/2}$	$\lvert \det W_l \rvert^{-1/2}$
Activation layer l	$\sup_z \lvert \det D\sigma_l^{-1}(z) \rvert^{1/2}$ ($= \infty$ for $\tanh$)	$\sqrt{S_l}$ (data-dependent)
Readout and sample	$\|v\|_{L^2}$, $1/\sqrt{n}$	$B\sqrt{r}$ or $B\sqrt{D_k}$, $1/\sqrt{n}$
Data dependence	None	S_l

Hashimoto et al. (2026) subsequently reformulate this framework algebraically in an RKHS, removing the smoothness requirement and allowing genuinely saturating activation functions such as $\tanh$. In this setting, the boundary singularity is controlled by assuming bounded inputs, $X \in [-A, A]$, which makes the corresponding quantity finite. This assumption, however, comes at a cost. Over the

preactivation range $[-A, A]$, a $\tanh$ layer incurs a layerwise constant of $\cosh^2(A)$, which diverges as e^{2A} when the discarded tail probability tends to zero.

Our bound follows the same Koopman framework, but differs in its choice of reference measure. By Lemma C.3, $M(\theta) = \operatorname{ess\,sup}_y p_\theta(y)$ is exactly the squared norm of the Koopman operator associated with f_θ, evaluated with respect to the data measure μ rather than the Lebesgue measure. The two bounds can be matched factor by factor (see Theorem 3.1). The linear layers contribute the same determinant factor, and the readout terms are also identical. The key difference arises at the interfaces involving saturating activation functions. Under the Lebesgue reference measure, the inverse Jacobian factor diverges. Under the data measure, however, the same singular factor is coupled with the transformed input density $p_X \circ \Psi_\theta$, which decays precisely in the region where the Jacobian diverges, as established in Theorem 3.1. Thus, the singularity that causes the Lebesgue-based bound to diverge is compensated by the input density under the data measure.

C Auxiliary Analytical Lemmas

This appendix collects auxiliary lemmas used in the main text.

C.1 Setup details and proofs for Section 2

Network factorization. For $l = 1, \dots, L-1$, the intermediate inverse preactivation is $u_l(y) := W_{l+1}^{-1} \circ \sigma_{l+1}^{-1} \circ \cdots \circ W_L^{-1}(y)$, with $u_{L-1}(y) = W_L^{-1}(y)$. The chain rule gives $|\det D\Psi_\theta(y)| = C_W(\theta)\, \rho_\theta(y)$. For elementwise activations $\sigma_l(x) = (\phi_l(x_1), \dots, \phi_l(x_d))$, $\rho_\theta(y) = \prod_{l=1}^{L-1} \prod_{j=1}^{d} |(\phi_l^{-1})'([u_l(y)]_j)|$, separating the contribution of the invertible linear layers from the nonlinear activation inverses.

The relaxation hierarchy. Recall from Section 2 the data-induced pushforward density

$$p_\theta(y) = C_W(\theta)\, p_X(\Psi_\theta(y))\, \rho_\theta(y)\, \mathbf{1}_{Y_\theta}(y).$$

The data-dependent bound sits inside a relaxation hierarchy. With $P_X := \|p_X\|_\infty$:

$$B_{\mathrm{sup,data}}(v,\theta) := \|v\|_{L^2}^2 \operatorname*{ess\,sup}_y p_\theta(y) \qquad \text{(data-dependent)}$$

$$B_{\mathrm{wc}}^{\mathrm{glob}}(v,\theta) := C_W(\theta)\, P_X\, \|v\|_{L^2}^2 \operatorname*{ess\,sup}_y \rho_\theta(y) \qquad \text{(global worst-case)}$$

$$B_{\mathrm{wc}}^{\mathrm{layer}}(v,\theta) := C_W(\theta)\, P_X\, \|v\|_{L^2}^2 \prod_{l=1}^{L-1} \operatorname{ess\,sup}_y \left|\det D\sigma_l^{-1}(u_l(y))\right| \qquad \text{(layerwise worst-case)}$$

These satisfy $B_{\mathrm{path}} \le B_{\mathrm{sup,data}} \le B_{\mathrm{wc}}^{\mathrm{glob}} \le B_{\mathrm{wc}}^{\mathrm{layer}}$, where $B_{\mathrm{path}} := \|v \circ f_\theta\|_{L^2(\mu)}^2$ (Proposition C.2). Worst-case bounds arise by replacing p_θ with increasingly coarse suprema; they can diverge even when the data-dependent quantity remains finite, which is the phenomenon this paper studies.

Lemma C.1 (Pushforward identity)**.** *For any measurable $v : Y_\theta \to \mathbb{C}$,*

$$\|v \circ f_\theta\|_{L^2(\mu)}^2 = \int_{Y_\theta} |v(y)|^2\, d\nu_\theta(y) = C_W(\theta) \int_{Y_\theta} |v(y)|^2 p_X(\Psi_\theta(y)) \rho_\theta(y)\, dy,$$

with both sides possibly $+\infty$.

The first identity is the pushforward measure definition applied to $g(y) = |v(y)|^2$: $\int_{X_0} |v(f_\theta(x))|^2\, d\mu(x) = \int_{Y_\theta} |v(y)|^2\, d\nu_\theta(y)$. If μ has density p_X, the change-of-variables formula for the C^1 diffeomorphism f_θ gives $\int_{X_0} |v(f_\theta(x))|^2 p_X(x)\, dx = \int_{Y_\theta} |v(y)|^2 p_X(\Psi_\theta(y)) |\det D\Psi_\theta(y)|\, dy$. Substituting $|\det D\Psi_\theta(y)| = C_W(\theta)\rho_\theta(y)$ from the chain-rule factorization yields the third form. Both sides may be $+\infty$ by convention.

Proposition C.2 (Relaxation hierarchy)**.** *For any nonzero $v \in L^2(Y_\theta)$,*

$$B_{\mathrm{path}} \le B_{\mathrm{sup,data}} \le B_{\mathrm{wc}}^{\mathrm{glob}} \le B_{\mathrm{wc}}^{\mathrm{layer}}.$$

If $P_X = \infty$, the last two bounds are interpreted as $+\infty$.

By Lemma C.1, $B_{\mathrm{path}}(v,\theta;\mu) = \int_{Y_\theta} |v(y)|^2 p_\theta(y)\,dy$. Since $p_\theta(y) \le \operatorname{ess\,sup}_{z\in Y_\theta} p_\theta(z)$ a.e., $B_{\mathrm{path}} \le B_{\mathrm{sup,data}}$. From equation C.1, $p_\theta(y) \le C_W(\theta) P_X \rho_\theta(y)$; taking essential suprema gives $B_{\mathrm{sup,data}} \le B_{\mathrm{wc}}^{\mathrm{glob}}$. For the final inequality, write $\rho_\theta(y) = \prod_{l=1}^{L-1} a_l(y)$ with $a_l(y) := |\det D\sigma_l^{-1}(u_l(y))|$. For nonnegative measurable functions, $\operatorname{ess\,sup}_y \prod_l a_l(y) \le \prod_l \operatorname{ess\,sup}_y a_l(y)$, yielding $B_{\mathrm{wc}}^{\mathrm{glob}} \le B_{\mathrm{wc}}^{\mathrm{layer}}$.

Lemma C.3 (The Koopman interface)**.** *For a C^1 diffeomorphism $g : \mathbb{R}^d \to \mathbb{R}^d$ and a probability measure μ, let $\mathcal{K}_g : L^2(\mathbb{R}^d,\lambda) \to L^2(\mathbb{R}^d,\mu)$, $\mathcal{K}_g v := v \circ g$. If $\nu := g_{\#}\mu$ has density p_ν, then $\|\mathcal{K}_g\|^2 = \operatorname{ess\,sup}_y p_\nu(y)$.*

$\|\mathcal{K}_g v\|^2_{L^2(\mu)} = \int |v|^2 p_\nu\,dy \le (\operatorname{ess\,sup} p_\nu)\|v\|^2_{L^2}$, with optimality by concentrating v near a maximizer of p_ν.

Thus $M(\theta) := \operatorname{ess\,sup}_y p_\theta(y)$ is the squared norm of the Koopman operator of f_θ measured against the data measure; Koopman-based analyses (Hashimoto et al., 2024; 2026) bound the same object through Lebesgue-space factorizations whose per-layer constants are infinite for saturating activations.

