Algebraic and combinatorial geometry in statistics and machine learning

Yulia Alexandr (UCLA)

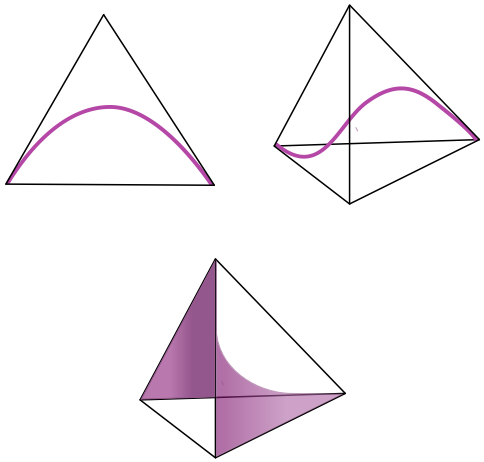
The University of Melbourne
January 21, 2026

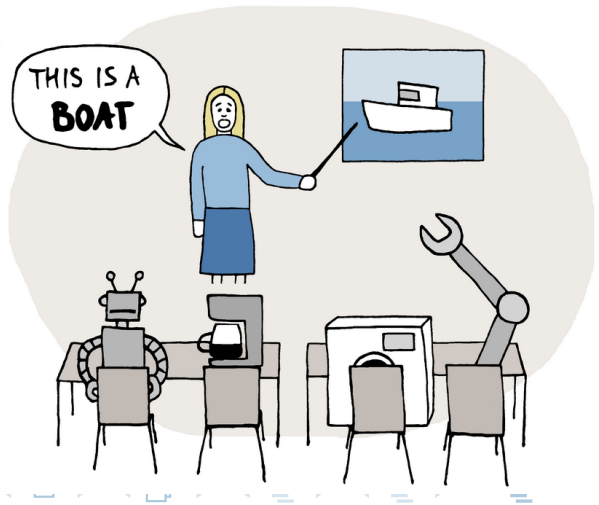
Overview

I use *algebraic geometry* and *combinatorics* to study the structure and properties of models that arise in applications.

- Statistics
  - Data privacy
  - Causal inference
  - Information geometry
  - Phylogenetics

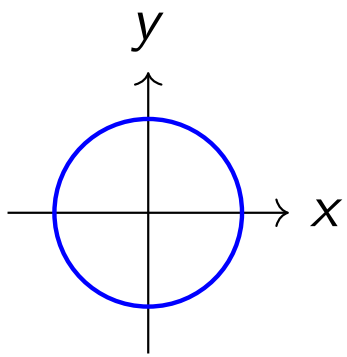
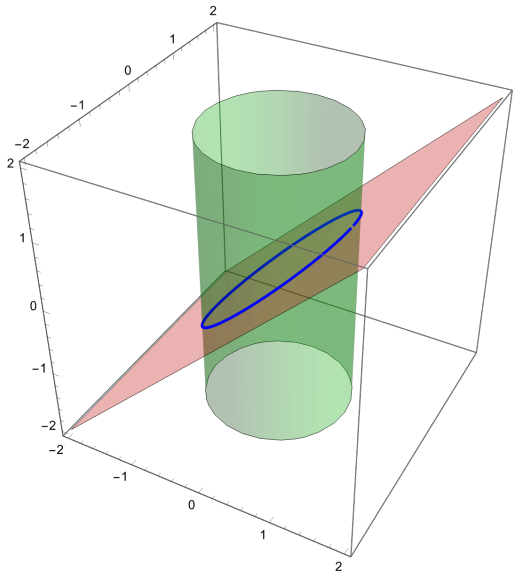
- Machine learning
  - Guarantees for model outputs
  - Neural network verification
  - Safety-critical and trustworthy AI





