

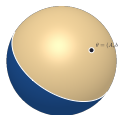
The Geometry of Neural Network Spaces

VC dimension, lower bounds, non-convexity, non-closedness

Johannes Herter

Columbia University · Theory of Deep Neural Networks

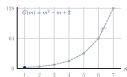
parameters $(A, b) \in \mathbb{S}^2$, one circle per point



each cell: one labeling (left: the cell of θ)



labeling of $\theta: 1$



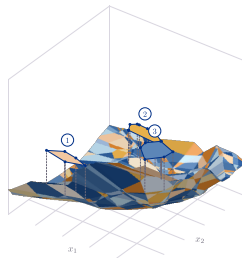
first value of $m = 1 = d + 3$

Last lecture

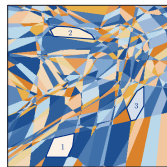
ReLU networks with $L \lesssim \log(1/\varepsilon)$ layers and $M \lesssim \varepsilon^{-d/(s-\delta)}$ weights approximate the unit ball of $C^s([0, 1]^d)$ to accuracy ε (Thm. 3.19).

Today, two questions.

- **Is that optimal?** Measure the richness of a class: VC dimension (Ch. 6).
- **What does $\mathcal{RNN}_\rho(S)$ look like?** Star-shaped, not convex, not closed (Ch. 7).



$L = 6$ 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, $\nabla R(\Phi) = (17.5, -12.0)$

Last lecture: a deep ReLU network cuts its input space into convex pieces and is affine on each.

6.1 *The growth function and the VC dimension*

Definitions (Sect. 6.1)

$H \subset \{h : X \rightarrow \{0, 1\}\}$, $S \subset X$ finite, $H_S := \{h|_S : h \in H\}$. **Growth function** $G_H(m) := \max\{|H_S| : S \subset X, |S| = m\} \leq 2^m$. S is **shattered** if $|H_S| = 2^{|S|}$, and $\text{VCdim}(H) := \max\{m \in \mathbb{N} : G_H(m) = 2^m\}$.

Example 6.1 ($X = \mathbb{R}^2$).

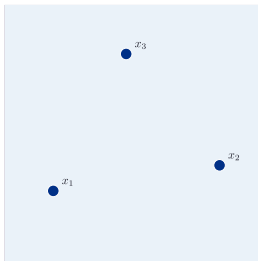
1. $H = \{0, 1\}$: $G_H(m) = 2$, so $\text{VCdim}(H) = 1$.
2. $H = \{0, \chi_\Omega, \chi_{\Omega^c}, 1\}$: two points $x_1 \in \Omega$, $x_2 \in \Omega^c$ are shattered, three never: $\text{VCdim}(H) = 2$.
3. Linear classifiers $h_{\theta, t} = \chi_{\mathbb{R}_+}(\langle (\cos \theta, \sin \theta), \cdot \rangle - t)$: three points in general position are shattered, no four points are: $\text{VCdim}(H) = 3$.

A non-asymptotic notion, on finitely many points; still, it will bound approximation rates from below (Sect. 6.2).

Shattering three points, and the lift to $\mathbb{R}^3$

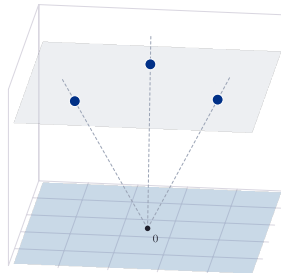


$\mathbf{1}_{\mathbb{R}_+}(Ax + b)$: labelling 111



labelling 8 of $2^3 = 8$

lifted: $(x_i, 1) \in \mathbb{R}^3$



plane $\{(A, b) \cdot z = 0\}$ through 0

All 2^3 labellings by one Heaviside neuron $\mathbf{1}_{\mathbb{R}_+}(Ax + b)$. Right: $\mathbf{1}_{\mathbb{R}_+}(Ax + b) = \mathbf{1}_{\mathbb{R}_+}(\langle(A, b), (x, 1)\rangle)$, so each labelling is a plane through 0 separating the lifted points $(x_i, 1)$; for 000 and 111 it is the floor.