Lemma C.4 (Gaussian regularization of the tanh boundary)**.** *For $\beta > 0$,*

$$\sup_{s\in(-1,1)} \frac{\exp\!\big(-\operatorname{arctanh}(s)^2/(2\beta^2)\big)}{1-s^2} \le \max\Big\{\tfrac{4}{3},\, e^{2\beta^2}\Big\} < \infty.$$

Proof. On $|s| \le 1/2$, $1-s^2 \ge 3/4$ and the exponential is at most 1, so the ratio is at most $4/3$. For $s = 1-t$ with $0 < t < 1/2$, the expansion $\operatorname{arctanh}(1-t) = \frac{1}{2}\log(2/t) + O(t)$ gives $\operatorname{arctanh}(1-t) \ge \frac{1}{2}\log(1/t)$ and $1-s^2 = t(2-t) \ge t$, so the ratio is at most $t^{-1}\exp(-\log^2(1/t)/(8\beta^2))$; with $u = \log(1/t)$ this is $\exp(u - u^2/(8\beta^2))$, and the concave quadratic $u - u^2/(8\beta^2)$ has supremum $2\beta^2$. The left boundary $s = -1+t$ is identical, since $\operatorname{arctanh}$ is odd and enters quadratically. □

C.2 Asymptotics of tanh iterates

Lemma C.5 (Tanh iterates)**.** *Let $a_0 = 1$, $a_{m+1} = \tanh(a_m)$. Then $a_m \sim \sqrt{3/(2m)}$ as $m \to \infty$. Moreover, $\sum_{m=1}^K a_m^2 = \frac{3}{2}\log K + O(1)$ and $\sum_{m=1}^\infty a_m^4 < \infty$.*

Since $0 < \tanh(x) < x$ for $x > 0$, (a_m) decreases to some $\ell \ge 0$; continuity gives $\ell = \tanh(\ell)$, hence $\ell = 0$. For small x, $\tanh(x) = x - x^3/3 + O(x^5)$, so $a_{m+1} = a_m - a_m^3/3 + O(a_m^5)$. With $u_m = a_m^{-2}$, $u_{m+1} = a_m^{-2}(1 - a_m^2/3 + O(a_m^4))^{-2} = u_m(1 + 2a_m^2/3 + O(a_m^4))$ using $(1-z)^{-2} = 1 + 2z + O(z^2)$. Since $a_m^2 = u_m^{-1}$, $u_{m+1} = u_m + \frac{2}{3} + O(a_m^2)$, and $a_m \to 0$ gives $u_{m+1} - u_m \to \frac{2}{3}$. The Stolz–Cesàro theorem with $y_m = m$ gives $u_m/m \to \frac{2}{3}$, hence $a_m^{-2} \sim 2m/3$, i.e. $a_m \sim \sqrt{3/(2m)}$. Then $a_m^2 \sim 3/(2m)$, so $\sum_{m=1}^K a_m^2 = \frac{3}{2}\sum_{m=1}^K (1/m) + O(1) = \frac{3}{2}\log K + O(1)$. Finally $a_m^4 = O(m^{-2})$, so $\sum_{m=1}^\infty a_m^4 < \infty$.

C.3 Logarithmic tail perturbations

In the shallow boundary asymptotics, we frequently replace $p_X(\alpha\log(1/t) + O(t))$ by $p_X(\alpha\log(1/t))$, or by similar expressions with additive $O(1)$ shifts. The following lemma justifies this for the tail families used in the paper.

Lemma C.6 (Stability of exponential and polynomial tails under logarithmic perturbations)**.** *Let $\alpha > 0$, $u = \log(1/t)$. (i) If $p_X(x) \asymp e^{-cx^s}$ as $x \to +\infty$ for some $c, s > 0$, then for any bounded $\delta_t = O(1)$, $p_X(\alpha u + \delta_t) \asymp \exp(-c\alpha^s u^s)$; in particular, $p_X(\alpha u + O(t)) \asymp p_X(\alpha u)$. (ii) If $p_X(x) \asymp x^{-r}$ as $x \to +\infty$ for some $r > 0$, then $p_X(\alpha u + \delta_t) \asymp u^{-r}$. Analogous statements hold for left tails as $x \to -\infty$.*

We prove the right-tail statements; the left-tail versions are identical after replacing x by $|x|$. For the exponential-tail case, assume $c_1 e^{-cx^s} \le p_X(x) \le c_2 e^{-cx^s}$ for large x. Let δ_t be bounded. For large u, $\alpha u/2 \le \alpha u + \delta_t \le 2\alpha u$, so $p_X(\alpha u + \delta_t) \le c_2\exp(-c(\alpha u/2)^s)$ and $p_X(\alpha u + \delta_t) \ge c_1\exp(-c(2\alpha u)^s)$, both of the form $\exp(-Cu^s)$. For the sharper equivalence, expand $(\alpha u + \delta_t)^s = \alpha^s u^s + s\alpha^{s-1}\delta_t u^{s-1} + O(u^{s-2})$ when $s \ge 1$ (for $0 < s < 1$ the remainder is $o(1)$); in all cases

lower-order terms absorb into multiplicative constants, giving $p_X(\alpha u + \delta_t) \asymp \exp(-c\alpha^s u^s)$. For polynomial tails, $c_1 x^{-r} \le p_X(x) \le c_2 x^{-r}$ for large x and $\alpha u + \delta_t \asymp u$ give $p_X(\alpha u + \delta_t) \asymp u^{-r}$.

D Additional Details for Shallow Tail Calculus

D.1 The boundary asymptotics

Theorem D.1 (Tanh boundary asymptotics)**.** *If p_X is ultimately regular on logarithmic scales (in particular, of exponential or polynomial tail), then as $t \downarrow 0$,*

$$p_w(\pm 1 + t) = \frac{1 + o(1)}{2w\,t}\, p_X\big(\pm \tfrac{1}{2w} \log \tfrac{2}{t}\big) \asymp \tfrac{1}{t}\, p_X\big(\pm \tfrac{1}{2w} \log \tfrac{1}{t}\big)\,.$$

D.2 On the regularity assumption in Theorem D.1

The boundary asymptotic in Theorem D.1 was stated under the assumption that p_X is ultimately regular on logarithmic scales. Lemma C.6 makes this precise for the two main tail families used in the paper, $p_X(x) \asymp e^{-cx^s}$ and $p_X(x) \asymp x^{-r}$. For these families, the additive $O(t)$ perturbation in $\operatorname{arctanh}(1-t) = \frac{1}{2}\log(2/t) + O(t)$ does not affect the leading asymptotic after applying p_X: $O(t)$ is exponentially small relative to $\log(1/t)$, and Lemma C.6 shows that even bounded additive perturbations inside the argument of p_X only change constants. Thus the simplified asymptotic $p_w(1-t) \asymp t^{-1} p_X((2w)^{-1}\log(1/t))$ is valid for all exponential and polynomial tail classes considered in the main text.

D.3 Proofs for Section 3

Proof of Theorem 3.1. From Theorem D.1, $p_w(1-t) \asymp t^{-1} p_X((2w)^{-1}\log(1/t))$. Let $u = \log(1/t)$, so $t = e^{-u}$. For exponential tails $p_X((2w)^{-1}u) \asymp \exp(-c(2w)^{-s}u^s)$, hence $p_w(1-t) \asymp \exp(u - c(2w)^{-s}u^s)$. For integrability, $\int_0^\delta p_w(1-t)\,dt \asymp \int_{\log(1/\delta)}^\infty \exp(-c(2w)^{-s}u^s)\,du$, finite for every $s > 0$. For boundedness, the exponent $\Phi(u) = u - c(2w)^{-s}u^s$ satisfies: $\Phi(u) \to -\infty$ if $s > 1$; $\Phi(u) = (1 - c/(2w))u$ if $s = 1$, so p_w is bounded iff $c \ge 2w$; $\Phi(u) \to +\infty$ if $0 < s < 1$, so $p_w \to \infty$, but integrability is preserved. For polynomial tails, $p_X((2w)^{-1}\log(1/t)) \asymp (\log(1/t))^{-r}$, so $p_w(1-t) \asymp t^{-1}(\log(1/t))^{-r}$; with $u = \log(1/t)$, $\int_0^\delta p_w(1-t)\,dt \asymp \int_{\log(1/\delta)}^\infty u^{-r}\,du$, finite iff $r > 1$. Since $t^{-1}(\log(1/t))^{-r} \to \infty$ for every r, the density is never bounded near $y = 1$. □

