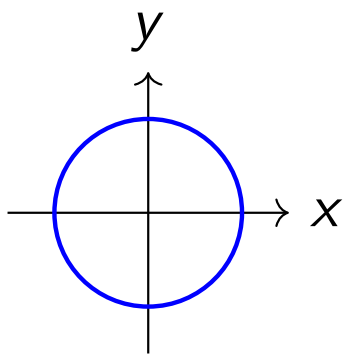
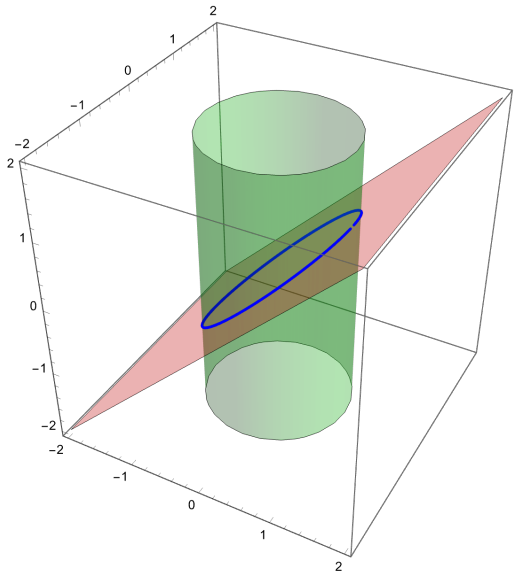
Can algebra be applied?

Yulia Alexandr (UCLA and Harvard)

Department of Mathematical Sciences Colloquium
Worcester Polytechnic Institute
August 29, 2025

Algebraic varieties

An *algebraic variety* is the set of all points that satisfy a system of polynomial equations.

Ideal	Variety
$\langle x^2 + y^2 - 1 \rangle$	
$\langle x^2 + y^2 - 1,\ x - z \rangle$	

Outline of talk

Table of Contents

Statistical models

Many statistical models are made up of distributions whose coordinates satisfy polynomial equations.

Example

Let X_1 and X_2 be binary random variables. Let $p_{ij} = \mathbb{P}(X_1 = i, X_2 = j)$. Then X_1 and X_2 are independent if and only if

$$\det \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = p_{11}p_{22} - p_{12}p_{21} = 0.$$

Exercise: think about why this is true!

Recall that X_1 and X_2 are *independent* if

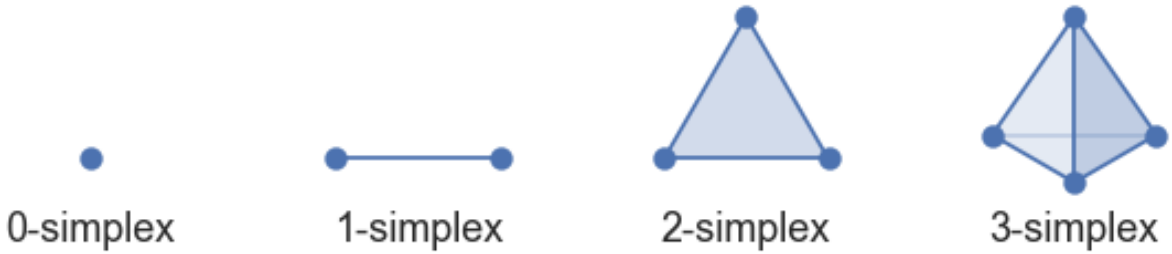
$$\mathbb{P}(X_1 = i | X_2 = j) = \mathbb{P}(X_1 = i) \cdot \mathbb{P}(X_2 = j).$$

Statistical models

How do algebraic statisticians think of statistical models?

- A *probability simplex* is defined as

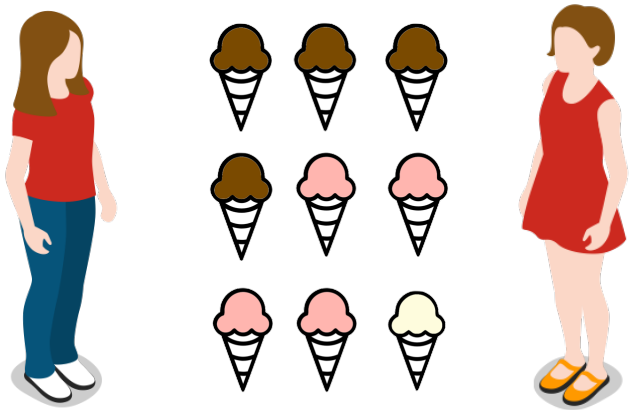
$$\Delta_{n-1} = \{(p_1, \ldots, p_n) : p_1 + \cdots + p_n = 1, p_i \geq 0 \text{ for } i \in [n]\}.$$



- A *statistical model* is a subset of Δ_{n-1}.
- A *variety* is the set of solutions to a system of polynomial equations.
- An *algebraic statistical model* is a subset $\mathcal{M} = \mathcal{V} \cap \Delta_{n-1}$ for some variety $\mathcal{V} \subseteq \mathbb{C}^n$.

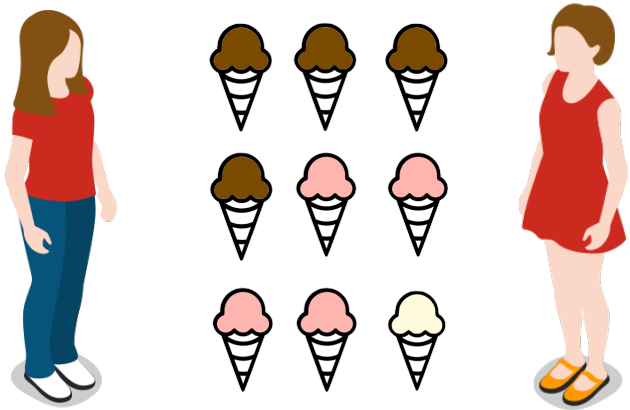
Maximum likelihood estimation

(p_1, p_2, p_3)



