Can algebra be applied?

Yulia Alexandr (UCLA)

UMSA Professor Talk
March 3, 2025

About me

I am a postdoc in the math department at UCLA.
Before this, I was...

- A math PhD student at UC Berkeley (2023).
  My thesis was in *algebraic statistics*.
- A math undergrad at Wesleyan University (2019).
  My undergrad thesis was in *algebraic graph theory*.
- An undergrad at Kingsborough Community College.

As an undergradudate, I participated in three REUs:

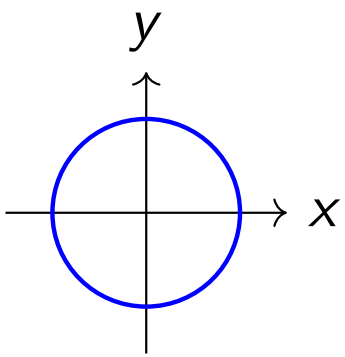
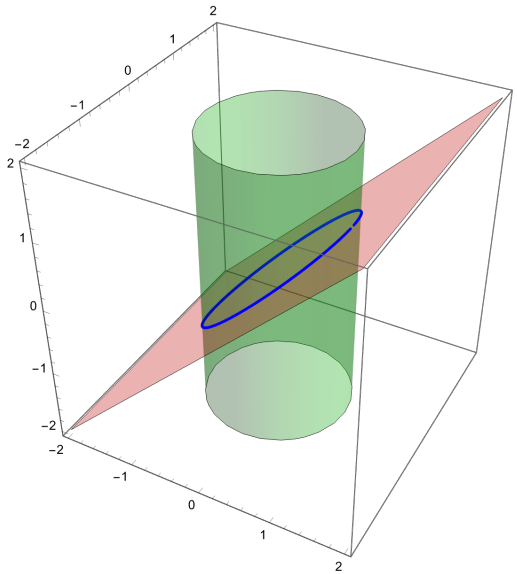
- Treespace REU at Lehman College (CUNY)
- DIMACS REU at Rutgers University
- Twin Cities REU at the University of Minnesota

Research

- My research is in *algebraic statistics* and *algebraic machine learning*.
  - I use algebraic, geometric, and combinatorial techniques to tackle problems in statistics and machine learning.
- Pure math? Applied math? Whatever the grant proposal calls for.
- I write code to generate examples and form conjectures. Then I try to prove theorems. Sometimes computational results are OK.
- Being able to code well is a useful skill in my research field.
  Some programming languages I use:
  - `Macaulay2`, `Singular`, `Bertini` (very niche!)
  - `Julia`, `Python`, `SageMath`
  - `Mathematica`, occasionally

Algebraic varieties

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

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

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

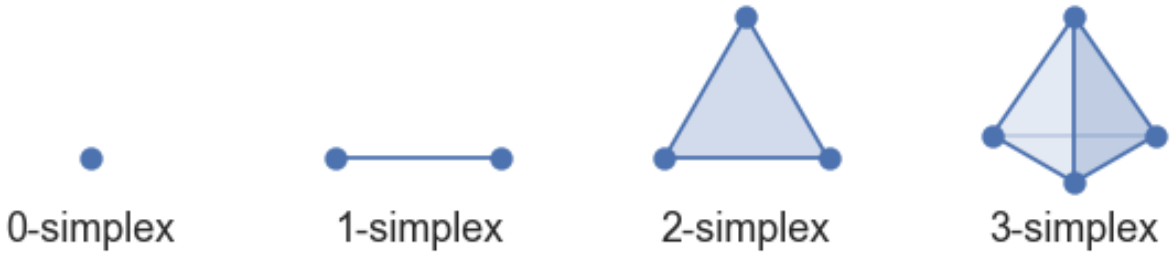
$$\mathbb{P}(X_1 = i | X_2 = j) = \mathbb{P}(X_1 = i) \cdot \mathbb{P}(X_2 = j).$$

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 $\mathcal{V} \subseteq \mathbb{C}^n$.