Proof of Theorem 3.2. The standard Gaussian density is $\varphi(x) = (2\pi)^{-1/2}e^{-x^2/2}$, giving $p_w(y) = [w(1-y^2)]^{-1}\varphi(w^{-1}\operatorname{arctanh}(y))$. Since $w \ge a$ and $w \le b$,

$$p_w(y) \le \frac{1}{a\sqrt{2\pi}}\, \frac{\exp\big(-\operatorname{arctanh}(y)^2/(2b^2)\big)}{1-y^2}.$$

Applying Lemma C.4 with $\beta = b$ gives

$$\sup_{w\in[a,b],\, y\in(-1,1)} p_w(y) \le \frac{1}{a\sqrt{2\pi}} \max\Big\{\tfrac{4}{3},\, e^{2b^2}\Big\} < \infty.$$

□

Proposition D.2 (Compactly supported inputs)**.** *If $\operatorname{supp}(p_X) \subseteq [-A, A]$ and $p_X \in L^\infty$, then with $r := \tanh(bA) < 1$, p_w is supported in $[-r, r]$ and $\sup_{w\in[a,b],\, y} p_w(y) \le \|p_X\|_\infty/[a(1-r^2)]$.*

Proof of Proposition D.2. If $|X| \le A$, then $|Y| = |\tanh(wX)| \le \tanh(bA) =: r < 1$, so p_w is supported in $[-r, r]$. On this interval, $1 - y^2 \ge 1 - r^2 > 0$, hence $p_w(y) = [w(1-y^2)]^{-1} p_X(w^{-1}\operatorname{arctanh}(y)) \le \|p_X\|_\infty/[a(1-r^2)]$, independent of $w \in [a, b]$. □

Proof of Theorem 3.3. Write $p_X = \sum_{j=1}^m \pi_j \varphi_j$ with φ_j the $\mathcal{N}(\mu_j, \sigma_j^2)$ density, so $p_w(y) = [w(1-y^2)]^{-1} \sum_j \pi_j \varphi_j(w^{-1}\operatorname{arctanh}(y))$, and bound each component separately. *Middle region*

$|y| \le 1/2$: $1 - y^2 \ge 3/4$ and $\varphi_j \le (\sqrt{2\pi}\sigma_j)^{-1}$, so the j-th component contributes at most $4\pi_j/(3a\sqrt{2\pi}\sigma_j)$. *Right boundary* $y = 1 - t$, $0 < t < 1/2$: as in the proof of Lemma C.4, $\operatorname{arctanh}(1-t) \ge \frac{1}{2}\log(1/t)$ and $1 - y^2 \ge t$. Let $v = \log(1/t)$. If $v \ge 2b\mu_j$ (automatic when $\mu_j \le 0$), then $w^{-1}\operatorname{arctanh}(y) - \mu_j \ge v/(2b) - \mu_j \ge 0$, so the j-th component is at most

$$\frac{\pi_j}{a\sqrt{2\pi}\sigma_j}\,\exp\Big(v - \frac{(v/(2b) - \mu_j)^2}{2\sigma_j^2}\Big).$$

The exponent is a concave quadratic in v maximized at $v^* = 2b(\mu_j + 2b\sigma_j^2)$, which lies in the region $v \ge 2b\mu_j$, with maximum $2b\mu_j + 2b^2\sigma_j^2$. If instead $v < 2b\mu_j$, then $t^{-1} = e^v < e^{2b\mu_j}$, and dropping the negative exponential gives the smaller bound $\pi_j e^{2b\mu_j}/(a\sqrt{2\pi}\sigma_j)$, dominated by the same expression. The left boundary is identical with μ_j replaced by $-\mu_j$, so each component contributes at most $\pi_j(a\sqrt{2\pi}\sigma_j)^{-1}\max\{4/3,\, e^{2b|\mu_j| + 2b^2\sigma_j^2}\}$ across the whole interval. Summing over j gives the claim. □

D.4 Sigmoid boundary asymptotics

Proposition D.3 (Sigmoid boundary asymptotics)**.** *For $f_w(x) = \sigma(wx)$, $\Psi_w(y) = w^{-1}\log(y/(1-y))$, $\Psi_w'(y) = [w\,y(1-y)]^{-1}$, the pushforward density on $(0,1)$ is $p_w(y) = [w\,y(1-y)]^{-1}p_X(w^{-1}\log(y/(1-y)))$, with $p_w(t) = \frac{1+o(1)}{w\,t}p_X(-w^{-1}\log\frac{1}{t})$ and $p_w(1-t) = \frac{1+o(1)}{w\,t}p_X(w^{-1}\log\frac{1}{t})$ as $t \downarrow 0$. For $p_X(x) \asymp e^{-cx^s}$, p_w is bounded near $y = 1$ iff $s > 1$, or $s = 1$ and $c \ge w$; analogous at $y = 0$ via the left tail.*

Proof. The density formula follows from change of variables. Near $y = t$, $\log(t/(1-t)) = -\log(1/t) + O(t)$ and $t(1-t) = t(1 + O(t))$, giving $p_w(t) = \frac{1+o(1)}{w\,t}p_X(-w^{-1}\log(1/t))$. Near $y = 1 - t$, $\log((1-t)/t) = \log(1/t) + O(t)$ and $(1-t)t = t(1 + O(t))$, giving $p_w(1-t) = \frac{1+o(1)}{w\,t}p_X(w^{-1}\log(1/t))$. The tail consequences follow by the same substitution $u = \log(1/t)$ used in the proof of Theorem 3.1, with the coefficient w^{-s} in place of $(2w)^{-s}$ due to the halved logarithmic scale. □

D.5 Softplus boundary asymptotics

Proposition D.4 (Softplus boundary asymptotics)**.** *For $f_w(x) = \mathrm{sp}(wx)$, $\Psi_w(y) = w^{-1}\log(e^y - 1)$, $\Psi_w'(y) = [w(1 - e^{-y})]^{-1}$, the pushforward density on $(0,\infty)$ is $p_w(y) = [w(1 - e^{-y})]^{-1}p_X(w^{-1}\log(e^y - 1))$. As $t \downarrow 0$, $p_w(t) = \frac{1+o(1)}{w\,t}p_X(-w^{-1}\log\frac{1}{t})$, so boundedness near 0 (controlled by the left tail) holds iff $s > 1$, or $s = 1$ and $c \ge w$. As $y \to +\infty$, $\Psi_w'(y) \to 1/w$ and $p_w(y) \sim w^{-1}p_X(y/w)$, so Gaussian p_X yields bounded p_w on $(0,\infty)$.*

Proof. The density formula follows from change of variables. As $y = t \downarrow 0$, $e^t - 1 = t + O(t^2)$, so $\log(e^t - 1) = \log t + O(t)$ and $1 - e^{-t} = t + O(t^2)$; hence $p_w(t) = \frac{1+o(1)}{w\,t}p_X(w^{-1}\log t) = \frac{1+o(1)}{w\,t}p_X(-w^{-1}\log(1/t))$. As $y \to +\infty$, $e^y - 1 \sim e^y$, so $\log(e^y - 1) = y + o(1)$ and $1 - e^{-y} \to 1$, giving $p_w(y) \sim w^{-1}p_X(y/w)$. For Gaussian p_X, the boundary singularity at 0 is regularized by the left Gaussian tail, and the density decays at infinity. □

E Additional Details for Deep Tanh Networks

E.1 Constants in the density growth theorem

By Proposition E.1, $D_K = \Theta(K^{3/2})$, so there exist $0 < c_D < C_D < \infty$ with $c_D K^{3/2} \le D_K \le C_D K^{3/2}$ for all sufficiently large K. The lower bound in Theorem 3.6 uses $p_K(f_K(1)) = \varphi(1)D_{K+1} \ge \varphi(1)c_D(K+1)^{3/2} \ge cK^{3/2}$; the upper bound uses $p_K(f_K(x)) \le C_0 D_K \le C_0 C_D K^{3/2}$. Combining gives $\|p_K\|_\infty = \Theta(K^{3/2})$.

