Constraining the outputs of ReLU neural networks

Yulia Alexandr (UCLA and Harvard)
joint work with **Guido Montúfar**

New Directions in Algebraic Statistics
The Institute for Mathematical and Statistical Innovation
July 21, 2025

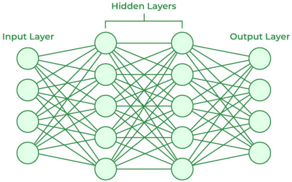
Neural networks

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

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

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.



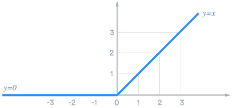
For a dataset $X = [x_1, x_2, \ldots, x_m]$ and unknown parameters θ we are interested in describing the constraints between the coordinates of the array of model outputs $F_X(\theta) = [f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_m)]$.

ReLU networks

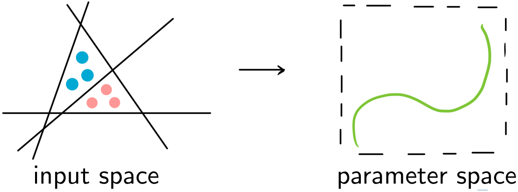
A *ReLU network* is given by the activation function

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

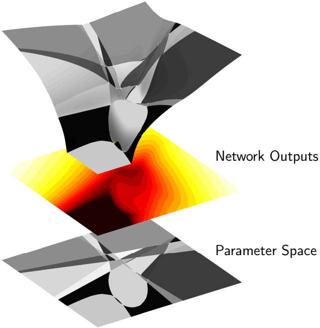
at each layer of the neural network.



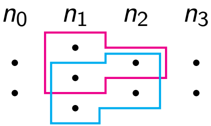
- this makes $f_\theta(x)$ piece-wise linear
  - natural subdivision of the input space into regions
  - $f_\theta(x)$ is a linear function of x in each region
- now consider multiple data points $X = [x_1, \ldots, x_m]$
  - this subdivision extends to the parameter space
  - $F_X(\theta)$ is multi-linear in θ in each activation region

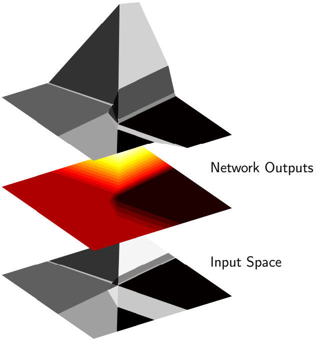


Fixed data vs. fixed parameters



Fixed Input Data





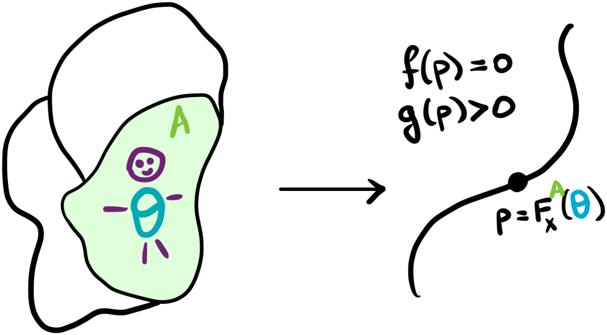
Fixed Parameters

$$X = [x_1, x_2]$$
$$A_1 = [(1,1,0),(1,0)]$$
$$A_2 = [(0,1,1),(1,1)]$$

The main question

Problem

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



Implicitization

Given a model, parametrized by

$$\varphi : \theta = (\theta_1, \ldots, \theta_n) \mapsto \big(f_1(\theta), f_2(\theta), \ldots f_m(\theta)\big),$$

we are interested in describing the polynomials defining $\overline{\text{image}}(\varphi)$. This process is called *implicitization*.

Implicitization

Given a model, parametrized by

$$\varphi : \theta = (\theta_1, \ldots, \theta_n) \mapsto \big(f_1(\theta), f_2(\theta), \ldots f_m(\theta)\big),$$

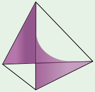
we are interested in describing the polynomials defining $\overline{\text{image}}(\varphi)$. This process is called *implicitization*.

Example (The independence model.)