Maximum likelihood estimation

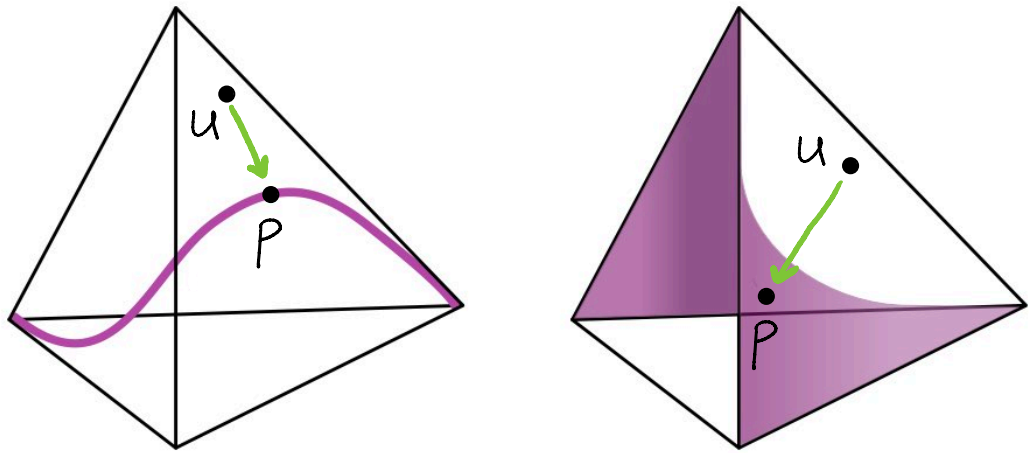
(p_1, p_2, p_3)



$$L = c \cdot p_1^{4/9} p_2^{4/9} p_3^{1/9}$$

$$\ell_u(p) = 4/9 \cdot \log(p_1) + 4/9 \cdot \log(p_2) + 1/9 \cdot \log(p_3) + \log(c).$$

Maximum likelihood estimation



Let $\mathcal{M} \subseteq \Delta_{n-1}$ be a statistical model.

For an empirical data point $u = (u_1, ..., u_n) \in \Delta_{n-1}$, the *log-likelihood function* with respect to u assuming distribution $p = (p_1, ..., p_n) \in \mathcal{M}$ is

$$\boxed{\ell_u(p) = u_1 \log p_1 + u_2 \log p_2 + \cdots + u_n \log p_n.}$$

Maximum likelihood estimation

Fix an algebraic statistical model $\mathcal{M} \subseteq \Delta_{n-1}$

1. The maximum likelihood estimation problem (MLE):

Given a sampled empirical distribution $u \in \Delta_{n-1}$, which point $p \in \mathcal{M}$ did it most likely come from? In other words, we wish to maximize $\ell_u(p)$ over all points $p \in \mathcal{M}$.

Maximum likelihood estimation

Fix an algebraic statistical model $\mathcal{M} \subseteq \Delta_{n-1}$

1. The maximum likelihood estimation problem (MLE):

Given a sampled empirical distribution $u \in \Delta_{n-1}$, which point $p \in \mathcal{M}$ did it most likely come from? In other words, we wish to maximize $\ell_u(p)$ over all points $p \in \mathcal{M}$.

2. Computing logarithmic Voronoi cells:

Given a point $q \in \mathcal{M}$, what is the set of all points $u \in \Delta_{n-1}$ that have q as a global maximum when optimizing the function $\ell_u(p)$ over $\mathcal{M}$?

The set of all such elements $u \in \Delta_{n-1}$ is the *logarithmic Voronoi cell* at q.

Logarithmic Voronoi cells

Proposition (A., Heaton)

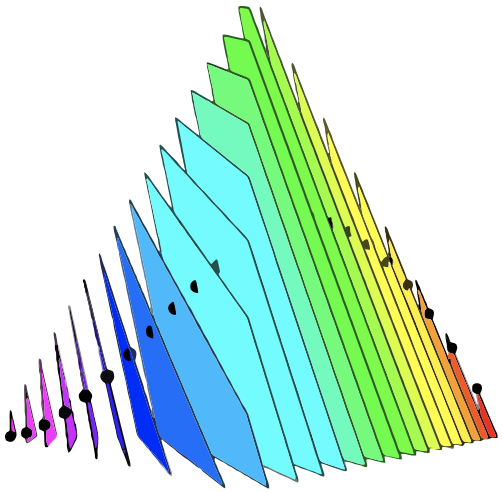
Logarithmic Voronoi cells are convex sets.

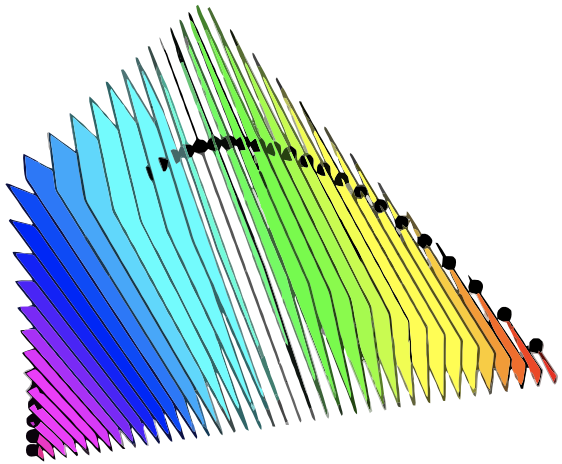
Theorem (A., Heaton)

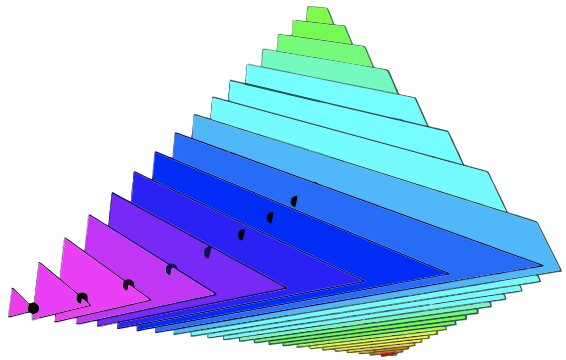
If $\mathcal{M}$ is a finite model, a linear model, or a toric model, the logarithmic Voronoi cell at any point $p \in \mathcal{M}$ is a polytope.

Example (The twisted cubic.)

The curve is given by $\theta \mapsto \left(\theta^3, 3\theta^2(1-\theta), 3\theta(1-\theta)^2, (1-\theta)^3\right)$.