Models as algebraic varieties

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

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

Part 1: Algebraic statistics

The independence model

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

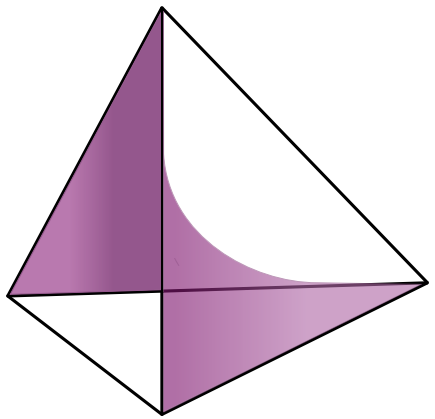
Example

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

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

Recall that X_1 and X_2 are *independent* if

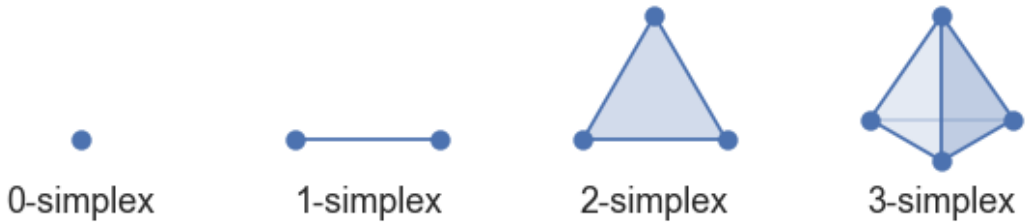
$$\mathbb{P}(X_1 = i, X_2 = j) = \mathbb{P}(X_1 = i) \cdot \mathbb{P}(X_2 = j).$$



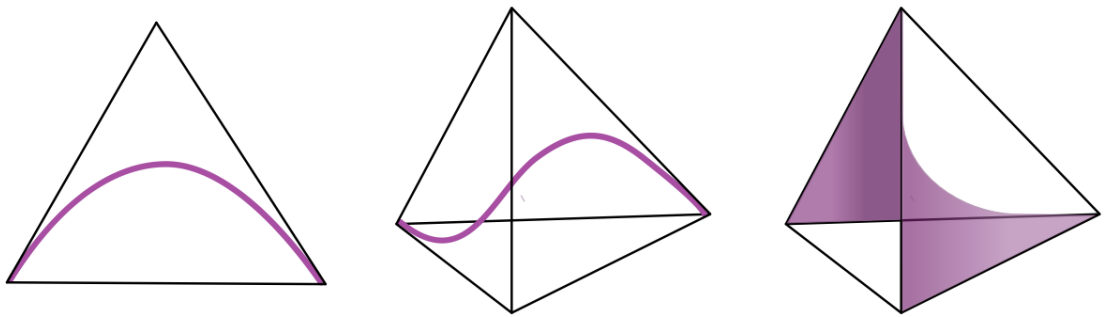
Statistical models

A *probability simplex* is defined as

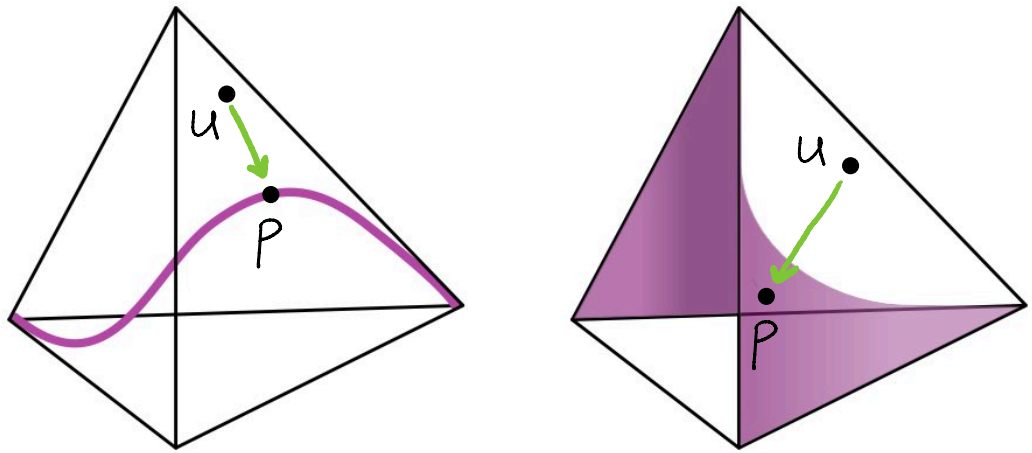
$$\Delta_{n-1} = \{(p_1, \ldots, p_n) : p_1 + \cdots + p_n = 1, p_i \geq 0 \text{ for } i \in [n]\}.$$



An *algebraic statistical model* is an intersection of some variety $\mathcal{V}$ with the probability simplex.



