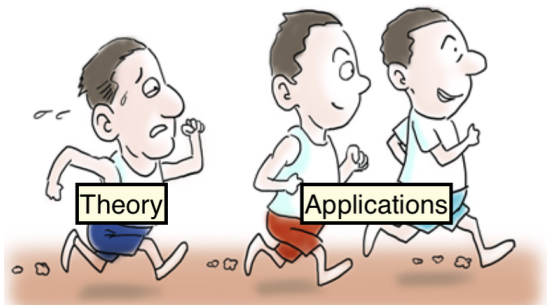


Algebraic invariants in machine learning

Yulia Alexandr (UCLA)

Boston College Math & Machine Learning seminar
April 29, 2026

Algebraic machine learning



DARPA AIQ: Artificial Intelligence Quantified



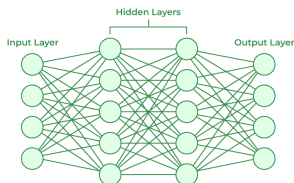
Neural networks

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 $\theta_{\ell} = W^{(\ell)}$.

- $n_0 =$ dimension of the input space
- $n_{\ell} =$ width of the ℓ th layer



For a dataset $X = [x_1, x_2, \dots, x_n]$ and unknown parameters θ we want to describe the **constraints** between the coordinates of the model outputs

$$F_X(\theta) = [f_{\theta}(x_1), f_{\theta}(x_2), \dots, f_{\theta}(x_n)].$$

ReLU networks

Constraining the outputs of ReLU neural networks
joint work with Guido Montúfar

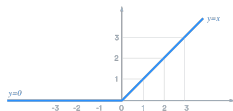
ReLU networks

A *ReLU network* is given by the activation function

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

at each layer of the neural network.

- $f_\theta(x)$ is piece-wise linear
 \implies subdivision of the **input space** into linear regions
- $F_X(\theta)$ is piece-wise multi-linear for fixed data X
 \implies subdivision of the **parameter space** into **activation regions**



Takeaway:

restricting to an activation region A turns $F_X(\theta)$ into a polynomial map!

The main question

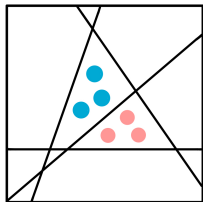
Fix X to be an arbitrary dataset.

Problem

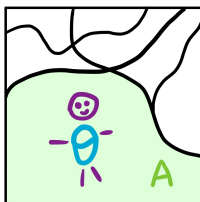
Describe the equations and inequalities that define the image of the map

$$\varphi_X^A : \theta \mapsto F_X^A(\theta)$$

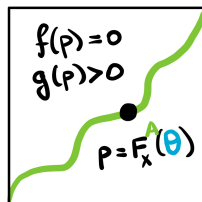
as θ varies over an arbitrary activation region A in the parameter space.



input space



parameter space



prediction space

Why do we care?

- Characterization of reachable outputs
 - ▶ does a point lie in the image of a neural network map?
- Safety verification
 - ▶ can the network ever produce an unsafe or undesired output?
 - ▶ let S be the set of unsafe outputs
 - ▶ if $V_X^A \cap S = \emptyset$, the network is **provably safe**
- Symbolic extrapolation
 - ▶ use invariants to reason about outputs even far from the training data
 - ▶ guarantee behavior without relying on over-approximation



Takeaway: invariants = provable guarantees!

Implicitization

Given a model, parametrized by

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

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

Implicitization

Given a model, parametrized by

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

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

Example (The independence model.)

Parametrization:

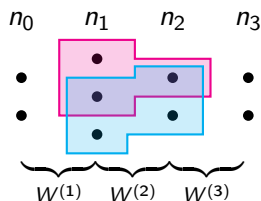
$$(\theta_1, \theta_2) \mapsto (\underbrace{\theta_1\theta_2}_{p_1}, \underbrace{\theta_1(1-\theta_2)}_{p_2}, \underbrace{(1-\theta_1)\theta_2}_{p_3}, \underbrace{(1-\theta_1)(1-\theta_2)}_{p_4}).$$

Implicit ideal: $I = \langle p_1p_4 - p_2p_3, p_1 + p_2 + p_3 + p_4 - 1 \rangle$.



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

The parametrization



$$X = [x_1, x_2, x_3, x_4]$$

$$a_1 = [(1, 1, 0), (1, 0)]$$

$$a_2 = [(0, 1, 1), (1, 1)]$$

$$A = [a_1, a_1, a_2, a_2]$$

$$M_1(\theta) = W^{(3)} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} W^{(2)} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} W^{(1)}$$

$$M_2(\theta) = W^{(3)} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} W^{(2)} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} W^{(1)},$$

$$\varphi_X^A : \theta = (W^{(1)}, W^{(2)}, W^{(3)}) \mapsto [M_1(\theta) [x_1 \ x_2] \mid M_2(\theta) [x_3 \ x_4]]$$

The parametrization

In general, for k blocks:

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

ReLU output variety: $V_X^A = \overline{\text{im}(\varphi_X^A)}$ with ideal I_X^A .

$$\varphi^A : \theta \mapsto [M_1(\theta) \mid M_2(\theta) \mid \cdots \mid M_k(\theta)].$$

ReLU pattern variety: $U^A = \overline{\text{im}(\varphi^A)}$ with ideal J^A .

