Neuromanifolds

The Geometry of Attention Networks and Polynomial Networks

Kathlén Kohn

based on joint works with

Joan Bruna — NYU

Nathan Henry — Univ. of Toronto

Giovanni Marchetti — KTH

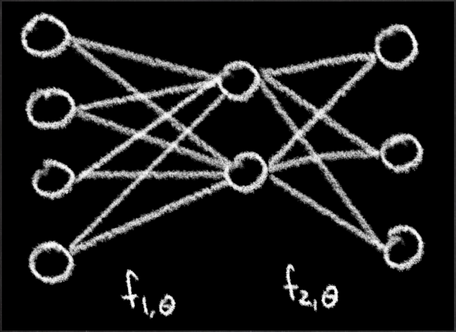
Stefano Mereta — KTH

Guido Montúfar — UCLA, MPI MiS Leipzig

Vahid Shahverdi — KTH

Matthew Trager — Amazon

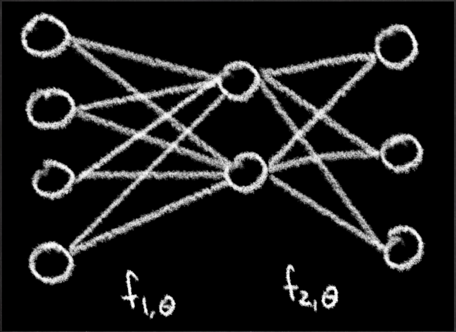
feedforward neural networks



are parametrized families of functions

$$\begin{aligned}\mu : \mathbb{R}^N &\longrightarrow \mathcal{M}, \\ \theta &\longmapsto f_{L,\theta} \circ \ldots \circ f_{1,\theta}\end{aligned}$$

feedforward neural networks

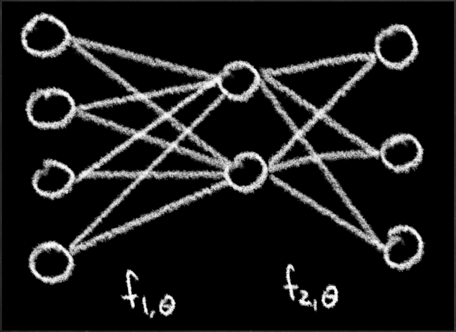


are parametrized families of functions

$$\begin{aligned} \mu : \mathbb{R}^N &\longrightarrow \mathcal{M}, \\ \theta &\longmapsto f_{L,\theta} \circ \ldots \circ f_{1,\theta} \end{aligned}$$

$L = \#$ layers, $f_{i,\theta} = (\sigma_i, \ldots, \sigma_i) \circ \alpha_{i,\theta}$,

feedforward neural networks

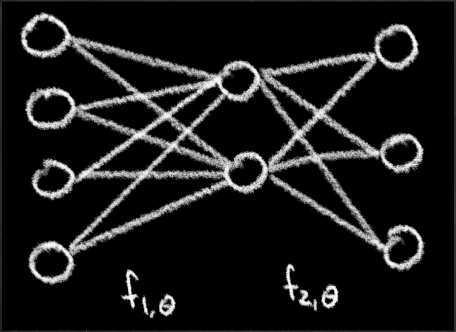


are parametrized families of functions

$$\begin{aligned}\mu : \mathbb{R}^N &\longrightarrow \mathcal{M}, \\ \theta &\longmapsto f_{L,\theta} \circ \ldots \circ f_{1,\theta}\end{aligned}$$

$L = \#$ layers, $f_{i,\theta} = (\sigma_i, \ldots, \sigma_i) \circ \alpha_{i,\theta}$,
$\sigma_i : \mathbb{R} \to \mathbb{R}$ activation, $\alpha_{i,\theta}$ affine linear

feedforward neural networks



$\mathcal{M} = \mathrm{im}(\mu) =$ neuromanifold

it is a manifold with boundary and singularities

are parametrized families of functions

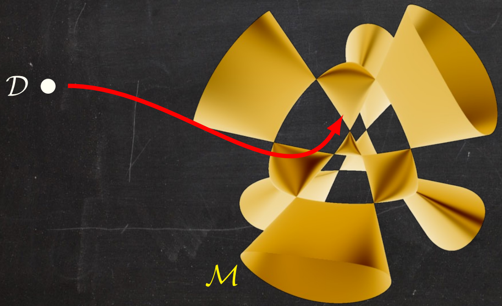
$$\mu : \mathbb{R}^N \longrightarrow \mathcal{M}, \quad \theta \longmapsto f_{L,\theta} \circ \ldots \circ f_{1,\theta}$$