Theorem 6.2: shattering by one neuron



Theorem 6.2 [1, Thm. 3.4]

$\varrho = \mathbf{1}_{\mathbb{R}_+}$, $S_N(d)$ the realisations of $((A_1, b_1), (1, 0))$, $A_1 \in \mathbb{R}^{1 \times d}$. Then $S_N(d)$ shatters $(x_i)_{i=1}^m \subset \mathbb{R}^d$ if and only if

$$(x_1, 1), (x_2, 1), \dots, (x_m, 1) \text{ are linearly independent.} \quad (6.1)$$

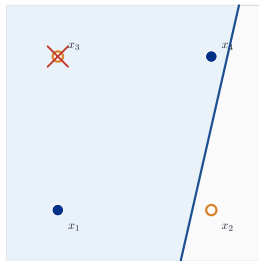
In particular $\text{VCdim}(S_N(d)) = d + 1$.

Reading. $\mathbf{1}_{\mathbb{R}_+}(Ax + b) = \mathbf{1}_{\mathbb{R}_+}(\langle(A, b), (x, 1)\rangle)$: a labelling is a hyperplane through 0 in $\mathbb{R}^{d+1}$ that separates the lifted points. Independent lifts can be sent to any signs by one linear functional; dependent lifts cannot, because the value at one of them is forced by the others.

Four points: XOR



best line at angle θ : 1 error(s)



$$(x_4, 1) = -(x_1, 1) + (x_2, 1) + (x_3, 1)$$

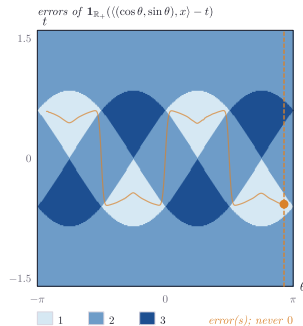
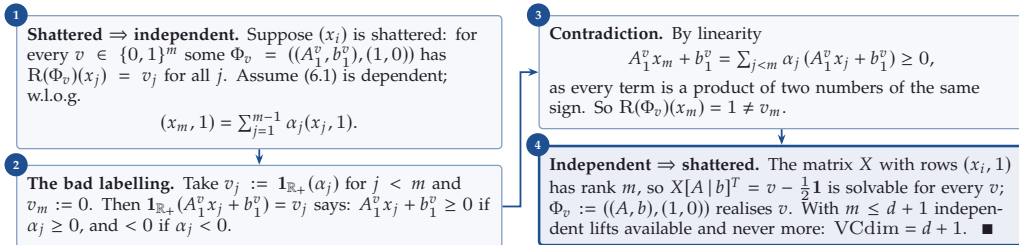


Fig. 6.2. Four points in $\mathbb{R}^2$ lift to four dependent vectors in $\mathbb{R}^3$; the labelling built from the dependence in the proof is XOR. Right: errors of every line (θ, t) ; the minimum over the whole parameter plane is 1.

Proof of Theorem 6.2



The notes take the complementary labelling; with $\mathbf{1}_{\mathbb{R}_+}(0) = 1$ the choice above avoids the tie $\sum = 0$.

Lemma 6.3: one parameter, infinite VC dimension

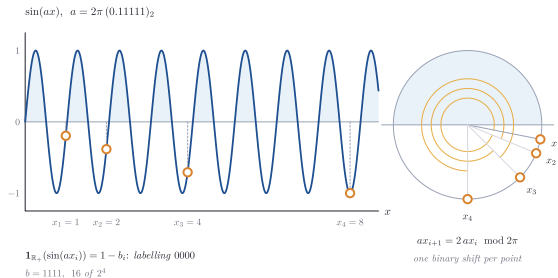


Lemma 6.3 [1, Lemma 7.2]

$H := \{x \mapsto \mathbf{1}_{\mathbb{R}_+}(\sin(ax)) : a \in \mathbb{R}\}$
has $\text{VCdim}(H) = \infty$.

