ReLU Networks and Discrete Geometry

Ben Smith

28th May 2025

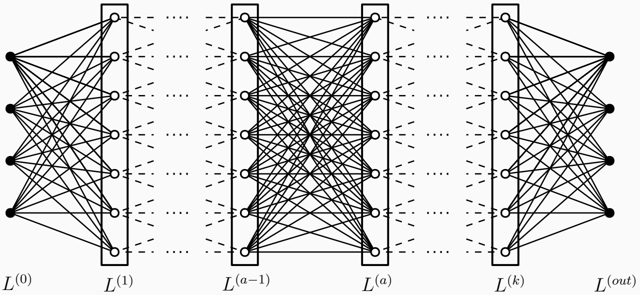
ReLU networks

ReLU Networks

Consider feedforward network with k hidden layers:

- Layer $L^{(a)}$ has n_a nodes,
- Layers connected by affine function $\rho^{(a)} \colon \mathbb{R}^{n_{a-1}} \to \mathbb{R}^{n_a}$,
- ReLU activation function $\sigma^{(a)} \colon \mathbb{R}^{n_a} \to \mathbb{R}^{n_a}$ sends $x \mapsto \max(x, 0)$.

$f = \rho^{(out)} \circ \sigma^{(n_k)} \circ \rho^{(n_k)} \circ \cdots \circ \sigma^{(n_1)} \circ \rho^{(n_1)}$ is a piecewise linear (PL) function.

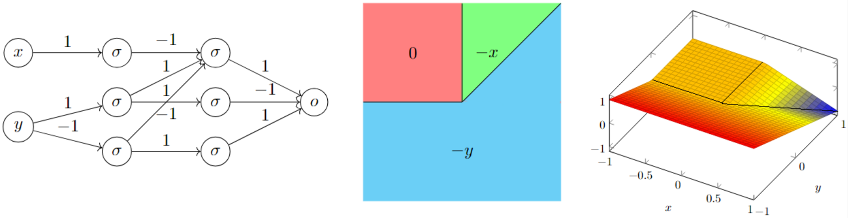


Functions from ReLU Networks

$$\mathcal{F}_d(n_1, \ldots, n_k) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \mid f \text{ representable by network with } n_a \text{ nodes in layer } L^{(a)} \right\}$$

Example

$f \in \mathcal{F}_2(3,3)$, $\quad f(x,y) = \max(-y, \min(0, -x))$



Question

Given some fixed architecture, what can we deduce about the resulting function?

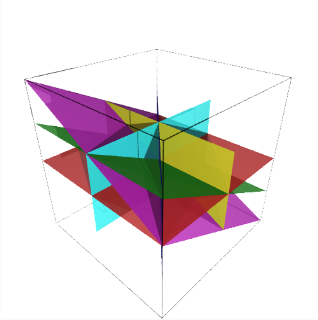
1. Complexity of a function,
2. Quantify decision boundaries in binary classification,
3. Bound depth required to represent a function.

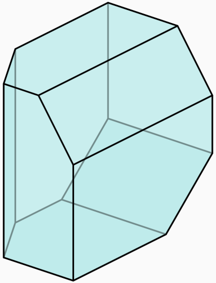
Discrete Geometry

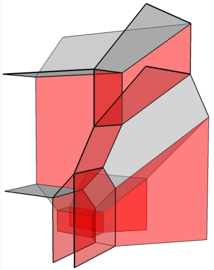
Discrete geometry $\approx$ 'combinatorial properties of geometric objects'

$\approx$ 'geometric properties of combinatorial objects'

E.g. Hyperplane arrangements, polyhedral geometry, tropical geometry, etc.