Model equations

Given a model, parametrized by

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

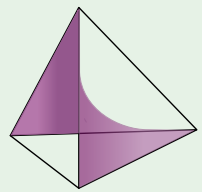
we are interested in describing the polynomials defining $\overline{\text{image}}(\varphi)$. This process is called *implicitization*.

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

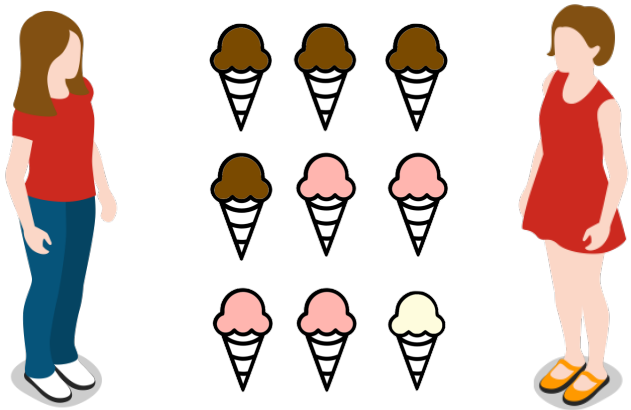
Model invariants capture core properties of the model that are independent of parameterization and remain unchanged under the model's symmetries.

- 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

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

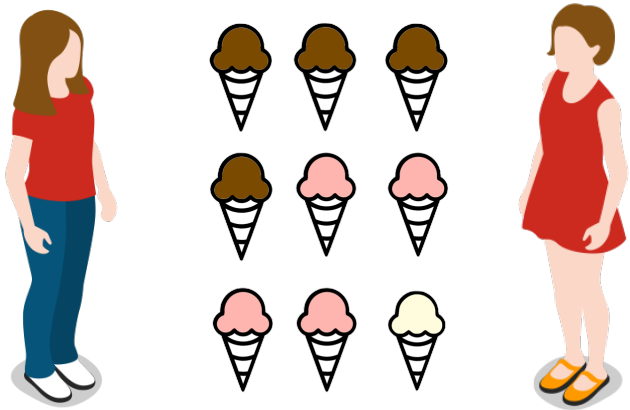
Maximum likelihood estimation

(p_1, p_2, p_3)



Maximum likelihood estimation

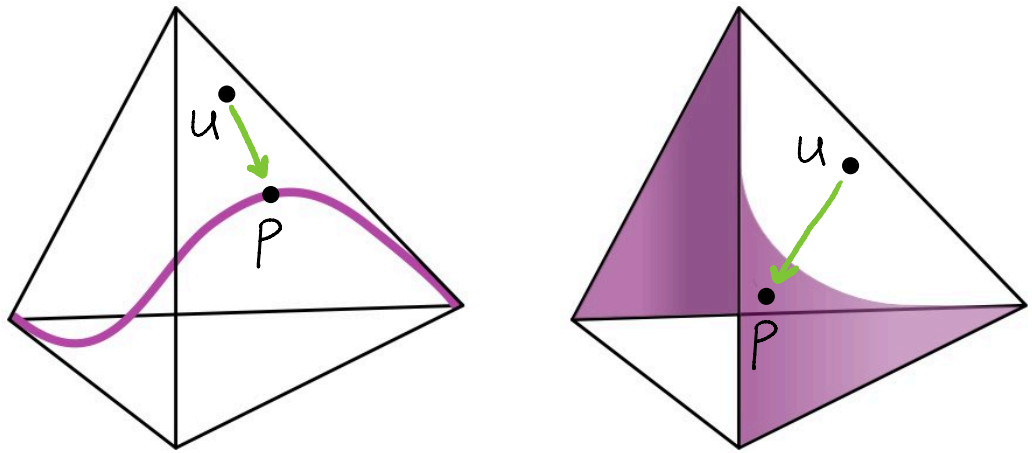
(p_1, p_2, p_3)



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

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

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 $\mathcal{M}$?