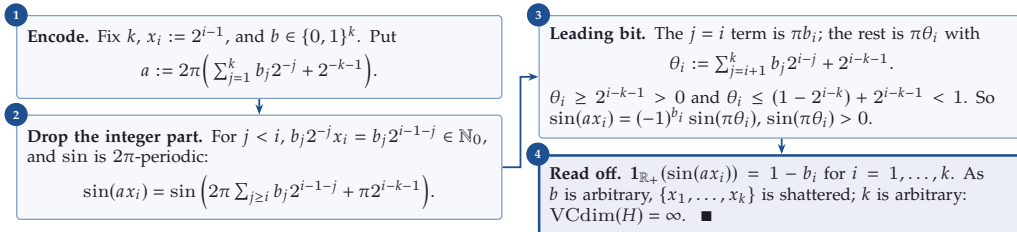
The trick: bit extraction. With $x_i = 2^{i-1}$, multiplying by x_i shifts the binary expansion of $a/2\pi$ by $i-1$ places; the sign of $\sin$ reads off the leading bit.

Moral. Parameter counts alone do not bound the VC dimension; the activation matters.



All 16 labellings of $\{1, 2, 4, 8\}$ with a single frequency a . Right: the angles $ax_i \bmod 2\pi$; each point doubles the previous angle.

Proof of Lemma 6.3



Theorem 6.4: piecewise polynomial networks



Theorem 6.4 [1, Thm. 8.8]

ϱ piecewise polynomial with $\leq \ell$ pieces of degree $\leq p$, $H := \{\mathbf{1}_{\mathbb{R}_+} \circ R(\Phi) : \Phi \in \mathcal{NN}_{d,L,M}\}$. Then for all $L, M \in \mathbb{N}$: $\text{VCdim}(H) \lesssim ML \log_2(M) + ML^2$.

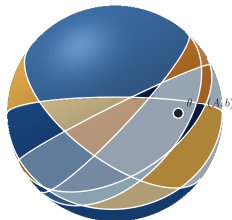
Stated without proof in the notes. The mechanism, in miniature (beyond the notes):

- One neuron on $\mathbb{R}^2$: parameters (A, b) , up to scale a point of S^2 . Each x_i cuts S^2 along the great circle $\langle (A, b), (x_i, 1) \rangle = 0$, and a labelling of $x_1, \dots, x_m$ is a **cell** of these circles.
- There are only $m^2 - m + 2$ cells: polynomially many, while 2^m labellings exist.
- For networks: on a fixed activation pattern $R(\Phi)(x_i)$ is a polynomial in the weights. Counting the sign patterns of polynomials (Warren's bound), pattern by pattern and layer by layer, gives Thm. 6.4.

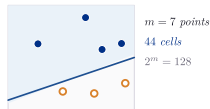
Labellings are cells



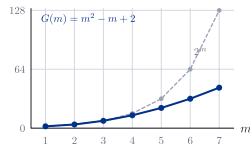
parameters $(A, b) \in S^2$, one circle per point



each cell: one labelling (light: the cell of θ)



labelling of θ : 1010110



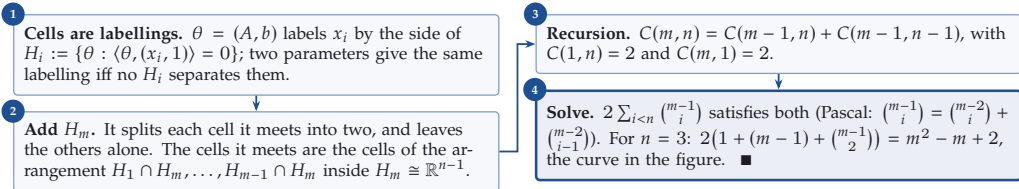
first miss at $m = 4 = d + 2$

Points are added one by one; every new circle splits only the cells it crosses. The growth function $m^2 - m + 2$ falls below 2^m at $m = 4 = d + 2$, one past $\text{VCdim} = d + 1$.

Counting cells: the growth function of a neuron



Claim (Cover, beyond the notes). m hyperplanes through 0 in $\mathbb{R}^n$ in general position cut $\mathbb{R}^n$ into $C(m, n) = 2 \sum_{i=0}^{n-1} \binom{m-1}{i}$ cells; hence for one Heaviside neuron on $\mathbb{R}^d$ ($n = d + 1$), $G_{S_N(d)}(m) = C(m, d + 1) \leq 2(m - 1)^d + 2$.



