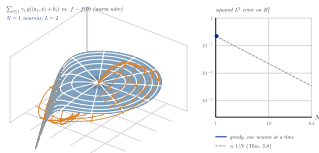


Breaking the Curse of Dimensionality

Compositional structure, manifolds, and Barron's theorem

Johannes Herter

Columbia University · Theory of Deep Neural Networks

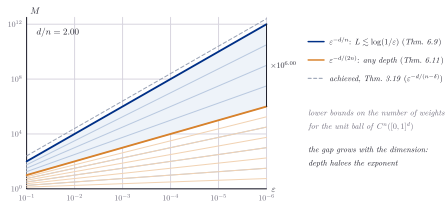


Lectures I and II

ReLU networks reach accuracy ε on the unit ball of $C^s([0, 1]^d)$ with $M \lesssim \varepsilon^{-d/(s-\delta)}$ weights (Thm. 3.19), and no piecewise polynomial network does essentially better: $M \gtrsim \varepsilon^{-d/s}$ up to logs (Thm. 6.9).

The exponent d/s is the problem. Images have $d \approx 10^5$ pixels; $\varepsilon^{-d/s}$ is astronomically large.

Today. Which extra structure lets networks escape, and why they exploit it so well.



Lecture II: lower bounds on the number of weights, growing with d/n .

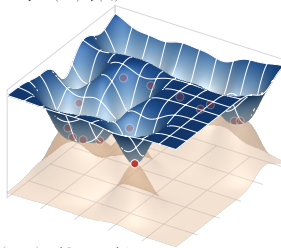
5.1 *Curse of dimensionality*

A sharp example [21]

$F_d := \{f \in C^\infty([0, 1]^d) : \sup_{|\alpha|=1} \|D^\alpha f\| \leq 1\}$, $e(n, d)$ the best error of *any* algorithm reconstructing every $f \in F_d$ from n point values. Then $e(n, d) = 1$ for all $n \leq 2^{\lfloor d/2 \rfloor} - 1$.

- **Why:** samples only reveal $f(x_i)$. A smooth f with $f(x_i) = 0$, $\max |f| \approx 1$ gives the same data as 0 and $-f$: some answer is off by ≈ 1 ; in high d for up to $2^{\lfloor d/2 \rfloor} - 1$ samples.
- Every bound $e(n, d) \leq C_{d,r} n^{-r}$ has $C_{d,r}$ exponential in d (Bellman).
- $\varepsilon^{-d/s}$ is optimal for C^s [9], also for networks (Thm. 6.9).
Hence: escaping needs an *additional assumption* on f .

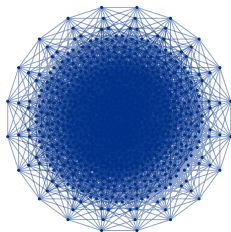
$n = 14$ samples (red): $f(x_i) = 0$



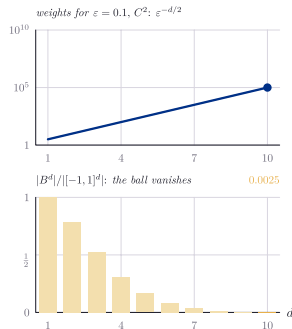
$f, -f$ (warm) and 0: same data

$d = 10$

the cube $[0, 1]^{10}$: 2^{10} corners



grid of spacing $\frac{1}{4}$: $5^{10} = 9,765,625$ points



Left: a grid of spacing $\frac{1}{4}$ needs 5^d points; for $d \geq 4$ the cube $[0, 1]^d$ itself, its 2^d corners projected to the plane.
Right: the weights $\varepsilon^{-d/2}$ needed for C^2 at $\varepsilon = 0.1$, and the volume of the inscribed ball, which goes to 0: in high dimension almost everything is far from the centre.

Structure (5.2) f is a composition of functions of at most k variables, along a graph.

Rate $\varepsilon^{-k/s}$: the **indegree** replaces d (Thm. 5.3).

Data (5.3) Only values on a d' -dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$ matter.

Rate $\varepsilon^{-d'/k}$: the **intrinsic dimension** replaces d (Thm. 5.6).