Maximum likelihood estimation



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

For an empirical distribution $u = (u_1, ..., u_n) \in \Delta_{n-1}$, 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? 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? 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)

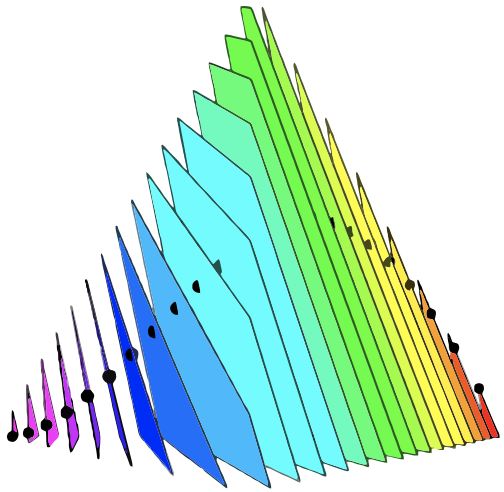
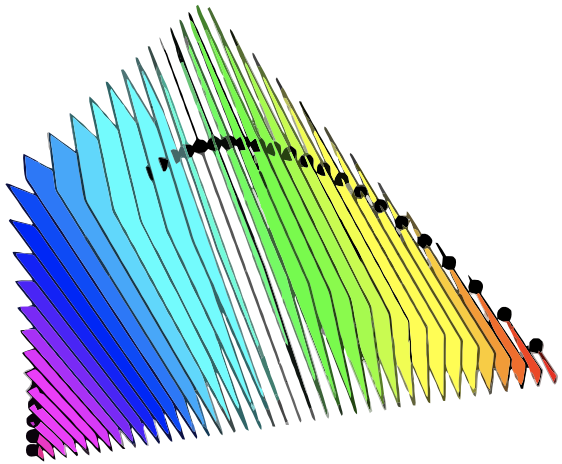
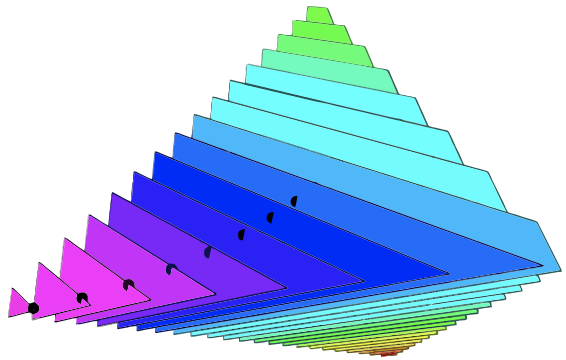
Logarithmic Voronoi cells are convex sets.

Theorem (A.-Heaton)

If $\mathcal{M}$ is a finite model, a linear model, a toric model, or a model of ML degree 1, the logarithmic Voronoi cell at any point $p \in \mathcal{M}$ is a polytope.

Example (The twisted cubic.)

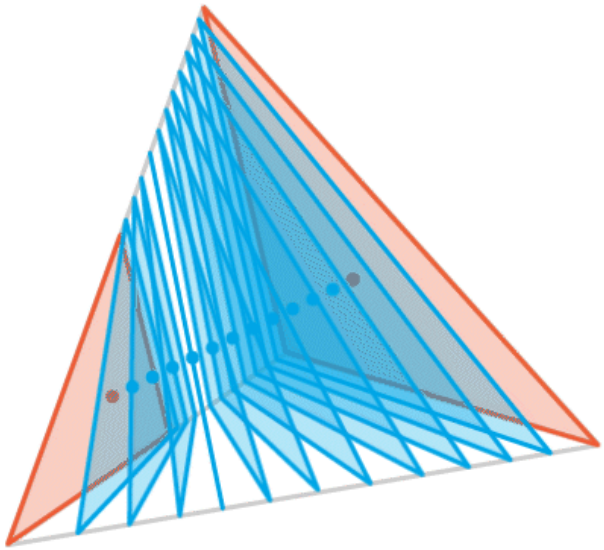
The curve is given by $\theta \mapsto \left(\theta^3, 3\theta^2(1-\theta), 3\theta(1-\theta)^2, (1-\theta)^3\right)$.

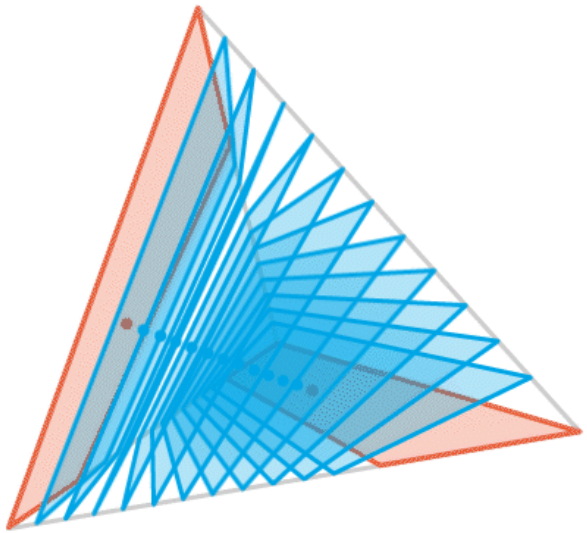




Linear models

Theorem (A.)

