DATA GEOMETRY AND THE METRIC ENTROPY OF INFINITE-WIDTH RELU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

The complexity of an infinite-width network is measured in $L^2(P_X)$, so it is a property of the data distribution and not of the architecture alone. For ReLU^k ridge features that dependence runs through a single quantity: the smoothness of the map from parameters to features, which is fixed by how much probability mass sits near moving activation boundaries. An anti-concentration, or margin, exponent for those boundaries therefore determines the metric entropy of the model class. We first show that this exponent can never exceed one, whatever the data, so a margin law imposed uniformly over the parameter space can only degrade the classical smoothness of ReLU^k features and never improve it. We then show that the degradation it predicts can be an artifact of that uniformity, and is so in the cases that motivate the question. Our main result is a *stratified* entropy law, in which the parameter smoothness is allowed to fall off on a set of positive codimension at a controlled rate; the resulting entropy exponent is governed by the better of two competing terms, and reduces to the smooth-dictionary bound of Siegel & Xu (2024) when the bad set is everything. The motivating case is data concentrated on a curved low-dimensional manifold, where the margin exponent does drop, but only at parameters whose activation boundary is tangent to the data — a set of codimension one. There the stratified law returns the undegraded classical exponent, reconciling the margin mechanism with intrinsic-dimension theory, and we delimit when a margin exponent below one arises at all. The predicted local exponents, and the rate at which the smoothness degrades near the bad set, are confirmed numerically to four digits.

1 INTRODUCTION

A two-layer network with an unbounded number of hidden units can be written as $f(x) = \int_\Theta \phi_\theta(x)\, d\mu(\theta)$ with $\|\mu\|_{\mathrm{TV}} \le R$, so that the model class is the radius-R variation ball

$$\mathcal{F}_R = R\,\overline{\mathrm{aco}}\{\phi_\theta : \theta \in \Theta\}.$$

This representation underlies Barron spaces, neural variation spaces, and several kernel and atomic models (Bach, 2017; E et al., 2022; Siegel & Xu, 2023; 2024). Its complexity is measured in $L^2(P_X)$, and this is the point of departure for the present paper: the ambient metric depends on the data distribution, so the same architecture need not have the same complexity on every dataset.

The mechanism by which P_X enters is geometric. For a ReLU^k ridge feature $\phi_\theta(x) = (\theta^\top \widetilde{x})_+^k$, a small change in θ moves an activation hyperplane, and the feature changes appreciably only on the thin slab swept out by that motion. Consequently the $L^2(P_X)$ modulus of the parameter map is governed by the *margin law*

$$P_X\big(|\theta^\top \widetilde{X}| \le t\big) \le C t^\alpha, \tag{1}$$

and a $(k+1)$st parameter difference of size $|h|$ has amplitude $O(|h|^k)$ on a tube of mass $O(|h|^\alpha)$, giving feature regularity $s = k + \alpha/2$. Feeding this into the smooth-dictionary entropy theory of Siegel & Xu (2024) yields $e_m(\mathcal{F}_R) \lesssim R\, m^{-1/2-(k+\alpha/2)/d}$.

This is where a naive reading goes wrong, in two ways.

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.

First: the uniform law can only degrade. We show (Proposition 7) that for *every* bounded input distribution, $\sup_\theta P_X(|\theta^\top \widetilde{X}| \le t) \gtrsim t$, simply because the bias is a free parameter. Hence $\alpha \le 1$ necessarily, the classical value $\alpha = 1$ is the best possible, and (1) can never buy a better rate. Any distributional statement built on the uniform margin law is a statement about *loss* of regularity.

Second: that loss can be spurious. The motivating case for $\alpha < 1$ is data concentrated on a curved low-dimensional manifold. There, (1) degrades to $\alpha = \frac{1}{2}$ — but only at *tangential* parameters, which form a set of codimension one in Θ. Charging that worst case across the whole parameter space predicts a slower entropy decay, in direct conflict with the intrinsic-dimension literature (Nakada & Imaizumi, 2020; Suzuki, 2019), which finds that manifold data makes estimation easier, not harder.

Our main contribution resolves this. We prove a stratified entropy law in which the parameter space carries two regularity exponents — a good one off a thin bad set B, and a worse one on it — together with a rate κ at which the good estimate blows up on approach to B. The resulting entropy exponent (Theorem 3) is

$$\tfrac{1}{2} + \gamma^\star, \qquad \gamma^\star = \min\Big\{ \tfrac{s_1}{p},\ \tfrac{s_0}{p-\nu} \Big\}, \qquad \text{valid when } \kappa < \tfrac{\nu s_1}{p},$$

where ν is the codimension of B. For curved data this returns $\gamma^\star = (k + \frac{1}{2})/d$: the undegraded classical exponent. The apparent conflict with intrinsic-dimension theory disappears.

Contributions.

1. $\alpha \le 1$ **is forced** (Proposition 7): a uniform margin law can only degrade ReLU^k feature regularity below $k + \frac{1}{2}$, which delimits what any uniform distributional theory can deliver.
2. **A stratified entropy law** (Theorem 3), localising the smoothness hypothesis over the parameter domain. It strictly refines the smooth-dictionary bound of Siegel & Xu (2024), recovered at $\nu = 0$.
3. $\mathbf{ReLU}^k$ **under a stratified margin law** (Theorem 10), which reads the exponents and the blow-up rate off the data geometry, with two examples in which the classical exponent survives: curve data (Example 11, verified numerically in Figure 1) and a density singularity on one hyperplane (Example 14). Proposition 12 delimits when a margin exponent below one arises at all.
4. **Learning rates and their sharpness** (Corollary 16, Theorem 17): squared risk $n^{-(p+2s)/(2p+2s)}$ for a net-based estimator, and a translation family showing the exponent minimax sharp as a function of (p, s) alone.

2 Setup, Background, and What Is New

Variation balls and entropy. Let $(\mathcal{X}, P_X)$ be the input space, $H = L^2(P_X)$, and $\Phi : \Theta \to H$, $\Phi(\theta) = \phi_\theta$, a feature map on a compact p-dimensional parameter domain. Write $B_1(A) = \overline{\mathrm{aco}}(A)$ and $\mathcal{F}_R = R\, B_1(\Phi(\Theta))$. For compact $A \subset H$ we use dyadic entropy numbers $e_m(A; H) = \inf\{\varepsilon > 0 : N(\varepsilon, A; H) \le 2^{m-1}\}$.

For θ ranging in a cube Q, an integer $M \ge 1$ and $t > 0$, the local M-th order modulus of smoothness is

$$\begin{aligned} \omega_M(\Phi, t; Q)_H &= \sup_{0<|h|\le t} \sup\big\{ \|\Delta_h^M \Phi(\theta)\|_H : [\theta, \theta + Mh] \subset Q \big\}, \\ \Delta_h^M \Phi(\theta) &= \sum_{i=0}^{M} (-1)^{M-i} \tbinom{M}{i} \Phi(\theta + ih). \end{aligned} \tag{2}$$

The unstratified theory, and its exact reach. The following is the state of the art and the baseline against which everything below should be read. Recall that $\mathrm{Lip}(s, L^\infty)$ denotes the classical Lipschitz space of order $s = r + \beta$ ($r \in \mathbb{Z}_{\ge 0}$, $0 < \beta \le 1$), extended to Banach-valued maps and to manifolds chartwise; Siegel & Xu (2024) write this class $\mathrm{Lip}^{\mathcal{M}}_\infty(s, X)$.

Theorem 1 (Siegel & Xu, 2024, Thm. 4)**.** *Let $s > 0$, let X be a type-2 Banach space, let $\mathcal{M}$ be a compact d-dimensional smooth manifold, let $P \in \mathrm{Lip}^{\mathcal{M}}_\infty(s, X)$, and let $\mathbb{D} \subset P(\mathcal{M})$. Then $e_m(B_1(\mathbb{D}))_X \lesssim m^{-1/2-s/d}$.*

	Siegel & Xu (2024, Thm. 4)	Theorem 3 (this paper)
parameter domain	compact d-dim. smooth manifold	$\mathbb{T}^p$; manifolds via App. B
target space	any type-2 Banach X	Hilbert $H = L^2(P_X)$
smoothness	$\mathrm{Lip}_\infty^{\mathcal{M}}(s, X)$, *uniform* on $\mathcal{M}$	s_1 off B, s_0 on B, blow-up rate κ
side condition	—	$\kappa < \nu s_1/p$ (†)
conclusion	$e_m \lesssim m^{-1/2-s/d}$	$e_m \lesssim m^{-1/2-\gamma^\star}$, $\gamma^\star = \min\{\frac{s_1}{p}, \frac{s_0}{p-\nu}\}$

Table 1: The delta. We are less general in the target space and recover their result at $\nu = 0$; the gain is that the smoothness hypothesis is no longer uniform.

Two facts about Theorem 1 matter for positioning. First, for non-integer s the class $\mathrm{Lip}(s, L^\infty)$ coincides with the Besov–Zygmund space $B^s_{\infty,\infty}$; only at integer s is it strictly smaller, a point Siegel & Xu (2024) note themselves. A restatement of Theorem 1 in Besov form is therefore not a substantive improvement, and we do not claim one. Second, Theorem 1 already covers compact manifolds and general type-2 targets, both more general than we need. Applied with $\mathrm{Lip}_\infty^{\mathcal{M}}(k + \frac{1}{p}, L^p)$ regularity of the ReLU^k dictionary (Siegel & Xu, 2024, Lemma 1), it gives the sharp $e_m(B_1(\mathbb{P}_k^d))_{L^2} \asymp m^{-1/2-(2k+1)/(2d)}$. Table 1 sets out exactly where our hypotheses differ from theirs.

What Theorem 1 does *not* do is localise: the hypothesis $P \in \mathrm{Lip}_\infty^{\mathcal{M}}(s, X)$ is uniform over $\mathcal{M}$, so a single rough sliver of parameter space fixes s everywhere. That is precisely the situation created by curved data, and Section 3 removes it. The Lebesgue-measure hypothesis in their Lemma 1 is likewise what Section 4 replaces by a margin law.

2.1 Related work

Variation and Barron spaces. The infinite-width model $\mathcal{F}_R$ goes under several names. Bach (2017) develops the $\mathcal{F}_1$ variation-norm viewpoint directly in $L^2(d\rho)$, and is the closest precedent for the present paper in one specific respect: the data measure is part of the geometry rather than a fixed reference. His analysis assumes ρ comparable to uniform measure on the sphere, the case $\alpha = 1$ here. E et al. (2022) construct the Barron space, Klusowski & Barron (2018) give ℓ^1-controlled rates for ReLU and ReLU^2 ridge combinations, and Siegel & Xu (2023) characterise the variation spaces of common shallow architectures. None of these lets the smoothness of the parameterisation vary over the parameter domain.

Convex-hull entropy. The half-order gain from absolute convexification is classical (Carl, 1997; Carl et al., 1999; 2014): a dictionary with covering radius $O(m^{-\sigma})$ has hull with $e_m \lesssim m^{-1/2-\sigma}$. That route alone cannot produce Theorem 1, because covering the image of a nonconstant C^1 map on a p-dimensional domain sees only first-order displacement and returns $m^{-1/2-1/p}$ however smooth the map. The extra s/p comes from multiscale cancellation, which is what Proposition 4 exploits and what survives being localised in θ.

Statistical rates for shallow networks. Parhi & Nowak (2023) prove near-minimax optimal rates $n^{-(d+3)/(2d+3)}$ for shallow ReLU estimation over second-order bounded-variation classes, which is exactly what Corollary 16 returns at $\alpha = 1$; Yang & Zhou (2025) and Yang & Zhou (2024) obtain related optimal rates for ReLU^k variation spaces and for over-parameterised shallow networks. We use the entropy-to-least-squares route in the standard form (Han & Wellner, 2019; Wainwright, 2019) and claim nothing new there: the statistical content of this paper is entirely in which entropy exponent is supplied to it.

Low-dimensional data structure. That data concentrated near a low-dimensional set makes estimation easier is established in several forms — Nakada & Imaizumi (2020) obtain rates governed by the Minkowski dimension of the support, and Suzuki (2019) shows deep ReLU networks adapt to Besov smoothness with the corresponding intrinsic exponents. This is the line our mechanism appears to contradict, since a uniform margin law degrades on precisely such data, and Section 4 is where the contradiction is resolved: the degradation is confined to a thin set of parameters (Example 11), and for curved hypersurface data in three or more dimensions it does not occur at all (Proposition 12).