Regularity (5.4)
 $\int |\xi| |\hat{f}(\xi)| d\xi \leq C$: a smoothness that grows with d .

Rate N^{-1} in L^2 , **no d at all** (Thm. 5.8).

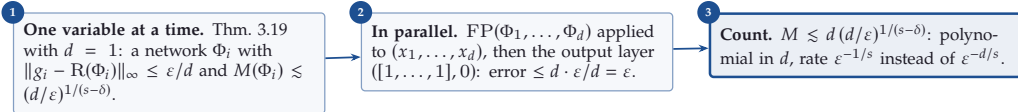
The common thread. Each assumption is matched by a network operation we already have: composition (Def. 2.9), localisation by hat functions (Prop. 3.3), and superposition of ridge functions (two-layer networks).

5.2 *Hierarchy assumptions*

Warm-up: additive functions



Example. $f(x) = \sum_{i=1}^d g_i(x_i)$ with $g_i \in C^s([0, 1])$, $\|g_i\|_{C^s} \leq 1$.



Question. Which richer class of “functions built from low-dimensional pieces” keeps this property, and can networks emulate the building process? $\Rightarrow$ compositional functions [27].

Definition 5.1

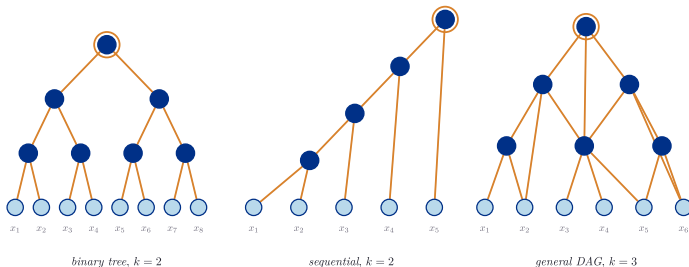
$\mathcal{G}(d, k, N)$: directed acyclic graphs with N vertices, indegree $\leq k$, exactly d vertices of indegree 0 (the inputs), and outdegree ≥ 1 for all but one vertex. Topological order $\eta_1, \dots, \eta_N$; $T_i := \{j : \eta_j \eta_i \text{ an edge}\}$, $d_i := \#T_i \leq k$.

Definition 5.2

For $f_i \in C^s(\mathbb{R}^{d_i})$, $\|f_i\|_{C^s} \leq 1$: $v_i := x_i$ ($i \leq d$), $v_i := f_i(v_{j_1}, \dots, v_{j_{d_i}})$ ($i > d, j_\ell \in T_i$). Then $f := v_N$ is a **compositional function**; the class is $\text{CF}(d, k, N; s)$.

Why this class? It mirrors the visual cortex (local features combined into more abstract ones [20, 27]), sensor networks that aggregate nearby readings first, and Boolean circuits made of small gates.

Fig. 5.1: three graphs



evaluate in topological order: $v_i = f_i((v_j)_{j \in T_i})$, $f = v_N$; each $f_i \in C^s(\mathbb{R}^{d_i})$, $d_i \leq k$

The vertices are evaluated in topological order; navy: computed, warm: the edges just used. Each internal vertex is a function of at most k variables, even though f depends on all d inputs.

Theorem 5.3: the indegree counts



Theorem 5.3

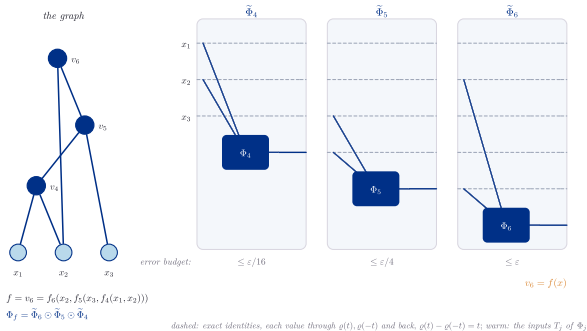
There is $C > 0$ such that for every $f \in \text{CF}(d, k, N; s)$ and $\varepsilon \in (0, \frac{1}{2})$ a ReLU network Φ_f has

$$L(\Phi_f) \leq CN^2 \log_2(k/\varepsilon), \quad M(\Phi_f) \leq CN^4(2k)^{\frac{kN}{s}} \varepsilon^{-\frac{k}{s}} \log_2(k/\varepsilon), \quad \|f - \text{R}(\Phi_f)\|_\infty \leq \varepsilon.$$

Remark 5.4. The rate is $\varepsilon^{-k/s}$: it depends on the maximal indegree k , not on d . Every vertex costs as much as a k -variate C^s function.