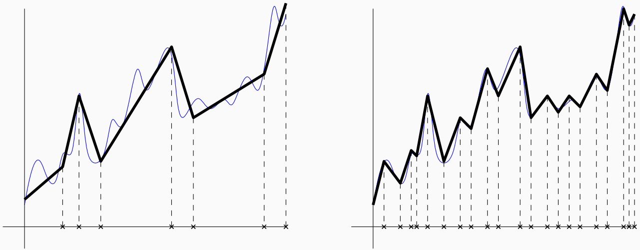


Geometric Complexity

Geometric complexity

Definition
Given PL function $f : \mathbb{R}^d \to \mathbb{R}$, the **geometric complexity** of f is

$$N(f) = \#\text{ regions of } \mathbb{R}^d \text{ on which } f \text{ linear} \quad .$$



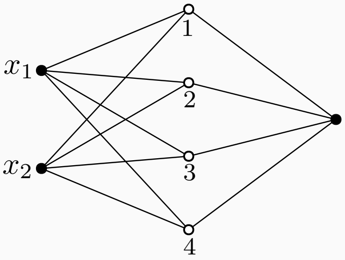
Fix some architecture $\mathcal{F} = \mathcal{F}_d(n_1, \ldots, n_k)$.

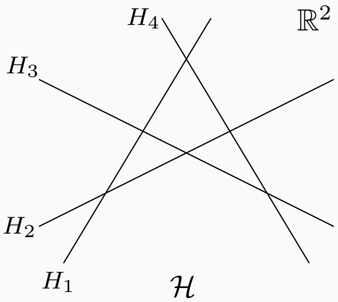
1. What is $\max_{f \in \mathcal{F}}(N(f))$, the maximum geometric complexity of a function in $\mathcal{F}$?
2. What is the ‘expected’ geometric complexity of a function in $\mathcal{F}$?

Geometric complexity of shallow networks

Consider $f \in \mathcal{F}_d(n)$, i.e. one hidden layer with n nodes.

- Function at node i is $x \mapsto \max(\langle a_i, x\rangle + b_i, 0)$,
- Nonlinear only on hyperplane $H_i = \{x \in \mathbb{R}^d \mid \langle a_i, x\rangle + b_i = 0\}$,
- $N(f)$ is the number of regions in hyperplane arrangement $\mathcal{H} = \{H_i\}_{i=1}^n$.



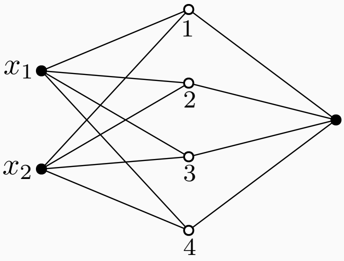


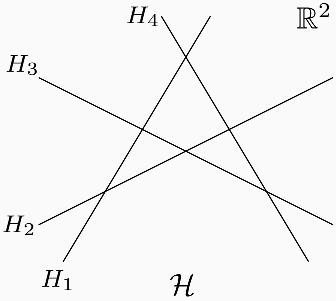
Geometric complexity of shallow networks

Theorem (Zaslavsky '75)
There exists a polynomial $\chi_{\mathcal{H}}(t)$ such that the number of regions of $\mathcal{H}$ is $|\chi_{\mathcal{H}}(-1)|$.

Corollary (Pascanu et. al. '13)

$$\max_{f \in \mathcal{F}_d(n)} (N(f)) = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d-1} + \binom{n}{d} \approx O(n^d)\,.$$

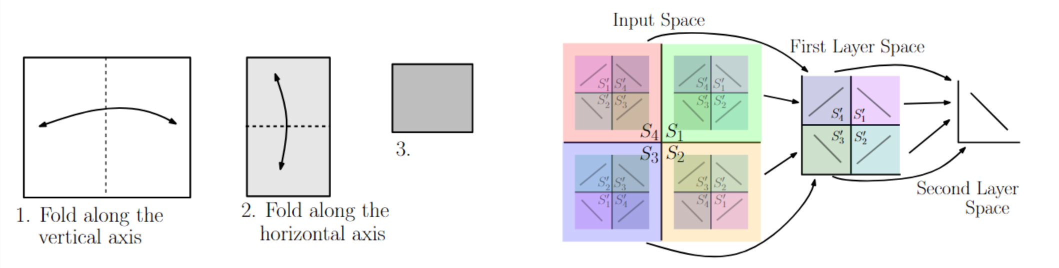




Geometric complexity of deep networks

Theorem (Montafur et. al. '14)
Consider $\mathcal{F} = \mathcal{F}_d(n_1, \ldots, n_k)$ *where* $n_i \geq n \geq d$*. Then*