6.2 *Lower bounds on approximation rates*

Lemma 6.5: approximation forces shattering



Lemma 6.5

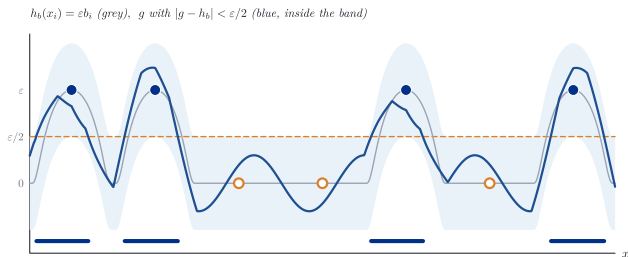
$K \subset \mathbb{R}^d$, $\varepsilon > 0$, $H \subset \{h : K \rightarrow \mathbb{R}\}$ with points $x_1, \dots, x_k$ such that every $b \in \{0, 1\}^k$ has an $h \in H$ with $h(x_i) = \varepsilon b_i$ (6.2). If every $h \in H$ has a $g \in G$ with $\sup_K |g - h| < \varepsilon/2$ (6.3), then

$$\text{VCdim}(\{\mathbf{1}_{\mathbb{R}_+} \circ (g - \varepsilon/2) : g \in G\}) \geq k. \quad (6.4)$$

Proof. For $b \in \{0, 1\}^k$ pick h_b by (6.2) and g_b by (6.3). Then $|g_b(x_i) - \varepsilon b_i| < \varepsilon/2$, so $g_b(x_i) - \varepsilon/2 > 0$ if $b_i = 1$ and < 0 if $b_i = 0$: $\mathbf{1}_{\mathbb{R}_+}(g_b - \varepsilon/2)(x_i) = b_i$, and $\{x_1, \dots, x_k\}$ is shattered. ■

Remark 6.6. If H satisfies (6.2) and networks with M weights, L layers approximate all of H to $\varepsilon/2$, Thm. 6.4 forces $ML \log_2 M + ML^2 \gtrsim k$. Next: a large k inside the C^n unit ball.

Thresholding recovers the labels



$b = 1100101$; bars: $\{\mathbf{1}_{\mathbb{R}_+}(g - \varepsilon/2) = 1\}$, recovered 1100101

Lemma 6.5 for six labellings b : h_b takes the values εb_i ; any g inside the band of width $\varepsilon/2$ around h_b , thresholded at $\varepsilon/2$, returns b exactly (navy bars: where $g > \varepsilon/2$).

Lemmas 6.7 and 6.8: many disjoint bumps

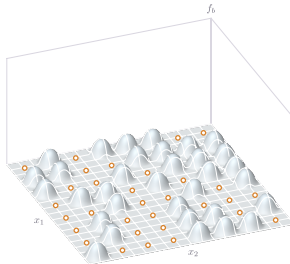


Lemma 6.7

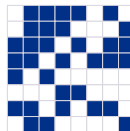
There is $C > 0$: for every $\varepsilon \in (0, 1]$ some $f \in C^n(\mathbb{R}^d)$ has $\text{supp } f \subset [-C\varepsilon^{1/n}, C\varepsilon^{1/n}]^d$ (6.5), $f(0) = \varepsilon$ and $\|f\|_{C^n} \leq 1$.

Lemma 6.8

There is $C > 0$: for every ε there are $x_1, \dots, x_k$, $k \geq C\varepsilon^{-d/n}$, and for every $b \in \{0, 1\}^k$ an $f_b \in C^n([0, 1]^d)$, $\|f_b\|_{C^n} \leq 1$, with $f_b(x_i) = \varepsilon b_i$.