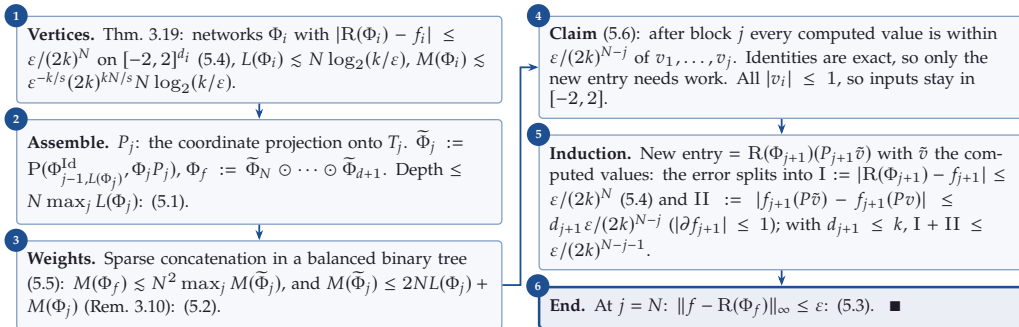
The price. The constant $(2k)^{kN/s}$ grows super-exponentially in k : errors are amplified along the graph, so each vertex must be approximated to $\varepsilon/(2k)^N$.

A network of networks



The proof of Thm. 5.3 turns the graph into *one* ReLU network: a block per vertex copies all values exactly and appends $v_j \approx f_j(v_{T_j})$. Example $f_4 = x_1 x_2$, $f_5 = \sin(x_3 + v_4)$, $f_6 = x_2 v_5$; the error budget grows by $2k$ per block.

Proof of Theorem 5.3



5.3 *Manifold assumptions*

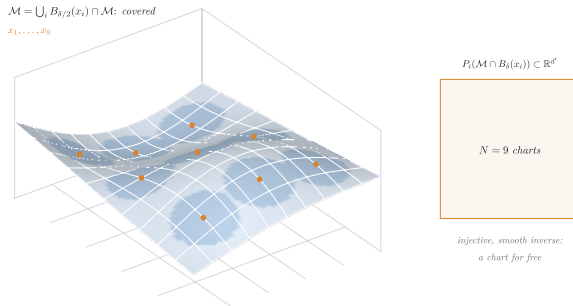
Natural images are not arbitrary pixel arrays: plausibly they fill a low-dimensional submanifold $\mathcal{M} \subset \mathbb{R}^d$, and we only need $|f(x) - \mathbf{R}(\Phi)(x)| \leq \varepsilon$ for $x \in \mathcal{M}$.

Definition 5.5

A smooth d' -dimensional submanifold $\mathcal{M} \subset \mathbb{R}^d$ is **(N, δ) -covered** if there are $x_1, \dots, x_N \in \mathcal{M}$ with $\bigcup_i B_{\delta/2}(x_i) \supset \mathcal{M}$, such that the orthogonal projection $P_i : \mathcal{M} \cap B_\delta(x_i) \rightarrow T_{x_i}\mathcal{M} \cong \mathbb{R}^{d'}$ is injective and smooth with smooth inverse.

Smoothness on $\mathcal{M}$: $\|f\|_{C^k, \delta, N} := \sup_i \|f \circ P_i^{-1}\|_{C^k(P_i(\mathcal{M} \cap B_\delta(x_i)))}$, i.e. C^k in the charts given by the tangent projections.