E.2 Proofs for the deep tanh theorems

Proof of Theorem 3.4. The inverse chain $A^{\circ K}$ is well-defined at positive y as long as every intermediate value remains in $(-1, 1)$; the maximal positive output occurs when the innermost preactivation approaches 1, giving $b_K = T^{\circ(K-1)}(1)$. The asymptotic $a_m \sim \sqrt{3/(2m)}$ is Lemma C.5, hence $b_K = a_{K-1} \sim \sqrt{3/(2(K-1))}$. □

Proof of Theorem 3.5. For $j = 0, \dots, K$, let $z_j(y) = A^{\circ j}(y)$, so $z_0(y) = y$, $z_K(y) = \Psi_K(y)$. At $y = b_K = a_{K-1}$, $z_j(b_K) = a_{K-1-j}$ for $j \le K-1$, with $z_{K-1}(b_K) = a_0 = 1$. Let $G_K(y) = z_{K-1}(y) = A^{\circ(K-1)}(y)$. For $K = 1$, G_1 is the identity; for $K \ge 2$, all intermediate boundary values a_{K-1-j} ($j \le K-2$) lie strictly inside $(-1, 1)$, so G_K is C^1 near b_K with $G_K'(b_K) = \prod_{j=0}^{K-2} A'(a_{K-1-j}) = \prod_{i=0}^{K-2} \cosh^2(a_i) = D_K$, where the last equality reindexes $i = K-1-j$ and uses $A'(a_m) = 1/(1-a_m^2) = \cosh^2(a_{m-1})$ since $1 - \tanh^2 = \operatorname{sech}^2$. For $y = b_K - t$, $G_K(b_K - t) = 1 - D_K t + o(t)$; letting $s = 1 - G_K(b_K - t) = D_K t + o(t)$, $\Psi_K(b_K - t) = A(1-s) = \frac{1}{2}\log((2-s)/s) = \frac{1}{2}\log(2/(D_K t)) + o(1) = \frac{1}{2}\log(1/t) + c_K + o(1)$. Differentiating, $\Psi_K'(y) = G_K'(y) A'(G_K(y))$; as $y \to b_K$, $G_K'(y) \to D_K$ and $A'(1-s) = 1/(1-(1-s)^2) = 1/(2s + O(s^2)) \sim 1/(2D_K t)$, so $\Psi_K'(b_K - t) \sim D_K \cdot (2D_K t)^{-1} = 1/(2t)$. The density formula follows from $p_K(y) = p_X(\Psi_K(y))\Psi_K'(y)$. □

Proposition E.1 (The depth constant)**.** *Let* $D_K := \prod_{i=0}^{K-2} \cosh^2(a_i)$. *Then* $c_K = \frac{1}{2}\log(2/D_K) = -\frac{3}{4}\log K + O(1)$, *and* $D_K = \Theta(K^{3/2})$.

Proof of Proposition E.1. For small x, $\log\cosh^2(x) = x^2 + O(x^4)$, so $\log D_K = \sum_{i=0}^{K-2} \log\cosh^2(a_i) = \sum_{i=0}^{K-2} a_i^2 + O\big(\sum_{i=0}^{K-2} a_i^4\big)$. Lemma C.5 gives $\sum_{i=1}^{K} a_i^2 = \frac{3}{2}\log K + O(1)$ and $\sum_{i=1}^{\infty} a_i^4 < \infty$, hence $\log D_K = \frac{3}{2}\log K + O(1)$, i.e. $D_K = \Theta(K^{3/2})$, and $c_K = \frac{1}{2}\log(2/D_K) = -\frac{3}{4}\log K + O(1)$. □

Proof of Theorem 3.6. Since f_K is increasing and symmetric, it suffices to analyze $x \ge 0$. With the forward orbit $x_0 = x$, $x_{j+1} = T(x_j)$, $f_K(x) = x_K$ and $f_K'(x) = \prod_{j=0}^{K-1} \operatorname{sech}^2(x_j)$, so $\Psi_K'(f_K(x)) = \prod_{j=0}^{K-1} \cosh^2(x_j)$ and $p_K(f_K(x)) = \varphi(x) \prod_{j=0}^{K-1} \cosh^2(x_j)$. *Lower bound.* At $x = 1$, $x_j = a_j$, so $p_K(f_K(1)) = \varphi(1) \prod_{j=0}^{K-1} \cosh^2(a_j) = \varphi(1) D_{K+1} = \Theta(K^{3/2})$ by Proposition E.1. *Upper bound.* Factor $p_K(f_K(x)) = [\varphi(x)\cosh^2(x)] \prod_{j=1}^{K-1} \cosh^2(x_j)$. The first factor is uniformly bounded: as $x \to \infty$, $\varphi(x)\cosh^2(x) \sim (4\sqrt{2\pi})^{-1} e^{-x^2/2+2x} \to 0$, so $C_0 := \sup_{x \ge 0} \varphi(x)\cosh^2(x) < \infty$. Since $x_1 = T(x) \in [0, 1)$ and T is increasing, $x_j = T^{\circ(j-1)}(x_1) \le T^{\circ(j-1)}(1) = a_{j-1}$ for $j \ge 1$; as $\cosh^2$ is increasing on $[0, \infty)$, $\prod_{j=1}^{K-1} \cosh^2(x_j) \le \prod_{j=1}^{K-1} \cosh^2(a_{j-1}) = D_K = \Theta(K^{3/2})$. Hence $p_K(f_K(x)) \le C_0 D_K = O(K^{3/2})$. The case $x \le 0$ follows by symmetry. Combining, $\|p_K\|_\infty = \Theta(K^{3/2})$. □

F Additional Details for Deep Softplus Networks

F.1 Proofs for deep softplus

Theorem F.1 (Deep softplus, exact form)**.** *Let* $S(x) = \log(1 + e^x)$. *Then* $S^{\circ K}(x) = \log(e^x + K)$, *taking values in* $(\log K, \infty)$ *with inverse derivative* $\Psi_K'(y) = e^y/(e^y - K)$. *Near the lower boundary,* $p_K(\log K + t) = \frac{1+o(1)}{t}\, p_X(\log K + \log t + o(1)),\ t \downarrow 0$.

Theorem F.2 (Deep softplus under Gaussian inputs)**.** *If* $X \sim \mathcal{N}(0, 1)$, *then* $\|p_K\|_\infty = \Theta(K)$.

Proof of Theorem F.1. We prove $S_K(x) = \log(e^x + K)$ by induction. For $K = 1$, $S_1(x) = \log(1+e^x) = \log(e^x+1)$. Assuming $S_K(x) = \log(e^x+K)$, then $S_{K+1}(x) = S(S_K(x)) = \log(1 + e^{S_K(x)}) = \log(1+e^x+K) = \log(e^x+K+1)$. Since $e^x > 0$, the image is $(\log K, \infty)$. The inverse is obtained from $y = \log(e^x + K)$: $e^y = e^x + K$, so $x = \log(e^y - K)$, and differentiating gives $\Psi_K'(y) = e^y/(e^y - K)$. The density formula $p_K(y) = p_X(\Psi_K(y))\Psi_K'(y) = \frac{e^y}{e^y - K} p_X(\log(e^y -$

K)) follows by change of variables. For the equivalent x-coordinate form, $S'_K(x) = e^x/(e^x+K) = 1/(1+Ke^{-x})$, so $\Psi'_K(S_K(x)) = 1/S'_K(x) = 1+Ke^{-x}$ and $p_K(S_K(x)) = p_X(x)(1+Ke^{-x})$. For the boundary, let $y = \log K + t$; then $e^y = Ke^t = K(1+t+O(t^2))$, so $e^y - K = K(e^t-1) = K(t+O(t^2))$, giving $\log(e^y-K) = \log K + \log t + o(1)$ and $e^y/(e^y-K) = e^t/(e^t-1) \sim 1/t$. □

Proof of Theorem F.2. By Theorem F.1, $\|p_K\|_\infty = \sup_{x\in\mathbb{R}} \varphi(x)(1+Ke^{-x})$. *Lower bound*: at $x = 0$, $\|p_K\|_\infty \geq \varphi(0)(1+K) = (1+K)/\sqrt{2\pi} = \Theta(K)$. *Upper bound*: $\varphi(x)(1+Ke^{-x}) \leq \varphi(x) + Ke^{-x}\varphi(x)$. The first term satisfies $\sup_x \varphi(x) = \varphi(0) = 1/\sqrt{2\pi}$. For the second, $e^{-x}\varphi(x) = (2\pi)^{-1/2}\exp(-x^2/2 - x) = (2\pi)^{-1/2}\exp(-\frac{1}{2}(x+1)^2 + \frac{1}{2})$, so $\sup_x e^{-x}\varphi(x) = e^{1/2}/\sqrt{2\pi}$. Therefore $\|p_K\|_\infty \leq (1+Ke^{1/2})/\sqrt{2\pi} = O(K)$. Combining, $\|p_K\|_\infty = \Theta(K)$. □

G Multidimensional and Affine Extensions

This appendix extends the one-dimensional results in three directions: multidimensional product networks (where the pushforward density factorizes coordinatewise), a genuinely multidimensional two-layer affine tanh network (where the singular boundary is no longer axis-aligned), and deep chains with arbitrary mixing (where a cut-norm recursion gives finite, explicit, and computable density bounds). Throughout, all densities are with respect to Lebesgue measure and all suprema are essential.