Proposition (A.-Montúfar)

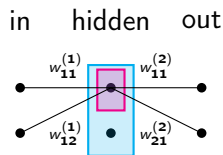
Any polynomial $f \in J^A$ gives rise to a unique polynomial $g = \psi^{-1}f \in I_X^A$, where ψ is a linear change of coordinates dependent on X .

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 $\theta = (W^{(1)}, W^{(2)})$

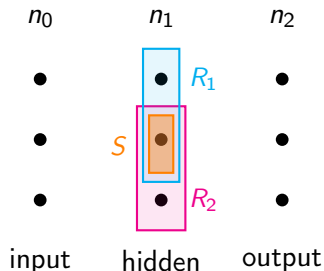
$$M_1(\theta) = \begin{pmatrix} w_{11}^{(1)} & w_{11}^{(2)} \\ w_{11}^{(1)} & w_{21}^{(2)} \end{pmatrix}, M_2(\theta) = \begin{pmatrix} w_{11}^{(1)} w_{11}^{(2)} + w_{21}^{(1)} w_{12}^{(2)} & w_{12}^{(1)} w_{11}^{(2)} + w_{22}^{(1)} w_{12}^{(2)} \\ w_{11}^{(1)} w_{21}^{(2)} + w_{21}^{(1)} w_{22}^{(2)}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

ReLU pattern variety: $\theta \mapsto [M_1(\theta) \mid M_2(\theta)]$

$$J^A = \langle \det(M_1), \det(M_2 - M_1) \rangle.$$

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)

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

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

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

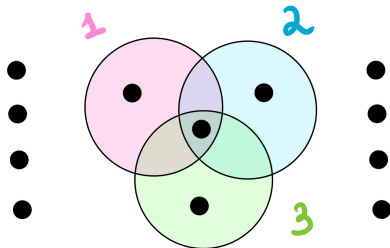
Question: Given two matrices $M_1, M_2 \in \mathbb{R}^{n_2 \times n_0}$ satisfying:

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

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

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

- Polynomial constraints from minors:

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

Question: Which λ give rise to minimal generators?

Many blocks, shallow networks

Linear combinations:

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

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

- Polynomial constraints from minors:

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

Question: Which λ give rise to minimal generators?

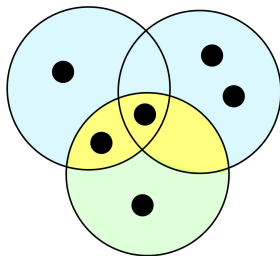
Blocks of linear combinations...

Shallow networks, dimension

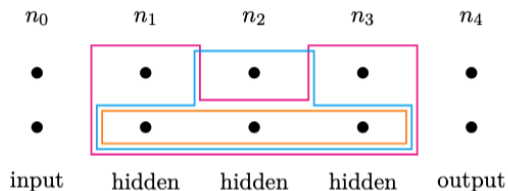
Two blocks: If $n_0 \geq n_1 \leq n_2$ then the ideal J^A has the expected dimension, namely

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

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



Two blocks, deep networks



$$R_1 = \{(1, 2, 1), (2, 2, 1), (1, 2, 2), (2, 2, 2)\}$$
$$R_2 = \{(2, 1, 2), (2, 2, 2)\}$$
$$S = \{(2, 2, 2)\}$$

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)

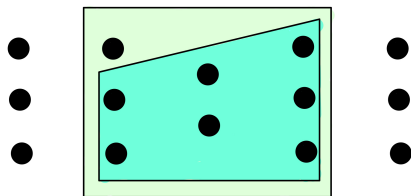
The ideal J^A contains:

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 $[M_1 \mid M_2]$.

Attention modules

Algebraic invariants of lightning self-attention
joint work with Hao Duan and Guido Montúfar

Lightning self-attention

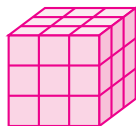
Let the weight matrices be $Q, K \in \mathbb{R}^{a \times d}$ and $V \in \mathbb{R}^{d' \times d}$.

For a data matrix $X \in \mathbb{R}^{d \times t}$ and $A = K^\top Q$, the *lightning self-attention* is:

$$\varphi_W(X) = VX(X^\top AX) \in \mathbb{R}^{d' \times t}$$

Observation:

For fixed parameters $W = (Q, K, V)$, each output coordinate is a **homogeneous cubic polynomial** in the entries of the input matrix X !



3-way tensor

$$\sum_{i,j,k} \underbrace{y_{i,j,k}}_{\text{coefficients}} x_i x_j x_k$$

We study the algebraic relations satisfied by the **polynomial coefficients** of this map. This defines the *attention variety*.

