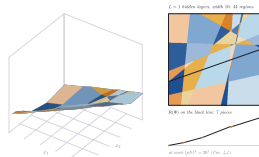


ReLU Networks and the Role of Depth

Finite elements, the square function, linear pieces

Johannes Herter

Columbia University · Theory of Deep Neural Networks

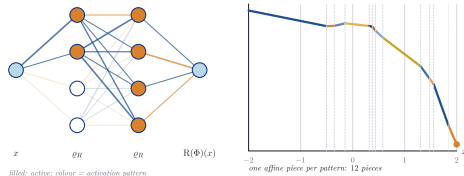


So far

Two-layer networks are universal (Thm. 2.4);
higher-order sigmoids give $M \lesssim \varepsilon^{-d/(s-\delta)}$
for C^s (Cor. 2.19).

Today: the ReLU $\varrho_R(x) = \max\{0, x\}$. Still:

- ReLU networks **are** linear finite elements (3.1);
- depth emulates x^2 cheaply (3.2), hence C^s rates (3.3);
- depth is **necessary** (4).



Architecture $(1, 4, 4, 1)$: as x sweeps $[-2, 2]$ neurons switch on and off; one affine piece per activation pattern.

Definition 2.8 (Neural network)

$\Phi = ((A_1, b_1), \dots, (A_L, b_L)), A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}, b_\ell \in \mathbb{R}^{N_\ell}, N_0 := d$, with *realisation* $R(\Phi) : \mathbb{R}^d \rightarrow \mathbb{R}^{N_L}$,

$$x_0 := x, \quad x_\ell := \varrho(A_\ell x_{\ell-1} + b_\ell) \ (\ell < L), \quad x_L := A_L x_{L-1} + b_L.$$

$N(\Phi) := d + \sum_j N_j$ neurons, $L(\Phi) := L$ layers, $M(\Phi) := \sum_j \|A_j\|_0 + \|b_j\|_0$ weights.

Definition 2.9 (Concatenation)

Merge the last affine map of Φ^2 with the first of Φ^1 : $(A_1^1 A_{L_2}^2, A_1^1 b_{L_2}^2 + b_1^1); L_1 + L_2 - 1$ layers, $R(\Phi^1 \bullet \Phi^2) = R(\Phi^1) \circ R(\Phi^2)$.

Definition 2.10 (Parallelisation)

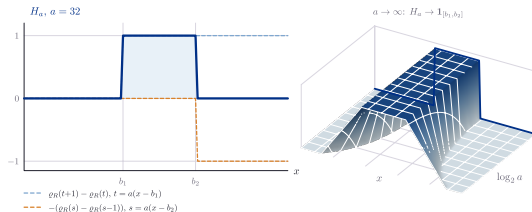
Block-diagonal stacking: $P(\Phi^1, \Phi^2)$ realises $x \mapsto (R(\Phi^1)(x), R(\Phi^2)(x))$; FP without shared input; $M = M(\Phi^1) + M(\Phi^2)$.

The ReLU is discriminatory



For $a \in \mathbb{R}, b_1 < b_2$:

$$H_a(x) := \varrho_R(ax - ab_1 + 1) - \varrho_R(ax - ab_1) - \varrho_R(ax - ab_2) + \varrho_R(ax - ab_2 - 1) \longrightarrow \mathbf{1}_{[b_1, b_2]} \quad (a \rightarrow \infty).$$



Case by case: $H_a = 0$ for $x < b_1 - \frac{1}{a}$; $= a(x - b_1 + \frac{1}{a}) \leq 1$ on $[b_1 - \frac{1}{a}, b_1]$; $= 1$ on $[b_1, b_2]$; $= 1 - a(x - b_2)$ on $[b_2, b_2 + \frac{1}{a}]$; $= 0$ beyond. So $\int \varrho_R(ax + b) d\mu = 0 \forall a, b$ gives $\int \mathbf{1}_{[b_1, b_2]}(ax) d\mu = 0$, and the proof of Prop. 2.6 concludes $\mu = 0$: **ReLU networks are universal**.