Location-dependent smoothness, and weighted variation. Letting smoothness vary with distance from a set is the defining feature of the *2-microlocal* spaces $B^{s,s'}_{p,q}(\mathbb{R}^n; U)$, whose embedding entropy is computed by Leopold & Skrzypczak (2011); Definition 2 is a coarse analogue, with κ for the second index and B for the singular set U. The settings differ in a way that forces a separate argument: 2-microlocal results concern embeddings between function spaces on a domain, whereas here Φ is Hilbert-valued on the *parameter* domain and the set covered is the convex hull of its image, so the half-order convexification gain, absent from the embedding problem, is what the budget must be built around. A genuine 2-microlocal version of Theorem 3 would be a natural sharpening; we do not attempt it. A different localisation is that of DeVore et al. (2025), whose *weighted* variation spaces reweight the ℓ_1 mass assigned to atoms and so change the class, whereas we leave the class unchanged and localise the parameterisation's smoothness.

Anti-concentration. Conditions of the form $P_X(|\theta^\top \widetilde{X}| \le t) \lesssim t^\alpha$ are small-ball or anti-concentration hypotheses, and the same flavour appears in the small-ball method of Mendelson (2015) for lower-isometry bounds and in the margin condition of Tsybakov (2004) in classification. Both of those control a regression function or an empirical process; here the condition acts instead on the *dictionary*, fixing the smoothness of $\theta \mapsto \phi_\theta$ as a map into $L^2(P_X)$, and Proposition 7 shows the achievable range of α is one-sided.

3 A STRATIFIED ENTROPY LAW

We now allow the regularity of Φ to vary over Θ. Throughout this section $\Theta = \mathbb{T}^p$. Corollary 22 in Appendix B transfers everything to a compact parameter manifold, under hypotheses imposed chartwise; Remark 23 explains why that is the right form in which to assume them.

Definition 2 (Stratified smoothness)**.** Let $B \subseteq \Theta$ be closed, $\nu \in [0, p]$, $s_1 \ge s_0 > 0$, $\kappa \ge 0$, let $M > s_1$ be an integer, and let $c_0 \ge 1$ be a separation constant. We write $\Phi \in \mathcal{S}^{s_1,s_0}_{\kappa,\nu}(\Theta, B; H)$ with constants (A, C_B, c_0) if, for all $0 < t \le 1$,

(S0) $\sup_{\theta\in\Theta} \|\Phi(\theta)\|_H \le A$;

(S1) $\big|\{\theta \in \Theta : \operatorname{dist}(\theta, B) \le \delta\}\big| \le C_B\, \delta^\nu$ for all $0 < \delta \le 1$;

(S2) $\omega_M(\Phi, t; Q)_H \le A\, t^{s_0}$ for every cube $Q \subset \Theta$ of side $3Mt$;

(S3) $\omega_M(\Phi, t; Q)_H \le A\, t^{s_1} \operatorname{dist}(Q, B)^{-\kappa}$ for every such cube with $\operatorname{dist}(Q, B) \ge c_0 t$.

Condition (S0) cannot be dropped: a constant map $\Phi \equiv Lv$ has vanishing moduli of every order, so (S2)–(S3) hold for every $A > 0$, yet $\overline{\operatorname{aco}}\,\Phi(\Theta) = [-Lv, Lv]$ has entropy numbers proportional to $L\|v\|$. The three constants play separate roles: A bounds the size of Φ and its moduli, C_B is purely geometric, and c_0 lets an application verify (S3) on cubes separated from B by a fixed multiple of t, which is what Section 4 produces. It costs nothing: in the proof the dyadic shells run over $i < j - L_0$ with $L_0 = \lceil \log_2(2c_0) \rceil$ rather than over $i < j$, the L_0 displaced shells being absorbed into an enlarged core whose cardinality grows by the constant factor $(4c_0)^\nu$. Constants change; no exponent does.

Condition (S2) is the uniform hypothesis of Theorem 1 at the pessimistic exponent s_0; (S3) says that away from B the map is in fact s_1-regular, with a constant that may blow up at rate κ on approach. The two are consistent at $\operatorname{dist}(Q, B) \asymp t$ precisely when $s_1 - \kappa \ge s_0$, which holds in all our examples.

Theorem 3 (Stratified entropy)**.** *Let $\Phi \in \mathcal{S}^{s_1,s_0}_{\kappa,\nu}(\Theta, B; H)$ with constants (A, C_B, c_0) and suppose the* stratification condition

$$\kappa < \frac{\nu s_1}{p} \tag{†}$$

holds. Then for every $R > 0$,

$$e_m(\mathcal{F}_R; H) \;\le\; C\, R\, A\, m^{-\frac{1}{2}-\gamma^\star}, \qquad \gamma^\star = \min\Big\{\frac{s_1}{p},\; \frac{s_0}{p-\nu}\Big\},$$

with the convention $s_0/(p-\nu) = +\infty$ when $\nu = p$. The bound is linear in the size constant A; the constant C depends only on $p, \nu, s_0, s_1, \kappa, M, c_0$ and on the geometric constant C_B, on which it

depends polynomially — the shell cardinalities carry a factor C_B, so Proposition 4 contributes $C_B^{s_1/p}$ and $C_B^{s_0/(p-\nu)}$ rather than a linear factor. If $B = \Theta$ then (S3) is vacuous, every dyadic shell in the proof is empty, and the conclusion $\gamma^\star = s_0/p$ holds without (†). Consequently $\log N(\varepsilon, \mathcal{F}_R; H) \lesssim (RA/\varepsilon)^q$ *with* $q = (\frac{1}{2} + \gamma^\star)^{-1} < 2$.

The proof is in Appendix A. It rests on a single reusable estimate, which we state here because both halves of $\gamma^\star$ are instances of it.

Proposition 4 (Multiscale dictionaries)**.** *Let $\mu, \sigma > 0$, $N \geq 1$, $\rho > 0$, and for each $j \geq 0$ let $\mathbb{D}_j \subset H$ be finite with $\#\mathbb{D}_j \leq N 2^{j\mu}$ and $\max_{v \in \mathbb{D}_j} \|v\|_H \leq \rho\, 2^{-j\sigma}$. Then*

$$e_m\Big(\sum_{j \geq 0} B_1(\mathbb{D}_j); H\Big) \;\leq\; C(\mu, \sigma)\, \rho\, N^{\sigma/\mu}\, m^{-1/2-\sigma/\mu}.$$

Theorem 3 follows by splitting the scale-j wavelet indices of Φ into dyadic shells around B and applying Proposition 4 once per shell. Writing $\Lambda_{j,i}$ for the scale-j wavelets at distance $\asymp 2^{-i}$ from B:

$$\begin{aligned} &\text{shells } i < j - L_0: && \#\Lambda_{j,i} \lesssim C_B 2^{jp} 2^{-i\nu}, && \text{radius} \lesssim 2^{-js_1} 2^{i\kappa}, \\ &\text{core}: && \#\Lambda_{j,*} \lesssim C_B (4c_0)^\nu 2^{j(p-\nu)}, && \text{radius} \lesssim 2^{-js_0}. \end{aligned}$$

The core is a multiscale family with $(\mu, \sigma) = (p - \nu, s_0)$ and yields the second term of $\gamma^\star$. Reindexing shell i by $j = i + l$ turns it into a family with $(\mu, \sigma) = (p, s_1)$ of base size $2^{i(p-\nu)}$ and base radius $2^{-i(s_1-\kappa)}$, whose contribution carries the factor $2^{-i\tau}$ with

$$\tau = (s_1 - \kappa) - \tfrac{(p-\nu)s_1}{p} = \tfrac{\nu s_1}{p} - \kappa.$$

Condition (†) is exactly $\tau > 0$, i.e. convergence of the sum over shells.

Remark 5 (Relation to Theorem 1). Take $B = \{\theta_0\}$ a single point, so $\nu = p$, together with $s_0 = s_1 = s$ and $\kappa = 0$; then (†) reads $0 < s$ and $\gamma^\star = s/p$, giving $e_m \lesssim m^{-1/2-s/p}$: Theorem 1 specialised to $X = H$ Hilbert. Take instead $B = \Theta$, so $\nu = 0$ and (S3) is vacuous; then $\gamma^\star = s_0/p$, the uniform law at the pessimistic exponent. The content of Theorem 3 lies strictly between these two degenerate choices: $\gamma^\star > s_0/p$ whenever $\nu > 0$ and $s_1 > s_0$.

Remark 6 (When (†) fails). If $\kappa \geq \nu s_1/p$ the shell sum no longer converges and the argument returns only an intermediate exponent obtained by truncating the shell range. We do not pursue the optimal exponent in that regime; see Problem 19.

4 ReLU features: reading the exponents off the data

4.1 The uniform margin law can only degrade

Proposition 7 ($\alpha \leq 1$ is forced)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be bounded with $\operatorname{diam}(\mathcal{X}) \leq L$, let P_X be any Borel probability measure on $\mathcal{X}$, and let $\Theta \supseteq \{(w, b) : \|w\| = 1,\ b \in I\}$ where I contains the range of $-w^\top x$. Then for every $0 < t \leq L$,*

$$\sup_{\theta \in \Theta} P_X\big(|\theta^\top \widetilde{X}| \leq t\big) \;\geq\; \frac{2t}{3L}.$$

Consequently, if $\sup_\theta P_X(|\theta^\top \widetilde{X}| \leq t) \leq C t^\alpha$ for all $0 < t \leq t_0$, then $\alpha \leq 1$.

Proof. Fix any unit w. The pushforward of P_X under $x \mapsto w^\top x$ is a probability measure on an interval J of length at most L. Partition J into $N = \lceil L/(2t) \rceil$ intervals of length $2t$; some interval $J_\star$ has mass at least $1/N \geq 2t/(L + 2t) \geq 2t/(3L)$ for $t \leq L$. Taking $b = -\operatorname{mid}(J_\star)$ gives $P_X(|w^\top X + b| \leq t) \geq 2t/(3L)$. The final claim follows by letting $t \downarrow 0$ in $Ct^\alpha \geq 2t/(3L)$. □

So $\alpha = 1$ — which holds whenever the one-dimensional projections of P_X have uniformly bounded densities — is the *best* case, and $s = k + \frac{1}{2}$ is the largest regularity the mechanism can produce. This is a structural limit, not an artifact of the proof, though it does require a free bias; see Remark 8.

Remark 8 (Scope of Proposition 7). The pigeonhole uses the bias: it selects b so that the slab $\{|w^\top x + b| \le t\}$ sits where the projected mass is. The conclusion therefore applies to parameterisations with a free bias ranging over the projection range, and *not* to bias-free (positively homogeneous) features, nor to models whose bias is confined to a subinterval that the argument cannot exploit. For such restricted parameterisations $\alpha > 1$ is not excluded by this argument, and the stratified machinery of Section 3 would then be improving rather than repairing a bound. We do not pursue that case.

4.2 Stratified margin laws

Definition 9 (Stratified margin law)**.** Let $\Theta \subset \mathbb{R}^{d+1}$ be a compact d-dimensional C^{k+1} parameter manifold with a finite atlas of uniformly bounded charts, let $\widetilde{x} = (x, 1)$, and let $B \subseteq \Theta$ be closed with $|\{\mathrm{dist}(\cdot, B) \le \delta\}| \le C_B \delta^\nu$. We say P_X satisfies a stratified margin law with data $(B, \nu, \alpha_1, \alpha_0, \kappa')$ if, for $0 < t \le t_0$,

$$\text{(M0)} \quad \sup_{\theta \in \Theta} P_X\big(|\theta^\top \widetilde{X}| \le t\big) \;\le\; C_0\, t^{\alpha_0},$$

$$\text{(M1)} \quad P_X\big(|\theta^\top \widetilde{X}| \le t\big) \;\le\; C_1\, t^{\alpha_1}\, \mathrm{dist}(\theta, B)^{-\kappa'} \quad \text{whenever } \mathrm{dist}(\theta, B) \ge t,$$

with $0 < \alpha_0 \le \alpha_1 \le 1$ and $\kappa' \ge 0$.