$L = \#$ layers, $\quad f_{i,\theta} = (\sigma_i, \ldots, \sigma_i) \circ \alpha_{i,\theta}$,
$\sigma_i : \mathbb{R} \to \mathbb{R}$ activation, $\quad \alpha_{i,\theta}$ affine linear

training a network

Given training data $\mathcal{D}$, the goal is to minimize the **loss**

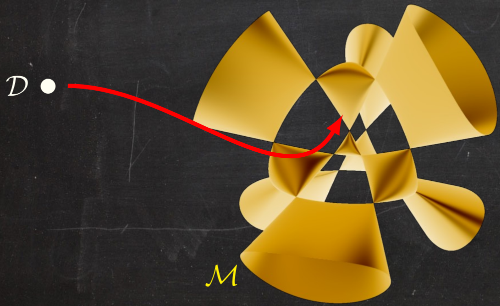
$$\mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}.$$



training a network

Given training data $\mathcal{D}$, the goal is to minimize the **loss**

$$\mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}.$$



Geometric questions:

- How does the network architecture affect the geometry of the function space?
- How does the geometry of the function space impact the training of the network?

understanding networks via algebraic optimization

For piecewise algebraic activation, the neuromanifold is a semi-algebraic set (defined by polynomial equalities and inequalities).

understanding networks via algebraic optimization

For piecewise algebraic activation, the neuromanifold is a semi-algebraic set (defined by polynomial equalities and inequalities).

Examples:

activation	loss
identity	
ReLU	
polynomial	

understanding networks via algebraic optimization

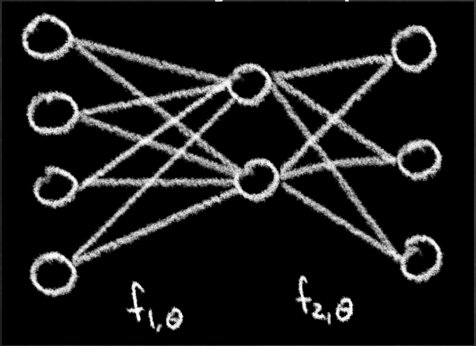
For piecewise algebraic activation, the neuromanifold is a semi-algebraic set (defined by polynomial equalities and inequalities).

Examples:

activation	loss	
identity	squared-error loss	= Euclidean dist
ReLU	Wasserstein distance	= polyhedral dist.
polynomial	cross-entropy	≅ KL divergence

If the loss is also algebraic (or has at least algebraic derivatives), network training is an algebraic optimization problem.

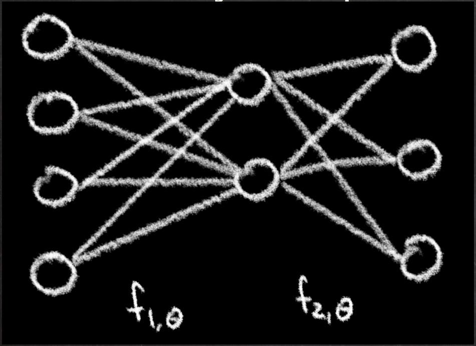
baby example: linear dense networks



In this example:

$$\mu : \mathbb{R}^{2\times 4} \times \mathbb{R}^{3\times 2} \longrightarrow \mathbb{R}^{3\times 4},$$
$$(W_1, W_2) \longmapsto W_2 W_1.$$

baby example: linear dense networks

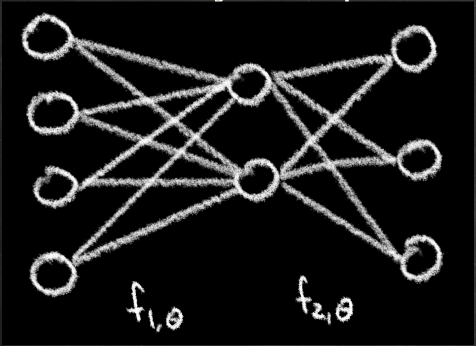


In this example:

