Constraining the outputs of ReLU neural networks

Yulia Alexandr (UCLA)
joint work with Guido Montúfar

AMS Sectional Meeting
California Polytechnic, San Luis Obispo, CA
May 3, 2025

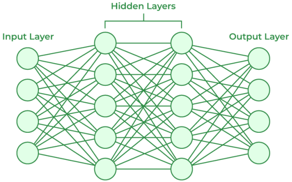
Neural networks

Any feedforward neural network with an activation function σ gives rise to

$$f_\theta : x \mapsto g_L \circ \sigma \circ g_{L-1} \dots \sigma \circ g_1(x)$$

where each layer has linear map $g_\ell : y \mapsto W_\ell y$ with parameter $\theta_\ell = W_\ell$.

The dimension of the input space n_0 and the layer widths n_ℓ determine the network's architecture.



For a dataset $X = [x_1, x_2, \dots, x_m]$ and unknown parameters θ we are interested in describing the constraints between the coordinates of the array of model outputs $F_X(\theta) = [f_\theta(x_1), f_\theta(x_2), \dots, f_\theta(x_m)]$.

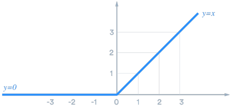
ReLU networks

A *ReLU network* is given by the activation function

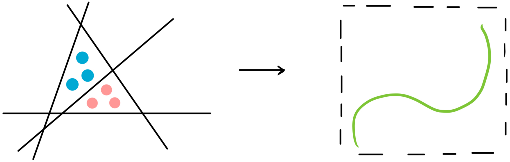
$$\sigma : y = (y_1, \ldots, y_{n_\ell}) \mapsto (\max\{0, y_1\}, \ldots, \max\{0, y_{n_\ell}\})$$

at each layer of the neural network.

- this makes $f_\theta(x)$ piece-wise linear
  - natural subdivision of the input space into regions
  - $f_\theta(x)$ is a linear function of x in each region



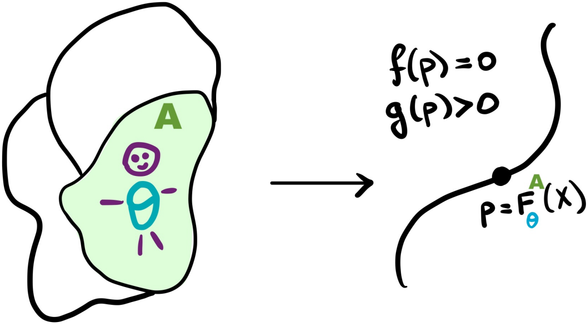
- now consider multiple data points $X = [x_1, \ldots, x_m]$
  - this subdivision extends to the parameter space
  - $F_X(\theta)$ is multi-linear in θ in each activation region



In general...

Problem

Describe the equations and inequalities that define the image of $F_X(\theta)$ as the parameter θ varies over an arbitrary activation region A in the parameter space.



Mathematical setup

Question: What constraints do the outputs of a ReLU network satisfy?

- Let $X = [x_1, \ldots, x_m]$ define the activation region $A = [a_1, \ldots, a_m]$.
- Split X into blocks $[X_1, \ldots X_k]$ such where X_i contains data points that follow the same activation pattern.
- Consider the parametrization $\varphi_X^A : \mathbb{R}^p \to \mathbb{R}^{n_L \times m} : \theta \mapsto F_X^A(\theta)$.
- Within each block, this parametrization can be written as $\theta \mapsto M(\theta)X$, where $M(\theta)$ is a matrix dependent on the activation pattern and θ.
- So, over all blocks, the parametrization is

$$\boxed{\varphi_X^A : \theta \mapsto [M_1(\theta)X_1 \mid M_2(\theta)X_2 \mid \cdots \mid M_k(\theta)X_k].}$$

Define the *ReLU output variety* as $\overline{\mathrm{im}(\varphi_X^A)}$. Denote it by V_X^A.

Question: What are the generators of $I_X^A := I(V_X^A)$?