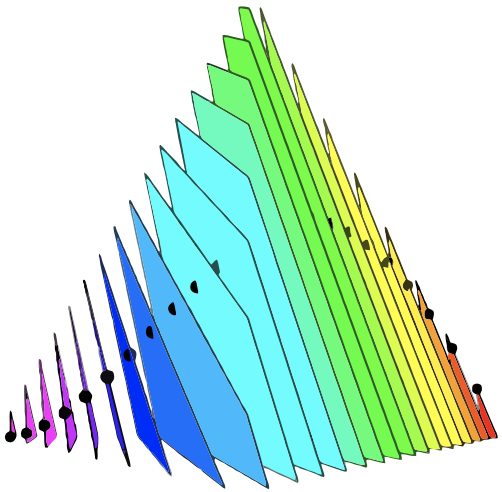
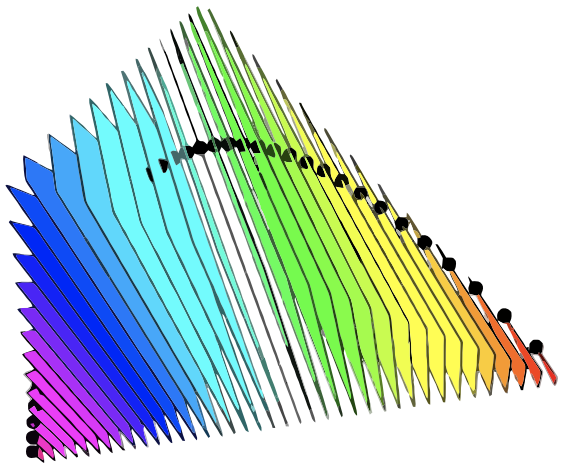
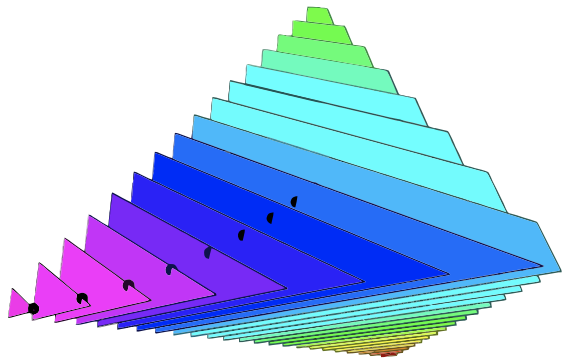
The set of all such elements $u \in \Delta_{n-1}$ is the *logarithmic Voronoi cell* at q.

Logarithmic Voronoi cells

- Logarithmic Voronoi cells are always convex sets.
- For nice statistical models, they are **polytopes**!
- They divide experimental data based on which point in the statistical model each sample most likely came from.
- They partition the probability simplex.

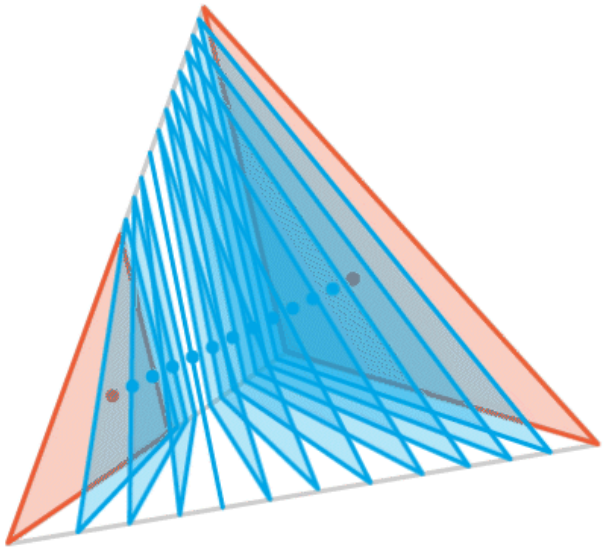
Example (The twisted cubic.)