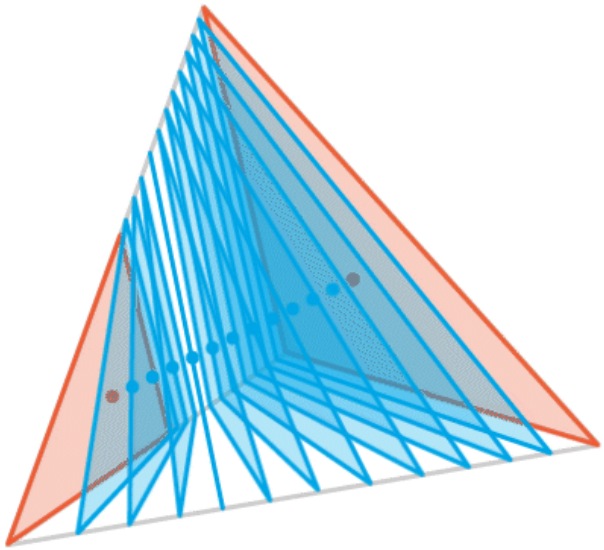


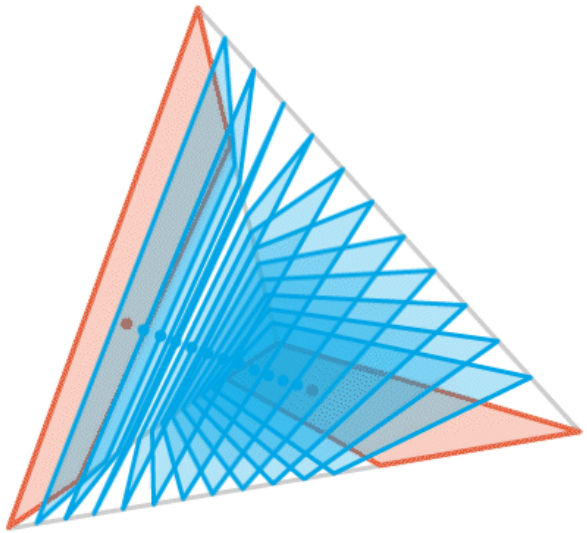


Linear models

Theorem (A.)

For linear models, logarithmic Voronoi cells at all interior points on the model have the same combinatorial type. This type can be described via Gale diagrams.



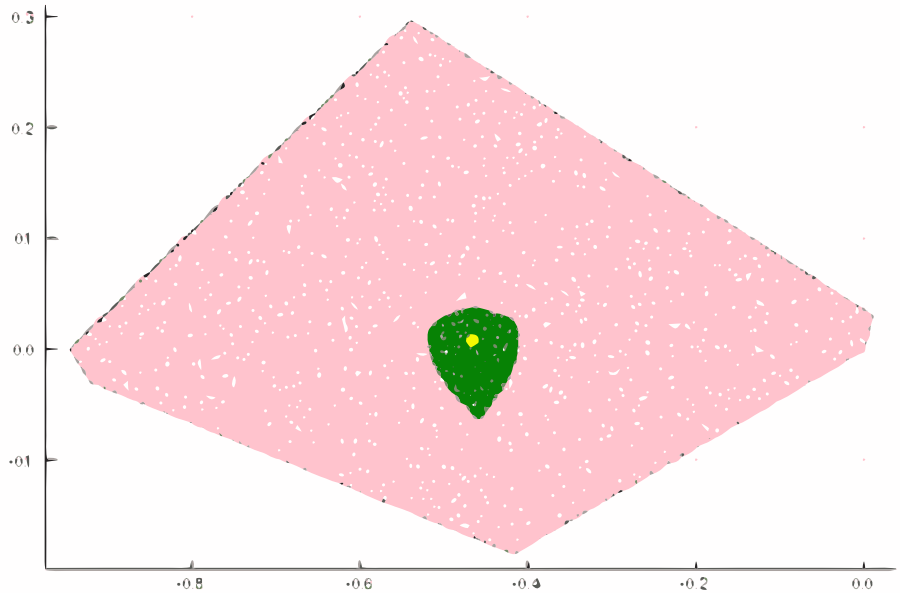


Why do we care?

Logarithmic Voronoi cells can also be useful in data privacy, particularly for statistical disclosure limitation.

- If a logarithmic Voronoi cell contains only one point then releasing the model estimate will also release the observed data to the public, even if it was intended to be private.

For models with complicated geometry, numerical methods are necessary to analyze logarithmic Voronoi cells.



Maximizing divergence

For two distributions $p, q \in \Delta_{n-1}$, the *Kullback-Leibler (KL) divergence* is

$$D(p||q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right).$$

For fixed $u \in \Delta_{n-1}$ maximizing $\ell_u(p)$ = minimizing $D(u||p)$ over $p \in \mathcal{M}$.

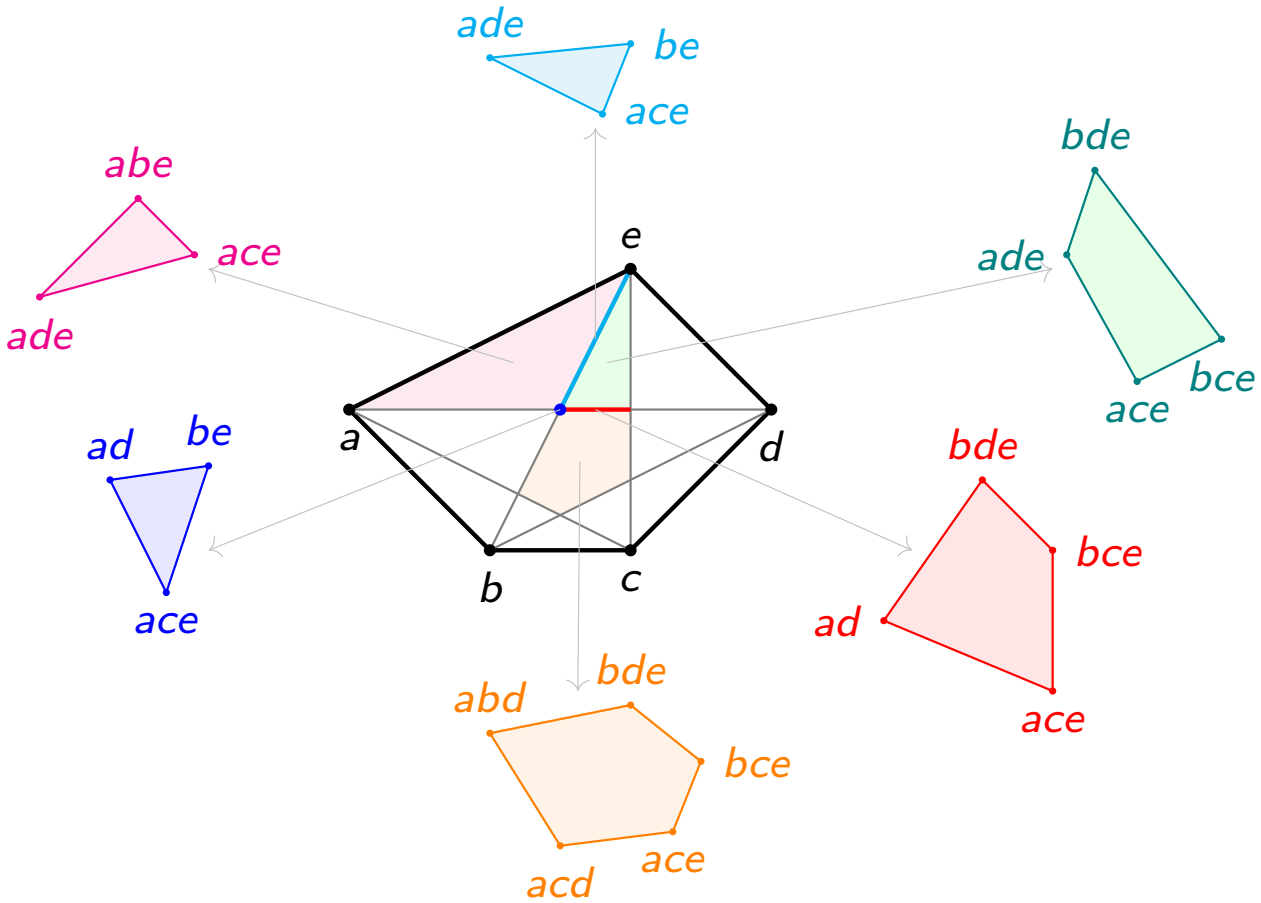
What is the maximum and the maximizers of $\max_{u \in \Delta_{n-1}} \min_{p \in \mathcal{M}} D(u||p)$**?**