G.1 Product networks

Let $X = (X_1, \ldots, X_d)$ have independent coordinates with densities p_j^X, let $f_j : \mathbb{R} \to Y_j \subseteq \mathbb{R}$ be C^1 diffeomorphisms, and let $f(x) = (f_1(x_1), \ldots, f_d(x_d))$. Denote by q_j the pushforward density of $Y_j = f_j(X_j)$.

Proposition G.1 (Factorization of product pushforward densities)**.** *$p_f(y) = \prod_{j=1}^d q_j(y_j)$, hence $\|p_f\|_\infty = \prod_{j=1}^d \|q_j\|_\infty$ (with the convention that the product may be infinite).*

Proof of Proposition G.1. Since the coordinates of X are independent, $\mu = \bigotimes_{j=1}^d \mu_j$. Because f is a product map, $f_\#\mu = \bigotimes_{j=1}^d (f_j)_\#\mu_j$, which has density $p_f(y) = \prod_{j=1}^d q_j(y_j)$. The inequality $\|p_f\|_\infty \leq \prod_j \|q_j\|_\infty$ is immediate. For the reverse, assume each $\|q_j\|_\infty < \infty$. Given $\varepsilon > 0$, for each j there exists a measurable set $A_j \subseteq Y_j$ of positive measure with $q_j(y_j) \geq \|q_j\|_\infty - \varepsilon$ on A_j. Then $A_1 \times \cdots \times A_d$ has positive product measure and on it $p_f(y) \geq \prod_j(\|q_j\|_\infty - \varepsilon)$; letting $\varepsilon \downarrow 0$ gives the reverse inequality. The case where some $\|q_j\|_\infty = \infty$ follows similarly. □

Corollary G.2 (Gaussian product networks)**.** *Let $X_j \sim \mathcal{N}(0,1)$ independently. (i) Shallow: if $f_j(x_j) = \tanh(w_j x_j)$ with $w_j \in [a_j, b_j] \subset (0,\infty)$, then $\sup_{w_1,\ldots,w_d} \|p_f\|_\infty \leq \prod_j M_j < \infty$, where $M_j = \sup_{w_j\in[a_j,b_j]} \|q_j\|_\infty$ is finite by Theorem 3.2. (ii) Deep coordinatewise tanh with depths K_j: $\|p_{K_1,\ldots,K_d}\|_\infty = \Theta\Big(\prod_j K_j^{3/2}\Big)$ by Theorem 3.6. (iii) Deep coordinatewise softplus with depth K: $\|p_{K,d}\|_\infty = \Theta(K^d)$ by Theorem F.2.*

G.2 Two-layer affine tanh networks

Theorem G.3 (Uniform pushforward density bound for two-layer affine tanh)**.** *Let $X \sim \mathcal{N}(0, I_d)$, $f(x) = W_2 T(W_1 x)$ with T applied elementwise, $W_1, W_2 \in \mathbb{R}^{d\times d}$ invertible. If the singular values of W_1 lie in $[\alpha, \beta] \subset (0,\infty)$ and $|\det W_2| \geq \delta > 0$, then*

$$\|p_f\|_\infty \leq \frac{M_\beta^d}{\delta(2\pi)^{d/2}\alpha^d}, \qquad M_\beta := \sup_{s\in(-1,1)} \frac{\exp(-\operatorname{arctanh}(s)^2/(2\beta^2))}{1-s^2} < \infty.$$

The bound depends only on α, β, δ, d, so $\sup_{W_1,W_2} \|p_f\|_\infty < \infty$ over the constrained class. The determinant condition on W_2 rules out near-singular compression of the output volume; the singular-value bounds on W_1 keep the Gaussian input nondegenerate after the first linear map.

Proof of Theorem G.3. Let $Z = W_1 X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = W_1 W_1^\top$. Since the singular values of W_1 lie in $[\alpha, \beta]$, the eigenvalues of Σ lie in $[\alpha^2, \beta^2]$, so $|\Sigma|^{1/2} \geq \alpha^d$ and $z^\top \Sigma^{-1} z \geq \|z\|^2/\beta^2$, giving

$$p_Z(z) \leq (2\pi)^{-d/2} \alpha^{-d} \exp(-\|z\|^2/(2\beta^2)).$$

Let $U = T(Z)$, $Y = W_2 U$. With $A = \operatorname{arctanh}$ applied elementwise, $p_U(u) = p_Z(A(u)) \prod_j (1 - u_j^2)^{-1}$, and $p_f(y) = |\det W_2|^{-1} p_U(W_2^{-1} y)$, so $\|p_f\|_\infty \leq \delta^{-1} \|p_U\|_\infty$. Substituting,

$$p_U(u) \leq (2\pi)^{-d/2} \alpha^{-d} \prod_{j=1}^{d} h_\beta(u_j), \qquad h_\beta(s) := (1 - s^2)^{-1} \exp(-\operatorname{arctanh}(s)^2/(2\beta^2)).$$

Lemma C.4 gives $M_\beta = \sup_{s \in (-1,1)} h_\beta(s) \leq \max\{4/3, e^{2\beta^2}\} < \infty$, and combining yields $\|p_f\|_\infty \leq M_\beta^d / (\delta (2\pi)^{d/2} \alpha^d)$, independent of W_1, W_2. □

G.3 Deep chains with arbitrary mixing: a cut-norm recursion

The results above cover compositions without mixing (W_l absent or product-structured) and a single mixing layer. We now show that finite, explicit density bounds exist for deep square chains with *arbitrary* invertible mixing. The key structural observation is that the tail-versus-singularity competition happens only at the first layer: in the chain $Z_0 = X$, $U_l = W_l Z_{l-1}$, $Z_l = T(U_l)$ with coordinatewise $\tanh$ and invertible $W_l \in \mathbb{R}^{d \times d}$, every $Z_{l-1} \in (-1, 1)^d$ is bounded for $l \geq 2$, so only deterministic Jacobian accumulation remains. If moreover $\operatorname{supp} p_X = \mathbb{R}^d$, induction gives $\operatorname{supp} Z_l = (-1, 1)^d$.

Proof of Proposition 4.4. Change of variables gives $p_{U_l}(u) = p_{Z_{l-1}}(W_l^{-1} u) \, |\det W_l|^{-1} \leq \|p_{Z_{l-1}}\|_\infty |\det W_l|^{-1}$ pointwise, with $\operatorname{supp} U_l \subseteq W_l (-1, 1)^d$. Then $\|p_{Z_l}\|_\infty = \sup_{u \in \operatorname{supp} U_l} p_{U_l}(u) \prod_j \cosh^2 u_j$, and $\prod_j \cosh^2 u_j \leq e^{\Phi_{\mathrm{sat}}(W_l)}$ on $W_l(-1, 1)^d$, since the maximum of the convex function $\sum_j \log \cosh^2 u_j$ over the box is attained at a vertex. Iterating the recursion gives the summed bound. The sandwich claim and computability of Φ_{sat} are proved in Appendix I.3. □

For the first layer, direct computation gives the Gaussian-tanh cancellation: with $\Sigma_1 = W_1 \Sigma_0 W_1^\top$, coordinatewise use of $u^\top \Sigma_1^{-1} u \geq \|u\|^2 / \lambda_{\max}(\Sigma_1)$ gives $\log \|p_{Z_1}\|_\infty \leq 2d \, \lambda_{\max}(\Sigma_1) - \frac{1}{2} \log \det \Sigma_1 - O(d)$, with equality in the leading term for diagonal Σ_1.

The recursion is nearly tight: lower bounds from corner paths (Proposition I.2) match the first-layer term and differ only in whether the vertex maximizing each layer's potential can be realized by a single orbit of the deterministic map $z \mapsto \tanh(Wz)$. The remaining gap is thus a combinatorial question about this discrete dynamics, independent of density estimation; we state it as an open problem.