$$\max_{f \in \mathcal{F}}(N(f)) \geq \left(\prod_{i=1}^{k-1} \left\lfloor \frac{n_i}{d} \right\rfloor^d \right) \sum_{j=0}^{d} \binom{d}{j} \approx O(n^{kd}).$$

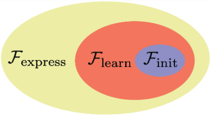


Theorem (Raghu et. al. '17)
This bound is asymptotically tight: $\max_{f \in \mathcal{F}}(N(f)) \leq O(n^{kd})$.

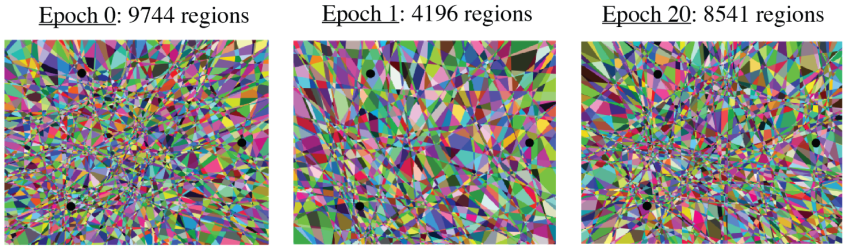
Expected complexity of deep networks

- Maximum complexity $\neq$ expected complexity
- (Hanin, Rolnick '19) At initialization, the complexity of $f \in \mathcal{F}_d(n_1, \ldots, n_k)$ is bounded above by

$$N(f) \leq (C \cdot \sum_{i=1}^{k} n_i)^d \,, \quad C \text{ constant} \;.$$



- Empirically, function stays closer to this bound during training rather than maximum.



Decision boundaries

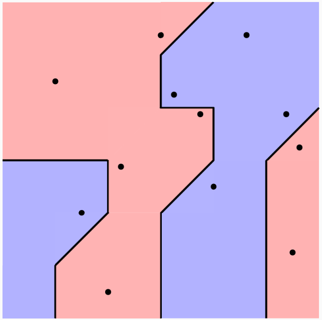
Decision boundaries in binary classification

$f : \mathbb{R}^d \to \mathbb{R}$ PL function, c decision threshold.

- If $f(x) < c$, then $x \in \mathit{RED}$,
- If $f(x) > c$, then $x \in \mathit{BLUE}$.

Question
Can we describe the decision boundary
$\mathcal{B}_f = \{x \in \mathbb{R}^d \mid f(x) = c\}$?

We'll attack this question with **tropical geometry**.



What is Tropical Geometry?

Geometry over the **tropical semiring** $\mathbb{T} = (\mathbb{R}, \oplus, \odot)$ with operations

$$a \oplus b = \max(a, b)\,, \quad a \odot b = a + b\,.$$

A **tropical polynomial** $g \in \mathbb{T}[X_1, \dots, X_d]$ is

$$\begin{aligned} g \colon \mathbb{R}^d &\to \mathbb{R} \\ x &\mapsto \bigoplus_{a \in \mathbb{N}^d} b_a \odot x_1^{\odot a_1} \odot \cdots \odot x_d^{\odot a_d} = \max_{a \in \mathbb{N}^d}(\langle a, x\rangle + b_a)\,. \end{aligned}$$