In other words, which point in the simplex is the farthest to its MLE?

- problem formulated by Ay '02 when $\mathcal{M}$ is a discrete exponential family
- many information-theoretic results by Ay, Matus, Montufar, Rauh, etc.
- neural networks develop in such a way to maximize the mutual information between the input and output of each layer.

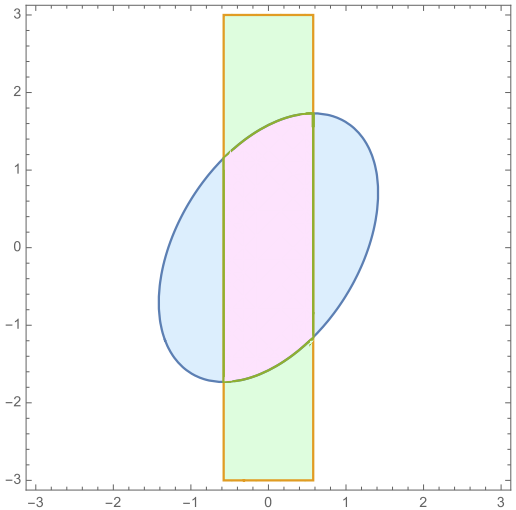
Toric models

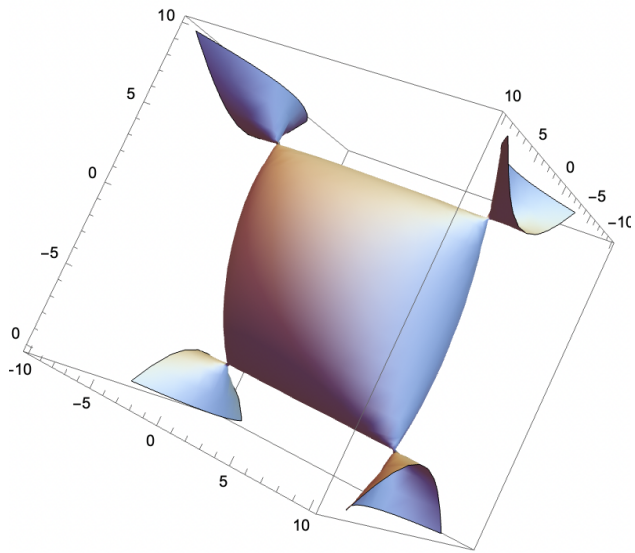
With Serkan Hoşten, we revisit this problem from a new perspective using logarithmic Voronoi polytopes. We present an *algorithm* that combines the combinatorics of the chamber complex with numerical algebraic geometry.



Extensions

- Continuous models (done for Gaussian models with Serkan Hoşten).
- Mixture models and other models with singularities.
  - Statistical disclosure limitation.
  - Data privacy.
  - Study Logarithmic Voronoi cells at singular and boundary points!





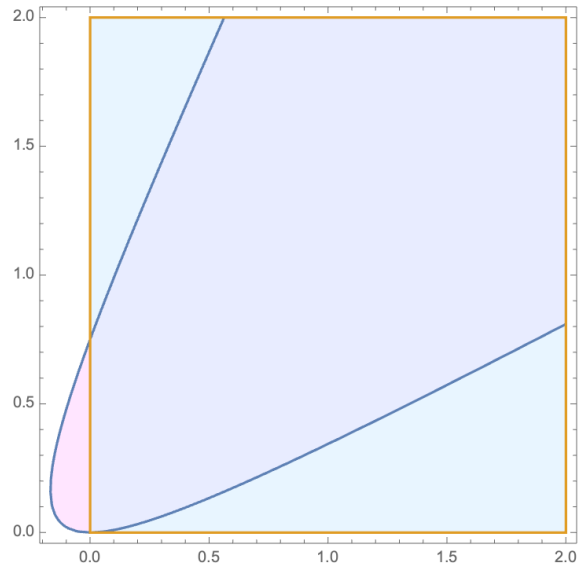
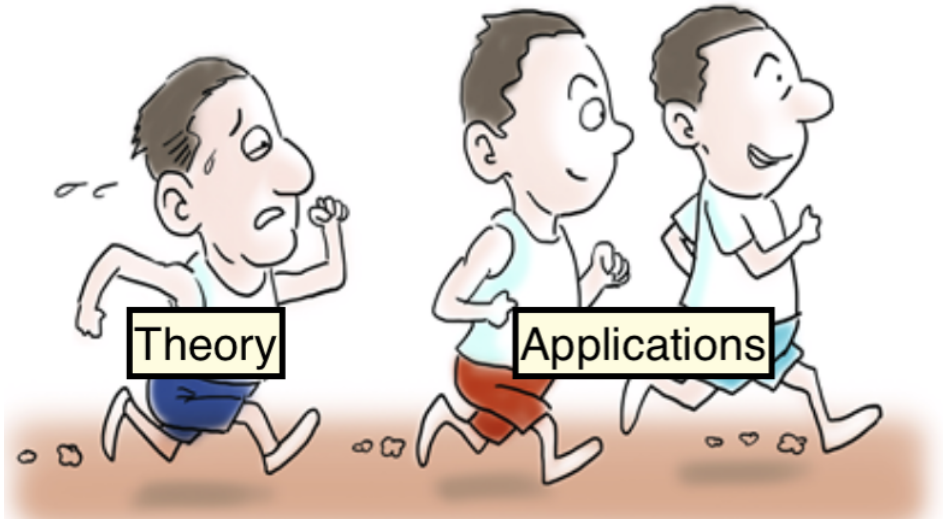