H Additional Details for Data-Dependent Complexity Consequences

H.1 Proofs for Section 4

Proof of Proposition 4.1(iii). Fix $\theta \in \Theta$. For a sample $X_1, \ldots, X_n$, let $\Phi_i = \Phi(f_\theta(X_i))$. The empirical Rademacher complexity of $\mathcal{F}_\theta = \{x \mapsto \langle a, \Phi(f_\theta(x)) \rangle : \|a\|_2 \leq B\}$ satisfies, by Cauchy–Schwarz,

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_\theta) = \mathbb{E}_\epsilon \sup_{\|a\|_2 \leq B} \Big\langle a, \tfrac{1}{n} \sum_i \epsilon_i \Phi_i \Big\rangle = \frac{B}{n} \mathbb{E}_\epsilon \Big\| \sum_i \epsilon_i \Phi_i \Big\|_2.$$

Jensen and $\mathbb{E}_\epsilon[\epsilon_i \epsilon_j] = \delta_{ij}$ give $\mathbb{E}_\epsilon \| \sum_i \epsilon_i \Phi_i \|_2 \leq (\sum_i \|\Phi_i\|_2^2)^{1/2}$. Taking expectation over the sample and applying Jensen again,

$$\mathfrak{R}_n(\mathcal{F}_\theta) \leq \frac{B}{n} \Big(\sum_i \mathbb{E} \|\Phi(f_\theta(X_i))\|_2^2 \Big)^{1/2} = B \sqrt{\frac{\mathbb{E} \|\Phi(f_\theta(X))\|_2^2}{n}}.$$

Now $\mathbb{E}\|\Phi(f_\theta(X))\|_2^2 = \sum_{j=1}^r \int_\Omega \psi_j(y)^2 p_\theta(y)\,dy$. Since $p_\theta \le M_\theta$ and the ψ_j are orthonormal, this is at most $M_\theta \sum_j \int_\Omega \psi_j^2\,dy = M_\theta r$. Hence $\mathfrak{R}_n(\mathcal{F}_\theta) \le B\sqrt{M_\theta r/n}$. This is a per-$\theta$ statement; controlling the union class $\bigcup_{\theta\in\Theta}\mathcal{F}_\theta$ requires the covering argument of Theorem 4.2, proved in Appendix I. □

Proof of Proposition 4.1(RKHS case). Let $y_i = f_\theta(X_i)$. By the RKHS representer identity, $\sup_{\|v\|_{\mathcal{H}_k}\le B}\sum_i a_i v(y_i) = B(\sum_{i,j} a_i a_j k(y_i,y_j))^{1/2}$. Taking $a_i = \epsilon_i/n$,

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_\theta) = \frac{B}{n}\mathbb{E}_\epsilon(\epsilon^\top K_\theta \epsilon)^{1/2},$$

where $(K_\theta)_{ij} = k(y_i,y_j)$. Jensen gives $\mathbb{E}_\epsilon(\epsilon^\top K_\theta\epsilon)^{1/2} \le (\mathbb{E}_\epsilon\,\epsilon^\top K_\theta\epsilon)^{1/2} = (\operatorname{tr} K_\theta)^{1/2} = (\sum_i k(y_i,y_i))^{1/2}$, yielding the empirical bound. Taking expectation over the sample and applying Jensen,

$$\mathfrak{R}_n(\mathcal{F}_\theta) \le \frac{B}{n}\mathbb{E}\Big(\sum_i k(f_\theta(X_i),f_\theta(X_i))\Big)^{1/2} \le \frac{B}{\sqrt{n}}\big(\mathbb{E}\,k(f_\theta(X),f_\theta(X))\big)^{1/2}.$$

Using the pushforward density, $\mathbb{E}\,k(f_\theta(X),f_\theta(X)) = \int_\Omega k(y,y)p_\theta(y)\,dy$. If $p_\theta \le M$, this is at most $M\int_\Omega k(y,y)\,dy = MD_k$, giving $\mathfrak{R}_n(\mathcal{F}_\theta) \le B\sqrt{MD_k/n}$. □

H.2 The pointwise log-density identity behind the estimator

For the $\tanh$ chain of Section 2 ($Z_0 = X$, $U_l = W_l Z_{l-1}$, $Z_l = \tanh(U_l)$), iterating the change of variables formula along a single sample path gives the pointwise identity

$$\log p_{Z_L}(z_L(x)) = \log p_X(x) - \sum_{l=1}^L \log|\det W_l| + \sum_{l=1}^L\sum_{j=1}^d \log\cosh^2 u_{lj}(x),$$

since $(\tanh^{-1})'(z) = 1/(1-z^2)$, so each inverse activation factor contributes $\log\cosh^2 u_{lj}$. Tracking these terms layer by layer is numerically fragile, which motivates the Gaussian-fit estimator of Section 4.4.

I Proofs for the Improved Complexity Results

I.1 Proof of Theorem 4.2 (full-parameter covering bound)

Recall the setting: $\theta = (W_1,\dots,W_L)$, $\|W_l\|_{\mathrm{op}} \le \rho$, Θ contained in a ball of radius R_Θ in $\mathbb{R}^P$, $P = Ld^2$, metric $\|\theta-\theta'\| = \max_l \|W_l - W_l'\|_{\mathrm{op}}$; $f_\theta = T\circ W_L\circ\cdots\circ T\circ W_1$ with coordinatewise $\tanh$; Φ is L_Φ-Lipschitz on $(-1,1)^d$ with orthonormal components; $p_\theta \le M$ uniformly.

Step 1: Lipschitzness in the parameters. Write h_l for the l-th layer activation ($h_0 = x$). Since $\tanh$ is coordinatewise 1-Lipschitz and $\|h_l\|_2 \le \sqrt{d}$,

$$\begin{aligned}\|h_l - h_l'\|_2 &= \|T(W_l h_{l-1}) - T(W_l' h_{l-1}')\|_2 \le \|W_l h_{l-1} - W_l' h_{l-1}'\|_2 \\ &\le \|W_l\|_{\mathrm{op}}\|h_{l-1}-h_{l-1}'\|_2 + \|W_l - W_l'\|_{\mathrm{op}}\|h_{l-1}'\|_2 \le \rho\,\|h_{l-1}-h_{l-1}'\|_2 + \sqrt{d}\,\|\theta-\theta'\|.\end{aligned}$$

Unrolling the recursion,

$$\|f_\theta(x) - f_{\theta'}(x)\|_2 \le \sqrt{d}\sum_{l=1}^L \rho^{L-l}\|W_l - W_l'\|_{\mathrm{op}} \le \sqrt{d}\,L\rho^{L-1}\|\theta-\theta'\| =: \ell_f\|\theta-\theta'\|,$$

uniformly in x. Consequently $\|\Phi(f_\theta(x)) - \Phi(f_{\theta'}(x))\|_2 \le \kappa\|\theta-\theta'\|$ with $\kappa := L_\Phi\ell_f$.

Step 2: subgaussian increments. Fix the sample and write $\Phi_i^\theta := \Phi(f_\theta(x_i))$. Since the readout is a Euclidean ball,

$$Z_\theta := \sup_{\|a\|_2 \le B} \frac{1}{n} \sum_i \epsilon_i \langle a, \Phi_i^\theta \rangle = \frac{B}{n} \Big\| \sum_i \epsilon_i \Phi_i^\theta \Big\|_2.$$

By the triangle inequality and Step 1,

$$Z_\theta - Z_{\theta'} \le \frac{B}{n} \Big\| \sum_i \epsilon_i \big(\Phi_i^\theta - \Phi_i^{\theta'} \big) \Big\|_2.$$

Flipping ϵ_i changes the right-hand side by at most $\frac{2B}{n} \|\Phi_i^\theta - \Phi_i^{\theta'}\|_2$, so by McDiarmid's inequality $Z_\theta - Z_{\theta'}$ is subgaussian (conditional on its mean) with variance proxy

$$\frac{B^2}{n^2} \cdot 4 \sum_i \|\Phi_i^\theta - \Phi_i^{\theta'}\|_2^2 \le \frac{4B^2\kappa^2}{n} \|\theta - \theta'\|^2.$$

The same bound holds for $Z_{\theta'} - Z_\theta$, and $\mathbb{E}_\epsilon[Z_\theta - Z_{\theta'}] = 0$ up to the centering absorbed in the variance proxy, so the process $\{Z_\theta\}$ has subgaussian increments in the metric $\rho_Z(\theta, \theta') = \frac{2B\kappa}{\sqrt{n}} \|\theta - \theta'\|$.