Theorem 10 (ReLU^k under a stratified margin law)**.** *Let $\mathcal{X}$ be bounded, let $\Phi_k(\theta)(x) = (\theta^\top \widetilde{x})_+^k$ with $k \ge 1$, and let P_X satisfy a stratified margin law as in Definition 9. Then*

$$\Phi_k \in \mathcal{S}^{s_1, s_0}_{\kappa, \nu}(\Theta, B; L^2(P_X)), \qquad s_i = k + \frac{\alpha_i}{2}, \qquad \kappa = \frac{\kappa'}{2},$$

with $M = k + 1$. If in addition $\kappa' < \nu(2k + \alpha_1)/d$, then

$$e_m\big(\mathcal{F}_R; L^2(P_X)\big) \;\lesssim\; R\, m^{-\frac{1}{2} - \gamma^\star}, \qquad \gamma^\star = \min\Big\{\frac{k + \alpha_1/2}{d},\; \frac{k + \alpha_0/2}{d - \nu}\Big\}.$$

The proof (Appendix C) is a chartwise boundary-crossing estimate, and it is worth seeing where α enters. For a chart increment h the stencil $\theta(u + ih)$, $i \le k + 1$, either stays on one side of the activation boundary, where the composed feature is C^{k+1} in u with derivatives bounded by $\|\theta\|_{C^{k+1}}$ alone and the difference is $O(|h|^{k+1})$; or it crosses, forcing $|\theta(u)^\top \widetilde{x}| \le C|h|$, where homogeneity of $t \mapsto t_+^k$ gives only $O(|h|^k)$. Splitting the $L^2(P_X)$ integral over the crossing tube $A_h = \{x : |\theta(u)^\top \widetilde{x}| \le C|h|\}$ and its complement gives

$$\big\|\Delta_h^{k+1} \Phi_k(\theta(u))\big\|^2_{L^2(P_X)} \;\le\; C|h|^{2k}\, P_X(A_h) \;+\; C|h|^{2k+2}. \tag{3}$$

The distributional content of the paper sits entirely in the factor $P_X(A_h)$: inserting (M0) gives $|h|^{2k+\alpha_0}$ and inserting (M1) gives $|h|^{2k+\alpha_1}\, \mathrm{dist}(\theta, B)^{-\kappa'}$; since $\alpha_i \le 1 < 2$ the crossing term dominates the smooth term in both cases, and taking square roots yields (S2) and (S3) with $s_i = k + \alpha_i/2$ and $\kappa = \kappa'/2$. Taking $B = \Theta$ (so $\nu = 0$ and (S3) is vacuous) recovers the uniform statement $\Phi_k \in B^{k+\alpha_0/2}_{\infty,\infty}$ and the uniform entropy bound $m^{-1/2-(k+\alpha_0/2)/d}$; taking $\alpha_0 = 1$ recovers the classical $s = k + \frac{1}{2}$ of Siegel & Xu (2024, Lemma 1).

4.3 The motivating example: data on a curve

Example 11 (Curved data in the plane)**.** Let $d = 2$, let $\Gamma \subset \mathbb{R}^2$ be a compact C^2 curve with nonvanishing curvature, and let P_X be normalised arclength on Γ. Parameterise $\Theta = \{(w, b) : \|w\| = 1\} \cong \mathbb{T} \times I$, so $p = d = 2$, and let $B = \{\theta : \{x : \theta^\top \widetilde{x} = 0\}$ is tangent to $\Gamma\}$, a curve in Θ, so $\nu = 1$. Then P_X satisfies a stratified margin law with

$$\alpha_1 = 1, \qquad \kappa' = \tfrac{1}{2}, \qquad \alpha_0 = \tfrac{1}{2}.$$

Hence $s_1 = k + \frac{1}{2}$, $s_0 = k + \frac{1}{4}$, $\kappa = \frac{1}{4}$, and (†) reads $\frac{1}{4} < \frac{k+1/2}{2}$, i.e. $k \ge 1$. Since $\min\{(k + \frac{1}{2})/2,\; k + \frac{1}{4}\} = (k + \frac{1}{2})/2$, Theorem 10 gives

$$e_m\big(\mathcal{F}_R; L^2(P_X)\big) \;\lesssim\; R\, m^{-\frac{1}{2} - \frac{k+1/2}{2}},$$

the *undegraded* classical exponent, whereas the uniform margin law with $\alpha_0 = \frac{1}{2}$ would give only $m^{-1/2-(k+1/4)/2}$.

The mechanism is visible on the unit circle: with $\theta^\top \widetilde{x} = \cos(\tau - \varphi) - c$ and $\delta := 1 - |c| \asymp \operatorname{dist}(\theta, B)$, one has $P_X(|\theta^\top \widetilde{X}| \le t) \asymp \min\{t\delta^{-1/2}, \sqrt{t}\}$ (Appendix C) — a transversal crossing of width $t/\sqrt{\delta}$ away from tangency, and a $\sqrt{t}$ cap at it.

It is worth delimiting how far this goes, because the answer is not "manifold data": what governs α_0 is the maximal order of contact between the data manifold and a hyperplane.

Proposition 12 (Where the margin degrades)**.** *Let* $\Theta = \{(w, b) : \|w\| = 1\}$.

(i) Let $\Gamma \subset \mathbb{R}^d$ *be a compact* C^d *curve with nondegenerate Frenet frame, i.e.* $\gamma', \dots, \gamma^{(d)}$ *linearly independent at every point, and let* P_X *be normalised arclength. Then* $\sup_\theta P_X(|\theta^\top \widetilde{X}| \le t) \asymp t^{1/d}$*, so* $\alpha_0 = 1/d$*.*

(ii) Let $S \subset \mathbb{R}^d$ *with* $d \ge 3$ *be a compact, connected, embedded* C^2 *hypersurface with nowhere-vanishing Gaussian curvature, and let* P_X *be normalised surface measure. Then* $\sup_\theta P_X(|\theta^\top \widetilde{X}| \le t) \asymp t$*, so* $\alpha_0 = 1$ *and no degradation occurs.*

(iii) Let $M \subset \mathbb{R}^d$ *be a compact* C^3 *submanifold of dimension* $r \ge 3$ *such that at every* $q \in M$ *the second fundamental form* H_n *is* nondegenerate *for every unit normal* n*, and let* P_X *be normalised* r*-dimensional Hausdorff measure. Then* $\alpha_0 = 1$*.*

Remark 13 (Why (iii) does not assume definiteness). Assuming H_n *definite* for every unit normal would give the tangential sublevel set measure $t^{r/2}$ at once, but that hypothesis is vacuous beyond codimension one. If the normal space has dimension at least two its unit sphere is connected, and $H_{-n} = -H_n$, so a path from n to $-n$ joins a positive definite form to a negative definite one and must pass through a singular one. Nondegeneracy survives such a path and is the right hypothesis in higher codimension; for $r \ge 3$ it suffices, since an indefinite nondegenerate form has sublevel sets of measure $\asymp t$ rather than $t^{r/2}$. At $r = 2$ an indefinite form gives $t \log(1/t)$ and the conclusion fails by a logarithm.

The same connectedness argument constrains (iii) itself. In codimension at least two the signature of H_n is constant along a path from n to $-n$, because nondegeneracy forbids an eigenvalue from crossing zero; but $H_{-n} = -H_n$ has the reversed signature, so if H_n has signature (a, b) then $a = b$ and $r = 2a$ is *even*. Hence in codimension at least two the hypothesis of (iii) is vacuous for odd r — in particular for $r = 3$ — and its content there begins at $r = 4$. In codimension one the normal sphere is the disconnected set $\{\pm n\}$, the argument does not apply, and (iii) reduces to (ii).

The proof is in Appendix C. Parts (ii) and (iii) say that for curved data of intrinsic dimension at least two — and, in codimension greater than one, at least four — there is nothing to repair, and the classical exponent already applies. Part (i) says the degradation is a phenomenon of *curve* data, and that it worsens with the ambient dimension, since a nondegenerate curve in $\mathbb{R}^d$ osculates a hyperplane to order d; the hypothesis of (iii) is exactly what such curves fail, their second fundamental form being degenerate in the binormal directions. Remark 13 explains why (iii) assumes nondegeneracy rather than definiteness. Example 11 is the case $d = 2$, where the two coincide and where we verify everything numerically; the stratified treatment of the higher-contact cases $d \ge 3$ is left open (Problem 19).

Geometry is not the only source of a small margin exponent, and the non-geometric source turns out to be the most favourable case of all.

Example 14 (A density singularity on one hyperplane)**.** Let $\mathcal{X} = [-1, 1]^d$ and let P_X have the integrable density $\rho(x) = c_\alpha |x_1|^{\alpha - 1}$, $0 < \alpha < 1$. Let $B = \{(\pm e_1, 0)\}$, two points of Θ, so that $\nu = p = d$. Then P_X satisfies a stratified margin law with

$$\alpha_1 = 1, \qquad \kappa' = 1 - \alpha, \qquad \alpha_0 = \alpha,$$

because a slab at parameter distance δ from B carries mass $\asymp t\,\delta^{\alpha - 1}$ when $t \le \delta$ and $\asymp t^\alpha$ when $t \ge \delta$ (Appendix C). Condition (†) reads $\frac{1-\alpha}{2} < k + \frac{1}{2}$ and so holds for every $k \ge 1$ and every $\alpha \in (0, 1)$; and since $\nu = p$ the second term of $\gamma^\star$ is absent, leaving

$$e_m\big(\mathcal{F}_R; L^2(P_X)\big) \lesssim R\, m^{-\frac{1}{2} - \frac{k + 1/2}{d}},$$

the classical exponent, for *every* $\alpha \in (0, 1)$, where the uniform margin law would give only $m^{-1/2 - (k + \alpha/2)/d}$.

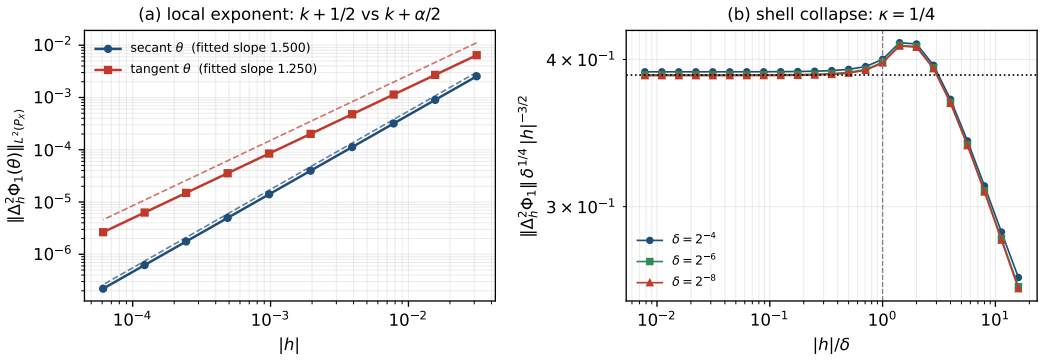


Figure 1: Numerical verification of Example 11 for $k = 1$ ($\Phi_1(\theta)(x) = (\theta^\top \widetilde{x})_+$, P_X uniform on the unit circle). **(a)** The second parameter difference obeys $\|\Delta_h^2 \Phi_1\|_{L^2(P_X)} \asymp |h|^{k+1/2}$ at a secant parameter and $\asymp |h|^{k+\alpha_0/2} = |h|^{5/4}$ at a tangent parameter; least-squares slopes over $|h| \in [2^{-14}, 2^{-5}]$ are 1.5000 and 1.2500. **(b)** At a near-tangent parameter with $\mathrm{dist}(\theta, B) = \delta$, the rescaled modulus $\|\Delta_h^2 \Phi_1\| \, \delta^{1/4} |h|^{-3/2}$ collapses onto a single curve for $\delta = 2^{-4}, 2^{-6}, 2^{-8}$ and is constant (≈ 0.388) throughout $|h| \ll \delta$, confirming both $s_1 = k + \frac{1}{2}$ and the blow-up rate $\kappa = \frac{1}{4}$ of (S3). Reproduction code is in Appendix F.

The contrast with the unstratified theory is instructive. Here a margin law of exponent $\alpha < 1$ holds only at a single reference hyperplane and not uniformly on an open family of shifted hyperplanes, which under a uniform hypothesis is exactly what makes such a density unusable as an example. Stratification says that a bad set this small should cost nothing — and at $\nu = p$ it costs nothing at all.

Remark 15 (Reconciliation with intrinsic-dimension theory). Example 11 is exactly the configuration in which the uniform margin law predicts slower entropy decay for manifold data, conflicting with Nakada & Imaizumi (2020); Suzuki (2019). The stratified law removes the conflict: the rough parameters are tangential, form a codimension-one set, and are dominated. The uniform degradation is an artifact of the supremum over θ.