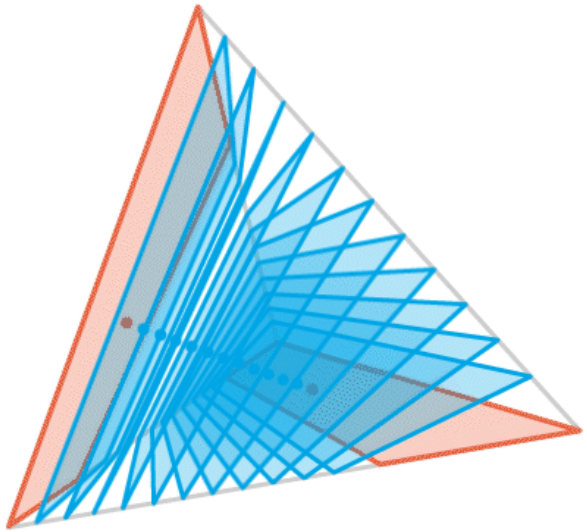
The curve is given by $\theta \mapsto (\theta^3, 3\theta^2(1-\theta), 3\theta(1-\theta)^2, (1-\theta)^3)$.

Linear models

For linear models, logarithmic Voronoi cells at all interior points on the model have the same combinatorial type.



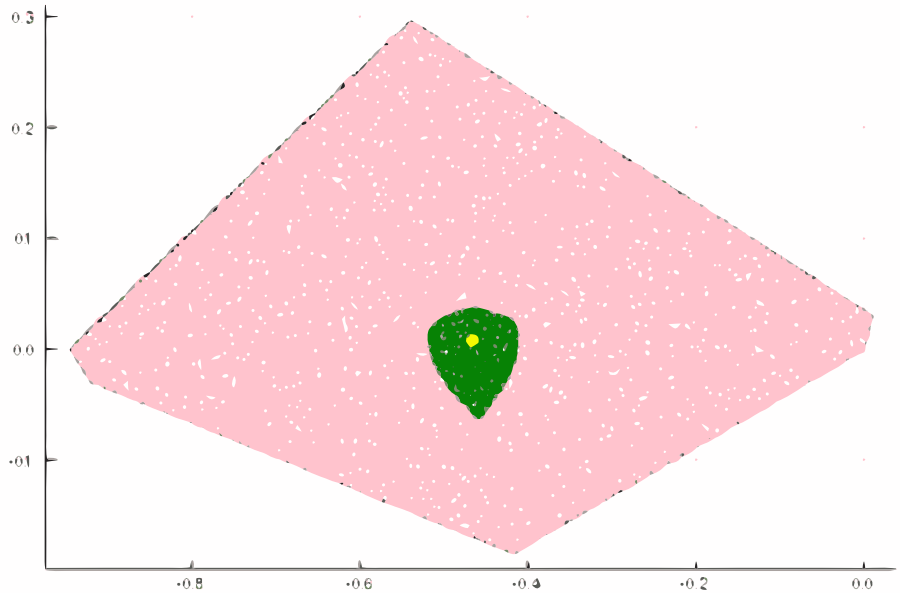


Why do we care?

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.