The parametrization

Expanding the (i, j) -th output coordinate yields a cubic polynomial:

$$\varphi_W(X)[i, j] = \sum_{k, m, \ell \in [d], n \in [t]} (a_{ml} v_{ik}) x_{kn} x_{mn} x_{\ell j}$$

Single-column monomials

All 3 variables come from the **target** column j ($n = j$):

$$x_{kj} x_{mj} x_{\ell j}$$

Cross-column monomials

2 variables from a **context** $n \neq j$,
1 variable from **target** j :

$$x_{kn} x_{mn} x_{\ell j}$$

Let y denote the ambient coefficient variables. Define the **parametrization**:

$$\mu : (A, V) \longmapsto (\text{cubic coefficients of } \varphi_W(X)[i, j])$$

Goal: Find the defining ideal $J \subseteq \mathbb{R}[y]$ of the **attention variety** $\mathcal{V} := \overline{\text{im}(\mu)}$.

Linear relations

Within a fixed output coordinate (i, j) , every **single-column** coefficient is a linear combination of **cross-column** coefficients.

Example: Fix the index set $\{1, 1, 2\}$.

Let $y_{(1,j),(1,j),(2,j)}$ be the coefficient for $x_{1j}x_{1j}x_{2j}$.

There are two ways to split $\{1, 1, 2\}$:

- Context gets $\{1, 1\}$, target gets 2
 \implies coefficient $y_{(1,n),(1,n),(2,j)}$ for $x_{1n}x_{1n}x_{2j}$
- Context gets $\{1, 2\}$, target gets 1
 \implies coefficient $y_{(1,n),(2,n),(1,j)}$ for $x_{1n}x_{2n}x_{1j}$

Summing over these valid splits yields a linear invariant:

$$y_{(1,j),(1,j),(2,j)} - y_{(1,n),(1,n),(2,j)} - y_{(1,n),(2,n),(1,j)} = 0$$

Single-column invariants

Let T be the symmetric tensor formed by the **single-column** coefficients, and $f_T(x)$ be its associated formal cubic polynomial.

Theorem (A.-Duan-Montúfar):

1. Factorization: f_T factors as $f_T(x) = (x^\top Ax)(v^\top x)$.

\implies it lies on the **Chow variety** of split type $(2,1)$.

\implies the **Lie algebra flattening matrix** $M_{\text{Lie}}(f_T)$ drops rank.

2. Low-rank constraints

Taking any $d \times d$ **slice** N of the tensor T yields the bound:

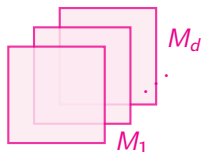
$$\text{rank}(N) \leq 2\text{rank}(A) + 1$$

\implies if $(2\text{rank}(A) + 1 < d)$, the maximal minors of these slices vanish!

When $d = 3$, $t = 1$, J is generated by 35 minimal octic polynomials!

\implies these are the 8×8 minors of $M_{\text{Lie}}(f_T)$.

Cross-column invariants



Fix an index ℓ . The **cross-column** coefficients form a symmetric matrix M_ℓ defined by

$$(M_\ell)_{k,m} = y_{(k,n),(m,n),(\ell,j)}$$

Under the parametrization, we have:

$$M_\ell = \frac{1}{2}(va_\ell^\top + a_\ell v^\top).$$

where a_ℓ is the ℓ -th column of A .

Theorem (A.-Duan-Montúfar):

1. Every matrix pencil $M(\lambda) = \sum_{\ell=1}^d \lambda_\ell M_\ell$ has **rank** ≤ 2 .

\implies 3×3 minors of $M(\lambda)$ yield cubic invariants in y .

2. The **unbalanced flattening matrix** $F_{n,j} = [\text{vec}(M_1) \dots \text{vec}(M_d)]:$

- If $a < d$, then $\text{rank}(F_{n,j}) \leq a$

\implies invariants of degree $a + 1$.

- If $a \geq d$, the $d \times d$ maximal minors of $F_{n,j}$ satisfy linear relations

\implies invariants of degree d .

More cross-column invariants

1. Veronese-type relations

Certain $r \times r$ minors of the cross-column coefficient matrices are proportional to Veronese monomials.

Example ($d = 2$):

Three such 2×2 minors satisfy $(d_1, d_2, d_3) \propto (v_1^2, v_2^2, v_1 v_2)$.

The quadratic relation

$$v_1^2 v_2^2 - (v_1 v_2)^2 = 0$$

becomes the **quartic** relation

$$d_1 d_2 - d_3^2 = 0.$$

In general: **degree- $2r$** polynomials.

2. Sylvester resultant relations

Cross-column slices \rightsquigarrow quadrics $q_1(y; x), \dots, q_m(y; x)$

On the attention variety, they have a common linear factor in x .

Restrict to a generic line

$$L = \lambda_1 u + \lambda_2 w:$$

$\implies q_1^L, \dots, q_m^L$ binary quadrics with a common root in \mathbb{P}^1

Hence, for every pair,

$$\text{Res}(q_r^L, q_s^L) = 0$$

giving additional **quartic** relations.

Summary