5 Learning rates and sharpness

Corollary 16 (Entropy to squared risk)**.** *Assume $Y_i = f^\star(X_i) + \xi_i$ with $X_i \sim P_X$ i.i.d., ξ_i centred sub-Gaussian and independent of the covariates, and $f^\star \in \mathcal{F}_R$. Assume the features are uniformly bounded, $\sup_\theta \|\phi_\theta\|_\infty \le b$, so that $\|f\|_\infty \le Rb$ on $\mathcal{F}_R$. If $\log N(\varepsilon, \mathcal{F}_R; L^2(P_X)) \lesssim (R/\varepsilon)^q$ with $q < 2$, then there is an estimator $\widehat{f}$ with values in $\mathcal{F}_R$ — a discrete minimum-contrast estimator over an ε_n-net, $\varepsilon_n \asymp n^{-1/(2+q)}$ — such that*

$$\|\widehat{f} - f^\star\|_{L^2(P_X)}^2 = O_{\mathbb{P}}\big(n^{-2/(2+q)}\big).$$

(A matching minimax lower bound requires a two-sided entropy estimate, which Theorem 3 does not provide in general; it is available for the translation family of Theorem 17.) In particular, under Theorem 3 with $s_0 = s_1 = s$ and $\kappa = 0$, $q = 2p/(p + 2s)$ and the squared risk is $O_{\mathbb{P}}\big(n^{-(p+2s)/(2p+2s)}\big)$.

The discretisation is deliberate. The hypothesis is an entropy bound in the *population* metric $L^2(P_X)$, which is what Theorem 3 supplies, whereas the localised analysis of the exact least-squares estimator is driven by empirical or bracketing entropy, and the former does not control the latter in general; the population-entropy formulation is due to Yang & Barron (1999). See Appendix E, where the theorem invoked is stated in full. The entropy-to-risk route itself is standard (Wainwright, 2019, Thm. 13.5); see also Han & Wellner (2019) and Appendix E. For ReLU ($k = 1$) with $\alpha = 1$ it returns $n^{-(d+3)/(2d+3)}$, matching the near-minimax optimal rate of Parhi & Nowak (2023) and the variation-space rates of Yang & Zhou (2025; 2024).

A sharp model: periodic translation features. Let P_X be uniform on $\mathbb{T}^d$ and $\phi_z(x) = \kappa(x - z)$, $z \in \mathbb{T}^d$, with $c(1 + |k|)^{-\beta} \le |\widehat{\kappa}_k| \le C(1 + |k|)^{-\beta}$ and $\beta > d/2$. The parameter map has exact regularity $s = \beta - d/2$.

Feature family	param. dim.	regularity	entropy exponent $\frac{1}{2}+\gamma^\star$
ReLU^k, bounded projected density	$p=d$	$s=k+\frac{1}{2}$	$\frac{1}{2}+\frac{k+1/2}{d}$
ReLU^k, uniform margin α	$p=d$	$s=k+\frac{\alpha}{2}$	$\frac{1}{2}+\frac{k+\alpha/2}{d}$
ReLU^k, curve data (Ex. 11)	$p=d=2$	$s_1=k+\frac{1}{2},\ s_0=k+\frac{1}{4}$	$\frac{1}{2}+\frac{k+1/2}{2}$
Periodic translations	$p=d$	$\beta-\frac{d}{2}$	$\frac{\beta}{d}$ (sharp)
Smooth compact features	p	all finite s	super-polynomial

Table 2: Feature regularity and data geometry determine variation-space entropy. Row 3 is the point: row 2 read uniformly at $\alpha=\frac{1}{2}$ predicts $\frac{1}{2}+\frac{k+1/4}{2}$, but stratification returns the row-1 exponent.

Theorem 17 (Sharp entropy and minimax rate)**.** *For the translation family above, $e_m(\mathcal{F}_R;L^2)\asymp R\,m^{-\beta/d}$. If in addition $\beta>d$, so that $\sum_k|\widehat{\kappa}_k|<\infty$ and the features are uniformly bounded as Corollary 16 requires, then in Gaussian random-design regression with X_i uniform on $\mathbb{T}^d$,*

$$\inf_{\widehat{f}}\sup_{f^\star\in\mathcal{F}_1}\mathbb{E}\|\widehat{f}-f^\star\|_{L^2}^2\asymp n^{-2\beta/(2\beta+d)}=n^{-\frac{p+2s}{2p+2s}},\qquad p=d,\ s=\beta-\tfrac{d}{2}.$$

The upper bound follows from Theorem 3 (with $\kappa=0$); the packing lower bound puts signed coefficients on a q^d translation grid and converts Hamming separation into L^2 separation by the normalised DFT identity; the minimax statement is Fano applied to that packing. Proofs are in Appendix D.

The hypothesis $\beta>d$ in the second half of Theorem 17 is needed and not cosmetic: $\beta>d/2$ places κ in L^2 but not in L^∞, and the uniform feature bound $\sup_z\|\phi_z\|_\infty=\|\kappa\|_\infty$ of Corollary 16 then fails. The entropy statement needs only $\beta>d/2$. The class $\{\kappa*\mu:\|\mu\|_{\mathrm{TV}}\le R\}$ is *not* contained in a Sobolev-β ball — taking $\mu=\delta_z$ gives $f=\kappa(\cdot-z)$ with $\|f\|_{H^\beta}^2\asymp\sum_k(1+|k|)^{2\beta}|\widehat{\kappa}_k|^2\asymp\sum_k 1=\infty$ — but it has the same entropy and minimax exponents as one, so Theorem 17 recovers a classical rate; its role is only to certify that the exponent in Corollary 16 cannot be improved *as a function of* (p,s) *alone*. It says nothing about sharpness for ReLU features — see Problem 18.

6 Discussion, Limitations, and Open Problems

What the paper claims, and what it does not. The entropy machinery for uniformly smooth dictionaries is due to Siegel & Xu (2024) and we do not improve on it in that setting. We claim only that ReLU^k feature regularity in $L^2(P_X)$ is governed by a margin law with exponent necessarily at most one, that the resulting uniform degradation can be spurious, and that stratification repairs it when it is; row 3 of Table 2 is the payoff. All entropy statements are upper bounds. Theorem 3 requires (†), which holds for the planar curve data of Example 11 for every $k\ge1$ but not for $d\ge3$ curves, whose contact order degrades with the ambient dimension by Proposition 12(i); and B is assumed to have a clean Minkowski dimension, so genuinely fractal bad sets are not covered.

Open problem 18 (Matching lower bounds). Local roughness of a feature manifold does not imply a matching entropy lower bound for its absolute convex hull. Under a two-sided stratified margin law with a transverse nondegeneracy condition and quantitative overlap control among parameter cells, is $e_m(\mathcal{F}_1;L^2(P_X))\asymp m^{-1/2-\gamma^\star}$? Note that Example 11 shows the naive conjecture at the uniform exponent $m^{-1/2-(k+\alpha_0/2)/d}$ is *false as an asymptotic*: the upper bound there is strictly better.

Open problem 19 (Beyond the stratification condition). When $\tau=\nu s_1/p-\kappa\le0$ the shell sum diverges, and optimising the budget against the truncation tail returns only $\gamma_A=\gamma_1 p(s_1-\kappa)/(s_1(p-\nu))$ with $\gamma_1=\frac{1}{2}+s_1/p$. This agrees with γ_1 at $\tau=0$ but degrades quickly, and can fall *below* the bound $\frac{1}{2}+s_0/p$ that (S2) gives for free — at $p=10$, $\nu=1$, $s_1=\frac{3}{2}$, $s_0=\frac{5}{4}$, $\kappa=\frac{1}{4}$ one gets $\gamma_A\approx0.602<0.625$. An argument that does not charge each shell separately is needed. What is the correct exponent?

Two further directions: anisotropic stratification, and whether training adapts to stratified regularity without knowing (B,ν,s_0,s_1).

Reproducibility Statement

This is a theoretical submission and all assumptions appear in the theorem statements; complete proofs are in Appendices A–E, and the unproved claims are isolated as Problems 18 and 19. Appendix F contains the complete code for Figure 1, which runs in under a minute on a laptop CPU and needs only NumPy and Matplotlib.

AI Use Statement

Generative AI tools assisted with literature discovery, proof development, numerical verification, and editing. All statements and citations were reviewed and verified by the authors, who take full responsibility for the content.

References

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. URL http://jmlr.org/papers/v18/14-546.html.

Bernd Carl. Metric entropy of convex hulls in Hilbert spaces. *Bulletin of the London Mathematical Society*, 29(4):452–458, 1997. doi: 10.1112/S0024609397003044.

Bernd Carl, Ioanna Kyrezi, and Alain Pajor. Metric entropy of convex hulls in Banach spaces. *Journal of the London Mathematical Society*, 60(3):871–896, 1999. doi: 10.1112/S0024610799008005.

Bernd Carl, Aicke Hinrichs, and Philipp Rudolph. Entropy numbers of convex hulls in Banach spaces and applications. *Journal of Complexity*, 30(5):555–587, 2014. doi: 10.1016/j.jco.2014.03.005.

Albert Cohen. *Numerical Analysis of Wavelet Methods*, volume 32 of *Studies in Mathematics and Its Applications*. Elsevier, Amsterdam, 2003. ISBN 9780444511249.

Ronald DeVore, Robert D. Nowak, Rahul Parhi, and Jonathan W. Siegel. Weighted variation spaces and approximation by shallow ReLU networks. *Applied and Computational Harmonic Analysis*, 74:101713, 2025. doi: 10.1016/j.acha.2024.101713.

Ronald A. DeVore and George G. Lorentz. *Constructive Approximation*, volume 303 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1993.

Manfredo P. do Carmo. *Differential Geometry of Curves and Surfaces*. Dover Publications, revised and updated second edition edition, 2016.

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. doi: 10.1007/s00365-021-09549-y.

Qiyang Han and Jon A. Wellner. Convergence rates of least squares regression estimators with heavy-tailed errors. *The Annals of Statistics*, 47(4):2286–2319, 2019. doi: 10.1214/18-AOS1748.

Tuomas Hytönen, Jan van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach Spaces, Volume I: Martingales and Littlewood–Paley Theory*. Springer, 2016. doi: 10.1007/978-3-319-48520-1.

Jason M. Klusowski and Andrew R. Barron. Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656, 2018.

Hans-Gerd Leopold and Leszek Skrzypczak. Entropy numbers of embeddings of some 2-microlocal Besov spaces. *Journal of Approximation Theory*, 163(4):505–523, 2011. doi: 10.1016/j.jat.2010.12.002.

Shahar Mendelson. Learning without concentration. *Journal of the ACM*, 62(3):21:1–21:25, 2015. doi: 10.1145/2699439.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(1):7018–7055, 2020. URL http://jmlr.org/papers/v21/20-002.html.

Rahul Parhi and Robert D. Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Transactions on Information Theory*, 69(2):1125–1140, 2023. doi: 10.1109/TIT.2022.3208653.

Jonathan W. Siegel and Jinchao Xu. Characterization of the variation spaces corresponding to shallow neural networks. *Constructive Approximation*, 57:1109–1132, 2023. doi: 10.1007/s00365-023-09626-4.

Jonathan W. Siegel and Jinchao Xu. Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *Foundations of Computational Mathematics*, 24(2):481–537, 2024. doi: 10.1007/s10208-022-09595-3.

Elias M. Stein. *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*, volume 43 of *Princeton Mathematical Series*. Princeton University Press, 1993.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In *International Conference on Learning Representations (ICLR)*, 2019.

Hans Triebel. *Theory of Function Spaces*, volume 78 of *Monographs in Mathematics*. Birkhäuser, Basel, 1983. doi: 10.1007/978-3-0346-0416-1.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. doi: 10.1214/aos/1079120131.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999. doi: 10.1214/aos/1017939142.

Yunfei Yang and Ding-Xuan Zhou. Nonparametric regression using over-parameterized shallow ReLU neural networks. *Journal of Machine Learning Research*, 25(165):1–35, 2024.

Yunfei Yang and Ding-Xuan Zhou. Optimal rates of approximation by shallow ReLU^k neural networks and applications to nonparametric regression. *Constructive Approximation*, 62:329–360, 2025. doi: 10.1007/s00365-024-09679-z.

A Proof of Theorem 3

A.1 Two standard entropy facts

Lemma 20 (Entropy of a Minkowski sum)**.** *Let $A_j \subset H$ be bounded sets containing the origin. For positive integers $m_1, \dots, m_J$,*

$$e_{m_1+\cdots+m_J-J+1}(A_1 + \cdots + A_J; H) \le \sum_{j=1}^{J} e_{m_j}(A_j; H).$$

The same holds for a norm-convergent infinite Minkowski sum.