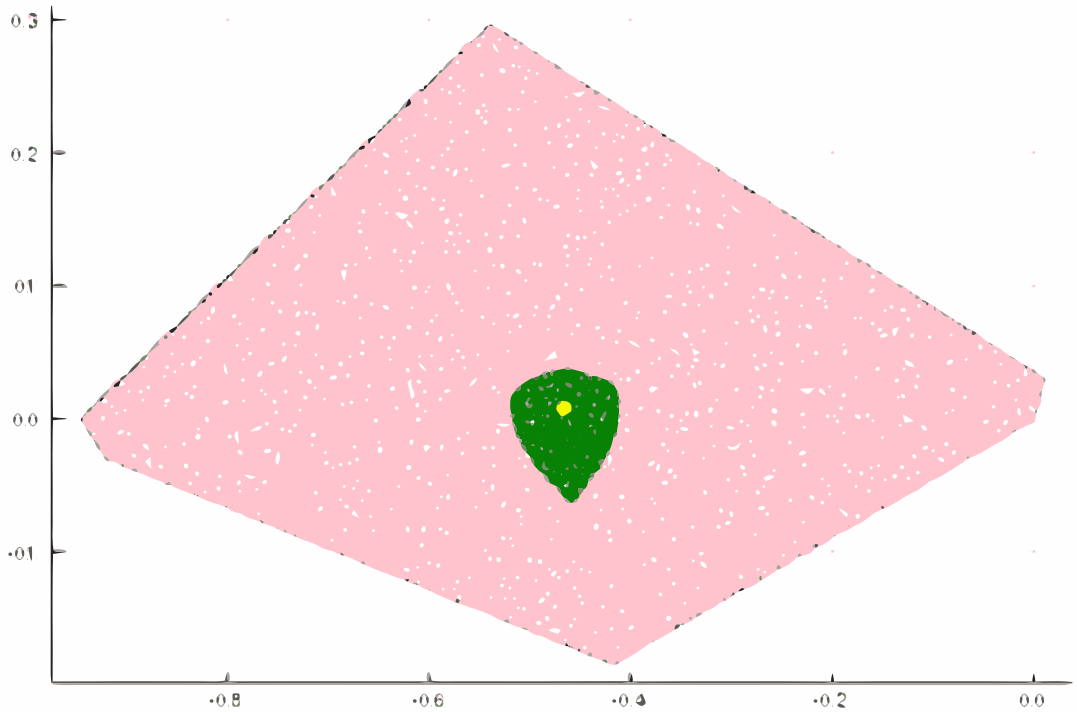
For linear models, logarithmic Voronoi cells at all interior points on the model have the same combinatorial type. This type can be described via Gale diagrams.





Why do we care?

- Data privacy, especially for statistical disclosure limitation
  - Logarithmic Voronoi cells at singular and boundary points of the model
- Estimate sensitivity to data perturbation
  - Boundaries of logarithmic Voronoi cells