3.1 *Linear finite elements and ReLU networks*

A **simplicial mesh** of $\Omega \subset \mathbb{R}^d$: $\mathcal{T} = (\tau_i)_{i=1}^{M_{\mathcal{T}}}$, d -simplices with $\bigcup \mathcal{T} = \Omega$, $\tau_i \cap \tau_j \subset \partial\tau_i \cap \partial\tau_j$ a lower-dimensional simplex. Nodes $(\eta_i)_{i=1}^{M_N}$. Locally convex: $\bigcup\{\tau_j : \eta_i \in \tau_j\}$ convex for every η_i .

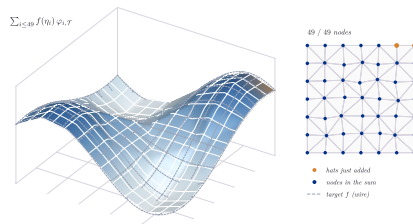
$$V_{\mathcal{T}} := \{f \in C(\Omega) : f|_{\tau_i} \text{ affine linear } \forall i\}.$$

Courant elements $\varphi_{i,\mathcal{T}} \in V_{\mathcal{T}}$: $\varphi_{i,\mathcal{T}}(\eta_j) = \delta_{ij}$.

Proposition 3.1

Every $f \in V_{\mathcal{T}}$ satisfies $f = \sum_{i=1}^{M_N} f(\eta_i) \varphi_{i,\mathcal{T}}$.

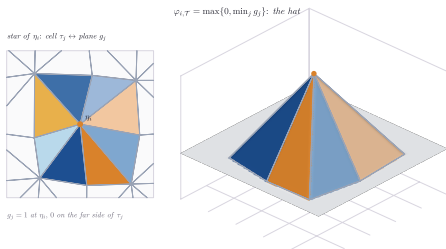
So: if every $\varphi_{i,\mathcal{T}}$ is a ReLU network, so is every $f \in V_{\mathcal{T}}$.



Building an element of $V_{\mathcal{T}}$ node by node,
 $\sum_i f(\eta_i) \varphi_{i,\mathcal{T}}$.

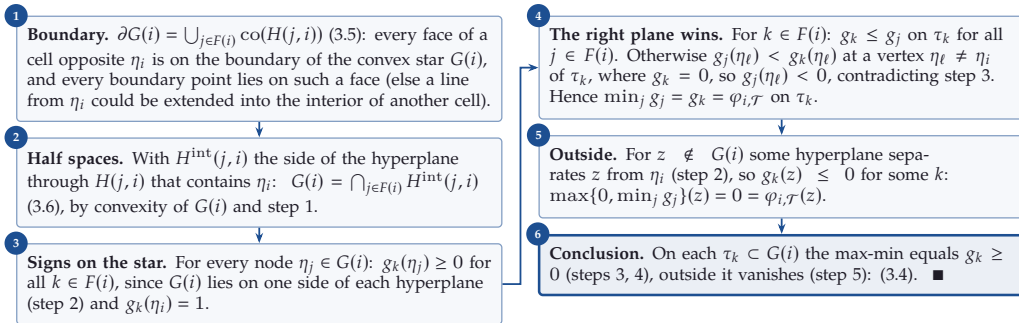
Proposition 3.2

Locally convex simplicial mesh, $F(i) := \{j : \eta_i \in \tau_j\}$, $H(j, i) := \{\eta_k \in \tau_j, \eta_k \neq \eta_i\}$. Then $\varphi_{i,\mathcal{T}} = \max \{0, \min_{j \in F(i)} g_j\}$ (3.4), g_j affine, $g_j = 0$ on $H(j, i)$, $g_j(\eta_i) = 1$.



The planes g_j of the cells around η_i enter one by one; each facet is coloured by the plane that is the minimum there; $G(i) := \bigcup_{j \in F(i)} \tau_j$ is the star of η_i . Finally $\max\{0, \cdot\}$ cuts the pyramid at zero: exactly the hat $\varphi_{i,\mathcal{T}}$.