$b \in \{0, 1\}^{64}$



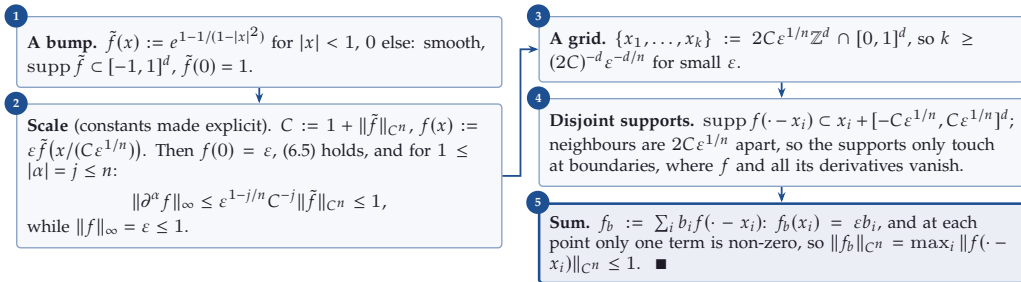
$k = 8^2$ points, spacing $1/8$

height $\varepsilon \sim (\text{spacing})^n$

$\Rightarrow k \gtrsim \varepsilon^{-d/n}$

$f_b = \sum_i b_i f(\cdot - x_i)$ on $[0, 1]^2$ (Fig. 6.3) for grids of spacing $1/3, \dots, 1/8$: smaller spacing allows only lower bumps, $\varepsilon \sim (\text{spacing})^n$.

Proofs of Lemmas 6.7 and 6.8



Theorem 6.9: a lower bound on weights



Theorem 6.9

ϱ piecewise polynomial. If for every $\varepsilon > 0$ networks in $\mathcal{NN}_{d,L(\varepsilon),M(\varepsilon)}$ reach $\sup_{\|f\|_{C^n} \leq 1} \inf_{\Phi} \|f - R(\Phi)\|_{\infty} \leq \varepsilon/2$, then

$$(M+1)L \log_2(M+1) + (M+1)L^2 \gtrsim \varepsilon^{-d/n}.$$

1

Rich class. $H := \{h \in C^n : \|h\|_{C^n} \leq 1\}$ satisfies (6.2) with $k \geq C\varepsilon^{-d/n}$ (Lemma 6.8).

2

Shatter. $G := \{R(\Phi) : \Phi \in \mathcal{NN}_{d,L,M}\}$ satisfies (6.3); Lemma 6.5: $\text{VCdim}(\{\mathbf{1}_{\mathbb{R}_+} \circ (g - \varepsilon/2)\}) \geq k$ (6.6).

3

Count. $g - \varepsilon/2$ is a network with $\leq M+1$ weights (one more bias). Thm. 6.4 bounds its VC dimension by the left side. ■

Remark 6.10. With $L \sim \log(1/\varepsilon)$: $M \gtrsim \varepsilon^{-d/n} / \log^2(1/\varepsilon)$. Thm. 3.19 is optimal up to logs and δ .

Theorem 6.11: depth can double the rate



With any depth,

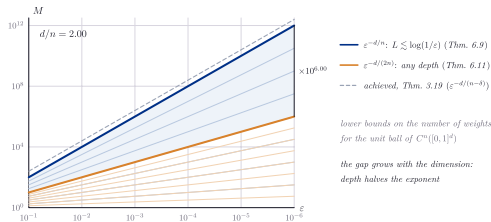
$\tilde{H} := \{1_{\mathbb{R}_+} \circ R(\Phi) : \Phi \in \bigcup_{\ell} \mathcal{NN}_{d,\ell,M}\}$ still has $\text{VCdim}(\tilde{H}) \lesssim M^2$ [1, Thm. 8.7] (6.8).

Theorem 6.11

ϱ piecewise polynomial. If networks with $M(\varepsilon)$ weights and any number of layers reach $\varepsilon/2$ uniformly on the unit ball of $C^n([0, 1]^d)$, then $M(\varepsilon) \gtrsim \varepsilon^{-d/(2n)}$.

Proof. As for Thm. 6.9, with (6.8) in place of Thm. 6.4: $(M + 1)^2 \gtrsim \varepsilon^{-d/n}$. ■

Remark 6.12. Sharp for ReLU [41].



Logarithmic depth cannot beat the navy line; unbounded depth reaches the orange one.