Fig. 5.2: tangent projections are charts



A surface $\mathcal{M} \subset \mathbb{R}^3$ ($d' = 2$, $d = 3$). For each centre x_i the patch $\mathcal{M} \cap B_{\delta}(x_i)$ (navy) is projected onto the tangent plane (warm); right: its image, a domain in $\mathbb{R}^{d'}$. The balls $B_{\delta/2}(x_i)$ (light) cover $\mathcal{M}$.

Theorem 5.6: *intrinsic dimension*



Theorem 5.6

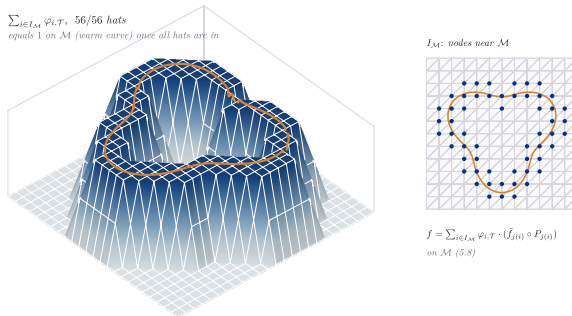
$\mathcal{M} \subset \mathbb{R}^d$ an (N, δ) -covered d' -dimensional manifold. There is $c > 0$: for every $\varepsilon > 0$ and $f \in C^k(\mathcal{M})$ with $\|f\|_{C^k, \delta, N} \leq 1$ a ReLU network Φ has

$$\sup_{x \in \mathcal{M}} |f(x) - \text{R}(\Phi)(x)| \leq \varepsilon, \quad M(\Phi) \leq c \varepsilon^{-d'/k} \log_2(1/\varepsilon), \quad L(\Phi) \leq c \log_2(1/\varepsilon).$$

Plan. (1) Write f as a sum of *localised* pieces, each a d' -variate function of an affine map: $f = \sum_{i \in I_{\mathcal{M}}} \varphi_{i, \mathcal{T}} \cdot (\tilde{f}_{j(i)} \circ P_{j(i)})$. (2) Every factor is a network: hats exactly (Prop. 3.3), affine maps exactly, $\tilde{f}_j$ by Thm. 3.19 in d' variables; products by Prop. 3.16.

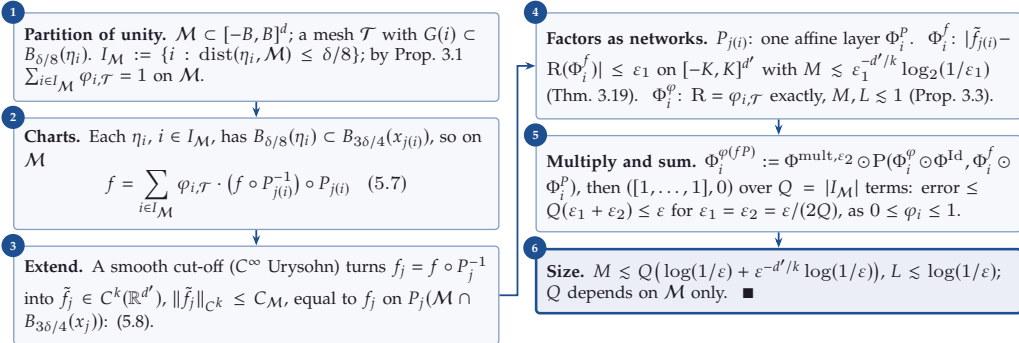
Remark 5.7. The constant still depends on d (e.g. through the number of cells around a node, exponential in d on a grid); the *rate* does not.

Fig. 5.3: hats near the manifold



A curve $\mathcal{M}$ in the plane and a simplicial mesh. The hats of the nodes I_M near $\mathcal{M}$ are added one by one; their sum equals 1 on $\mathcal{M}$ (Prop. 3.1), so they split f into pieces that each live in one chart.

Proof of Theorem 5.6



5.4 *Dimension dependent regularity*

Barron [2]

$$\Gamma_C := \left\{ f \in L^1(\mathbb{R}^d) : \|\hat{f}\|_1 < \infty, \int_{\mathbb{R}^d} |2\pi\xi| |\hat{f}(\xi)| d\xi < C \right\}.$$

$\|\hat{f}\|_1 < \infty$: f continuous (Fourier inversion); the first moment: C^1 , $\|\nabla f\|_\infty \leq C$.