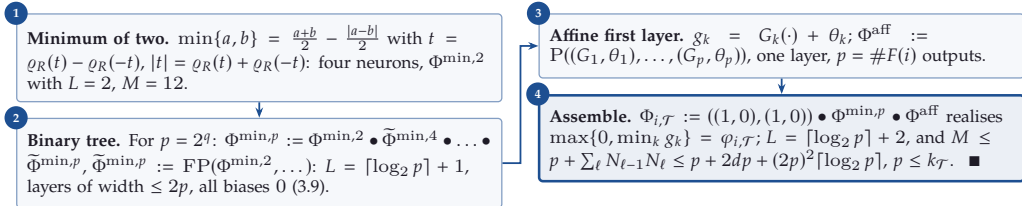
Proof of Proposition 3.2



Prop. 3.3: Courant elements are networks



Proposition 3.3. Locally convex mesh, $k_{\mathcal{T}} := \max_i \#\{j : \eta_j \in \tau_i\}$ (3.7). For every i there is a ReLU network Φ_i with $L(\Phi_i) = \lceil \log_2 k_{\mathcal{T}} \rceil + 2$, $M(\Phi_i) \leq C(k_{\mathcal{T}} + d)k_{\mathcal{T}} \log_2 k_{\mathcal{T}}$ and $R(\Phi_i) = \varphi_{i,\mathcal{T}}$ (3.8).



Theorem 3.4: linear finite elements



Theorem 3.4

Let $\mathcal{T}$ be a locally convex partition of $\Omega \subset \mathbb{R}^d$ with M_N nodes. For every $f \in V_{\mathcal{T}}$ there is a NN Φ_f with $L(\Phi_f) \leq \lceil \log_2 k_{\mathcal{T}} \rceil + 2$, $M(\Phi_f) \leq CM_N(k_{\mathcal{T}} + d)k_{\mathcal{T}} \log_2 k_{\mathcal{T}}$ and $R(\Phi_f) = f$.

Proof. Prop. 3.1 writes $f = \sum_i f(\eta_i)\varphi_{i,\mathcal{T}}$; parallelise the $\Phi_{i,\mathcal{T}}$ of Prop. 3.3 (padding shallower ones with identities) and combine with the output layer $([f(\eta_1), \dots, f(\eta_{M_N})], 0)$. ■

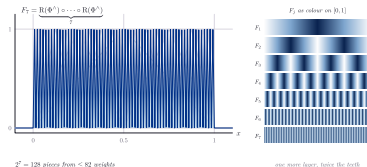
Remark 3.5. ReLU networks with $C \cdot N$ weights approximate as well as finite elements with N degrees of freedom, e.g. $\inf_{g \in V_{\mathcal{T}_n}} \|f - g\|_{L^2} \lesssim n^{-2/d} \|f\|_{W^{2,2d/(d+2)}}$. But piecewise *linear* approximation stops improving for $s \geq 2$.

3.2 *Approximation of the square function*

The sawtooth: Proposition 3.6



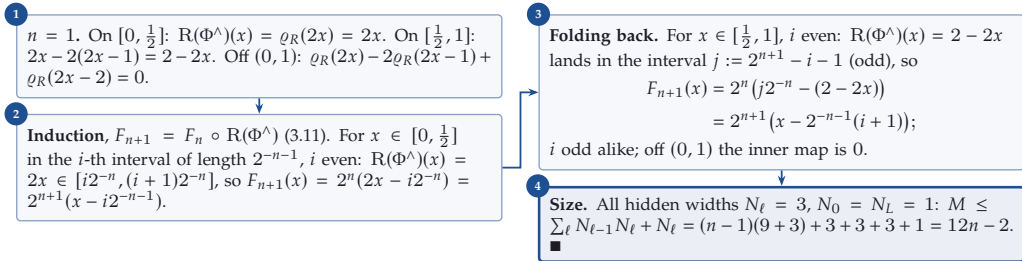
$\Phi^\wedge := ((A_1, b_1), (A_2, 0))$, $A_1 = (2, 2, 2)^T$, $b_1 = (0, -1, -2)^T$, $A_2 = [1, -2, 1]$:
 $R(\Phi^\wedge)(x) = \varrho_R(2x) - 2\varrho_R(2x - 1) + \varrho_R(2x - 2)$, a hat; $M(\Phi^\wedge) = 8$.