Proof. Choose 2^{m_j-1}-point covers of A_j with radii arbitrarily close to $e_{m_j}(A_j)$. Sums of the cover centres give at most $2^{m_1+\cdots+m_J-J}$ centres, and the triangle inequality gives the sum of the radii. The infinite case follows by truncation. □

Lemma 21 (Finite-dictionary convex-hull entropy; Carl, 1997, Prop. 1). *Let $\mathbb{D} = \{v_1, \dots, v_N\} \subset H$ with $\max_\nu \|v_\nu\|_H \le \rho$. Then, for absolute constants $C, c > 0$,*

$$e_m(B_1(\mathbb{D}); H) \le C\rho \cdot \begin{cases} 1, & m = 1, \\ m^{-1/2}\sqrt{1 + \log(N/m)}, & 2 \le m \le N, \\ N^{-1/2} 2^{-cm/N}, & m > N. \end{cases}$$

A.2 The multiscale dictionary estimate

We first prove Proposition 4, restated here for convenience: if $\#\mathbb{D}_j \le N 2^{j\mu}$ and $\max_{v \in \mathbb{D}_j} \|v\|_H \le \rho\, 2^{-j\sigma}$ for all $j \ge 0$, then $e_m(\sum_{j \ge 0} B_1(\mathbb{D}_j); H) \le C(\mu, \sigma)\rho N^{\sigma/\mu} m^{-1/2-\sigma/\mu}$.

Proof of Proposition 4. Reduction to $N = 1$, $\rho = 1$. Write $j_0 = \lceil (\log_2 N)/\mu \rceil$, an integer. Then $\#\mathbb{D}_j \le 2^{(j+j_0)\mu}$ and $\|v\| \le \rho\, 2^{-j\sigma} = \rho\, 2^{j_0\sigma} 2^{-(j+j_0)\sigma}$, so after reindexing by $l = j + j_0$ (and discarding the levels $l < j_0$, which only decreases the set) the family satisfies the hypotheses with $N = 1$ and ρ replaced by $\rho\, 2^{j_0\sigma} \asymp \rho N^{\sigma/\mu}$. Homogeneity in ρ finishes the reduction.

The case $N = \rho = 1$. By monotonicity it suffices to treat $m \asymp 2^{r\mu}$, $r \in \mathbb{Z}_{\ge 0}$. Put $a = \sigma + \mu/2$ and fix $0 < \delta < 2\sigma$. For $j < r$ set $m_j = \lceil A_1 (r - j + 1) 2^{j\mu} \rceil$ with A_1 large; for $r \le j < J_r := \lceil (1 + \frac{\mu}{2\sigma}) r \rceil$ set $m_j = \lceil A_2 2^{r\mu} 2^{-\delta(j-r)} \rceil$; cover the remaining tail by one ball. Then $\sum_j (m_j - 1) \lesssim 2^{r\mu} \asymp m$.

For $j < r$ we have $m_j \gg \#\mathbb{D}_j$, so the third branch of Lemma 21 applies and, after increasing A_1, gives $e_{m_j}(B_1(\mathbb{D}_j)) \lesssim 2^{-ar} 2^{-\eta(r-j)}$ for some $\eta > 0$. For $r \le j < J_r$ the middle branch gives

$$e_{m_j}(B_1(\mathbb{D}_j)) \lesssim 2^{-ar} 2^{-(\sigma - \delta/2)(j-r)} \sqrt{1 + (\mu + \delta)(j - r)},$$

which is summable in j because $\delta < 2\sigma$. Finally $\sum_{j \ge J_r} 2^{-j\sigma} \lesssim 2^{-\sigma J_r} \lesssim 2^{-(\sigma+\mu/2)r} = 2^{-ar}$. Lemma 20 now yields $e_{Cm} \lesssim 2^{-ar} \asymp m^{-1/2-\sigma/\mu}$, and monotonicity absorbs the constant in the entropy index. □

A.3 Wavelet decomposition adapted to B

Fix a compactly supported periodic wavelet system on $\mathbb{T}^p$ of regularity larger than s_1 and with more than M vanishing moments, with finitely many wavelet types and $O(2^{jp})$ translates at scale j; write $\psi^\ell_{j,\nu}$ for the L^2-normalised wavelets and Λ_j for the index set at scale j. Set $c^\ell_{j,\nu} = \langle \Phi, \psi^\ell_{j,\nu} \rangle \in H$, the integral taken in the Bochner sense; vector-valued wavelet theory for Hilbert-space-valued maps is standard (Hytönen et al., 2016; Cohen, 2003; Triebel, 1983).

Since $\psi^\ell_{j,\nu}$ has more than M vanishing moments, for any polynomial P of degree $< M$ we have $c^\ell_{j,\nu} = \langle \Phi - P, \psi^\ell_{j,\nu} \rangle$, and Whitney's inequality (DeVore & Lorentz, 1993, Ch. 6) bounds the best local polynomial approximation by the local modulus (2). Hence, with $Q_{j,\nu} \supset \operatorname{supp} \psi^\ell_{j,\nu}$ a cube of side $\asymp 2^{-j}$,

$$\|c^\ell_{j,\nu}\|_H \;\lesssim\; 2^{-jp/2}\, \omega_M\big(\Phi, 2^{-j}; Q_{j,\nu}\big)_H. \tag{4}$$

Partition Λ_j into dyadic shells around B. Because (S3) is available only on cubes separated from B by $c_0 t$, and here $t = 2^{-j}$, the shells that may use it are those at distance at least $c_0 2^{-j}$; we therefore set

$$L_0 := \lceil \log_2(2c_0) \rceil,$$

and for $0 \le i < j - L_0$ put

$$\Lambda_{j,i} = \big\{(\ell, \nu) \in \Lambda_j : 2^{-i-1} < \operatorname{dist}(Q_{j,\nu}, B) \le 2^{-i}\big\},\ \Lambda_{j,*} = \big\{(\ell, \nu) : \operatorname{dist}(Q_{j,\nu}, B) \le 2^{-(j-L_0)}\big\},$$

which is a partition of Λ_j: the shells exhaust $\{\operatorname{dist} > 2^{-(j-L_0)}\}$ and the enlarged core $\Lambda_{j,*}$ takes the rest, absorbing the L_0 innermost shells. For $i < j - L_0$ we then have $\operatorname{dist}(Q_{j,\nu}, B) > 2^{-i-1} \ge 2^{-(j-L_0)} = 2^{L_0} 2^{-j} \ge 2c_0 2^{-j}$, so (S3) applies literally. By (S1), $\#\Lambda_{j,i} \lesssim C_B 2^{jp} 2^{-i\nu}$ for $i < j - L_0$, while $\#\Lambda_{j,*} \lesssim C_B 2^{L_0\nu} 2^{j(p-\nu)}$; the factor $2^{L_0\nu} \le (4c_0)^\nu$ is a constant, so the core count

has the same order in j as before and only the constant changes. Applying (4) with (S3) on $\Lambda_{j,i}$ ($i < j - L_0$, where $\operatorname{dist} > 2^{-i-1} \geq 2c_0 2^{-j}$ as arranged above) and with (S2) on $\Lambda_{j,*}$:

$$\|c^\ell_{j,\nu}\|_H \lesssim A\, 2^{-jp/2} 2^{-js_1} 2^{i\kappa} \quad \text{on } \Lambda_{j,i}, \qquad \|c^\ell_{j,\nu}\|_H \lesssim A\, 2^{-jp/2} 2^{-js_0} \quad \text{on } \Lambda_{j,*}.$$

Set $v^\ell_{j,\nu} = 2^{jp/2} c^\ell_{j,\nu}$ and $a^\ell_{j,\nu}(\theta) = 2^{-jp/2} \psi^\ell_{j,\nu}(\theta)$, so that $\Phi(\theta) = \Phi_{\mathrm{c}}(\theta) + \sum_j \sum_{(\ell,\nu)\in\Lambda_j} a^\ell_{j,\nu}(\theta)\, v^\ell_{j,\nu}$. Bounded wavelet overlap gives $\sup_\theta \sum_{(\ell,\nu)\in\Lambda_j} |a^\ell_{j,\nu}(\theta)| \leq C$, and this bound is inherited by every sub-collection $\Lambda_{j,i} \subset \Lambda_j$. The coarse part Φ_{c} takes values in a fixed finite-dimensional space and is bounded there by $\|\Phi\|_{L^\infty(H)} \leq A$ by (S0) — this is the only place (S0) is used, and the only place a bound on the size of Φ, as opposed to its increments, is available. Absorbing it into the first scale, we obtain with $\mathbb{D}_{j,i} = \{v^\ell_{j,\nu} : (\ell,\nu) \in \Lambda_{j,i}\}$:

$$B_1\big(\Phi(\mathbb{T}^p)\big) \;\subset\; C \sum_{i\geq 0} \sum_{j>i} B_1(\mathbb{D}_{j,i}) \;+\; C \sum_{j\geq 0} B_1(\mathbb{D}_{j,*}). \tag{5}$$

A.4 Allocation across shells

The core. The family $\{\mathbb{D}_{j,*}\}_{j\geq 0}$ has $\#\mathbb{D}_{j,*} \lesssim 2^{j(p-\nu)}$ and radius $\lesssim A 2^{-js_0}$, so Proposition 4 with $(\mu,\sigma) = (p-\nu, s_0)$ gives $e_m \lesssim A\, m^{-1/2-s_0/(p-\nu)}$. (If $\nu = p$ then $\mu = p - \nu = 0$ and Proposition 4 does not apply. In that case each scale carries $O(1)$ atoms of radius $\asymp 2^{-js_0}$; splitting a budget m over the first J scales and covering the rest by one ball gives $2^{-cm/J} + 2^{-Js_0}$, and optimising in J yields $e_m \lesssim \exp(-c'\sqrt{m})$. This is root-exponential rather than exponential, but it still decays faster than every polynomial, which is all the convention $s_0/(p-\nu) = \infty$ is used for.)

Shell i. Reindex $j = i + l$; the constraint $i < j - L_0$ becomes $l > L_0$. Then $\#\mathbb{D}_{i+l,i} \lesssim C_B 2^{i(p-\nu)} 2^{lp}$ and the radius is $\lesssim A 2^{-i(s_1-\kappa)} 2^{-ls_1}$. Padding the family with empty dictionaries at the levels $l \leq L_0$, so that it is indexed by all $l \geq 0$ as Proposition 4 requires, that proposition applies with $\mu = p$, $\sigma = s_1$, $N_i = C_B 2^{i(p-\nu)}$ and $\rho_i = A 2^{-i(s_1-\kappa)}$, giving

$$e_{m_i}\Big(\sum_{j>i} B_1(\mathbb{D}_{j,i})\Big) \;\lesssim\; A\, 2^{-i(s_1-\kappa)} 2^{i(p-\nu)s_1/p}\, m_i^{-1/2-s_1/p} \;=\; A\, 2^{-i\tau}\, m_i^{-1/2-s_1/p},$$

where

$$\tau := (s_1 - \kappa) - \frac{(p-\nu)s_1}{p} = \frac{\nu s_1}{p} - \kappa \;>\; 0$$

precisely by the stratification condition (†).

Truncation and budget. Since $\kappa < \nu s_1/p \leq s_1$ we have $s_1 - \kappa > 0$, so $e_1\big(\sum_{j>i} B_1(\mathbb{D}_{j,i})\big) \leq \sum_{l\geq 1} A 2^{-i(s_1-\kappa)} 2^{-ls_1} \lesssim A 2^{-i(s_1-\kappa)}$, and therefore the shells with $i > I$ can be covered by a single ball of radius $\lesssim A 2^{-I(s_1-\kappa)}$. Choose

$$I = \Big\lceil \frac{(1/2 + s_1/p)}{s_1 - \kappa} \log_2 m \Big\rceil, \qquad \text{so that} \qquad A 2^{-I(s_1-\kappa)} \leq A\, m^{-1/2-s_1/p}.$$

For $i \leq I$ set $m_i = \lceil Z^{-1} m 2^{-\epsilon i} \rceil$ with $0 < \epsilon < \tau/(1/2 + s_1/p)$ and $Z = 2\sum_{i\geq 0} 2^{-\epsilon i}$; since $I = O(\log m)$ we get $\sum_{i\leq I} m_i \leq m/2 + I + 1 \lesssim m$, leaving budget for the core. Then

$$\sum_{i\leq I} A\, 2^{-i\tau} m_i^{-1/2-s_1/p} \;\lesssim\; A\, m^{-1/2-s_1/p} \sum_{i\geq 0} 2^{-i\tau} 2^{\epsilon i(1/2+s_1/p)} \;\lesssim\; A\, m^{-1/2-s_1/p},$$

the geometric series converging by the choice of ϵ. Combining the core bound, the shell bound, the truncated tail and (5) via Lemma 20, and scaling by R, gives $e_m(\mathcal{F}_R) \lesssim RA\, m^{-1/2-\gamma^\star}$ with $\gamma^\star = \min\{s_1/p,\ s_0/(p-\nu)\}$. Inverting the dyadic entropy-number definition gives the covering-number form. □

B Reduction from a manifold to the torus

Let Θ be a compact p-dimensional C^M manifold with a finite atlas $T_1, \ldots, T_J : [-1,1]^p \to \Theta$ as in Siegel & Xu (2024, Lemma 2), each a C^M diffeomorphism onto its image with

$\|DT_r\|_\infty, \|DT_r^{-1}\|_\infty \le L$, and let $\chi_1, \dots, \chi_J$ be a subordinate partition of unity with $\chi_r \in C_c^\infty((-1,1)^p)$ and $\|\chi_r\|_{C^M} \le \Lambda$.