Parametrization:

$$(\theta_1, \theta_2) \quad \mapsto \quad \big(\underbrace{\theta_1\theta_2}_{p_1}, \ \underbrace{\theta_1(1-\theta_2)}_{p_2}, \ \underbrace{(1-\theta_1)\theta_2}_{p_3}, \ \underbrace{(1-\theta_1)(1-\theta_2)}_{p_4}\big).$$

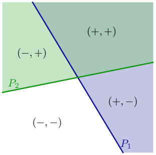
Implicit ideal: $I = \langle p_1 p_4 - p_2 p_3, \ \ p_1 + p_2 + p_3 + p_4 - 1\rangle.$

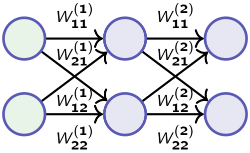


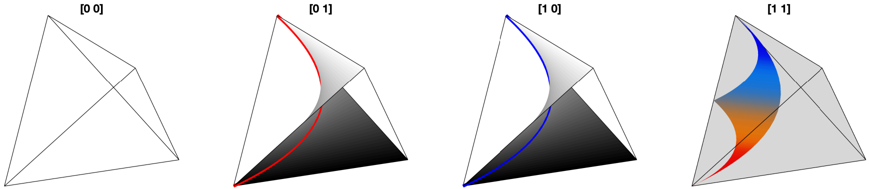
The generators of the ideal I are called *model invariants*.

Parametrization

- The number of linear pieces over the input space can be enormous.
- The linear pieces share parameters and are not independent.
- We investigate the relationships between the linear pieces.

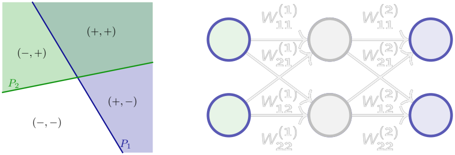


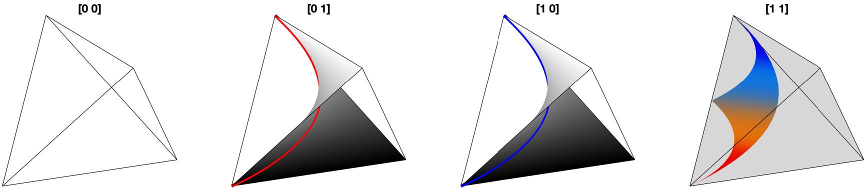




Parametrization

- The number of linear pieces over the input space can be enormous.
- The linear pieces share parameters and are not independent.
- We investigate the relationships between the linear pieces.

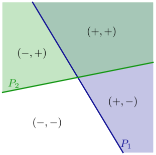


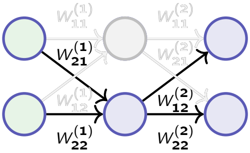


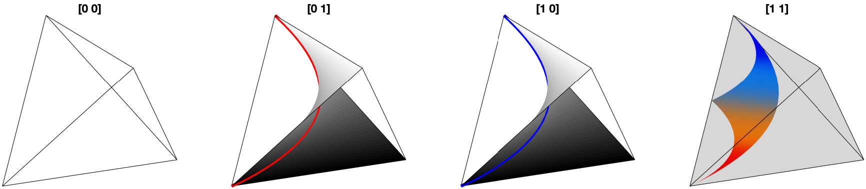
$M_{[00]} = 0$

Parametrization

- The number of linear pieces over the input space can be enormous.
- The linear pieces share parameters and are not independent.
- We investigate the relationships between the linear pieces.



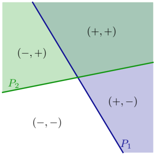


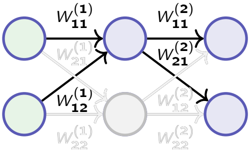


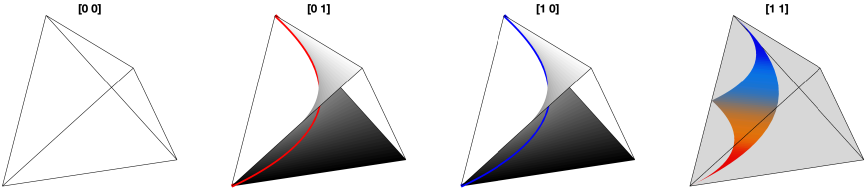
$$M_{[01]} = W^{(2)} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} W^{(1)}$$

Parametrization

- The number of linear pieces over the input space can be enormous.
- The linear pieces share parameters and are not independent.
- We investigate the relationships between the linear pieces.



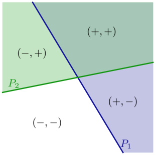


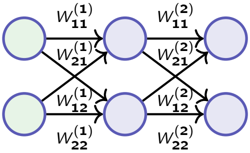


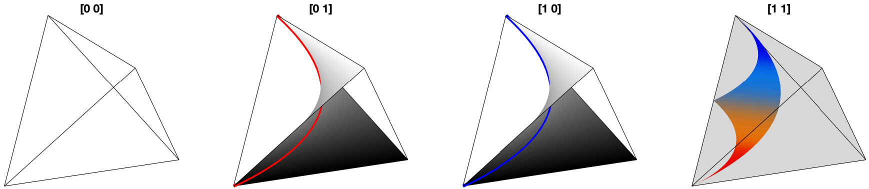
$$M_{[10]} = W^{(2)} \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right] W^{(1)}$$

Parametrization

- The number of linear pieces over the input space can be enormous.
- The linear pieces share parameters and are not independent.
- We investigate the relationships between the linear pieces.







$$M_{[11]} = W^{(2)}W^{(1)}$$

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 $\theta \mapsto M_i(\theta)X_i$, 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 $I_X^A := I(V_X^A)$? Dimension? Degree?