Tropical polynomials $\leftrightarrow$ Maximum of linear functions $\leftrightarrow$ **Convex** PL functions

The **tropical hypersurface** associated to $g : \mathbb{R}^d \to \mathbb{R}$ is

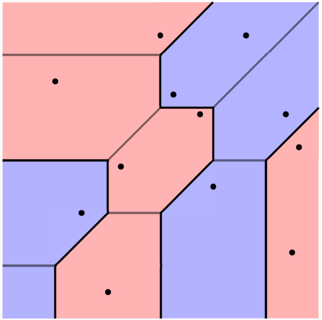
$$\mathcal{T}(g) = \{x \in \mathbb{R}^d \mid g \text{ non-linear at } x\}\,.$$

Decision boundaries as tropical hypersurfaces

Theorem (Zhang et. al. '18)
Let $f : \mathbb{R}^d \to \mathbb{R}$ PL function and c decision threshold. The decision boundary $\mathcal{B}_f$ is contained in a tropical hypersurface.

- (Melzer '86, Kripfganz, Schulze '87) Every PL function f is the difference $f = g - h$ of convex PL functions.
- Network structure gives concrete construction of one way to do this.
- $\mathcal{B}_f \subseteq \mathcal{T}(\tilde{f})$ where

$$\tilde{f} = g \oplus c \odot h = \max(g, h + c)\,.$$



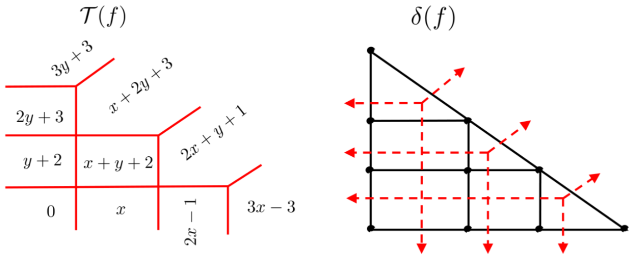
Benefits

Theorem (Zhang et. al. '18)
Let $f : \mathbb{R}^d \to \mathbb{R}$ PL function and c decision threshold. Then $\mathcal{B}_f \subseteq \mathcal{T}(\tilde{f})$ where

$$\tilde{f} = g \oplus c \odot h = \max(g, h + c)\,, \quad f = g - h \text{ where } g, h \text{ convex}\,.$$

Benefits:

- Tropical hypersurfaces are highly structured,
- Opens up host of new algebraic and polyhedral tools,
- Convex things easier to work with (see dc-optimization)



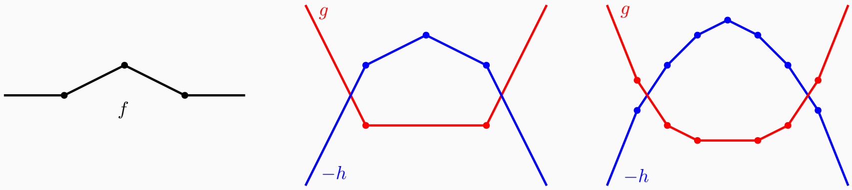
Drawbacks

Theorem (Zhang et. al. '18)
Let $f : \mathbb{R}^d \to \mathbb{R}$ PL function and c decision threshold. Then $\mathcal{B}_f \subseteq \mathcal{T}(\tilde{f})$ where

$$\tilde{f} = g \oplus c \odot h = \max(g, h + c)\,, \quad f = g - h \text{ where } g, h \text{ convex}\,.$$

Drawback: $\mathcal{T}(\tilde{f})$ much more complicated than $\mathcal{B}_f$ in general:

- We do not know how/if you can decompose $f = g - h$ 'efficiently',
- $\mathcal{B}_f$ 'tracks real solutions' while $\mathcal{T}(\tilde{f})$ 'tracks complex solutions'.