Proposition 3.6

$F_n := R(\Phi^\wedge \bullet \dots \bullet \Phi^\wedge)$ (n times) is, on $(0, 1)$, the sawtooth $F_n(x) = 2^n(x - i2^{-n})$ on $[i2^{-n}, (i+1)2^{-n}]$, i even, $= 2^n((i+1)2^{-n} - x)$, i odd (3.10); $F_n = 0$ off $(0, 1)$. $L = n + 1$, $M \leq 12n - 2$.

Proof of Proposition 3.6



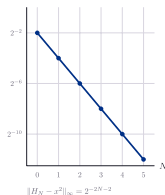
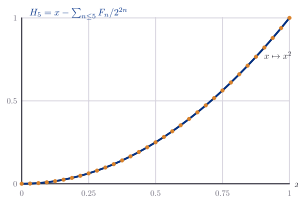
Remark 3.7 (depth vs. width). Two-layer nets have $\leq N+1 \leq M(\Phi)+1$ pieces, F_n has $2^n \geq 2^{(M+2)/12}$: exponentially many pieces from linearly many weights [Telgarsky].

Prop. 3.8 (Yarotsky): squares from sawtooths



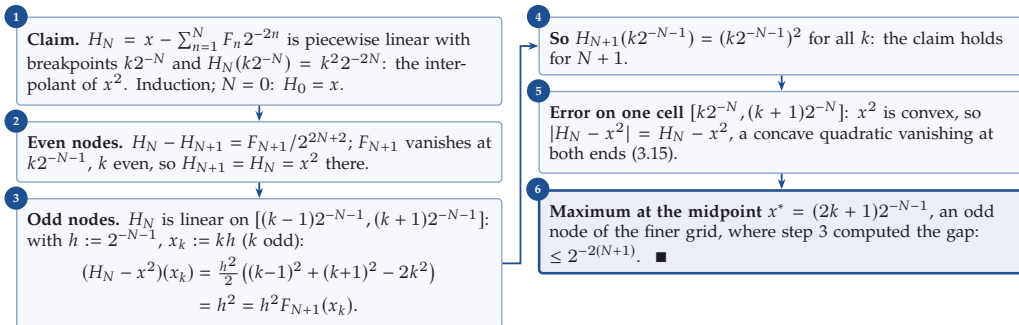
Proposition 3.8 [40]

For every $x \in [0, 1]$ and $N \in \mathbb{N}$:
$$\left| x^2 - x + \sum_{n=1}^N \frac{F_n(x)}{2^{2n}} \right| \leq 2^{-2N-2}. \quad (3.12)$$



H_N is the piecewise linear interpolant of x^2 on the grid $k2^{-N}$; each sawtooth fixes the new midpoints, the error falls by 4.

Proof of Proposition 3.8



Lemma 3.9

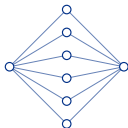
$\Phi^{\text{Id}} := (([\text{Id} \atop -\text{Id}], 0), ([\text{Id} \mid -\text{Id}], 0))$ realises $\text{Id}_{\mathbb{R}^d}$, as $\varrho_R(x) - \varrho_R(-x) = x$; with any $L \geq 2$ layers: $\Phi_{d,L}^{\text{Id}}$ (Rem. 3.10).

Definition 3.12 (Sparse concatenation)

$\Phi^1 \odot \Phi^2 := \Phi^1 \bullet \Phi^{\text{Id}} \bullet \Phi^2$: $L_1 + L_2$ layers, realisation $R(\Phi^1) \circ R(\Phi^2)$, $M \leq 2M(\Phi^1) + 2M(\Phi^2)$ (Rem. 3.13).

Why (Example 3.11). $\Phi = ((1_N, 0), (1_N^T, 0))$ has $M = 2N$, but $\Phi \bullet \Phi$ multiplies out the dense block $1_N 1_N^T$: $M = 2N + N^2$. The identity layer keeps the factors apart.