Single block

When all data points in X follow the same activation pattern A, the map is

$$\varphi_X^A : \theta \mapsto M(\theta)X.$$

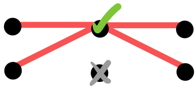
Example

Let $n_0 = n_1 = n_2 = 2$ and let $A = [1, 0]$. Then for any $X \in \mathbb{R}^{2\times m}$,

$$\varphi_X^A : (W^{(1)}, W^{(2)}) \mapsto MX = \begin{pmatrix} w_{11}^{(1)} w_{11}^{(2)} & w_{12}^{(1)} w_{11}^{(2)} \\ w_{11}^{(1)} w_{21}^{(2)} & w_{12}^{(1)} w_{21}^{(2)} \end{pmatrix} [x_1 \dots x_n].$$

The polynomials defining the image are:

1. one quadratic polynomial induced by $\det M$
2. linear polynomials induced by linear dependencies of X.



Single block

Let $r = \operatorname{rank} M$.

Proposition (A.-Montúfar, 2025+)

The ideal I_X^{A} is generated by $n_L \cdot \min\{m - n_0, 0\}$ linear polynomials and $\binom{n_L}{r+1}\binom{\min\{n_0,m\}}{r+1}$ homogeneous polynomials of degree $r+1$.

- linear polynomials $\rightarrow$ linear dependencies between data points in X
- degree $r+1$ polynomials $\rightarrow$ certain minors of MX, which do not depend on the dataset X

The pattern variety

We consider the parametrization $\varphi^A : \theta \mapsto [M_1(\theta) \mid M_2(\theta) \mid \cdots \mid M_k(\theta)]$.
Define the *pattern variety* to be $\overline{\mathrm{im}(\varphi^A)}$.

For each $i \in [k]$, we assume that:

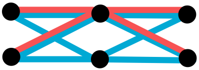
- $|X_i| = n_0$,
- all points in X_i follow the same activation pattern,
- all points in X_i are linearly independent.

Proposition (A.-Montúfar, 2025+)

Any polynomial $f \in J^{\mathrm{A}}$ gives rise to a unique polynomial $g = \psi^{-1} f \in I_X^{\mathrm{A}}$, where ψ is a linear change of coordinates dependent on X.

So, we can study the ideal of the pattern variety J^{A} instead!

Two blocks: example



Example: Consider a general dataset $X = [\ x_1, x_2,\ x_3, x_4\ ]$.

- $X_1 = [x_1, x_2]$ follow the activation pattern $(1, 0)$.
- $X_2 = [x_3, x_4]$ follow the activation pattern $(1, 1)$.

The image of $\varphi^{\mathsf{A}}(\theta)$ is $[M_1(\theta) \mid M_2(\theta)]$ where $\theta = (W^{(1)}, W^{(2)})$ and

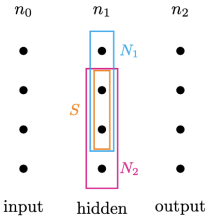
$$M_1 = \begin{pmatrix} w_{11}^{(1)} w_{11}^{(2)} & w_{12}^{(1)} w_{11}^{(2)} \\ w_{11}^{(1)} w_{21}^{(2)} & w_{12}^{(1)} w_{21}^{(2)} \end{pmatrix}, M_2 = \begin{pmatrix} w_{11}^{(1)} w_{11}^{(2)} + w_{21}^{(1)} w_{12}^{(2)} & w_{12}^{(1)} w_{11}^{(2)} + w_{22}^{(1)} w_{12}^{(2)} \\ w_{11}^{(1)} w_{21}^{(2)} + w_{21}^{(1)} w_{22}^{(2)} & w_{12}^{(1)} w_{21}^{(2)} + w_{22}^{(1)} w_{22}^{(2)} \end{pmatrix}.$$

The ideal of the image of $\theta \mapsto [M_1 \mid M_2] = \left(\begin{smallmatrix} m_1 & m_3 & m_5 & m_7 \\ m_2 & m_4 & m_6 & m_8 \end{smallmatrix} \right)$ is:

$$J^{\mathsf{A}} = \langle m_1 m_4 - m_2 m_3, \quad \det \left(\begin{smallmatrix} m_1 & m_3 \\ m_2 & m_4 \end{smallmatrix} \right) - \det \left(\begin{smallmatrix} m_1 & m_7 \\ m_2 & m_8 \end{smallmatrix} \right) - \det \left(\begin{smallmatrix} m_5 & m_3 \\ m_6 & m_4 \end{smallmatrix} \right) \rangle$$

The ideal I_X^{A} is obtained from J^{A} in terms of fixed but arbitrary data X_1, X_2!