Table of Contents

Algebra in machine learning

- AI is advancing faster than ever before, revolutionizing many fields
- These advances outpace the development of theoretical methods to understand its limits and uses
- Bridging this gap is crucial for ensuring the responsible and effective use of AI

  This is the goal of the mathematical machine learning community.



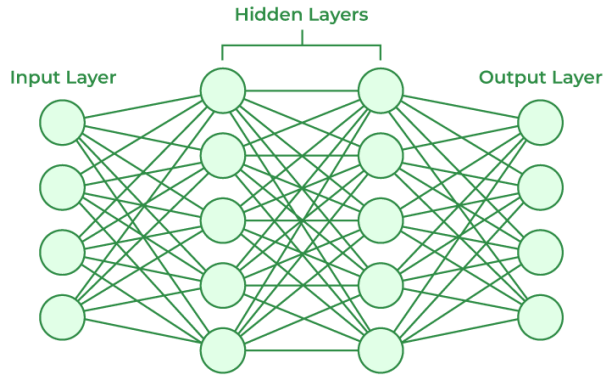
Neural networks

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.



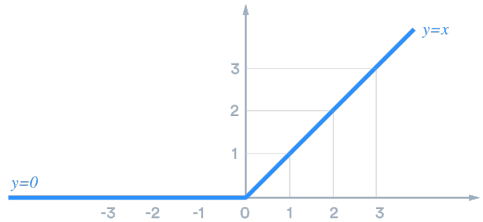
For a dataset $X = [x_1, x_2, \ldots, x_n]$ and unknown parameters θ we are interested in describing the constraints between the coordinates of the array of model outputs $F_X(\theta) = [f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_n)]$.

ReLU networks

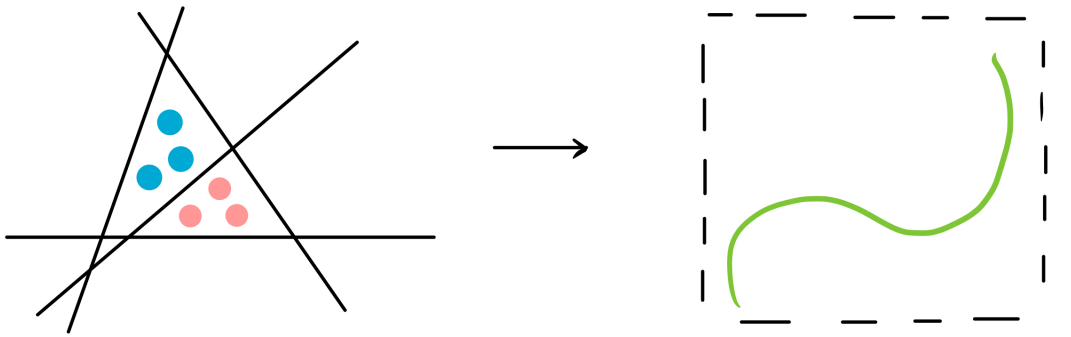
A *ReLU network* is given by the activation function

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

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

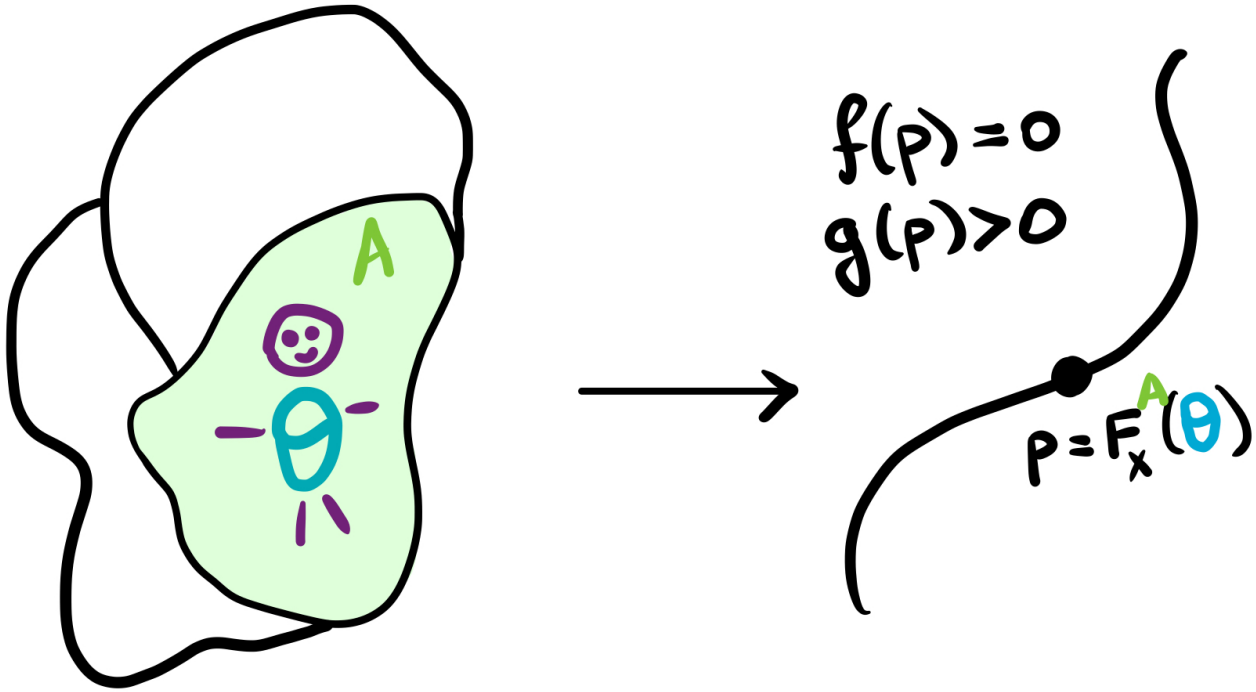


input space parameter space

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.