We state the reduction with *chartwise* hypotheses. This is not a weakening in practice: the ReLU estimate of Theorem 10 takes its finite differences in the chart variable, so it verifies exactly the hypothesis below. It does avoid a genuine technical difficulty, discussed in Remark 23.

Corollary 22 (Compact parameter manifolds)**.** *Suppose each $\Phi \circ T_r$ satisfies (S0), (S2) and (S3) on $[-1,1]^p$ with constants (A, c_0) and bad set $T_r^{-1}(B)$, and that B satisfies (S1) on Θ with constant C_B. If (†) holds, then $e_n(B_1(\Phi(\Theta)); H) \lesssim n^{-1/2-\gamma^*}$, with the implied constant depending in addition on J, L and Λ.*

Proof. Set $\Psi_r := \chi_r \cdot (\Phi \circ T_r)$, extended by zero to $\mathbb{T}^p$; the extension is legitimate because χ_r is compactly supported in the open cube, so no discontinuity is created. We check that Ψ_r lies in the stratified class on $\mathbb{T}^p$ with bad set $B_r := T_r^{-1}(B)$. It is important not to intersect with $\operatorname{supp}\chi_r$ here: a cube Q can be far from $T_r^{-1}(B) \cap \operatorname{supp}\chi_r$ while lying close to a point of $T_r^{-1}(B)$ just outside the support, and (S3) would then be applied where it is not available.

Geometry. T_r is bi-Lipschitz with constant L, so it distorts distances by a factor in $[L^{-1}, L]$ and Lebesgue measure by a factor in $[L^{-p}, L^p]$. Hence (S1) transfers from B on Θ to B_r on $\mathbb{T}^p$ with C_B replaced by $L^{p+\nu} C_B$. Nothing else is transported: the moduli conditions (S0), (S2), (S3) are hypotheses on $g := \Phi \circ T_r$ directly, so no composition estimate is required.

Cutoff. By (S0), $\|g\|_{L^\infty(H)} \le A$. Both (S2) and (S3) have the form

$$\omega_M(g,t;Q)_H \;\le\; A_Q\, t^\sigma, \qquad (\sigma, A_Q) = \big(s_0,\, A\big) \;\text{ or }\; \big(s_1,\, A\,\mathrm{dist}(Q, B_r)^{-\kappa}\big),$$

the distance factor sitting in the constant and not in the exponent. The Leibniz rule for finite differences gives

$$\Delta_h^M(\chi_r g)(\theta) = \sum_{i=0}^{M} \binom{M}{i} \Delta_h^i \chi_r(\theta)\, \Delta_h^{M-i} g(\theta + ih).$$

The term $i = 0$ is bounded by $\Lambda\, \omega_M(g,t;Q) \le \Lambda A_Q t^\sigma$. For $1 \le i \le M$ we have $\|\Delta_h^i \chi_r\|_\infty \le \Lambda |h|^i \le \Lambda t^i$, and Marchaud's inequality (DeVore & Lorentz, 1993, Ch. 2, Thm. 8.1) converts $\omega_M(g,t;Q) \le A_Q t^\sigma$ together with $\|g\|_\infty \le A$ into

$$\omega_{M-i}(g,t;Q) \;\lesssim\; (A_Q + A)\, t^{\min\{M-i,\,\sigma\}}, \qquad 1 \le i \le M,$$

with no logarithmic loss because $\sigma \notin \{1, \dots, M-1\}$ in our applications, where $\sigma \in (k, k+\frac{1}{2}]$ and $M = k+1$. Hence the i-th term is $\lesssim \Lambda(A_Q + A)\, t^{\,i+\min\{M-i,\sigma\}}$, and $i + \min\{M-i, \sigma\} \ge \sigma$ in both cases: if $M - i \ge \sigma$ the exponent is at least $i + \sigma$, and otherwise it equals $M > \sigma$. Summing, $\omega_M(\chi_r g, t; Q) \lesssim \Lambda(A_Q + A)\, t^\sigma$, so (S2) and (S3) are preserved with A replaced by $C(M)\Lambda A$ and the distance-dependent factor in A_Q carried through unchanged; (S0) is preserved with A replaced by ΛA.

Assembly. Since $\sum_r \chi_r \equiv 1$ on Θ, every $\theta \in \Theta$ has $\Phi(\theta) = \sum_r \chi_r(\theta)\Phi(\theta) = \sum_r \Psi_r(T_r^{-1}\theta)$, the r-th term vanishing when θ lies outside the r-th chart. Hence $\Phi(\Theta) \subset \sum_{r \le J} \Psi_r(\mathbb{T}^p)$ and therefore $B_1(\Phi(\Theta)) \subset \sum_{r \le J} B_1(\Psi_r(\mathbb{T}^p))$. Applying Theorem 3 on the torus to each summand and combining by Lemma 20 with $m_1 = \cdots = m_J = \lceil n/J \rceil$ gives the claim. □

Remark 23 (Why the hypothesis is chartwise). An intrinsic formulation — assume (S2), (S3) for Φ on Θ and deduce them for $\Phi \circ T$ — would require a composition estimate for moduli of smoothness under a nonlinear chart, and that is not elementary: equally spaced stencil points do not remain equally spaced under T. Nor does the obvious repair work. Writing $T = A_0 + \varrho$ with A_0 the linearisation at the centre of a cube of side t and $\|\varrho\|_\infty \lesssim t^2$, and letting P be a best approximating polynomial of degree $< M$, Markov's inequality gives $\|P \circ T - P \circ A_0\|_\infty \lesssim t^2 \cdot t^{-1} \|P\|_\infty \lesssim t\, \|\Phi\|_\infty$, which is not $O(t^\sigma)$ for $\sigma > 1$. The correct statement is the invariance of $B^\sigma_{\infty,\infty}$ under C^M diffeomorphisms (Triebel, 1983, §4.3), a genuine theorem proved by a multiscale argument. We do not need it: the estimate we actually prove for ReLU^k features is already chartwise, so nothing is lost by assuming the hypothesis in that form.

C ReLU features: proofs

C.1 Proof of Theorem 10

By Appendix B we work in a single chart $u \mapsto \theta(u)$ with derivatives through order $k+1$ uniformly bounded, and take $M = k+1 > s_1$ (valid since $\alpha_1 \le 1$ gives $s_1 = k + \alpha_1/2 \le k + \frac{1}{2} < k+1$).

Fix $x \in \mathcal{X}$ and a chart increment h with $|h| \le t$. Consider the stencil values $g_i := \theta(u+ih)^\top \widetilde{x}$, $i = 0, \dots, k+1$. Since θ is C^{k+1} with bounded derivatives and $\mathcal{X}$ is bounded,

$$|g_i - g_0| \le C_\Theta \, i \, |h|, \qquad i = 0, \dots, k+1. \tag{6}$$

Off the crossing tube. Suppose $|g_0| > C_\Theta(k+1)|h|$. By (6) all g_i have the same sign as g_0, so on the whole stencil the map $u \mapsto (\theta(u)^\top \widetilde{x})_+^k$ agrees either with 0 or with the C^{k+1} function $u \mapsto (\theta(u)^\top \widetilde{x})^k$, whose derivatives through order $k+1$ are bounded uniformly in terms of $\|\theta\|_{C^{k+1}}$ and $\operatorname{diam} \mathcal{X}$ only — there is no blow-up near the boundary, because the stencil never reaches it. Hence

$$\left|\Delta_h^{k+1} \Phi_k(\theta(u))(x)\right| \le C|h|^{k+1}.$$

On the crossing tube. Otherwise $|g_0| \le C_\Theta(k+1)|h|$, and then $|g_i| \le C|h|$ for all i by (6), so by homogeneity of $t \mapsto t_+^k$ each term of the difference is $O(|h|^k)$ and

$$\left|\Delta_h^{k+1} \Phi_k(\theta(u))(x)\right| \le C|h|^k.$$

Writing $A_h = \{x : |\theta(u)^\top \widetilde{x}| \le C|h|\}$ and splitting the $L^2(P_X)$ integral,

$$\left\|\Delta_h^{k+1} \Phi_k(\theta(u))\right\|_{L^2(P_X)}^2 \;\le\; C|h|^{2k} \, P_X(A_h) \;+\; C|h|^{2k+2}. \tag{7}$$

The tube radius is $C_\Theta(k+1)|h|$, so (M1) applies at that radius exactly when $\operatorname{dist}(\theta(u), B) \ge C_\Theta(k+1)t$. This is why Definition 2 carries the separation constant: we obtain (S3) with $c_0 = 2C_\Theta(k+1)$, the factor 2 absorbing the passage from $\operatorname{dist}(\theta(u), B)$ to $\operatorname{dist}(Q, B)$ over a cube of side $3Mt$. No rescaling of t is involved. Inserting (M0), $P_X(A_h) \le C_0(C|h|)^{\alpha_0}$, and using $\alpha_0 \le 1 \le 2$ so that the first term dominates, gives $\|\Delta_h^{k+1}\Phi_k\| \lesssim |h|^{k+\alpha_0/2}$ uniformly, which is (S2) with $s_0 = k + \alpha_0/2$. If instead $\operatorname{dist}(\theta(u), B) \ge t \ge |h|$, inserting (M1) gives $P_X(A_h) \le C_1(C|h|)^{\alpha_1} \operatorname{dist}(\theta(u), B)^{-\kappa'}$ and hence $\|\Delta_h^{k+1}\Phi_k\| \lesssim |h|^{k+\alpha_1/2} \operatorname{dist}(\theta(u), B)^{-\kappa'/2}$, which is (S3) with $s_1 = k + \alpha_1/2$ and $\kappa = \kappa'/2$. Taking suprema over the cube Q and over $|h| \le t$, and using $\operatorname{dist}(Q, B) \asymp \operatorname{dist}(\theta, B)$ for $\theta \in Q$ when $\operatorname{dist}(Q, B) \ge t$ and Q has side $3Mt$, gives Definition 2. The entropy bound is then Theorem 3, whose condition (†) reads $\kappa'/2 < \nu(k + \alpha_1/2)/d$, i.e. $\kappa' < \nu(2k + \alpha_1)/d$. □

C.2 Proof of the margin law in Example 11

Take Γ the unit circle and P_X normalised arclength; the general C^2 curve with nonvanishing curvature is identical after a bounded reparameterisation. Writing $x = (\cos\tau, \sin\tau)$ and $\theta = (n(\varphi), -c)$, we have $\theta^\top \widetilde{x} = \cos(\tau - \varphi) - c$, and $B = \{|c| = 1\}$, a one-dimensional subset of the two-dimensional Θ, so (S1) holds with some finite C_B and with $\nu = 1$. Put $\delta = 1 - |c| \asymp \operatorname{dist}(\theta, B)$ and $\upsilon = \tau - \varphi$.

If $t \le \delta$: the equation $\cos\upsilon = c$ has two solutions $\pm \arccos c$, at which $|\frac{d}{d\upsilon}\cos\upsilon| = \sqrt{1-c^2} \asymp \sqrt{\delta}$, so $\{|\cos\upsilon - c| \le t\}$ is contained in two intervals of length $\asymp t/\sqrt{\delta}$ and $P_X(|\theta^\top \widetilde{X}| \le t) \asymp t\,\delta^{-1/2}$. This is (M1) with $\alpha_1 = 1$, $\kappa' = \frac{1}{2}$.