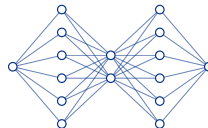
$\Phi, M(\Phi) = 2N$



$\Phi \bullet \Phi$: middle layer $A_1 A_2$, N^2 weights



$\Phi \odot \Phi = \Phi \bullet \Phi^{\text{Id}} \bullet \Phi$: $\leq 4N + \dots$

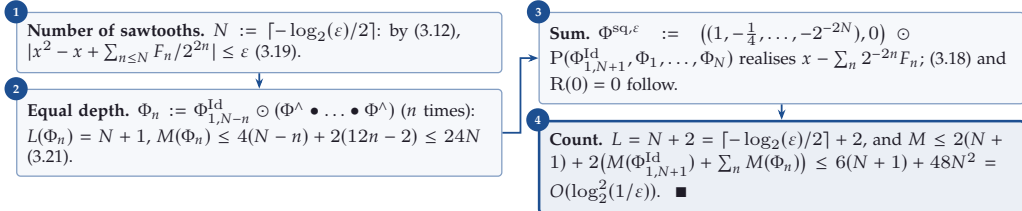


$$\varrho_R(t) - \varrho_R(-t) = t$$

Prop. 3.14: the square network



Proposition 3.14 [40]. For $\varepsilon \in (0, \frac{1}{2})$ there is $\Phi^{\text{sq}, \varepsilon}$ with $L = O(\log_2(1/\varepsilon))$ (3.16), $M = O(\log_2^2(1/\varepsilon))$ (3.17), $|\mathbf{R}(\Phi^{\text{sq}, \varepsilon})(x) - x^2| \leq \varepsilon$ on $[0, 1]$ (3.18), and $\mathbf{R}(\Phi^{\text{sq}, \varepsilon})(0) = 0$.



Remark 3.15: multiplication by bit extraction



A proof of Prop. 3.14 without (3.12) [31], from three ideas:

- **Semi-binary multiplication:**

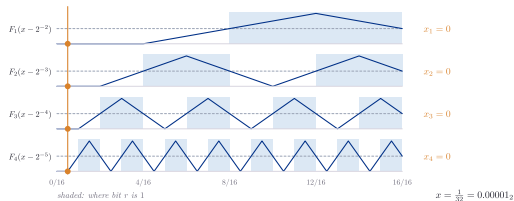
$$x = \sum_{\ell} x_{\ell} 2^{-\ell} \Rightarrow$$
$$xy = \sum_{\ell \leq N} 2^{-\ell} x_{\ell} y + O(2^{-N}).$$

- **Multiplying by a bit is one ReLU:**

$$\varrho_R(2^{-\ell} y + x_{\ell} - 1) = 2^{-\ell} x_{\ell} y \text{ for } y \in [0, 1].$$

- **Bits are cheap:** $F_N(x - 2^{-N-1}) > \frac{1}{2}$ iff $x_N = 1$, then a steep ReLU step.

N bit extractors and N multipliers: $O(N)$ neurons for error 2^{-N} .



Reading the binary digits of x off shifted sawtooths, for eight values of x .

3.3 *Approximation of smooth functions*

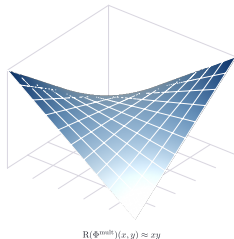
Prop. 3.16: multiplication from squares



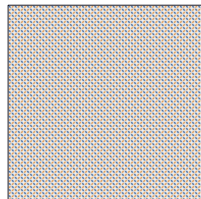
Proposition 3.16

For $p, K \in \mathbb{N}$, $\varepsilon \in (0, \frac{1}{2})$ there is $\Phi^{\text{mult}, p, \varepsilon}$ with $L = O(\log_2 K \cdot \log_2(1/\varepsilon))$, $M = O(\log_2 K \cdot \log_2^2(1/\varepsilon))$ and $|\mathbf{R}(\Phi^{\text{mult}, p, \varepsilon})(x) - \prod_{\ell} x_{\ell}| \leq \varepsilon$ on $[-K, K]^p$; $\mathbf{R}(\cdot)(x) = 0$ if some $x_{\ell} = 0$.