7

Spaces of realisations of neural networks

Definition 7.1

$S = (N_0, \dots, N_L)$ is the **architecture** of Φ if $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$. $\mathcal{NN}(S)$: all such networks, a vector space of dimension $\sum_\ell (N_{\ell-1} + 1)N_\ell$ with $\|\Phi\|_{\text{total}} := \max_\ell \|A_\ell\|_\infty + \max_\ell \|b_\ell\|_\infty$;

$$\mathcal{RNN}_\varrho(S) := \{\mathcal{R}(\Phi) : \Phi \in \mathcal{NN}(S)\} \subset L^p(K) \quad \text{for compact } K, \varrho \text{ continuous.}$$

Theorem 7.2 [23, Prop. 4]

Ω compact, ϱ continuous: $\mathcal{R} : \mathcal{NN}(S) \rightarrow L^\infty(\Omega)$ is continuous; locally Lipschitz if ϱ is.

Why (one layer at a time). On bounded sets of weights all hidden values stay in a compact set $[-B, B]$, where ϱ is uniformly continuous (Lipschitz with constant Lip_B). Changing (A_ℓ, b_ℓ) by δ moves $A_\ell x_{\ell-1} + b_\ell$ by $\lesssim \delta$; ϱ and the later layers propagate this with bounded factors.

7.1 *Network spaces are not convex*

Star-shaped, with 0 as a centre



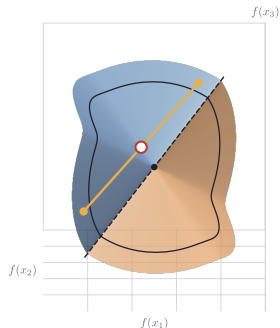
Definition 7.3

$x \in Z$ is a **centre** if $[x, y] \subset Z$ for all $y \in Z$; Z is **star-shaped** if it has one.

Proposition 7.4

$f \in \mathcal{RNN}_\rho(S)$, $\lambda \in \mathbb{R} \Rightarrow \lambda f \in \mathcal{RNN}_\rho(S)$ (scale A_L, b_L): 0 is a centre.

Constants are centres too.



$$\{(f(x_1), f(x_2), f(x_3)) : f = c \tanh(2(\cdot - t))\}$$

■ rays $cu(t)$, $c > 0$, shaded by t

■ the same for $c < 0$

— $u(t)$ on the unit sphere

--- constants $c1$: limits, not in the set

— a chord whose midpoint leaves the set

star-shaped with centre 0 (Prop. 7.4)

not convex (Thm. 7.5), not closed

$$x = (-1.2, -0.1, 1.4)$$

$c \tanh(2(x - t))$ at three points: rays through 0 stay in the set; chords leave it.

Theorem 7.5: few centres



Theorem 7.5 [23, Prop. C.4]

ϱ Lipschitz. Then $\mathcal{RNN}_\varrho(S)$ contains at most $M^* := \sum_{\ell=1}^L (N_{\ell-1} + 1)N_\ell$ linearly independent centres.

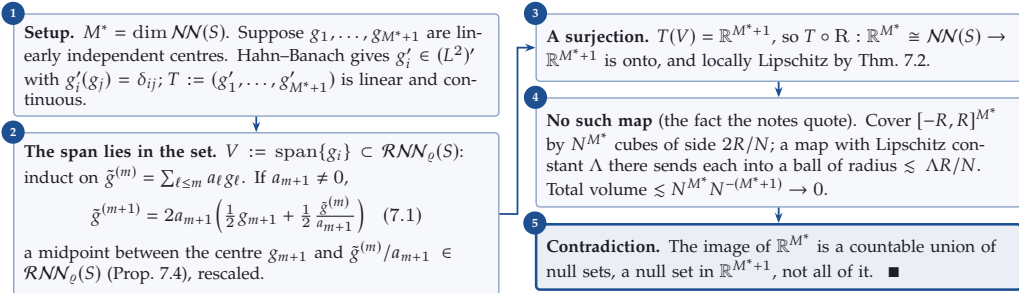
Corollary 7.6

If $\mathcal{RNN}_\varrho(S)$ contains more than M^* linearly independent functions, it is **not convex**: in a convex set every point is a centre.

Remark 7.7. The only Lipschitz ϱ for which $\mathcal{RNN}_\varrho(S)$ has $\leq M^*$ independent elements are affine [23]; for locally Lipschitz ϱ , all non-polynomials qualify.