Example 5.14 [2, Sect. IX]:

1. $g \in \Gamma_C \Rightarrow a^{-d}g(a(\cdot - b)) \in \Gamma_C$ for $a > 0, b \in \mathbb{R}^d$ (shifts and dilations);
2. $\sum_i c_i g_i \in \Gamma_{\|c\|_1 C}$ for $g_i \in \Gamma_C$ (superpositions);
3. the Gaussian $e^{-|x|^2/2}$ is in Γ_C with $C = O(d^{1/2})$;
4. if the derivatives up to order $\lfloor d/2 \rfloor + 2$ are square integrable, $g \in \Gamma_C$ with $C \lesssim \|g\|_{W^{\lfloor d/2 \rfloor + 2, 2}}$.

Theorem 5.8: a dimension-free rate



Theorem 5.8 [2, Thm. 1]

$f \in \Gamma_C$, ϱ sigmoidal, $N \in \mathbb{N}$, $c > 4C^2$. There is a network Φ with $L(\Phi) = 2$, $M(\Phi) \leq N(d + 2) + 1$ and

$$\frac{1}{|B_1^d|} \int_{B_1^d} |f(x) - R(\Phi)(x)|^2 dx \leq \frac{c}{N}.$$

Lemma 5.12 (Fourier analysis):
 $f - f(0) \in \overline{\text{co}}(G_C)$, G_C = scaled
Heaviside ridges $\gamma \mathbf{1}_{\mathbb{R}_+}(\langle a, x \rangle + b)$,
 $|\gamma| \leq 2C$.

Lemma 5.10 (probability): anything
in $\overline{\text{co}}(G)$ is within B^2/N of an aver-
age of N elements of G .

Theorem 5.8: N ridges with error² $\leq$
 $4C^2/N$; replace $\mathbf{1}_{\mathbb{R}_+}$ by $\varrho(\lambda \cdot)$, $\lambda \rightarrow$
 ∞ .

Remark 5.9. The rate cannot be improved beyond $N^{-(2+d)/d}$ [25]: for large d , essentially N^{-1} is the truth.

Lemma 5.10: approximate Carathéodory



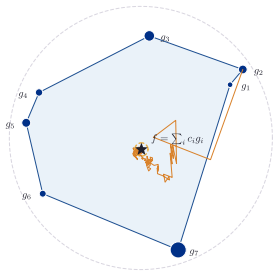
Lemma 5.10 [2, 26]

G a subset of a Hilbert space with $\|g\| \leq B$ for all $g \in G$, and $f \in \overline{\text{co}}(G)$. For every N and $c' > B^2$ there are $g_1, \dots, g_N \in G$ and $c_i \in [0, 1]$, $\sum c_i = 1$, with $\left\|f - \sum_{i=1}^N c_i g_i\right\|^2 \leq c'/N$ (5.12).

Intuition (Maurey's empirical method). Write f as an average $\sum_i c_i g_i$ (possibly over millions of g_i). Pick N of them *at random* with probabilities c_i and average. The average is unbiased, and its variance is at most B^2/N , in *any* Hilbert space: no dimension appears.

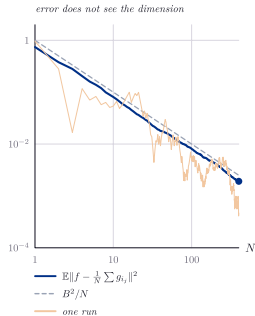
Carathéodory: f is an exact convex combination of $d + 1$ points, which depends on d . Lemma 5.10 trades exactness for an error $B/\sqrt{N}$, independent of the dimension.