Proof ($p = 2$): $xy = K^2((\frac{|x+y|}{2K})^2 - (\frac{|x-y|}{2K})^2)$,
 $|t| = \varrho_R(t) + \varrho_R(-t)$: two squares $\Phi^{\text{sq}, \varepsilon}/(2K^2)$ in parallel; $p > 2$ a binary tree. ■



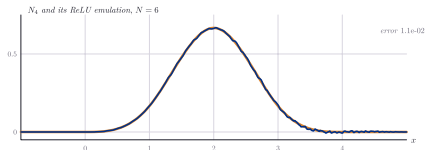
$\mathbf{R}(\Phi^{\text{mult}})(x, y) \approx xy$



error $\mathbf{R}(\Phi) - xy$, $N = 5$
 $\pm 2.4e - 04$ (colour scale)

$N = 0, \dots, 5$ sawtooths per square; the error is a checkerboard of chord errors.

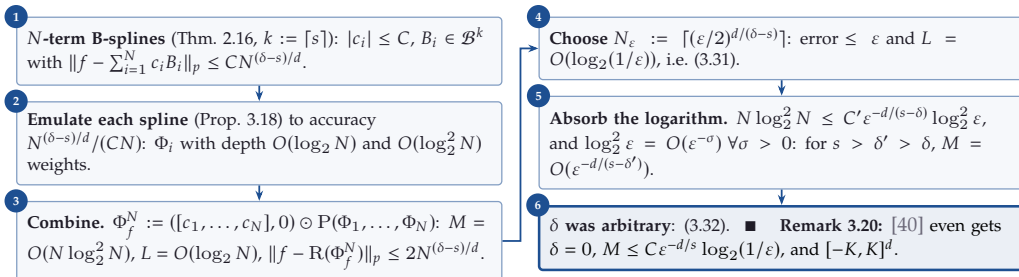
Cor. 3.17: x^p with the same complexity. **Prop. 3.18:** B-splines $N_{\ell,t,k}^d$ with $L = O(\log_2(1/\varepsilon))$, $M = O(\log_2^2(1/\varepsilon))$, from $N_k(x) = \frac{1}{(k-1)!} \sum_{\ell} (-1)^{\ell} \binom{k}{\ell} (x - \ell)_+^{k-1}$, a cut-off and a product.



Theorem 3.19

For $s > \delta > 0$, $p \in (0, \infty]$ there is C : every $f \in C^s([0, 1]^d)$, $\|f\|_{C^s} \leq 1$, has a ReLU network with $L \leq C \log_2(1/\varepsilon)$, $M \leq C \varepsilon^{-d/(s-\delta)}$, $\|f - R(\Phi)\|_{L^p} \leq \varepsilon$.

Proof of Theorem 3.19



4 *The role of depth*

Prop. 4.1: compact support needs depth

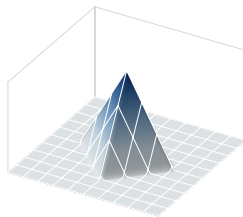


Proposition 4.1

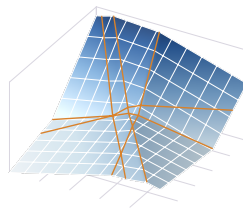
$d \geq 2$, ReLU.

- $L \geq 3$: some Φ has $\text{supp } R(\Phi) = B_{\|\cdot\|_1}(0)$.
- $L \leq 2$: $\text{supp } R(\Phi)$ compact $\Rightarrow R(\Phi) \equiv 0$.

First part: $\sum_{\ell} \varrho_R(x_{\ell}) + \varrho_R(-x_{\ell}) = \|x\|_1$, then $\varrho_R(1 - \|x\|_1)$: three layers.