Depth bounds on PL functions

Depth bounds on PL functions

$$\mathcal{PL}_d = \{f\colon \mathbb{R}^d \to \mathbb{R} \mid f \text{ piecewise linear } \}$$

$$\mathcal{F}_d(k) = \{f\colon \mathbb{R}^d \to \mathbb{R} \mid f \text{ representable with } k \text{ hidden layers } \} = \bigcup_{n_i \in \mathbb{N}} \mathcal{F}_d(n_1, \ldots, n_k)$$

These classes of functions are related by

$$\mathcal{F}_d(1) \subseteq \mathcal{F}_d(2) \subseteq \cdots \subseteq \mathcal{F}_d(k) \subseteq \cdots \subseteq \mathcal{PL}_d\,.$$

Question
How strict are these containments? What depth do we need to represent all PL functions?

$\mathbf{MAX}_d$

Define $\mathrm{MAX}_d \colon \mathbb{R}^d \to \mathbb{R}$ PL function takes the maximum of d inputs.

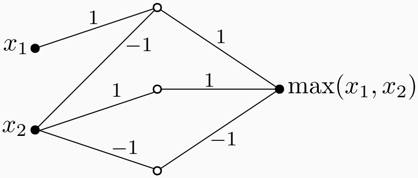
Theorem (Wang, Sun '05)
For each $f \in \mathcal{PL}_d$, there exists affine $A_1, \ldots, A_s \colon \mathbb{R}^d \to \mathbb{R}^{d+1}$ and $\sigma_1, \ldots, \sigma_s \in \{\pm 1\}$ such that

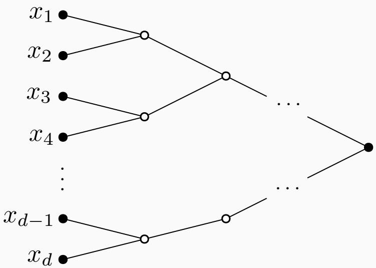
$$f(x) = \sum_{i=1}^{s} \sigma_i \cdot MAX_{d+1}(A_i(x))\,.$$

Corollary
If $MAX_{d+1} \in \mathcal{F}_{d+1}(k)$, then $\mathcal{PL}_d = \mathcal{F}_d(k)$.

Depth bounds on MAX_d

Theorem (Hertrich et. al. '21)
MAX_d can be represented in $\lceil \log_2(d) \rceil$ layers.



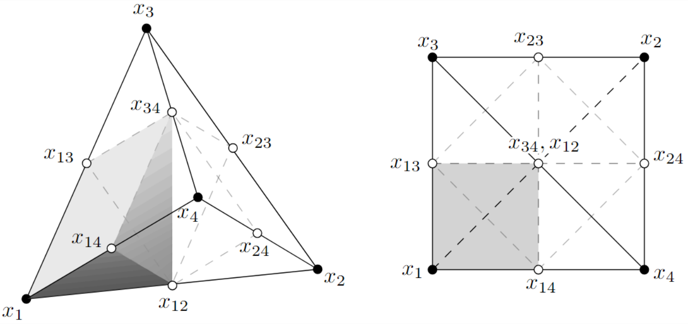


- Conjectured that this bound is sharp.
- Proved when only allowed integer weights...
- ...but false!

Improved depth bounds on MAX_d

Theorem (Hertrich et. al. '25)
MAX_d can be represented in $\lceil \log_3(d-2) \rceil + 1$ layers.

- Show MAX_5 can be represented in 2 layers,
- Proof strategy uses polyhedral algebra of neural networks.



Corollary
Every PL function $f \in \mathcal{PL}_d$ can be represented in $\lceil \log_3(d-1) \rceil + 1$ layers.

Conclusion

Discrete geometry is a powerful tool for analysing piecewise linear functions and ReLU networks:

- Can quantify the complexity of a given functions,
- Gives tools for describing decision boundaries in binary classification.
- Bounds depth required to represent a function.

Thank you for listening!