Single block

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

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

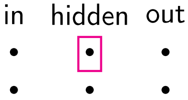
Example

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

$$\varphi_X^A : (W^{(1)}, W^{(2)}) \mapsto M(\theta)X = \begin{pmatrix} w_{11}^{(1)} w_{11}^{(2)} & w_{12}^{(1)} w_{11}^{(2)} \\ w_{11}^{(1)} w_{21}^{(2)} & w_{12}^{(1)} w_{21}^{(2)} \end{pmatrix} [x_1 \ldots x_m].$$

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 $r = \operatorname{rank} M(\theta)$ for generic θ.

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 $\to$ linear dependencies between data points in X
- degree $r+1$ polynomials $\to$ 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 *ReLU 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+)

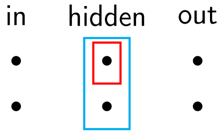
Any polynomial $f \in J^{\mathrm{A}}$ gives rise to a unique polynomial $g = \psi^{-1} f \in I_X^{\mathrm{A}}$, where ψ is a linear change of coordinates dependent on X.

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

Example: 2 blocks

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

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



ReLU output variety: $\theta \mapsto [M_1(\theta)X_1 \mid M_2(\theta)X_2]$ with $\theta = (W^{(1)}, W^{(2)})$

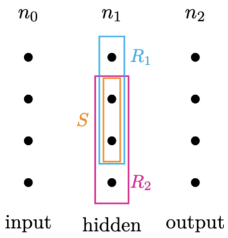
$$M_1(\theta) = \begin{pmatrix} w_{11}^{(1)} w_{11}^{(2)} & w_{12}^{(1)} w_{11}^{(2)} \\ w_{11}^{(1)} w_{21}^{(2)} & w_{12}^{(1)} w_{21}^{(2)} \end{pmatrix}, M_2(\theta) = \begin{pmatrix} w_{11}^{(1)} w_{11}^{(2)} + w_{21}^{(1)} w_{12}^{(2)} & w_{12}^{(1)} w_{11}^{(2)} + w_{22}^{(1)} w_{12}^{(2)} \\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

ReLU pattern variety: $\theta \mapsto [M_1(\theta) \mid M_2(\theta)] = \begin{pmatrix} m_1 & m_3 & m_5 & m_7 \\ m_2 & m_4 & m_6 & m_8 \end{pmatrix}$

$$J^A = \langle \det \begin{pmatrix} m_1 & m_3 \\ m_2 & m_4 \end{pmatrix}, \quad \det \begin{pmatrix} m_1 - m_5 & m_3 - m_7 \\ m_2 - m_6 & m_4 - m_8 \end{pmatrix} \rangle.$$

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 $|R_1| = r_1$, $|R_2| = r_2$, $|S| = s$.
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* $M_1 - M_2$.

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

Sufficiency

Consider the map

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

where $\mathcal{M}_r = \{X \in \mathbb{R}^{n_2 \times n_0} : \text{rank}(X) \leq r\}$.

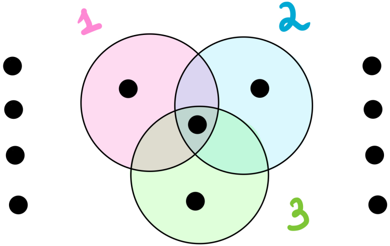
Question: Given two matrices $M_1, M_2 \in \mathbb{R}^{n_2 \times n_0}$ satisfying:

1. $\text{rank}\, M_1 \leq a + c$;
2. $\text{rank}\, M_2 \leq b + c$;
3. $\text{rank}[M_1 \mid M_2]$ and $\text{rank}[M_1^T \mid M_2^T] \leq a + b + c$;
4. $\text{rank}(M_1 - M_2) \leq a + b$,

can we find A, B, C such that:

- $M_1 = A + C$ and $M_2 = B + C$;
- $\text{rank}\, A \leq a, \quad \text{rank}\, B \leq b, \quad \text{rank}\, C \leq c$?