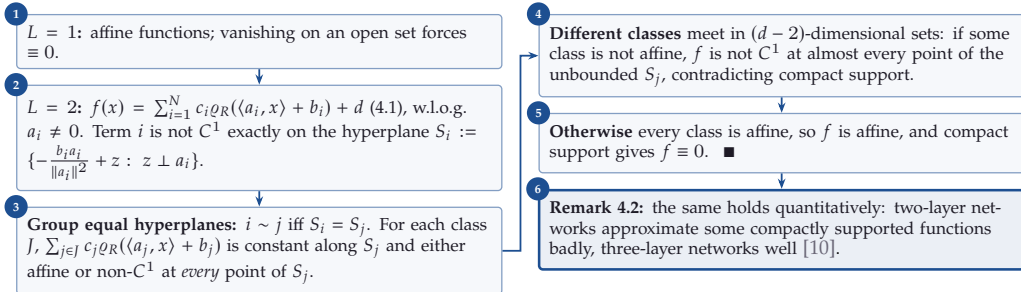
$\varrho_R(1 - \|x\|_1)$: $L = 3$



$L = 2$: kinks are whole lines

Left: $\varrho_R(1 - \|x\|_1)$. Right: a two-layer network; its kinks (orange) are entire lines, so it cannot vanish outside a bounded set unless it is affine.

Proof of Proposition 4.1, second part



4.2 Number of pieces: Theorem 4.3



Theorem 4.3 [37]

Let ϱ be piecewise affine with p pieces. Every Φ with $d = 1$, $N_L = 1$ and $N_1, \dots, N_{L-1} \leq N$ has $R(\Phi)$ with at most $(pN)^{L-1}$ affine pieces.

Corollary 4.4

For $d \in \mathbb{N}$: at most $(pN)^{L-1}$ pieces along every line. (Restrict to $t \mapsto x_0 + tv$: that is $R(\Phi \bullet \Phi_0)$, $\Phi_0 = ((v, x_0))$.)

1

$L = 2$: $\sum_{k=1}^{N_1} c_k \varrho(a_k x + b_k) + d$; sums add pieces: $\leq Np$.



2

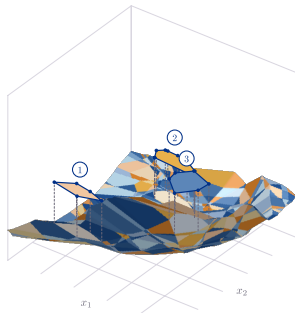
Composition. If f has $\tilde{p}$ pieces, $\varrho \circ f$ has $\leq p\tilde{p}$: on each piece f is constant or injective.



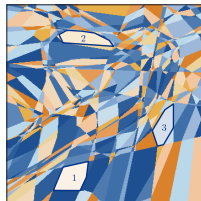
3

$L \rightarrow L + 1$: $R(\Phi_{L+1}) = \sum_{k \leq N_L} (A_{L+1})_k \varrho(h_k) + b_{L+1}$, each h_k an L -layer net: $\leq N \cdot p(pN)^{L-1} = (pN)^L$ pieces. ■

Linear regions of a deep ReLU network



$L = 6$ hidden layers, width 10: 881 regions



every region is a convex polygon,
and $R(\Phi)$ is affine on it:

- 1: 5 edges, 32/60 neurons on, $\nabla R(\Phi) = (3.4, -12.4)$
- 2: 10 edges, 27/60 neurons on, $\nabla R(\Phi) = (0.1, 3.0)$
- 3: 6 edges, 34/60 neurons on, $\nabla R(\Phi) = (17.5, -12.0)$

A random width-10 network $\mathbb{R}^2 \rightarrow \mathbb{R}$, one layer at a time; one colour per activation pattern. Below: $R(\Phi)$ along the black line, the 1D count of Cor. 4.4. Then three regions are lifted out: convex polygons, $R(\Phi)$ affine on each.

4.3 Approximation of non-linear functions



Proposition 4.5 [11]

$f \in C^2([a, b])$ not affine. There is $c = c(f) > 0$: $\|g - f\|_\infty > cp^{-2}$ for every piecewise affine g with $\leq p$ pieces.