Maximizing divergence

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

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

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

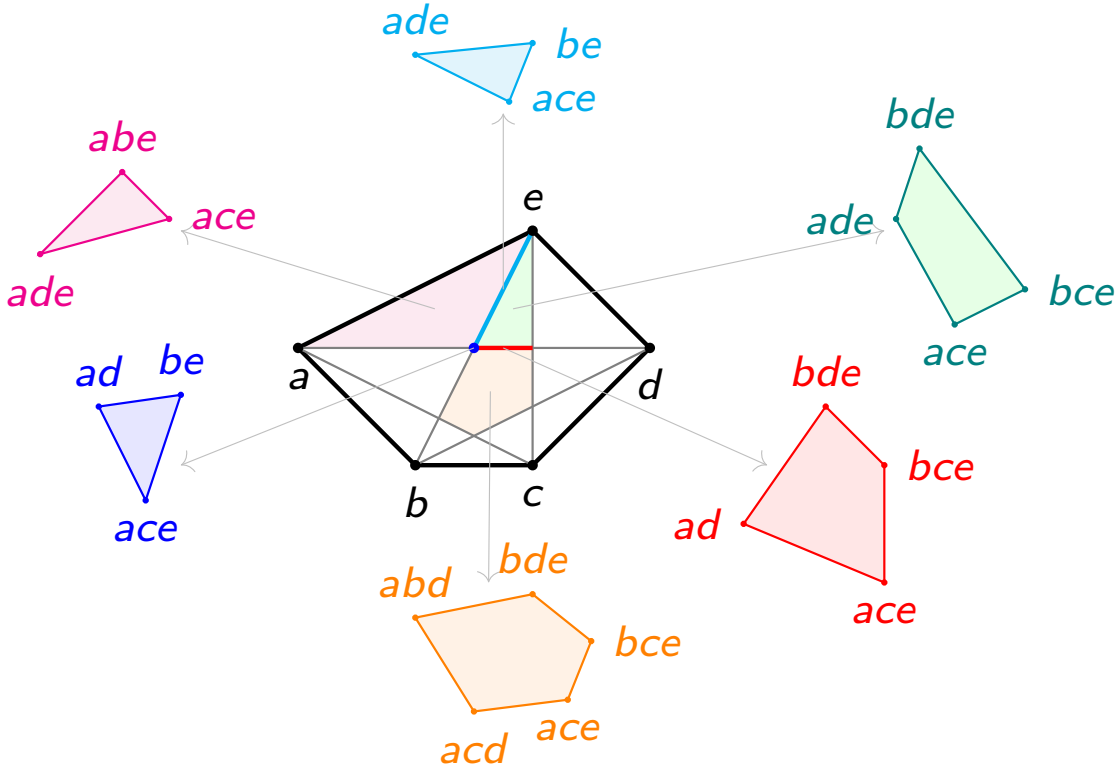
What is the maximum and the maximizers of $\max_{u \in \Delta_{n-1}} \min_{p \in \mathcal{M}} D(u||p)$**?**

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

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

Toric models

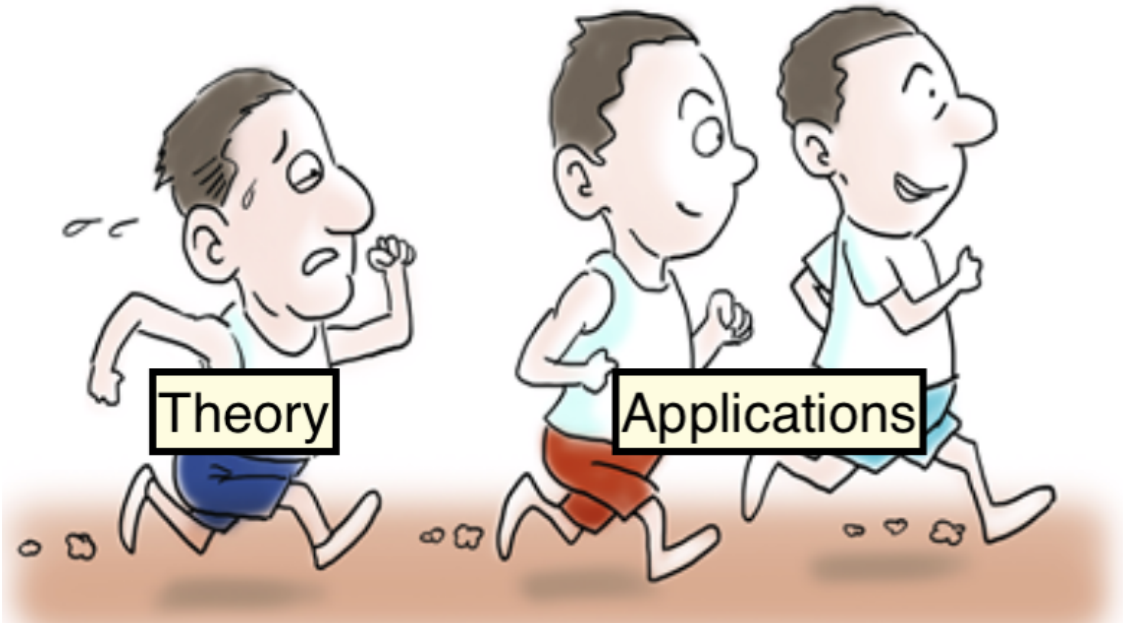
In joint work with Serkan Hoşten, we give an *algorithm* to compute the maximum and its maximizers.



Theorem (A.-Hoşten)

Logarithmic Voronoi cells at points in the same chamber have the same combinatorial type and the same vertex supports.