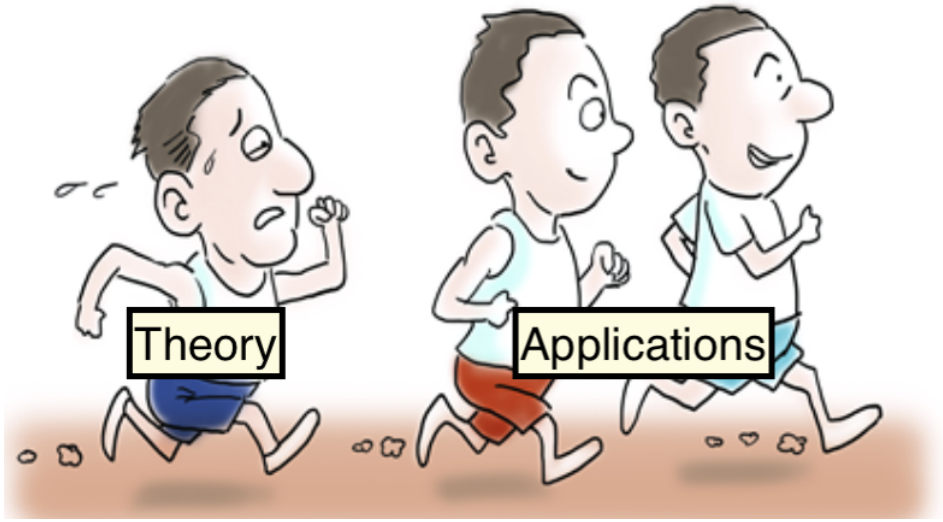
For models with complicated geometry, numerical methods are necessary to analyze logarithmic Voronoi cells.



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.



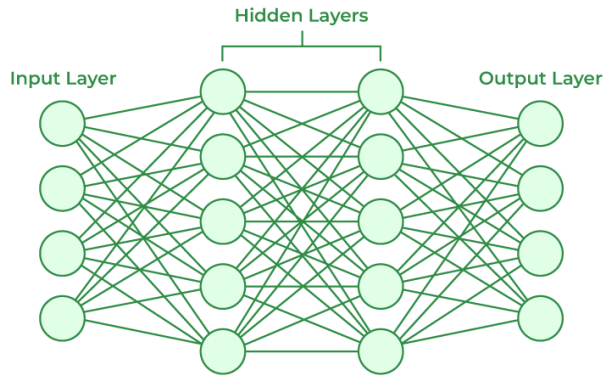
Neural networks

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.



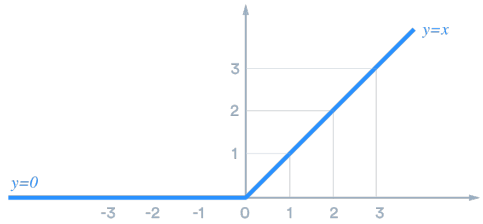
For a dataset $X = [x_1, x_2, \ldots, x_n]$ and unknown parameters θ we are interested in describing the constraints between the coordinates of the array of model outputs $F_X(\theta) = [f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_n)]$.

ReLU networks

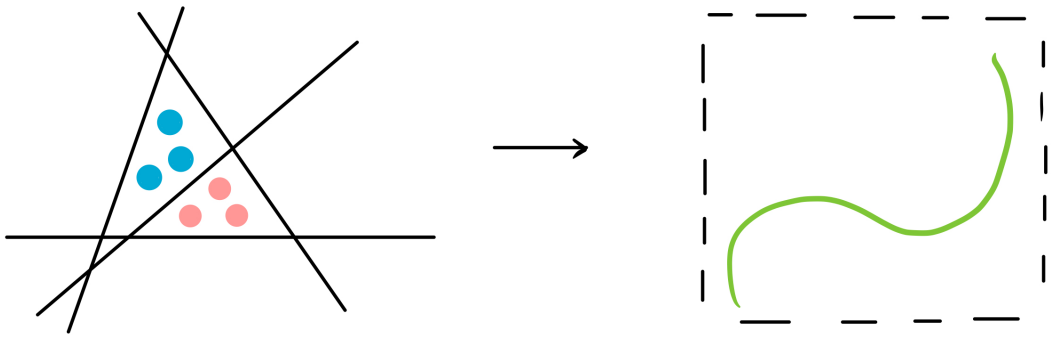
A *ReLU network* is given by the activation function

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

at each layer of the neural network.