Idea. Centres span a linear space inside $\mathcal{RNN}_\varrho(S)$; a space of dimension $M^* + 1$ cannot be the Lipschitz image of $\mathbb{R}^{M^*}$.

Proof of Theorem 7.5



Very non-convex: Proposition 7.9



Definition 7.8

X in a metric space is **r -convex** if $\bigcup_{x \in X} B_r(x)$ is convex.

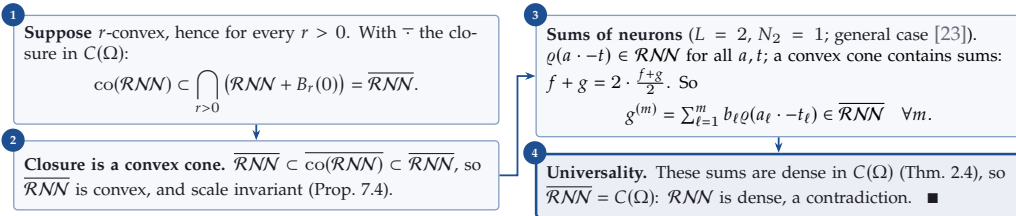
By Prop. 7.4, $\mathcal{RNN}_\varrho(S) + B_r(0) = \frac{r}{r'} (\mathcal{RNN}_\varrho(S) + B_{r'}(0))$: r -convex for one r iff for all r . No thickening ever fills the holes between the rays.

Proposition 7.9 [23, Thm. 2.2]

Ω compact, $\varrho \in C^1$ discriminatory, $\mathcal{RNN}_\varrho(S)$ not dense in $C(\Omega)$. Then $\mathcal{RNN}_\varrho(S)$ is r -convex for no $r > 0$.

Reading. If a fixed architecture is not already universal, it is far from convex at every scale: averaging two networks of architecture S typically leaves $\mathcal{RNN}_\varrho(S)$ by a fixed relative distance.

Proof of Proposition 7.9



7.2 *Network spaces are not closed*

Theorem 7.10: limits that are not networks

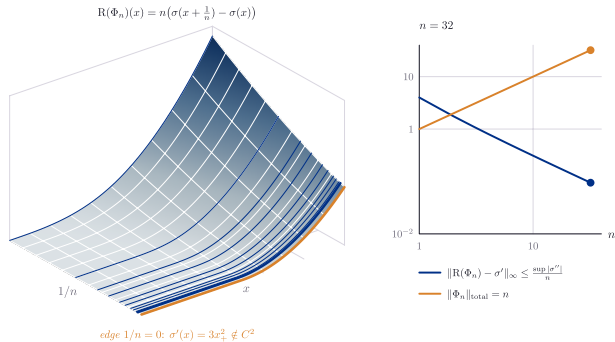


Theorem 7.10

$S = (N_0, N_1, \dots, N_{L-1}, 1)$ with $N_1 \geq 2$, $\Omega \subset \mathbb{R}^d$ compact with non-empty interior, $\varrho \in C^2 \setminus C^\infty$.
Then $\mathcal{RNN}_\varrho(S)$ is **not closed** in L^∞ .

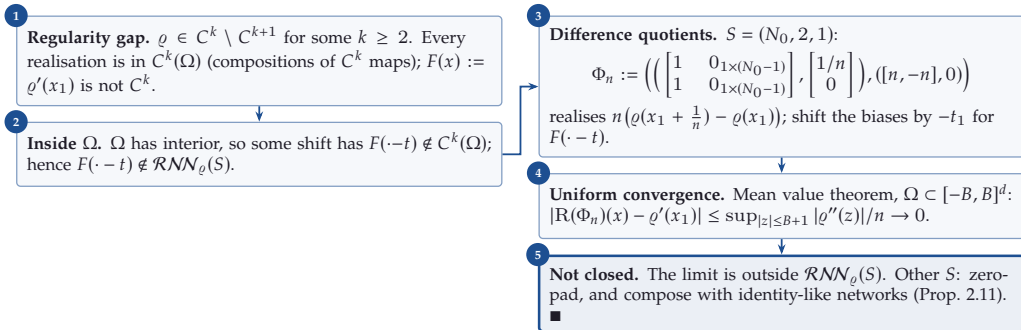
Idea. Two neurons form a difference quotient, $n(\varrho(x + \frac{1}{n}) - \varrho(x)) \rightarrow \varrho'(x)$. Networks are as smooth as ϱ , and ϱ' is one derivative rougher, so the limit is not a network.