Part 2: Algebraic machine learning



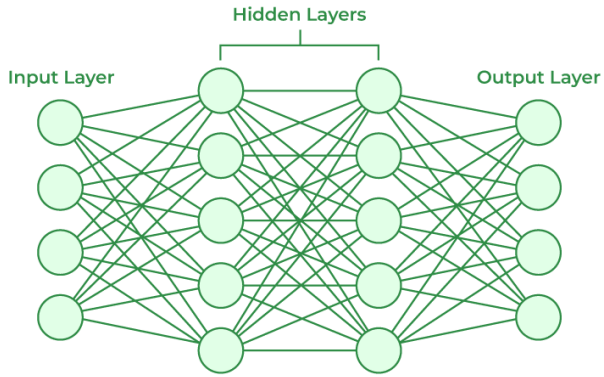
Neural networks

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

- $n_0 =$ dimension of the input space
- $n_\ell =$ width of the ℓth layer



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

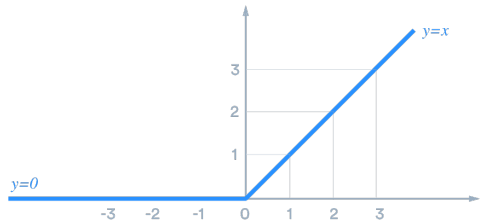
$$F_X(\theta) = [f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_n)].$$

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.



- $f_\theta(x)$ is piece-wise linear
  $\Longrightarrow$ subdivision of the input space into linear regions

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

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

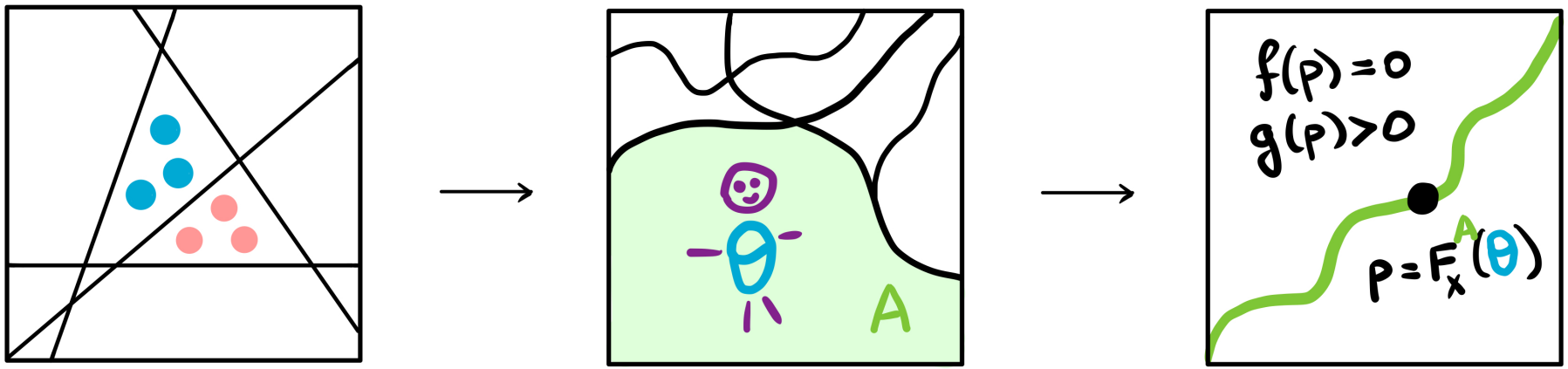
The main question

Problem

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

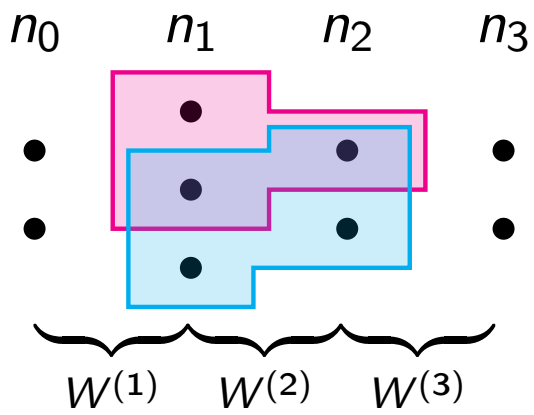
$$\varphi_X^A : \theta \mapsto F_X^A(\theta)$$