Model equations

Given a model, parametrized by

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

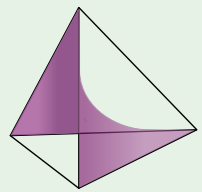
we are interested in describing the polynomials defining $\overline{\text{image}}(\varphi)$. This process is called *implicitization*.

Example (The independence model.)

Parametrization:

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

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*.

Implicitization

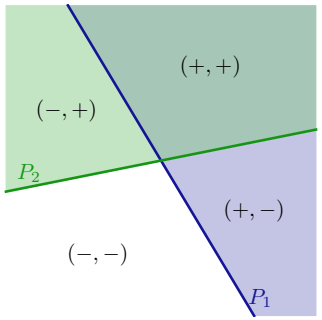
Model invariants capture core properties of the model that are independent of parameterization and remain unchanged under the model's symmetries.

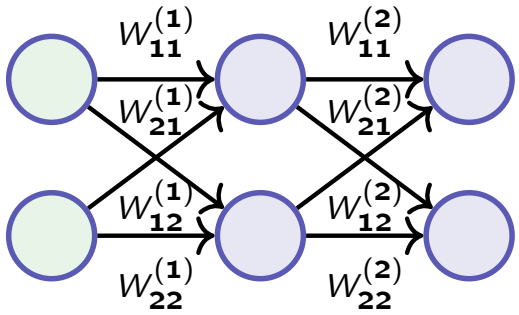
- Identifiability
  - Can model parameters be uniquely determined from observed data?
- Model selection
  - Invariants can serve as useful statistics for testing model fit and constraint-based model selection
- Inference
  - Polynomials encode information about independence
- Model predictions
  - Invariants provide reliable theoretical guarantees
- Neural network verification?

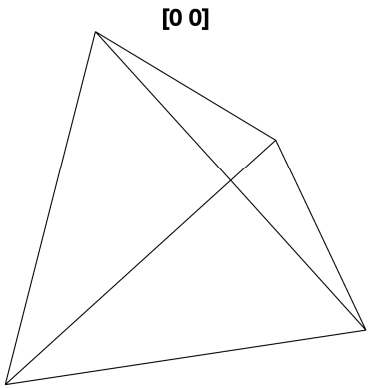
However, implicitization is also **very computationally expensive**!

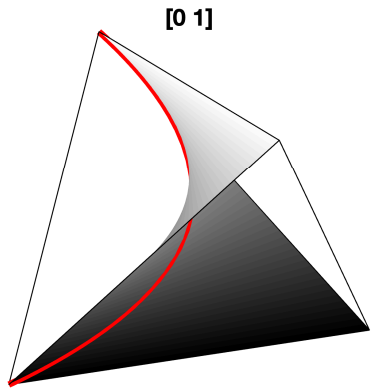
Parametrization

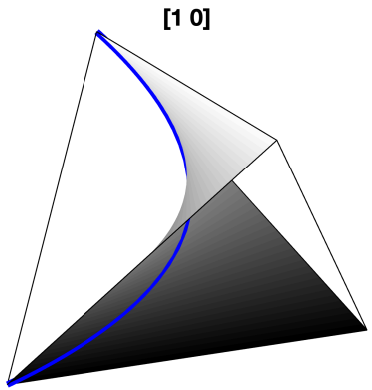
- The number of linear pieces over the input space can be enormous.
- The linear pieces share parameters and are not independent.
- We investigate the relationships between the linear pieces.

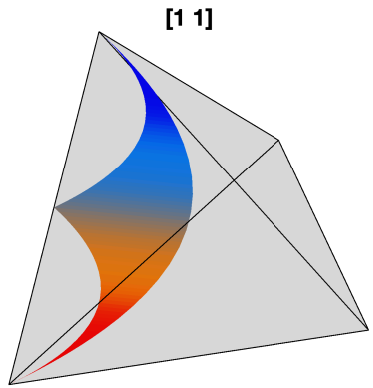






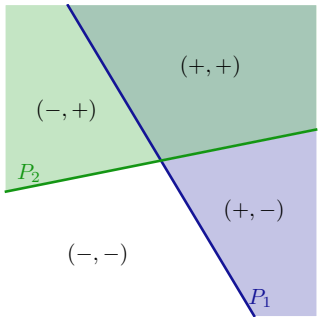


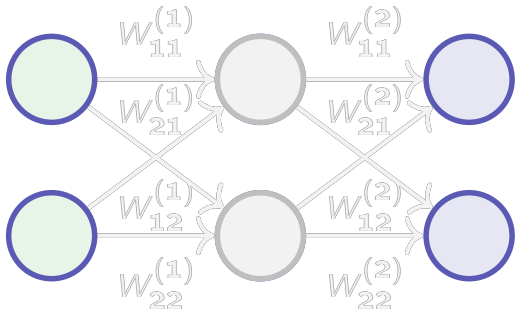


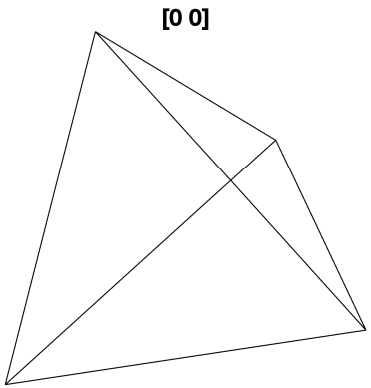


Parametrization

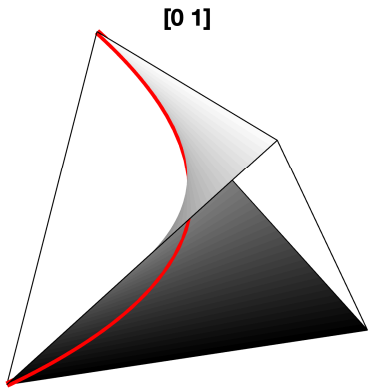
- The number of linear pieces over the input space can be enormous.
- The linear pieces share parameters and are not independent.
- We investigate the relationships between the linear pieces.