Random averages converge

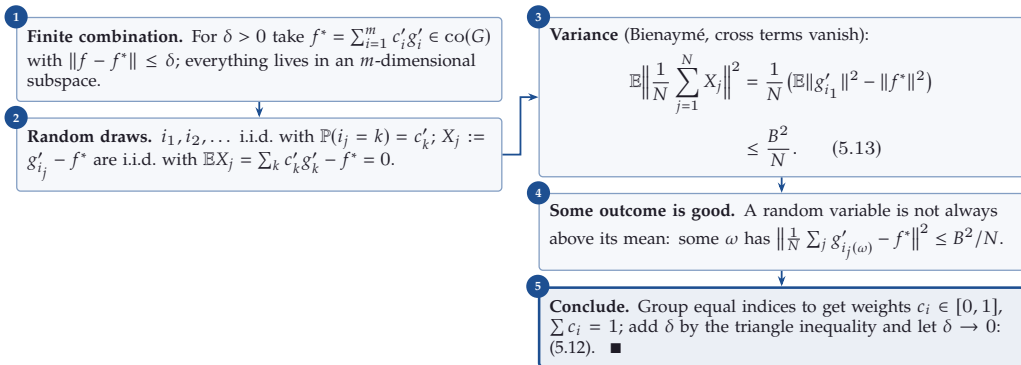


$$\frac{1}{N} \sum_{j \leq N} g_{i_j}, i_j \sim c \text{ i.i.d.}, N = 400$$

Seven points g_i with $\|g_i\| \leq B = 1$ (disc size: the weight c_i), $f = \sum c_i g_i$ (star). Warm: the running average of i.i.d. draws; gold: radius $B/\sqrt{N}$. Right: the mean squared error over 400 runs follows B^2/N .



Proof of Lemma 5.10



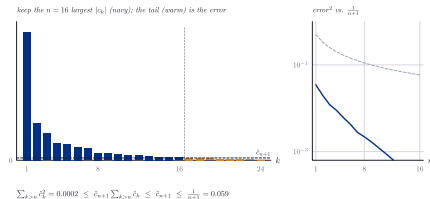
Special case: Stechkin's estimate



If $G = (\varphi_i)$ is an orthonormal basis and $f = \sum_i c_i \varphi_i$ with $\|c\|_1 = 1$, keep the n largest coefficients (indices Λ_n); with $\tilde{c}$ the decreasing rearrangement, by Parseval

$$\left\| f - \sum_{j \in \Lambda_n} c_j \varphi_j \right\|^2 = \sum_{k > n} \tilde{c}_k^2 \leq \tilde{c}_{n+1} \sum_{k > n} \tilde{c}_k \leq \tilde{c}_{n+1} \leq \frac{1}{n+1},$$

the last step since $(n+1)\tilde{c}_{n+1} \leq \sum_{k \leq n+1} \tilde{c}_k \leq 1$ (5.14–5.16). Deterministic, but only for orthonormal G .



Lemma 5.12: averages of ridge steps



The dictionary

$G_C := \{B_1^d \ni x \mapsto \gamma \mathbf{1}_{\mathbb{R}_+}(\langle a, x \rangle + b) : a \in \mathbb{R}^d, b \in \mathbb{R}, |\gamma| \leq 2C\}$: one Heaviside neuron, scaled.

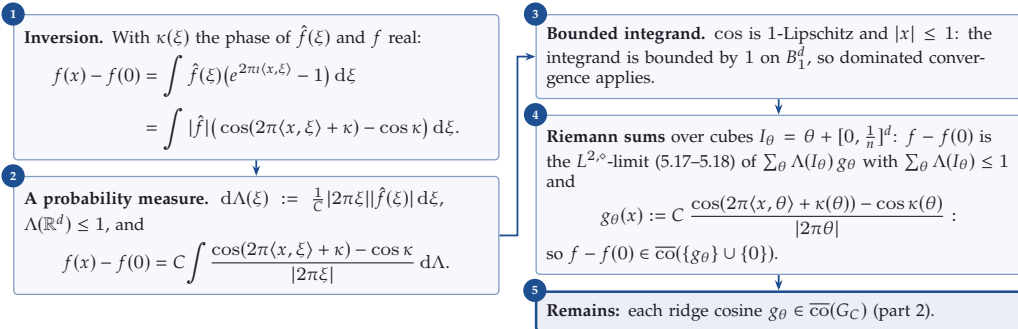
Lemma 5.12

For $f \in \Gamma_C$: $f|_{B_1^d} - f(0) \in \overline{\text{co}}(G_C)$, the closure in $L^{2,\diamond}(B_1^d)$, $\|g\|_{L^{2,\diamond}}^2 := \frac{1}{|B_1^d|} \int_{B_1^d} |g|^2$.