Step 3: Dudley plus reference point. For any reference point $\theta_0 \in \Theta$, Dudley's entropy integral gives

$$\mathbb{E}_\epsilon \sup_{\theta \in \Theta} Z_\theta \le \mathbb{E}_\epsilon Z_{\theta_0} + C \, \frac{2B\kappa}{\sqrt{n}} \int_0^{2R_\Theta} \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|, u)} \, du.$$

The P-dimensional ball satisfies $\mathcal{N}(\Theta, \|\cdot\|, u) \le (1 + 2R_\Theta/u)^P$, and substituting $t = u/(2R_\Theta)$,

$$\int_0^{2R_\Theta} \sqrt{P \log(1 + 2R_\Theta/u)} \, du = 2R_\Theta \sqrt{P} \int_0^1 \sqrt{\log(1 + 1/t)} \, dt =: c_1 R_\Theta \sqrt{P}$$

with $c_1 < \infty$ a universal constant. Taking expectation over the sample and applying Proposition 4.1(iii) at θ_0,

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_S \, \mathbb{E}_\epsilon \sup_\theta Z_\theta \le B \sqrt{\frac{Mr}{n}} + \frac{C \, B\kappa R_\Theta \sqrt{P}}{\sqrt{n}},$$

which is the claim once $\kappa = L_\Phi \sqrt{d} \, L \rho^{L-1}$ is inserted.

Remark I.1 (Layerwise covering refinement)**.** *Equipping Θ with the product of per-layer balls and covering each layer separately refines the parametrization term to $\frac{CBL_\Phi}{\sqrt{n}} \sum_{l=1}^L \rho^{L-l} \sqrt{d} \, R_l \sqrt{P_l}$, a genuinely layerwise sum in which deeper layers pay less. The proof is identical, with the metric ρ_Z decomposed layer by layer through Step 1.*

I.2 Proofs for Section 4.3 (layerwise structure)

Proof of Proposition 4.3. By definition of the pushforward, $\|\mathcal{K}_l v\|^2_{L^2(\mu_{l-1})} = \int |v(g_l(x))|^2 \, d\mu_{l-1}(x) = \int |v(y)|^2 \, d\mu_l(y) = \|v\|^2_{L^2(\mu_l)}$. Iterating the isometry gives $\|v \circ f_\theta\|_{L^2(\mu_0)} = \|v\|_{L^2(\mu_L)}$, and the final inequality is Lemma C.3 applied once at the interface between μ_L and λ. □

I.3 Proofs for Section G.3 (deep mixed chains)

Proof of the sandwich claim in Proposition 4.4. From $2|v| - 2\log 2 \le \log\cosh^2 v \le 2|v|$, summing over coordinates and maximizing over vertices,

$$2 \max_s \|Ws\|_1 - 2d \log 2 \le \Phi_{\mathrm{sat}}(W) \le 2 \max_s \|Ws\|_1,$$

and $\max_{s \in \{\pm 1\}^d} \|Ws\|_1 = \max_{s,t \in \{\pm 1\}^d} t^\top W s = \|W\|_{\infty \to 1}$. The cut norm is $\max_{S,T \subseteq [d]} \sum_{j \in S, k \in T} W_{jk} = \frac{1}{4} \|W\|_{\infty \to 1}$ up to the standard symmetrization. Grothendieck's theorem gives the SDP relaxation within $K_G < 1.783$. For the random-matrix scale, with i.i.d. $W_{jk} \sim \mathcal{N}(0, s_w^2)$, each $t^\top W s \sim \mathcal{N}(0, d^2 s_w^2)$, and a union bound over the 4^d pairs gives the upper bound $2\sqrt{\log 2} \, s_w d^{3/2}(1 + o(1))$; the matching lower bound is the standard cut-norm lower bound for random matrices. □

Proposition I.2 (Corner-path lower bound and the residual gap)**.** *Let $\Sigma_1 = W_1\Sigma_0 W_1^\top$ be the first-layer preactivation covariance and write $a(s) := s^\top \Sigma_1^{-1} s$ for $s \in \{\pm 1\}^d$. Then*

$$\log \|p_{Z_L}\|_\infty \;\geq\; \max_{s_1\in\{\pm1\}^d}\left[\frac{2d^2}{a(s_1)} + \sum_{l=2}^{L}\sum_{j=1}^{d} \log\cosh^2\big((W_l z_{l-1}(s_1))_j\big)\right] - \sum_{l=1}^{L} \log|\det W_l| - \tfrac{1}{2}\log\det\Sigma_0 - O(Ld).$$

where $z_1(s_1)$ is evaluated at the first-layer maximizer along the path $x = tW_1^{-1}s_1$ and $z_l = \tanh(W_l z_{l-1})$ for $l \geq 2$ is the deterministic orbit thereafter.

Proof sketch. Along the path $x = tW_1^{-1}s_1$, the first-layer objective is $-\frac{t^2 a(s_1)}{2} + d\log\cosh^2 t + O(d)$, maximized at $t^* = 2d/a(s_1)$ with value $2d^2/a(s_1) - 2d\log 2 + o(d)$; at this point $z_1 = \tanh(t^*)s_1$ converges to the vertex s_1. After the first layer the chain is deterministic, so the contribution of layers $l \geq 2$ is the saturation sum along the orbit. The lower bound evaluates the density at the output of this path. Comparing with Proposition 4.4, the residual gap is exactly the difference between maximizing each layer's potential independently (upper bound) and realizing the maxima along a single orbit (lower bound). □

I.4 The mollifier route: a structural bridge to the Koopman framework

The covering proof of Theorem 4.2 is the route we recommend; this subsection records an alternative whose structure parallels the Koopman-based analysis of Hashimoto et al. (2024; 2026) line by line, together with its costs. Take the Gaussian mollifier $p_{c,x}(x') = (c/\pi)^{d/2} e^{-c\|x'-x\|^2}$ on the input space and define the regularized model

$$G_c(\theta, v; x) := \big\langle v \circ f_\theta,\, p_{c,x}\big\rangle_{L^2(\mu)} = \int v(f_\theta(x'))\, p_{c,x}(x')\, p_X(x')\, dx',$$

which converges to the density-weighted predictor $p_X(x)\, v(f_\theta(x))$ as $c \to \infty$. For the class $\mathcal{G}_c = \{G_c(\theta, v; \cdot) : \theta \in \Theta, \|v\|_{L^2} \leq B\}$, Cauchy–Schwarz in $L^2(\mu)$ separates the supremum cleanly:

$$\widehat{\mathfrak{R}}_S(\mathcal{G}_c) \leq \frac{1}{n}\sup_{\theta,v}\|v\circ f_\theta\|_{L^2(\mu)} \cdot \mathbb{E}_\epsilon\Big\|\sum_i \epsilon_i p_{c,x_i}\Big\|_{L^2(\mu)}.$$

The first factor is $B\sqrt{M}$ by Lemma C.3; the second is computed from $\|p_{c,x}\|^2_{L^2(\lambda)} = (c/(2\pi))^{d/2}$ and $\|p_{c,x}\|^2_{L^2(\mu)} \leq P_X(c/(2\pi))^{d/2}$ with $P_X := \|p_X\|_\infty$, giving

$$\widehat{\mathfrak{R}}_S(\mathcal{G}_c) \leq \frac{B\sqrt{MP_X}\,(c/(2\pi))^{d/4}}{\sqrt{n}}.$$

The trade-offs mirror those of the Koopman route: c large brings G_c close to the true model but inflates the constant; and the $c \to \infty$ limit is the density-*weighted* predictor, not the predictor itself. Removing the weight with the de-weighted mollifier $q_{c,x} := p_{c,x}/p_X$ (requiring $p_X > 0$) restores the pointwise limit but the norm $\|q_{c,x}\|^2_{L^2(\mu)}$ grows like $\exp\big(\frac{c}{\sigma_X^2\lambda}\|x\|^2\big)$ for Gaussian inputs with $\lambda := 2c - \frac{1}{2\sigma_X^2} > 0$, so a finite bound requires restricting the *evaluation points* to $\|x\| \leq R$—a restriction on where the model is evaluated, strictly weaker than the bounded-input premise on the data distribution. Under the mollifier route the hypothesis class maps to a ball of radius $B\sqrt{M} = B\sqrt{M_0\prod_l \Lambda_l}$ in $L^2(\mu)$, against the worst-case radius $\prod_l \|\rho(g_l)\| \prod_l \|A_l\|\|v\|$ in the Koopman analysis: the structures align factor by factor, with the data measure replacing Lebesgue measure at exactly the interfaces where saturating activations make the latter infinite.