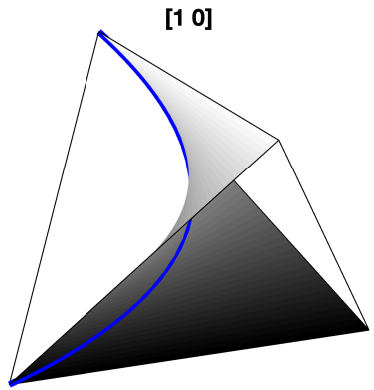


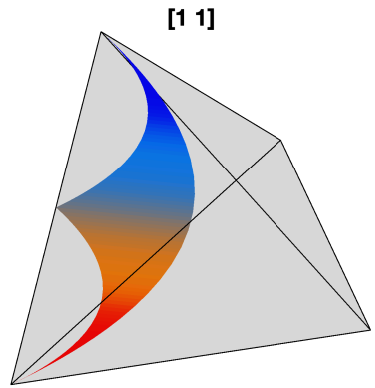




$M_{[00]} = 0$

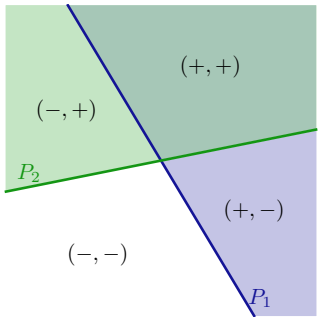


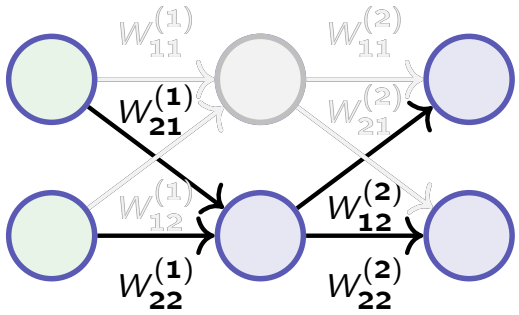


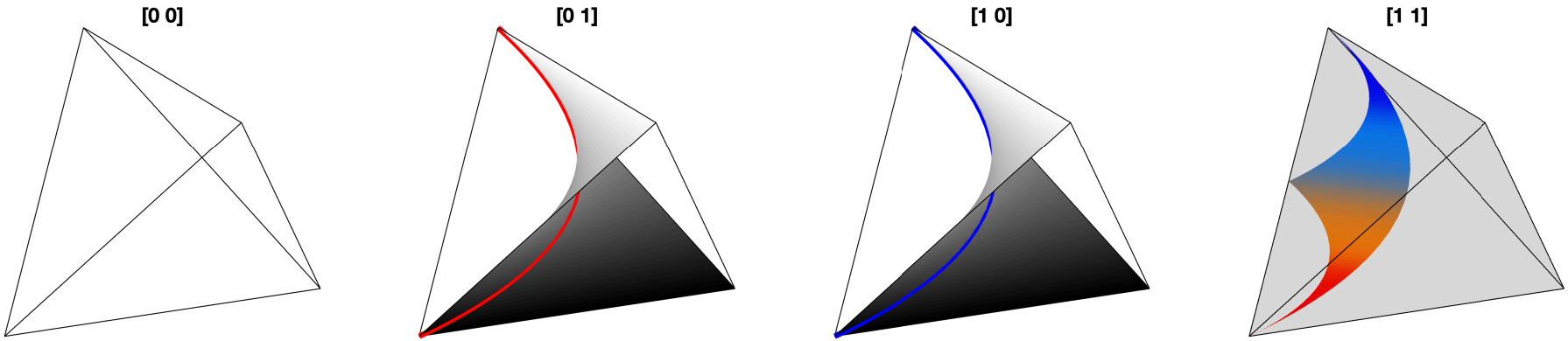


Parametrization

- 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.



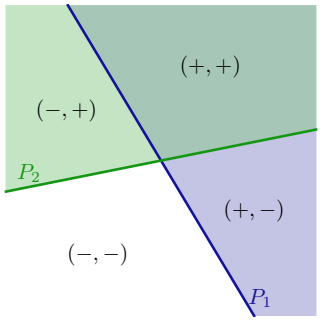


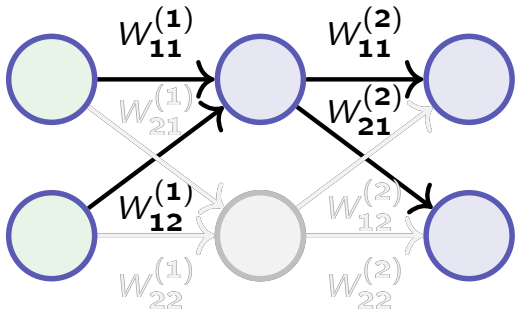


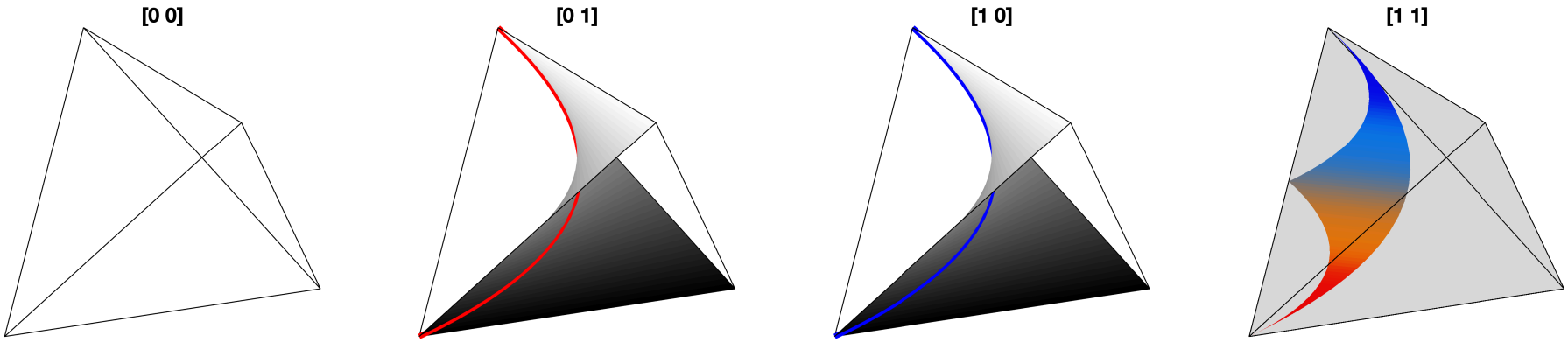
$$M_{[01]} = W^{(2)} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} W^{(1)}$$

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.