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



input space parameter space prediction space

The parametrization



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

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

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

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

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

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

$$\varphi_X^A : \theta = (W^{(1)}, W^{(2)}, W^{(3)}) \mapsto \left[M_1(\theta) \begin{bmatrix} x_1 & x_2 \end{bmatrix} \,\middle|\, M_2(\theta) \begin{bmatrix} x_3 & x_4 \end{bmatrix} \right]$$

The parametrization

In general, for k blocks:

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

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

$$\boxed{\varphi^{\mathsf{A}} : \theta \mapsto [\, M_1(\theta) \mid M_2(\theta) \mid \cdots \mid M_k(\theta) \,].}$$

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

Proposition (A.-Montúfar)

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

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

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

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

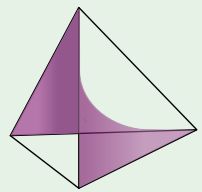
we are interested in describing the polynomials defining $\overline{\text{image}}(\varphi)$. This process is called *implicitization*.

Example (The independence model.)

Parametrization:

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

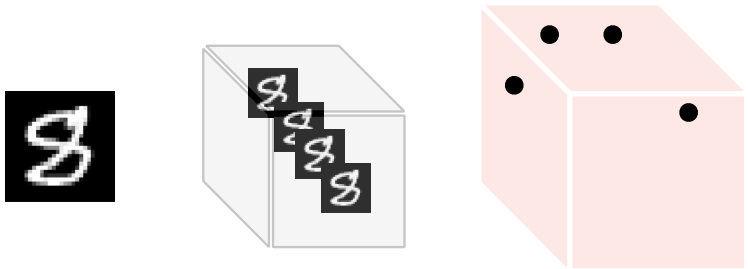
Implicit ideal: $I = \langle p_1 p_4 - p_2 p_3, \;\; p_1 + p_2 + p_3 + p_4 - 1 \rangle$.

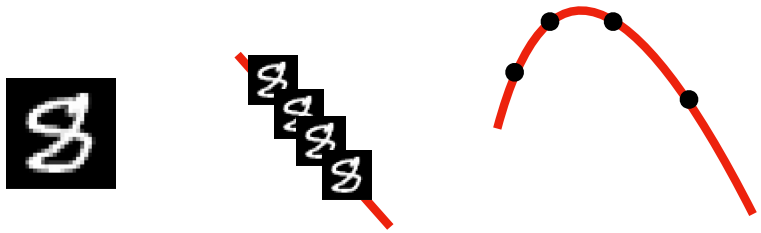


The generators of the ideal I are called *model invariants*.

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 $V_X^A \cap S = \varnothing$, 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

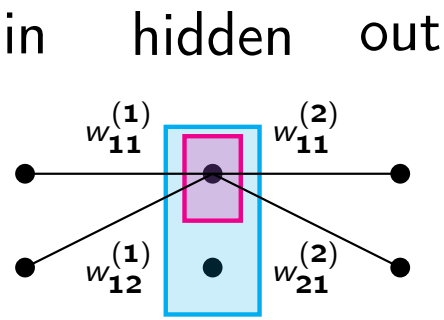




Example: 2 blocks

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

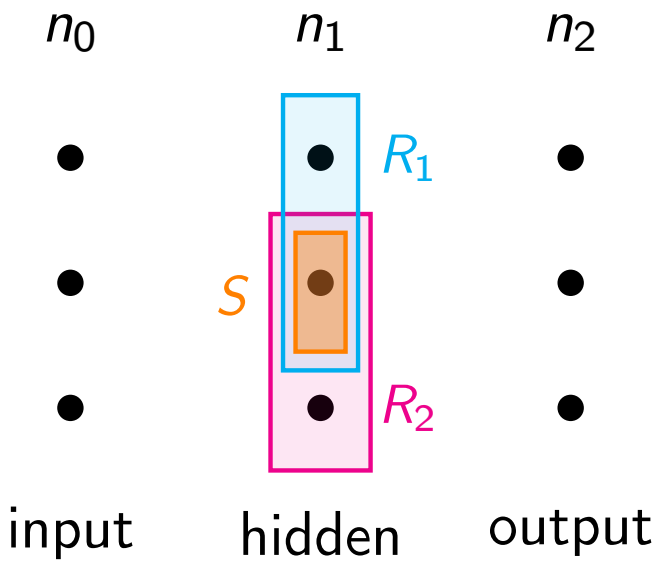
$$M_1(\theta) = \begin{pmatrix} w_{11}^{(1)} w_{11}^{(2)} & w_{12}^{(1)} w_{11}^{(2)} \\ w_{11}^{(1)} w_{21}^{(2)} & w_{12}^{(1)} w_{21}^{(2)} \end{pmatrix}, M_2(\theta) = \begin{pmatrix} w_{11}^{(1)} w_{11}^{(2)} + w_{21}^{(1)} w_{12}^{(2)} & w_{12}^{(1)} w_{11}^{(2)} + w_{22}^{(1)} w_{12}^{(2)} \\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