Proof (the notes cite it; here in full): $|f''| \geq \kappa > 0$ on some I , $|I| = \lambda$; a piece of g covers $[\alpha, \beta] \subset I$, $h := \beta - \alpha \geq \lambda/p$. With ℓ the chord of $e := f - g$ there, $e - \ell$ vanishes at α, β and $|(e - \ell)''| \geq \kappa$, so $\max |e - \ell| \geq \kappa h^2/8$; $|\ell| \leq \max |e|$ gives $\|f - g\|_\infty \geq \frac{\kappa \lambda^2}{16} p^{-2}$. ■

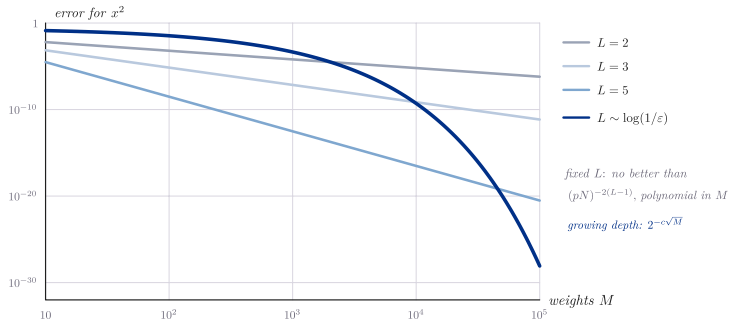
Theorem 4.6

$f \in C^2([0, 1]^d)$ not affine, ϱ piecewise affine with p pieces. Every NN with L layers and $< N$ neurons per layer has $\|f - R(\Phi)\|_\infty \geq c(pN)^{-2(L-1)}$.

Proof. A line $x_0 + tv$ on which f is not affine carries $\leq (pN)^{L-1}$ pieces of $R(\Phi)$ (Cor. 4.4); apply Prop. 4.5. ■

Remark 4.7. Fixed depth gives only polynomial rates; Prop. 3.14 needs growing depth.

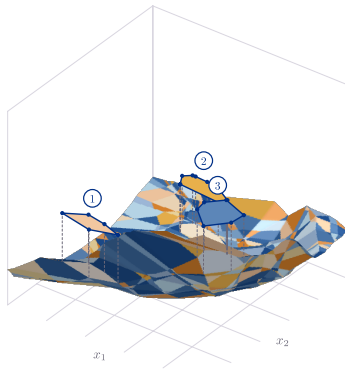
Fixed depth against growing depth



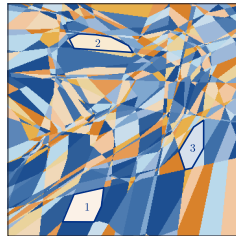
Grey: the lower bound of Thm. 4.6 at fixed L ($M \sim LN^2$), polynomial in M . Navy: Prop. 3.14, $\varepsilon = 2^{-2N-2}$ at $M \approx 48N^2$: eventually below every fixed depth.

- ReLU networks represent linear finite elements exactly (Prop. 3.2, Thm. 3.4): each Courant element is a max-min of planes.
- Compositions of a hat make 2^n pieces with $O(n)$ weights (Prop. 3.6), and x^2 is a sum of sawtooths (Prop. 3.8): $O(\log^2(1/\varepsilon))$ weights (Prop. 3.14).
- Squares give products, powers, B-splines and the rate $\varepsilon^{-d/(s-\delta)}$ for C^s (Thm. 3.19).
- Depth is necessary: compact support (Prop. 4.1), pieces $\leq (pN)^{L-1}$ (Thm. 4.3), only polynomial rates at fixed depth (Thm. 4.6).

Next lecture. How complex are sets of networks? VC dimension, matching lower bounds, and the geometry of network spaces.



$L = 6$ hidden layers, width 10: 881 regions



every region is a convex polygon,
and $R(\Phi)$ is affine on it:

- 1: 5 edges, 32/60 neurons on, $\nabla R(\Phi) = (3.4, -12.4)$
- 2: 10 edges, 27/60 neurons on, $\nabla R(\Phi) = (0.1, 3.0)$
- 3: 6 edges, 34/60 neurons on, $\nabla R(\Phi) = (17.5, -12.0)$

Thank you.
Johannes Herter