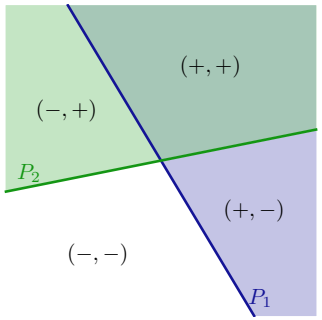


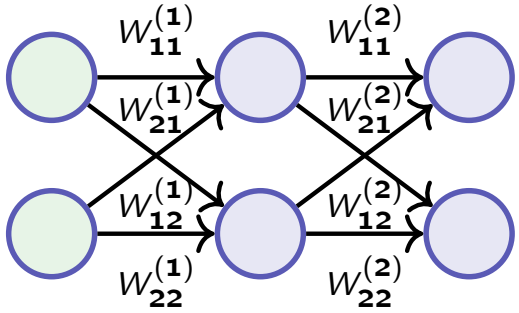


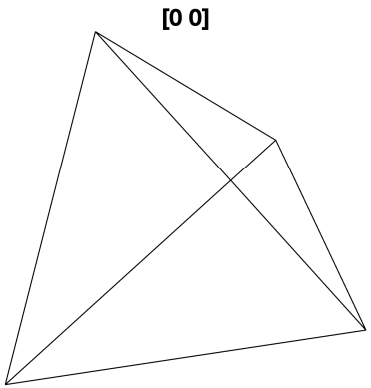
$$M_{[10]} = W^{(2)} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} W^{(1)}$$

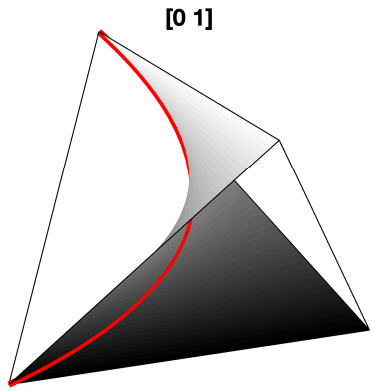
Parametrization

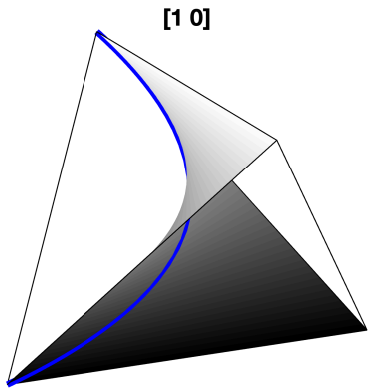
- The number of linear pieces over the input space can be enormous.
- The linear pieces share parameters and are not independent.
- We investigate the relationships between the linear pieces.

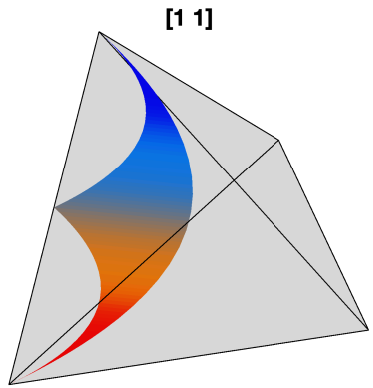












$$M_{[11]} = W^{(2)} W^{(1)}$$

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

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

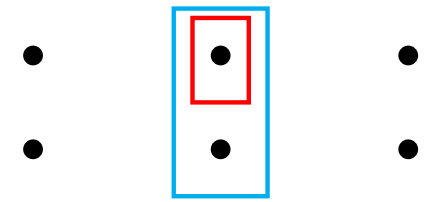
Define the *ReLU output variety* as $\overline{\text{im}(\varphi_X^A)}$. Denote it by V_X^A.

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

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

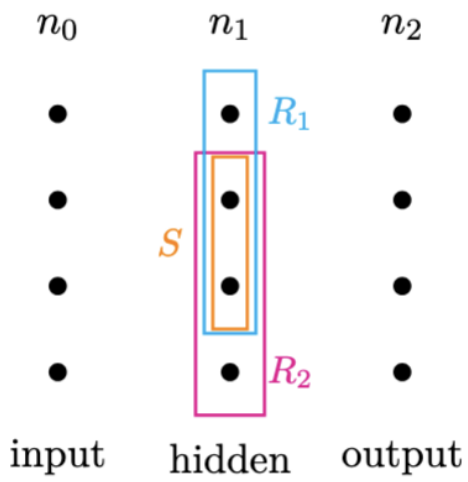
$$M_1(\theta) = \begin{pmatrix} w_{11}^{(1)} w_{11}^{(2)} & w_{12}^{(1)} w_{11}^{(2)} \\ w_{11}^{(1)} w_{21}^{(2)} & w_{12}^{(1)} w_{21}^{(2)} \end{pmatrix}, M_2(\theta) = \begin{pmatrix} w_{11}^{(1)} w_{11}^{(2)} + w_{21}^{(1)} w_{12}^{(2)} & w_{12}^{(1)} w_{11}^{(2)} + w_{22}^{(1)} w_{12}^{(2)} \\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

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

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

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 $|R_1| = r_1$, $|R_2| = r_2$, $|S| = s$.
Let $t = r_1 + r_2 - 2s$.

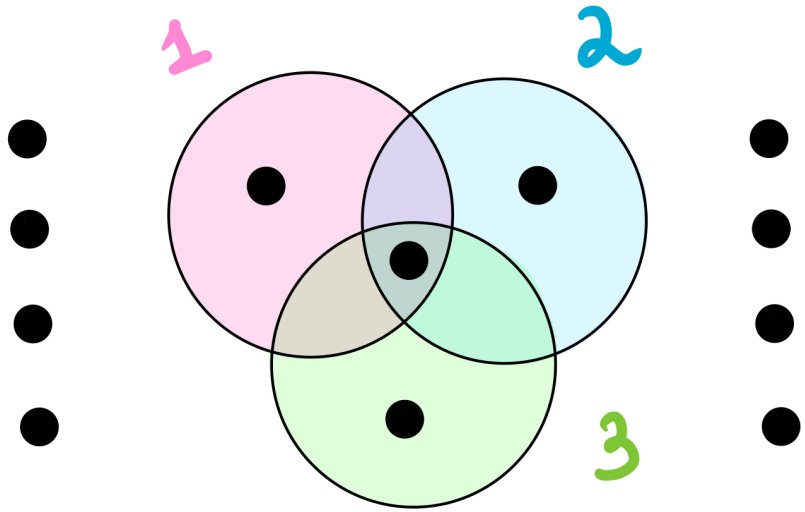
Theorem (A.-Montúfar, 2025+)

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.

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

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

Thank you!

Questions?