Converging quotients, exploding weights



$\sigma(x) = x_+^3 \in C^2 \setminus C^3$. Each slice $1/n$ of the surface is a two-neuron network $n(\sigma(x + \frac{1}{n}) - \sigma(x))$; the front edge $\sigma'(x) = 3x_+^2$ is not. Right: the error falls like $1/n$ while $||\Phi_n||_{\text{total}} = n$.

Proof of Theorem 7.10



Bounded weights are closed: Proposition 7.12



Proposition 7.12

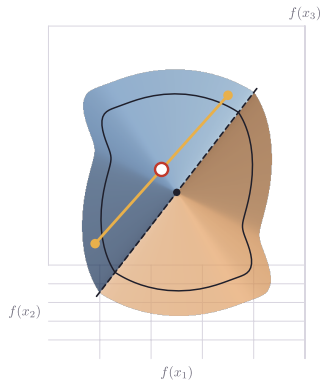
Ω compact, ϱ continuous, $C > 0$. Then $\mathcal{RNN}_C := \{R(\Phi) : \Phi \in \mathcal{NN}(S), \|\Phi\|_{\text{total}} \leq C\}$ is closed in $C(\Omega)$.

Proof. $\{\Phi : \|\Phi\|_{\text{total}} \leq C\}$ is compact (Heine–Borel), R is continuous (Thm. 7.2), and continuous images of compact sets are compact, hence closed. ■ **Remark 7.11.** Thm. 7.10 holds for many more activations [23], but *not* for the ReLU [23, Thm. 3.8].

Consequence: exploding weights. Let $g \in \overline{\mathcal{RNN}_\varrho(S)} \setminus \mathcal{RNN}_\varrho(S)$ and $R(\Phi_n) \rightarrow g$. If $\|\Phi_n\|_{\text{total}} \leq C$ along a subsequence, then $g \in \overline{\mathcal{RNN}_C} = \mathcal{RNN}_C \subset \mathcal{RNN}_\varrho(S)$, which is false. So $\|\Phi_n\|_{\text{total}} \rightarrow \infty$.

For training: minimising over $\mathcal{NN}(S)$ can send the weights to infinity; the infimum need not be attained.

- One Heaviside neuron shatters (x_i) iff the lifts $(x_i, 1)$ are independent: $\text{VCdim} = d + 1$ (Thm. 6.2); a single frequency of $\sin$ has $\text{VCdim} = \infty$ (Lemma 6.3).
- Piecewise polynomial networks: $\text{VCdim} \lesssim ML \log M + ML^2$ (Thm. 6.4), a count of cells.
- Disjoint bumps + thresholding: ε -approximation of the C^n ball needs $ML \log M + ML^2 \gtrsim \varepsilon^{-d/n}$ (Thm. 6.9); with any depth $M \gtrsim \varepsilon^{-d/(2n)}$ (Thm. 6.11).
- $\mathcal{RNN}_\varrho(S)$ is star-shaped (Prop. 7.4) with at most $\dim \mathcal{NN}(S)$ independent centres (Thm. 7.5), so not convex, and not r -convex at any scale (Prop. 7.9).
- For $\varrho \in C^2 \setminus C^\infty$ it is not closed (Thm. 7.10); the approximating weights must explode (Prop. 7.12).



$$\{(f(x_1), f(x_2), f(x_3)) : f = c \tanh(2(\cdot - t))\}$$

■ rays $c u(t)$, $c > 0$, shaded by t

■ the same for $c < 0$

— $u(t)$ on the unit sphere

--- constants c_1 : limits, not in the set

— a chord whose midpoint leaves the set

star-shaped with centre 0 (Prop. 7.4)

not convex (Thm. 7.5), not closed

$$x = (-1.2, -0.1, 1.4)$$

Thank you.

Johannes Herter