If $t \ge \delta$: near the near-tangential point $\cos\upsilon - c \approx \delta - \upsilon^2/2$, so $\{|\cos\upsilon - c| \le t\} \subset \{|\upsilon| \lesssim \sqrt{t}\}$ and $P_X(|\theta^\top \widetilde{X}| \le t) \lesssim \sqrt{t}$. Combining the two regimes, $\sup_\theta P_X(|\theta^\top \widetilde{X}| \le t) \asymp \sqrt{t}$, which is (M0) with $\alpha_0 = \frac{1}{2}$; the supremum is attained at tangential θ. □

C.3 The margin law of Example 14

Let $\rho(x) = c_\alpha |x_1|^{\alpha-1}$ on $[-1,1]^d$, $0 < \alpha < 1$; this is integrable since $\int_{-1}^1 |u|^{\alpha-1} du = 2/\alpha$. Let $\theta = (w, b)$ with $\|w\| = 1$, write η for the angle between w and $\pm e_1$, and put $\delta = \max\{\eta, |b|\} \asymp \operatorname{dist}(\theta, B)$ with $B = \{(\pm e_1, 0)\}$.

Aligned slabs ($\eta = 0$). Here the slab is $\{|x_1 + b| \le t\}$ and its mass is $\asymp \int_{|u+b|\le t} |u|^{\alpha-1} du$. If $|b| \le t$ the interval straddles the singularity and the integral is $\asymp t^\alpha$; if $|b| \ge t$ the integrand is $\asymp |b|^{\alpha-1}$ on an interval of length $2t$, giving $\asymp t|b|^{\alpha-1}$.

Tilted slabs. Write $w = (\cos\eta)e_1 + (\sin\eta)v$ with $v \perp e_1$. For fixed x' the slab constrains x_1 to an interval of length $2t/\cos\eta \asymp t$ centred at $x_1^\star(x') = -(b + (v \cdot x')\sin\eta)/\cos\eta$, and as x' ranges over the cube the centre $x_1^\star$ sweeps a range of length $\asymp \eta$. Suppose $t \le \eta$. A fraction $\asymp t/\eta$ of the x' have $|x_1^\star| \lesssim t$ and contribute $\asymp t^\alpha$ each; the x' with $|x_1^\star| \asymp r$ for $t \le r \le \eta$ occupy a fraction $\asymp r/\eta$ and contribute $\asymp t\, r^{\alpha-1}$. Hence the mass is

$$\asymp \frac{t}{\eta}\, t^\alpha + \frac{1}{\eta}\int_t^\eta t\, r^{\alpha-1}\, dr \asymp \frac{t^{1+\alpha}}{\eta} + t\,\eta^{\alpha-1} \asymp t\,\eta^{\alpha-1},$$

the second term dominating because $t^{1+\alpha}\eta^{-1}/(t\eta^{\alpha-1}) = (t/\eta)^\alpha \le 1$.

Combining the two computations, for $t \le \delta$ the mass is $\asymp t\,\delta^{\alpha-1}$, which is (M1) with $\alpha_1 = 1$ and $\kappa' = 1 - \alpha$; and taking the supremum over θ, which is attained when $\delta \lesssim t$, gives $\asymp t^\alpha$, which is (M0) with $\alpha_0 = \alpha$. The two agree at $t \asymp \delta$, as they must. Finally B consists of two points of the d-dimensional manifold Θ, so $|\{\mathrm{dist}(\cdot, B) \le \delta\}| \asymp \delta^d$ and $\nu = p = d$. □

C.4 Proof of Proposition 12

(i) Curves. We use the standard sublevel-set bound for functions of finite type, which supplies the uniformity that a pointwise Taylor expansion does not.

Lemma 24 (Stein, 1993, Ch. VIII, Prop. 2)**.** *Let $g \in C^r([a,b])$ satisfy $\max_{1\le j\le r} |g^{(j)}(s)| \ge c_1 > 0$ for every $s \in [a,b]$. Then $\big|\{s \in [a,b] : |g(s)| \le t\}\big| \le C_r\,(t/c_1)^{1/r}$, with C_r depending only on r and $b - a$.*

Parameterise Γ by arclength, $\gamma : [0, L] \to \mathbb{R}^d$, and for $\theta = (w, b)$ with $\|w\| = 1$ set $g(s) = \gamma(s)^\top w + b$, so that $g^{(j)}(s) = \gamma^{(j)}(s)^\top w$ for $j \ge 1$. The map $(s, w) \mapsto \max_{1\le j\le d} |\gamma^{(j)}(s)^\top w|$ is continuous and strictly positive on the compact set $[0, L] \times S^{d-1}$: it vanishes only if w is orthogonal to $\gamma'(s), \dots, \gamma^{(d)}(s)$, which span $\mathbb{R}^d$ by the nondegeneracy hypothesis. Hence there is $c_1 > 0$ with $\max_{1\le j\le d} |g^{(j)}(s)| \ge c_1$ for all s and all θ, and Lemma 24 gives $P_X(|g| \le t) \le C_d(t/c_1)^{1/d}$ uniformly in θ. No separate treatment of transversal and tangential θ, and no counting of zeros, is needed.

For the matching lower bound, fix s_0 and choose w orthogonal to $\gamma'(s_0), \dots, \gamma^{(d-1)}(s_0)$ — a one-dimensional choice, and $\gamma^{(d)}(s_0)^\top w \ne 0$ because the d vectors span $\mathbb{R}^d$. Taking $b = -\gamma(s_0)^\top w$ gives $r = d$ and hence $P_X(|g| \le t) \asymp t^{1/d}$.

(ii) Hypersurfaces. We first record that the second fundamental form is definite everywhere, which nonvanishing Gaussian curvature does *not* give pointwise — a saddle has nonzero curvature — but does give globally on a compact connected hypersurface. Indeed, at a point $q_\star$ of S at maximal distance from the origin the surface lies inside a ball tangent at $q_\star$, so the shape operator there is definite. The Gaussian curvature is $\det$ of the shape operator, so if it never vanishes no eigenvalue can change sign; by connectedness the signature is constant, hence definite throughout. (This is the local form of Hadamard's convexity theorem; see do Carmo (2016, Ch. 5).)

Transversal hyperplanes contribute $\asymp t$. For a tangent hyperplane at $q \in S$, write S locally as a graph $x_d = f(x')$ over T_qS with $f(0) = 0$, $\nabla f(0) = 0$ and Hessian $H = D^2 f(0)$, definite by the previous paragraph and uniformly so by compactness, giving $g(x') = \frac{1}{2}(x')^\top H x' + O(|x'|^3)$. Then $\{|g| \le t\}$ is contained in a ball of radius $\lesssim \sqrt{t}$, whose surface measure is $\asymp t^{(d-1)/2}$. Since $d \ge 3$ we have $(d-1)/2 \ge 1$ and therefore $t^{(d-1)/2} \lesssim t$. For a hyperplane at angle η from tangency the sublevel set is $\{|\frac{1}{2}(x')^\top H x' - \eta x_1' - c| \le t\}$; definiteness of H makes the level sets ellipsoids, and such a set has measure $\lesssim t$ uniformly in (η, c) once $d \ge 3$. Hence $\sup_\theta P_X(|g| \le t) \asymp t$, the supremum already being attained on transversal hyperplanes.

(iii) Higher codimension. Let $\dim M = r \ge 3$ and let $q \in M$. Transversal hyperplanes contribute $\asymp t$ as before, so only tangential ones matter. Write M near q as a graph over $T_qM \cong \mathbb{R}^r$: in normal coordinates $v \in \mathbb{R}^r$ the point is $q + v + \frac{1}{2}\mathrm{I\!I}(v, v) + O(|v|^3)$ with $\mathrm{I\!I}$ the second fundamental form. For

a unit normal n at q the tangent hyperplane with normal n gives $g(v) = \frac{1}{2}\langle \mathrm{II}(v,v), n\rangle + O(|v|^3) = \frac{1}{2} v^\top H_n v + O(|v|^3)$, where H_n is the shape operator in the direction n. Since M is C^3 and compact, the remainder in the graph expansion is $O(|v|^3)$ with a constant uniform in q, and by compactness of the unit normal bundle $\mathcal{N}M = \{(q,n) : q \in M,\ n \perp T_q M,\ |n| = 1\}$ the nondegeneracy is uniform too: there is $c_2 > 0$ with $|H_n v| \geq c_2 |v|$ for all $(q,n) \in \mathcal{N}M$ and all $v \in T_q M$. We estimate the sublevel set through the gradient rather than the value, which handles nearly tangent hyperplanes at the same time.

Let θ be any parameter whose hyperplane meets a fixed coordinate neighbourhood of q, and write the restriction of the affine functional as $g(v) = \frac{1}{2} v^\top H_n v + \langle \eta, v\rangle + c + O(|v|^3)$, where η measures the tilt away from tangency. Then $\nabla g(v) = H_n v + \eta + O(|v|^2)$. Two regimes occur. If $|\eta| > \eta_0$ for a threshold $\eta_0 > 0$ fixed below, then shrinking the neighbourhood radius ϱ so that $C\varrho + C\varrho^2 \leq \eta_0/2$ — possible uniformly, since C is uniform over $\mathcal{N}M$ — gives $|\nabla g(v)| \geq |\eta| - C\varrho - C\varrho^2 \geq \eta_0/2$ throughout, so g is transversal there and that neighbourhood contributes $\asymp t$. If instead $|\eta| \leq \eta_0$, then g has exactly one critical point in the neighbourhood, at $v_\star = -H_n^{-1}\eta + O(|\eta|^2)$, and recentring at $v_\star$ we may assume $\eta = 0$ and

$$|\nabla g(v)| \ \geq\ c_2|v| - C|v|^2 \ \geq\ \tfrac{1}{2} c_2 |v|$$

for $|v| \leq \varrho$, with ϱ uniform in (q,n,η). Finitely many such neighbourhoods cover M, by compactness, and outside them $|\nabla g|$ is bounded below, so the transversal estimate $\asymp t$ applies there.

On $\{|v| \leq \sqrt{t}\}$ the measure is $\asymp t^{r/2} \leq t$ because $r \geq 2$. On the dyadic annulus $|v| \asymp \rho$ with $\sqrt{t} \leq \rho \leq \varrho$ the coarea formula and the gradient bound give

$$\big|\{|g| \leq t\} \cap \{|v| \asymp \rho\}\big| \ \lesssim\ \frac{t\,\rho^{\,r-1}}{\rho} \ =\ t\,\rho^{\,r-2},$$

and summing over dyadic $\rho \leq \varrho$ gives $\lesssim t$ because $r \geq 3$ makes the geometric sum converge at its top end. (For $r = 2$ the sum is $t\log(1/t)$, which is the loss recorded in Remark 13.) Hence the tangential contribution is $\lesssim t$, so $\sup_\theta P_X(|g| \leq t) \asymp t$ and $\alpha_0 = 1$, the supremum being attained transversally.

Note that (i) is *not* a case of (iii): for a curve in $\mathbb{R}^d$ with $d \geq 3$ the shape operator in a binormal direction vanishes, so H_n is degenerate for those n, which is precisely how the higher-order contact of part (i) arises. □

Consistency check: at $t \asymp \delta$ the two bounds agree, $t\delta^{-1/2} \asymp \sqrt{t}$, and correspondingly $s_1 - \kappa = k + \frac{1}{2} - \frac{1}{4} = k + \frac{1}{4} = s_0$.

D Periodic translations: regularity, packing, minimax

Regularity. For $\Phi(z)(x) = \sum_k \widehat{\kappa}_k e^{2\pi i k\cdot x} e^{-2\pi i k\cdot z}$, a parameter-frequency block at $|k| \asymp 2^j$ satisfies, by Parseval, $\sup_z \|\Delta_j^z \Phi(z)\|_{L^2_x}^2 = \sum_{|k|\asymp 2^j} |\widehat{\kappa}_k|^2 \asymp 2^{jd} 2^{-2\beta j}$, so the exact regularity is $s = \beta - d/2$. Theorem 3 with $\kappa = 0$, $s_0 = s_1 = s$ gives the entropy upper bound $e_m \lesssim R m^{-\beta/d}$.