J Proofs for Dimensionality Reduction

J.1 Proof of Proposition 5.1 (coarea density)

For any test function $\phi \in C_c(\mathbb{R}^{d_1})$, apply the coarea formula with $u(x) = \phi(E(x))\, p_X(x)/J_E(x)$:

$$\begin{aligned}\mathbb{E}\,\phi(Z) &= \int_{\mathbb{R}^{d_0}} \phi(E(x))\, p_X(x)\, dx = \int_{\mathbb{R}^{d_0}} \frac{\phi(E(x))p_X(x)}{J_E(x)}\, J_E(x)\, dx \\ &= \int_{\mathbb{R}^{d_1}} \phi(z)\Big(\int_{E^{-1}(z)} \frac{p_X(x)}{J_E(x)}\, d\mathcal{H}^{d_0-d_1}(x)\Big) dz.\end{aligned}$$

Since ϕ is arbitrary, the inner integral is the density of Z with respect to Lebesgue measure on $\mathbb{R}^{d_1}$.

J.2 Proof of Theorem 5.2 (rectangular affine-tanh)

Let $U := WX \sim \mathcal{N}(0, \Sigma)$ with $\Sigma := WW^\top$, whose eigenvalues lie in $[\alpha^2, \beta^2]$. Since $Z = T(U)$ is a same-dimension diffeomorphism onto $(-1,1)^{d_1}$, coordinatewise change of variables gives

$$p_Z(z) = p_U(\operatorname{arctanh} z) \prod_{j=1}^{d_1} \frac{1}{1-z_j^2}, \qquad p_U(u) = (2\pi)^{-d_1/2} \det(\Sigma)^{-1/2} e^{-\frac{1}{2} u^\top \Sigma^{-1} u}.$$

Since $\lambda_{\max}(\Sigma) \le \beta^2$, we have $u^\top \Sigma^{-1} u \ge \|u\|^2/\beta^2$, and $\det(\Sigma)^{-1/2} \le \alpha^{-d_1}$, hence

$$p_Z(z) \le (2\pi)^{-d_1/2} \alpha^{-d_1} \prod_{j=1}^{d_1} \frac{\exp\big(-\operatorname{arctanh}(z_j)^2/(2\beta^2)\big)}{1-z_j^2} \le (2\pi)^{-d_1/2} \alpha^{-d_1} M_\beta^{d_1}.$$

Finiteness of M_β, with the explicit bound $M_\beta \le \max\{4/3, e^{2\beta^2}\}$, is Lemma C.4.

J.3 Reconstruction losses over bottleneck autoencoders

Take the truncated squared loss $\ell_C(x, \hat{x}) = \min(\|x - \hat{x}\|^2, C^2)$, with C the input diameter (e.g. $C = \sqrt{d_0}$ for inputs in $[0,1]^{d_0}$). The loss class $\mathcal{L} = \{x \mapsto \ell_C(x, D(E(x)))\}$ is $2C$-Lipschitz in the function $x \mapsto D(E(x))$, so vector contraction gives

$$\mathfrak{R}_n(\mathcal{L}) \le 2\sqrt{2}\, C \sum_{j=1}^{d_0} \mathfrak{R}_n(\mathcal{H}_j), \qquad \mathcal{H}_j = \{x \mapsto D_j(E(x))\}.$$

Each coordinate class $\mathcal{H}_j$ is controlled by Corollary 5.3 with the readout class given by the decoder's coordinate functions (for a linear decoder $D_j(z) = \langle a_j, z\rangle$ with $\|a_j\| \le B_D$, the features are the d_1 latent coordinates, bounded on the latent cube $(-1,1)^{d_1}$). Parameter variation over all layers is handled by the covering term of Theorem 4.2, whose proof uses only the 1-Lipschitzness and bounded range of $\tanh$, both preserved under dimension reduction. Assembling the pieces gives a generalization bound for the full bottleneck autoencoder whose only data-dependent quantity is M_E at the bottleneck.

K Experiment Setup

All experiments are conducted on MNIST (LeCun et al., 1998) using $\tanh$ networks. Unless otherwise specified, we train models with Adam (Kingma & Ba, 2015) (learning rate 10^{-3}, batch size 128) for 20 epochs on $S = 10{,}000$ training samples. Each configuration is evaluated over 3 independent random seeds. Since our reported statistic is the Spearman rank correlation between the bound and the empirical generalization gap *across configurations*, rather than a per-configuration statistic, we do not report per-configuration error bars.

Classification with label noise (Figure 2a, top). We use a three-layer classifier $\mathrm{Linear}(784 \to 256) \to \mathrm{Tanh} \to \mathrm{Linear}(256 \to 256) \to \mathrm{Tanh} \to \mathrm{Linear}(256 \to 10)$, trained under the global protocol described above. The label-noise level is varied by randomly relabeling a fraction of the training labels, with noise rates in $0, 0.1, 0.2, 0.4$. The generalization gap is defined as the difference between the training and test errors.

Width-variation autoencoders (Figure 2a, bottom). To vary model capacity without changing the dimensionality of individual layers, we use *square* autoencoders with three fixed layers: $\mathbb{R}^d \xrightarrow{\text{Linear}+b} \text{Tanh} \xrightarrow{\text{Linear}+b} \text{Tanh} \xrightarrow{\text{Linear}+b} \text{Tanh} \longrightarrow \mathbb{R}^d$, where all weight matrices lie in $\mathbb{R}^{d\times d}$ and the first weight matrix is initialized orthogonally. The square architecture is deliberate: some baseline bounds do not apply to dimension-reducing maps (in particular, the Koopman-based bound requires invertibility), so introducing a bottleneck would create an uncontrolled confound in the comparison. We instead vary the effective width by changing the input resolution. MNIST images are normalized to $[-1, 1]$, resized to $s \times s$ using area interpolation with $s = \text{round}(\sqrt{d})$, and flattened to $\mathbb{R}^d$, with $d \in 16, 64, 256, 784$ corresponding to $s \in 4, 8, 16, 28$; $d = 784$ retains the original resolution.

We compare our density bound against the Koopman–RKHS bound of Hashimoto et al. (2026), the Frobenius product-norm bound of Neyshabur et al. (2015), and the spectral-margin bound of Bartlett et al. (2017).

Bottleneck autoencoders and linear probes (Figure 1d and Figure 2b). We use a symmetric bottleneck autoencoder with encoder

$$\text{Linear}(784 \to 256) \to \text{Tanh} \to \text{Linear}(256 \to 64) \to \text{Tanh} \to \text{Linear}(64 \to d_1),$$

whose output is the bottleneck preactivation u, with latent representation $z = \tanh(u)$. The decoder mirrors the encoder:

$$\text{Linear}(d_1 \to 64) \to \text{Tanh} \to \text{Linear}(64 \to 256) \to \text{Tanh} \to \text{Linear}(256 \to 784),$$

with no output activation. We vary the bottleneck width over $d_1 \in 2, 4, 8, 16, 32, 64$. At the bottleneck, the density bound is estimated either using a single Gaussian fit or a BIC-selected Gaussian mixture, as described in Section 4.4.

For downstream evaluation, we freeze the encoder and train a linear probe $\text{Linear}(d_1 \to 10)$ with softmax cross-entropy on the latent representations. The probe is trained under the same global protocol, with downstream sample sizes $n \in 100, 500, 2000, 10000$.

Numerical verification (Figure 1a–c). For the depth-scaling experiments, we evaluate $f_K = \tanh^{\circ K}$ and $S^{\circ K}$ under $X \sim \mathcal{N}(0, 1)$ for depths $K \in [8, 256]$. The quantity $|p_K|_\infty$ is computed exactly from the closed-form densities in Theorems 3.6 and F.2. For the tail phase diagram, we evaluate the pushforward density of $\tanh(wX)$ near the boundary for Gaussian, Laplace, and Student-t inputs with $\nu = 3$, using $w \in 0.3, 0.5$. For the rectangular bound, we consider $E(x) = T(Wx)$ with singular values in $[\alpha, \beta]$, set $\beta = 1$, and compare the exact supremum with the bound in Theorem 5.2 over $\alpha \in 0.3, 0.4, 0.6, 0.8, 1.0$ and $d_1 \in 1, 2, 4, 8$. Finally, the density bounds for trained encoders (Figure 1d) are evaluated on the bottleneck autoencoders described in Appendix K.