- this makes $f_\theta(x)$ is 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 $X = [x_1, \ldots, x_n]$
  - this subdivision extends to the parameter space
  - $F_X(\theta)$ is multi-linear in θ in each activation region

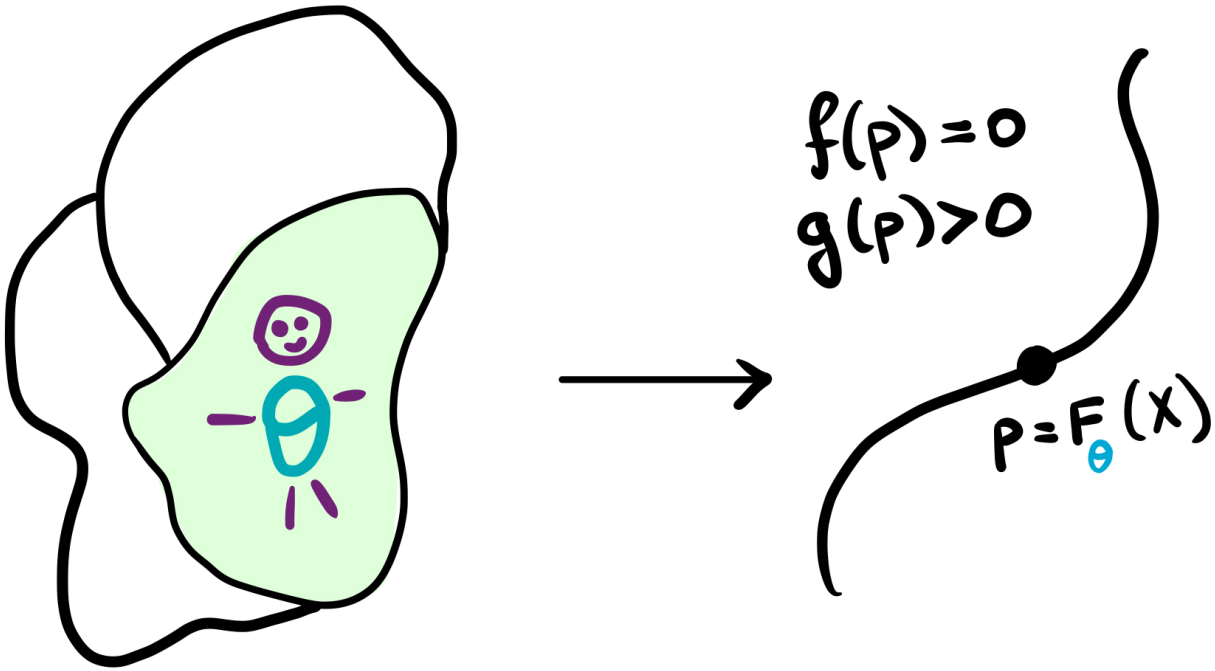


In general...

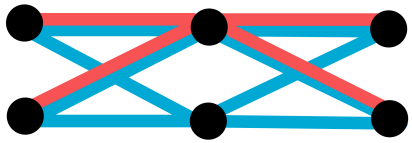
Together with Guido Montúfar, we are working on the following problem.

Problem

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



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 $(1, 1)$.

Then the image of $F_X(\theta)$ is parametrized as $[M_1 X_1 \mid M_2 X_2]$ where

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

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

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

The ideal of invariants of $\theta \mapsto [M_1 X_1 \mid M_2 X_2]$ can be obtained from I_M in terms of fixed but arbitrary data X_1 and X_2!

Implications

Our research suggests that for a given ReLU network we could:

- establish general theorems regarding its invariants
- characterize these invariants for general data in terms of the data itself
- determine dimensions and degrees of their corresponding ideals

Another perspective

Problem

Study the sets of all datasets X that map to a particular output $F_X(\theta)$ for some optimal θ.

- *what should "optimal" mean?*
- *how can we describe these ReLU Voronoi cells?*

Thank you!

Questions?