Example: 3 blocks



- 48 cubics: 3-minors of M_1, M_2, and M_3;
- 48 cubics: 3-minors of $M_1 - M_2$, $M_2 - M_3$, and $M_2 - M_3$;
- 120 quartics: 4-minors of $[M_i \mid M_j]$ and $[M_i^T \mid M_j^T]$;
- 40 quartics: 4-minors of $[M_1 - M_2 \mid M_2 - M_3]$ and $\begin{bmatrix} M_1 - M_2 \\ M_2 - M_3 \end{bmatrix}$;
- 2000 quintics: algebraically independent 5-minors of

$$\begin{bmatrix} M_1 & M_2 \\ M_3 & M_2 \end{bmatrix}, \begin{bmatrix} M_1 & M_2 \\ M_3 & M_3 \end{bmatrix}, \begin{bmatrix} M_2 & M_3 \\ M_1 & M_1 \end{bmatrix}, \begin{bmatrix} M_2 & M_3 \\ M_1 & M_3 \end{bmatrix}, \begin{bmatrix} M_3 & M_1 \\ M_2 & M_2 \end{bmatrix}, \begin{bmatrix} M_3 & M_1 \\ M_2 & M_1 \end{bmatrix}.$$

Many blocks, shallow networks

Linear combinations:

- Each $M_i(\theta) = W^{(2)}\,\mathrm{diag}(A_i) W^{(1)}$ is a sum of rank-one matrices.
- For $\lambda \in \mathbb{Z}^k$,

$$\mathrm{rank}\left(\sum_i \lambda_i M_i(\theta)\right) \leq \left|\mathrm{supp}\left(\sum_i \lambda_i A_i\right)\right|.$$

- Polynomial constraints from minors:

$$(|\,\mathrm{supp}(\sum_i \lambda_i A_i)| + 1)\text{-minors} \in J^{\mathrm{A}}.$$

Question: Which λ give rise to minimal generators?

Many blocks, shallow networks

Linear combinations:

- Each $M_i(\theta) = W^{(2)}\operatorname{diag}(A_i)W^{(1)}$ is a sum of rank-one matrices.
- For $\lambda \in \mathbb{Z}^k$,

$$\operatorname{rank}\left(\sum_i \lambda_i M_i(\theta)\right) \le \left|\operatorname{supp}\left(\sum_i \lambda_i A_i\right)\right|.$$

- Polynomial constraints from minors:

$$(|\operatorname{supp}(\sum_i \lambda_i A_i)| + 1)\text{-minors} \in J^{\mathsf{A}}.$$

Question: Which λ give rise to minimal generators?

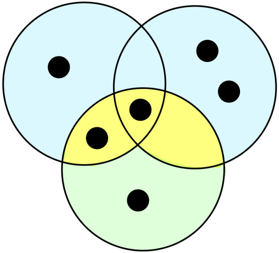
Blocks of linear combinations...

Shallow networks, dimension

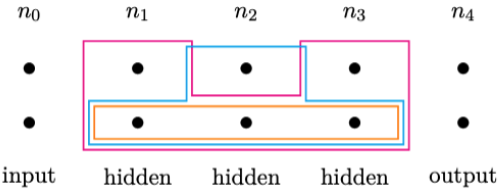
Two blocks: If $n_0 \geq n_1 \leq n_2$ then the ideal J^{A} has the expected dimension, namely

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

Many blocks: If $n_0 \geq n_1 \leq n_2$ then the ideal J^{A} has the expected dimension.



Two blocks, deep networks



$$
\begin{aligned}
R_1 &= \{(1,2,1),(2,2,1),\\
&\quad\ (1,2,2),(2,2,2)\}\\
R_2 &= \{(2,1,2),(2,2,2)\}\\
S &= \{(2,2,2)\}
\end{aligned}
$$

The *path network* determined by $R_1 \setminus S$ 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 $R_1 \setminus S$;
- $r_b =$ rank of the path network on $R_2 \setminus S$;
- $r_c =$ rank of the fully connected network on S.

Let $t = r_a + r_b$.

Deep networks

Two blocks:

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;

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

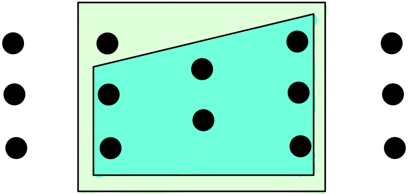
3b. $(n_{\min}+1)$-*minors of* $[M_1^T \mid M_2^T]$ *if* $A_1^\ell = A_2^\ell$ *for all* $\ell < \ell_{\min}$.

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

Many blocks: Similar to shallow networks, except:

- have to consider rank-1 matrices determined by *paths*;
- get looser rank bounds.

Example: 2 blocks, deep network



J^A is generated by:

- 9 quadratics: 2-minors of $M_1 - M_2$;
- 10 cubics: 3×3 minors of $[M_1 \mid M_2]$.

Thank you!

Questions?