Two blocks, shallow networks



Let $|N_1| = r_1$, $|N_2| = r_2$, $|S| = s$.
Let $t = r_1 + r_2 - 2s$.

Theorem (A.-Montúfar, 2025+)

The ideal J^{A} contains:

1. $(r_1 + 1)$*-minors of* M_1;
2. $(r_2 + 1)$*-minors of* M_2;
3. $(n_1 + 1)$*-minors of* $[M_1 \mid M_2]$ *and* $[M_1^T \mid M_2^T]$;
4. $(t + 1)$*-minors of* $M_2 - M_1$.

Conjecture: no other polynomials are needed to generate the ideal.

Equivalent statement

Consider the map

$$\mathcal{M}_a \times \mathcal{M}_b \times \mathcal{M}_c \to \mathbb{R}^{n_2 \times 2n_0} : (A, B, C) \mapsto [M_1 = A + B | M_2 = B + C]$$

where $\mathcal{M}_r = \{X \in \mathbb{R}^{n_2 \times n_0} : \mathrm{rank}(X) \leq r\}$. The implicitization problem becomes eliminating the variables associated with A, B, C from the ideal

$$\begin{aligned} I = &\langle M_1 - A - C, M_2 - B - C\rangle \\ &+ \langle a\text{-minors of } A\rangle + \langle b\text{-minors of } B\rangle + \langle c\text{-minors of } C\rangle. \end{aligned}$$

Question: is the resulting ideal in $\mathbb{C}[M_1, M_2]$ is generated by

1. $(a + b + 1)$-minors of M_1;
2. $(b + c + 1)$-minors of M_2;
3. $(a + b + c + 1)$-minors of $[M_1|M_2]$ and $[M_1^T|M_2^T]$;
4. $(a + c + 1)$-minors of $M_2 - M_1$?

Two blocks, shallow networks, dimension

Theorem (A.-Montúfar, 2025+)

Suppose $s \geq 1$, so that the two blocks overlap nontrivially. If either of the following conditions holds:

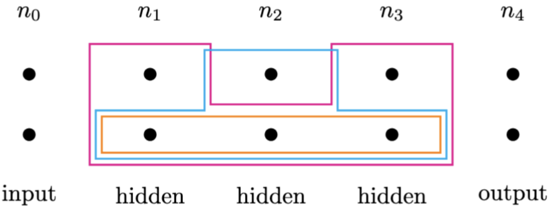
- $n_0 \geq n_1$ *and* $n_1 \leq n_2 + 1$;
- $n_2 \geq n_1$ *and* $n_1 \leq n_0 + 1$,

then the ideal J^{A} *has the expected dimension namely:*

$$\dim(\mathcal{M}_a) + \dim(\mathcal{M}_b) + \dim(\mathcal{M}_c).$$

Otherwise, there is a drop.

Two blocks, deep networks



$$
\begin{aligned}
N_1 &= \{(2,1,2),(2,2,2)\}\\
N_2 &= \{(1,2,1),(2,2,1),\\
&\qquad (1,2,2),(2,2,2)\}\\
S &= \{(2,2,2)\}
\end{aligned}
$$

The *path network* determined by $N_2 \setminus S$ has rank 2, even though all three paths pass through the same neuron in the middle layer. Let

- $r_a =$ rank of the path network on $N_1 \setminus S$;
- $r_b =$ rank of the fully connected network on S;
- $r_c =$ rank of the path network on $N_2 \setminus S$;

Let $t = r_a + r_c$.

Two blocks, deep networks

Theorem (A.-Montúfar, 2025+)

The ideal J^{A} contains:

1. $(r_a + r_b + 1)$*-minors of* M_1*;*
2. $(r_b + r_c + 1)$*-minors of* M_2*;*

3a. $(n_{\min} + 1)$*-minors of* $[M_1 \mid M_2]$ *if* $A_1^\ell = A_2^\ell$ *for all* $\ell > \ell_{\min}$.

3b. $(n_{\min} + 1)$*-minors of* $[M_1^T \mid M_2^T]$ *if* $A_1^\ell = A_2^\ell$ *for all* $\ell < \ell_{\min}$.

4. $(t + 1)$*-minors of* $M_2 - M_1$.

Thank you!

Questions?