Remark 5.11. With Lemma 5.10 this suffices: an *arbitrarily wide* two-layer network with convex output weights is approximated by N neurons with error² $\propto 1/N$.

Route. Fourier inversion writes $f - f(0)$ as an average of ridge cosines; each ridge cosine is an average of ridge steps.

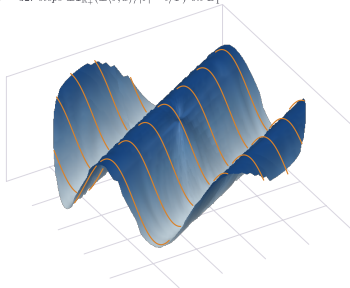
Lemma 5.12, part 1: Fourier to ridges



Ridge cosines as staircases

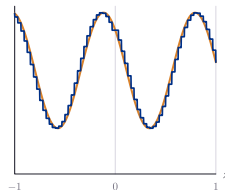


$T = 32$: steps $\pm \mathbf{1}_{\mathbb{R}_+}(\pm \langle \theta, x \rangle / |\theta| - i/T)$ on B_1^2



warm: the ridge $g_\theta(x) = \hat{g}_\theta(\langle x, \theta \rangle / |\theta|)$ along lines parallel to θ

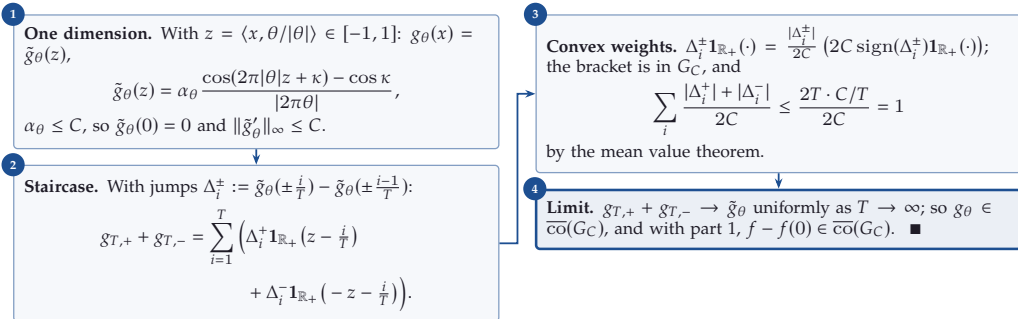
$\hat{g}_\theta(z)$ and its staircase



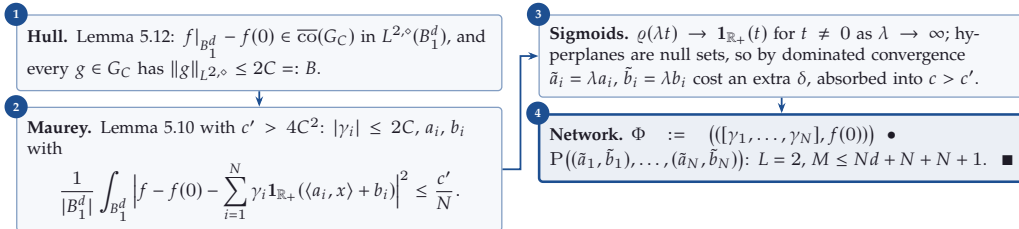
$\sum |\text{jump}| = 1.277 \leq 2 \|\hat{g}_\theta\|_\infty \leq 2C$
 $\Rightarrow$ a convex combination of
 $\pm 2C \mathbf{1}_{\mathbb{R}_+}(\cdot)$ ridge steps

g_θ depends on x only through $z = \langle x, \theta \rangle / |\theta|$. Its staircase with T steps on each side of 0 is a sum of ridge steps; the total jump is at most $2C$, so after dividing by $2C$ it is a convex combination of elements of G_C , converging as $T \rightarrow \infty$.