Packing lower bound. Let $G_q = \{\nu/q : \nu \in \{0,\dots,q-1\}^d\}$ and $\mu_a = q^{-d}\sum_\nu a_\nu \delta_{\nu/q}$ with $a_\nu \in \{-1,1\}$, so $\|\mu_a\|_{\mathrm{TV}} = 1$. For two codes a, b with $c = a - b$, the normalised DFT Parseval identity on the grid gives

$$\sum_{k\in\{0,\dots,q-1\}^d} \Big| q^{-d} \sum_\nu c_\nu e^{-2\pi i k\cdot\nu/q}\Big|^2 = q^{-d}\sum_\nu |c_\nu|^2.$$

Hence $\|\kappa * (\mu_a - \mu_b)\|_{L^2}^2 \geq \min_{|k|<q} |\widehat{\kappa}_k|^2 \cdot q^{-d}\sum_\nu |c_\nu|^2 \gtrsim q^{-2\beta}$ whenever the Hamming distance between a and b is $\Omega(q^d)$. A Gilbert–Varshamov code supplies $\exp(\Omega(q^d))$ such a, so $\log N(cq^{-\beta}, \mathcal{F}_1; L^2) \gtrsim q^d$, i.e. $\log N(\varepsilon) \gtrsim \varepsilon^{-d/\beta}$, matching the upper bound.

A frequency-localised packing. The grid packing above is adequate for the entropy bound, which needs only separation, but it is *not* adequate for Fano: the latter also requires an upper bound on the pairwise distances, and the grid construction does not supply one. Indeed the $k = 0$ mode alone

contributes $|\widehat{\kappa}_0|^2|\widehat{c}_0|^2$ with $\widehat{c}_0 = q^{-d}\sum_\nu c_\nu$ typically of size $q^{-d/2}$, so that $\|f_a - f_b\|_{L^2}^2 \gtrsim q^{-d}$, which for $\beta > d/2$ is far larger than $q^{-2\beta}$. We therefore use a second, frequency-localised packing, whose elements lie in a ball of radius $O(q^{-\beta})$ and so are automatically close as well as separated.

Let $\Lambda_q = \{k \in \mathbb{Z}^d : q \le |k| < 2q\}$, so $|\Lambda_q| \asymp q^d$, and choose a subset Λ_q^+ containing exactly one of each pair $\{k, -k\}$. For $\epsilon \in \{-1,1\}^{\Lambda_q^+}$ put

$$\mu_\epsilon = c_4\, q^{-d/2} \sum_{k\in\Lambda_q^+} \epsilon_k \big(e_k + e_{-k}\big), \qquad e_k(x) = e^{2\pi i k\cdot x},$$

which is real valued. Since $\mathbb{T}^d$ has unit measure, $\|\mu_\epsilon\|_{L^1} \le \|\mu_\epsilon\|_{L^2} = c_4 q^{-d/2}(2|\Lambda_q^+|)^{1/2} \le 1$ for a suitable absolute c_4, so $f_\epsilon := \kappa * \mu_\epsilon \in \mathcal{F}_1$. For two codes,

$$\|f_\epsilon - f_{\epsilon'}\|_{L^2}^2 = 8c_4^2 q^{-d} \sum_{k\in\Lambda_q^+:\,\epsilon_k\neq\epsilon'_k} |\widehat{\kappa}_k|^2,$$

and since $|\widehat{\kappa}_k| \asymp q^{-\beta}$ throughout Λ_q this is $\asymp q^{-d}q^{-2\beta}\, d_H(\epsilon,\epsilon')$. A Gilbert–Varshamov code on $\{-1,1\}^{\Lambda_q^+}$ gives $\exp(\Omega(q^d))$ points with $d_H \asymp q^d$, hence

$$c_5\, q^{-2\beta} \;\le\; \|f_\epsilon - f_{\epsilon'}\|_{L^2}^2 \;\le\; C_5\, q^{-2\beta} \qquad (\epsilon \neq \epsilon'), \tag{8}$$

the upper bound holding because at most $|\Lambda_q^+| \asymp q^d$ coordinates can differ. This packing also reproves the entropy lower bound.

Minimax lower bound. Apply Fano to the family of (8). In the Gaussian model with noise variance σ^2 and n samples the pairwise KL divergence is $n\|f_\epsilon - f_{\epsilon'}\|_{L^2}^2/(2\sigma^2) \le C_5 n q^{-2\beta}/(2\sigma^2)$ — now a genuine upper bound, by the right-hand inequality in (8). Fano's inequality therefore yields a minimax lower bound of order $q^{-2\beta}$ as soon as $C_5 n q^{-2\beta}/(2\sigma^2) \le \frac{1}{2}c_1 q^d$, i.e. $q \asymp n^{1/(2\beta+d)}$, for which $q^{-2\beta} \asymp n^{-2\beta/(2\beta+d)}$. The matching upper bound is Corollary 16 with $q = d/\beta$.

E Entropy to squared risk

Inverting $e_m(\mathcal{F}_R) \lesssim m^{-\gamma}$, $\gamma = \frac{1}{2} + \gamma^\star$, gives $\log N(\varepsilon, \mathcal{F}_R; L^2(P_X)) \lesssim \varepsilon^{-q}$ with $q = \gamma^{-1} < 2$, and solving $\varepsilon_n^{-q} \asymp n\varepsilon_n^2$ gives $\varepsilon_n \asymp n^{-1/(2+q)}$.

A word on which theorem converts this into a rate, since the point is easy to get wrong. Our entropy bound is for the *population* metric $L^2(P_X)$, whereas the localised empirical-process analysis of the exact least-squares estimator (Wainwright, 2019, Ch. 13) is driven by empirical, uniform or bracketing entropy, and population entropy does not in general control those. We therefore invoke a result whose hypothesis is population metric entropy directly, and state it precisely. Yang & Barron (1999) consider regression $Y = f^\star(X) + \xi$ with $X \sim P_X$ i.i.d., $f^\star$ in a class $\mathcal{F}$ uniformly bounded in L^∞, and Gaussian noise of known variance. They show that if ε_n solves $\log N(\varepsilon; \mathcal{F}, L^2(P_X)) \asymp n\varepsilon^2$, then the minimax squared risk over $\mathcal{F}$ is of order ε_n^2, the upper bound being attained by a minimum-contrast estimator over an ε_n-net of $\mathcal{F}$.

Each hypothesis holds here. The class $\mathcal{F}_R$ is uniformly bounded by Rb under the feature bound of Corollary 16; the design is P_X itself, so the relevant metric is the one Theorem 3 controls; and $\log N(\varepsilon) \lesssim (R/\varepsilon)^q$ with $q < 2$ gives $\varepsilon_n \asymp n^{-1/(2+q)}$.

For the upper bound the Gaussian assumption can be dropped, and it is worth saying how, since Fano supplies only the lower half of the Yang–Barron result. Let $f_1, \dots, f_N$ be an ε_n-net of $\mathcal{F}_R$ with $\log N = \log N(\varepsilon_n) \asymp n\varepsilon_n^2$, and let $\widehat{f}$ minimise the empirical risk $\frac{1}{n}\sum_i (Y_i - f_j(X_i))^2$ over the net. Write $\Delta_j = \|f_j - f^\star\|_{L^2(P_X)}$ and let $j^\star$ index a net point with $\Delta_{j^\star} \le \varepsilon_n$. Expanding the empirical risk, the comparison between $\widehat{f}$ and $f_{j^\star}$ reduces to controlling, over the net, the two centred sums

$$\frac{2}{n}\sum_{i\le n} \xi_i\big(f_j - f^\star\big)(X_i) \qquad \text{and} \qquad \frac{1}{n}\sum_{i\le n}\Big[(f_j - f^\star)^2(X_i) - \Delta_j^2\Big].$$

Both are handled by Bernstein's inequality, and it is worth being careful about the first: under random design its summands $\xi_i(f_j - f^\star)(X_i)$ have variance $\sigma^2\Delta_j^2$ but Orlicz parameter of order $Rb\,\sigma$, so a

pure sub-Gaussian bound at scale $\Delta_j\sigma/\sqrt{n}$ is not available; the summands are sub-exponential and Bernstein gives

$$\left|\tfrac{2}{n}\textstyle\sum_i \xi_i(f_j - f^\star)(X_i)\right| \;\lesssim\; \sigma\Delta_j\sqrt{\tfrac{\log N}{n}} \;+\; Rb\,\sigma\,\tfrac{\log N}{n}.$$

The second sum has summands bounded by $(2Rb)^2$ and variance $\lesssim (Rb)^2\Delta_j^2$, so Bernstein gives $\lesssim Rb\,\Delta_j\sqrt{\log N/n} + (Rb)^2\log N/n$. A union bound over the N net points costs only the $\log N$ already displayed. Since $\log N \asymp n\varepsilon_n^2$, both right-hand sides are $\lesssim (\sigma+Rb)\big(\Delta_j\varepsilon_n + (\sigma+Rb)\varepsilon_n^2\big)$, and inserting this into the basic inequality gives a quadratic in $\Delta = \|\widehat{f} - f^\star\|_{L^2(P_X)}$ whose solution is $\Delta \lesssim \varepsilon_n$, hence $\|\widehat{f} - f^\star\|^2_{L^2(P_X)} \lesssim \varepsilon_n^2$ with probability tending to one. Only uniform boundedness and sub-Gaussian noise are used; the implied constants depend on σ and Rb.

No claim is made for the *exact* least-squares estimator over $\mathcal{F}_R$, which would additionally require an empirical or bracketing entropy bound that population entropy does not supply. See Han & Wellner (2019) for refinements.

With $q = 2p/(p+2s)$ the rate is $n^{-(p+2s)/(2p+2s)}$; for ReLU with $p = d$, $s = \frac{3}{2}$ it is $n^{-(d+3)/(2d+3)}$.

F Numerical details

Figure 1 uses P_X uniform on the unit circle, $k = 1$, and the centred second difference in the bias coordinate, $\Delta_h^2\Phi_1(c) = \Phi_1(c-h) - 2\Phi_1(c) + \Phi_1(c+h)$, with the $L^2(P_X)$ norm computed by the midpoint rule on 2^{23} equispaced arclength nodes. Panel (a) fits $\log\|\Delta_h^2\Phi_1\|$ against $\log|h|$ over $|h| \in [2^{-14}, 2^{-5}]$ at the secant parameter $c = 0$ (slope 1.5000) and the tangent parameter $c = 1$ (slope 1.2500). Panel (b) evaluates $c = 1-\delta$ for $\delta \in \{2^{-4}, 2^{-6}, 2^{-8}\}$ and plots $\|\Delta_h^2\Phi_1\|\,\delta^{1/4}|h|^{-3/2}$ against $|h|/\delta$; over $|h| \le \delta/8$ the fitted slopes are 1.5001 and the prefactor lies in $[0.3876, 0.3906]$ across all three values of δ, confirming $\kappa = \frac{1}{4}$.

```
import numpy as np
import matplotlib; matplotlib.use("Agg")
import matplotlib.pyplot as plt

N  = 1 << 23           # quadrature nodes on the circle
t  = (np.arange(N) + 0.5) * (2*np.pi/N)
ct = np.cos(t)

def mod2(c, h, k=1):   #||Delta_h^2 Phi_k(c)||_{L^2(P_X)}, centred
    f = lambda cc: np.maximum(ct - cc, 0.0)**k
    g = f(c - h) - 2.0*f(c) + f(c + h)
    return np.sqrt(np.mean(g*g))

fig, ax = plt.subplots(1, 2, figsize=(9.0, 3.3))

# panel (a): local exponent at a secant and at a tangent parameter
hs = 2.0**(-np.arange(5, 15))
for c, lab in ((0.0, "secant"), (1.0, "tangent")):
    v = np.array([mod2(c, h) for h in hs])
    slope = np.polyfit(np.log(hs), np.log(v), 1)[0]
    ax[0].loglog(hs, v, "o-", ms=4, label="%s  (slope %.3f)"
    % (lab, slope)) ax[0].set_xlabel("|h|"); ax[0].legend();
    ax[0].grid(alpha=.25, which="both")

# panel (b): shell collapse, confirming kappa = 1/4
for delta in (2.0**-4, 2.0**-6, 2.0**-8):
    r  = 2.0**np.arange(-7, 4.01, 0.5)          # |h| / delta
    hh = delta*r
    v  = np.array([mod2(1.0 - delta, h) for h in hh])
    ax[1].loglog(r, v/(hh**1.5 * delta**-0.25), "o-", ms=3.5,
                 label="delta = 2^%d" % np.log2(delta))
```

```
ax[1].axhline(0.388, ls=":", c="k");
ax[1].axvline(1.0, ls="--", c="gray")
ax[1].set_xlabel("|h|/delta");
ax[1].legend();
ax[1].grid(alpha=.25, which="both")

plt.tight_layout();
plt.savefig("fig_circle.pdf", bbox_inches="tight")
```