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

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

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)

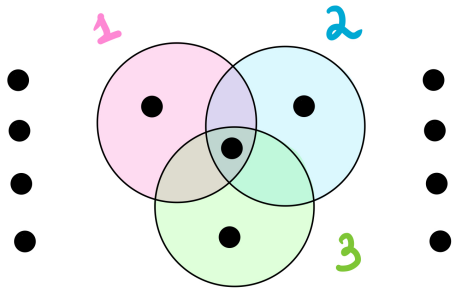
The ideal J^{A} contains:

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.

Generalizations

- Two blocks (deep networks):
  - similar to the shallow case, but now we count shared *paths*.
- Multiple blocks:
  - rank conditions appear on linear combinations $\sum_i \lambda_i M_i$,
  - and on block matrices built from these combinations.



2,256 invariants of degree 3, 4, and 5!

- Dimension:
  - expected dimension holds when n_1 is small

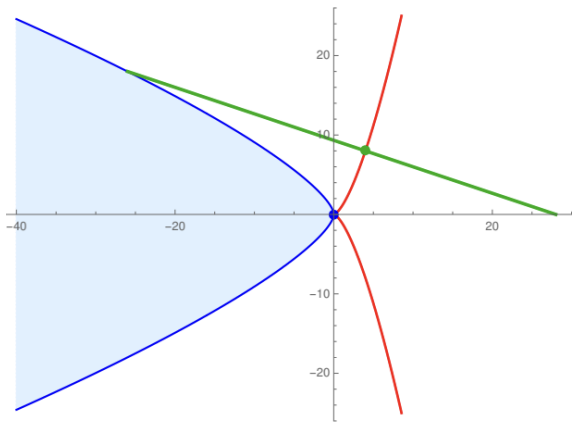
Open problems:

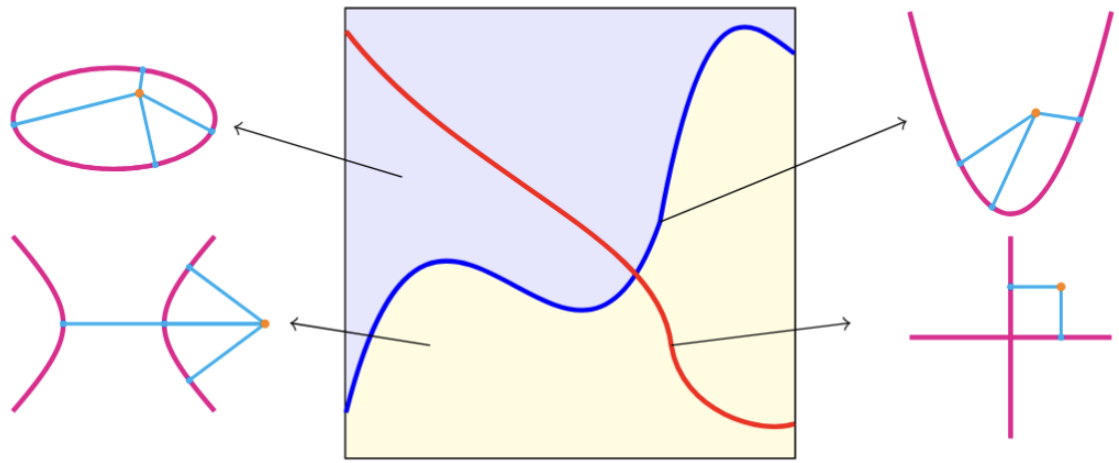
- Sufficiency of generators for the shallow case
- Finite representation for deep networks with multiple blocks
- Inequalities defining the image

Research program

Applied algebraic geometry, combinatorics, and optimization:

- Geometry in statistical inference
  - Describe (logarithmic) Voronoi cells at singular points
  - Use matroid theory to study models with hidden variables
- Algebraic foundations of deep learning
  - Derive invariants for self-attention mechanisms
  - Develop algebraic methods for robustness verification
- Impact and mentorship
  - Expand algebraic methods into [your favorite field]
  - Build a research group bridging theory and real-world impact





Thank you! Questions?