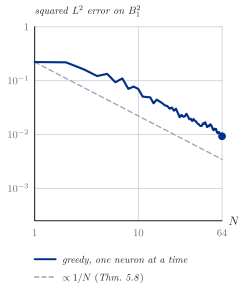
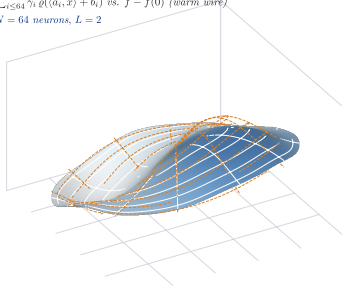
Lemma 5.12, part 2: ridges to steps



Proof of Theorem 5.8



$\sum_{i \leq 64} \gamma_i \varrho(\langle a_i, x \rangle + b_i)$ vs. $f - f(0)$ (warm wire)
 $N = 64$ neurons, $L = 2$



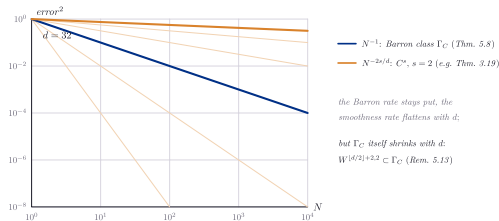
A constructive version of Lemma 5.10 (relaxed greedy, Jones/Barron): $f_N = (1 - \alpha_N)f_{N-1} + \alpha_N \gamma g$, one sigmoid ridge at a time from 6000 candidates, $\alpha_N = 2/(N + 1)$. Two Gaussian bumps plus a tilt on B_1^2 ; the squared error follows $1/N$.

Remark 5.13: no free lunch



Why the rate is plausible. Γ_C is a **dimension dependent** regularity class: it gets more restrictive as d grows. Necessary: bounded first derivatives. Sufficient: $W^{\lfloor d/2 \rfloor + 2, 2}(\mathbb{R}^d)$, which grows with d .

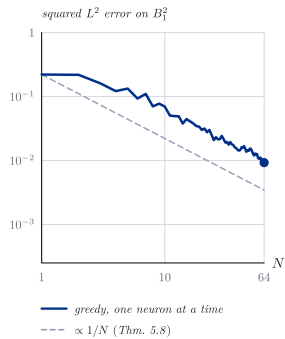
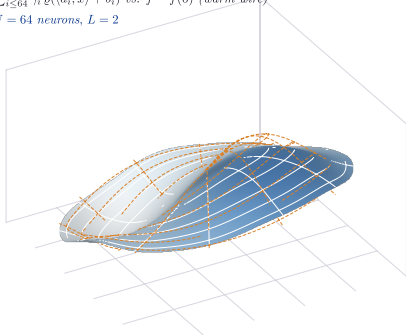
Compare. $W^{\lfloor d/2 \rfloor + 2, 2}$ also gives N^{-1} (squared L^2) for sums of N B-splines [22, 9]. The network does not beat classical methods *there*; its advantage is that Γ_C is larger and the network finds the right ridge directions.



The Barron rate is fixed; the C^2 rate flattens as d grows.

- No method beats $\varepsilon^{-d/s}$ on C^s ; escaping needs extra structure (5.1).
- **Compositional** f (indegree k): rate $\varepsilon^{-k/s}$, via a network of networks with exact identities; errors grow by $2k$ per vertex (Thm. 5.3).
- **Manifold** data ($\dim \mathcal{M} = d'$): rate $\varepsilon^{-d'/k}$, via hats near $\mathcal{M}$, tangent charts and products (Thm. 5.6).
- **Barron** class: squared L^2 error $\leq 4C^2/N$ with N neurons, two layers, no d in the rate (Thm. 5.8).
- Its two ingredients: Fourier inversion makes f an average of ridge steps (Lemma 5.12); random sampling of an average converges like $1/N$ in any Hilbert space (Lemma 5.10).

$\sum_{i \leq 64} \gamma_i \varrho(\langle a_i, x \rangle + b_i)$ vs. $f - f(0)$ (warm wire)
 $N = 64$ neurons, $L = 2$



Thank you.

Johannes Herter