$$\mu : \mathbb{R}^{2\times 4} \times \mathbb{R}^{3\times 2} \longrightarrow \mathbb{R}^{3\times 4},$$
$$(W_1, W_2) \longmapsto W_2 W_1.$$

$$\mathcal{M} = \{W \in \mathbb{R}^{3\times 4} \mid \operatorname{rank}(W) \leq 2\}$$

baby example: linear dense networks



In this example:

$$\mu : \mathbb{R}^{2\times 4} \times \mathbb{R}^{3\times 2} \longrightarrow \mathbb{R}^{3\times 4},$$
$$(W_1, W_2) \longmapsto W_2 W_1.$$

$$\mathcal{M} = \{W \in \mathbb{R}^{3\times 4} \mid \operatorname{rank}(W) \leq 2\}$$

In general:

$$\mu : \mathbb{R}^{k_1\times k_0} \times \mathbb{R}^{k_2\times k_1} \times \ldots \times \mathbb{R}^{k_L\times k_{L-1}} \longrightarrow \mathbb{R}^{k_L\times k_0},$$
$$(W_1, W_2, \ldots, W_L) \longmapsto W_L \cdots W_2 W_1.$$

$\mathcal{M} = \{W \in \mathbb{R}^{k_L\times k_0} \mid \operatorname{rank}(W) \leq \min(k_0, \ldots, k_L)\}$ is an algebraic variety and we know its singularities etc.

example: attention networks

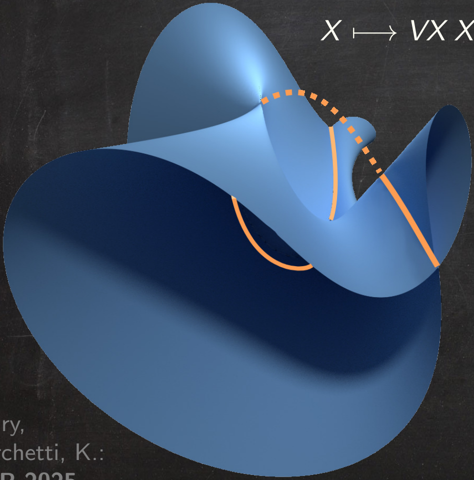
A single-layer lightning self-attention network with weights $Q, K \in \mathbb{R}^{a\times d}$ and $V \in \mathbb{R}^{d'\times d}$ is

$$\mathbb{R}^{d\times t} \longrightarrow \mathbb{R}^{d'\times t},$$
$$X \longmapsto VX\,X^\top K^\top QX.$$

example: attention networks

A single-layer lightning self-attention network with weights $Q, K \in \mathbb{R}^{a\times d}$ and $V \in \mathbb{R}^{d'\times d}$ is

$$\mathbb{R}^{d\times t} \longrightarrow \mathbb{R}^{d'\times t},$$
$$X \longmapsto VX\, X^\top K^\top QX.$$



A slice of the 5-dimensional neuromanifold $\mathcal{M}$ for $a = d = t = 2, d' = 1$.

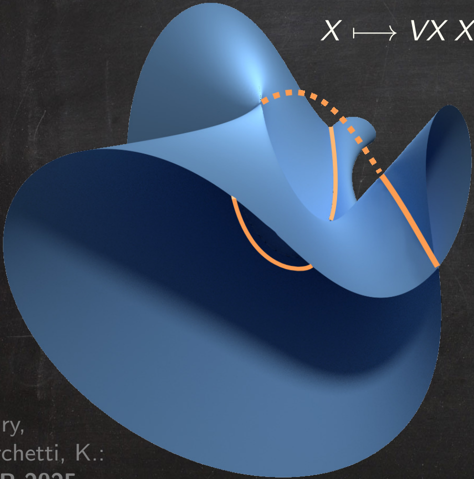
It is singular along the orange curve, and has boundary points where the curve leaves/enters $\mathcal{M}$.

Henry, Marchetti, K.: **ICLR 2025**

example: attention networks

A single-layer lightning self-attention network with weights $Q, K \in \mathbb{R}^{a \times d}$ and $V \in \mathbb{R}^{d' \times d}$ is

$$\mathbb{R}^{d \times t} \longrightarrow \mathbb{R}^{d' \times t},$$
$$X \longmapsto VX\,X^\top K^\top QX.$$



A slice of the 5-dimensional neuromanifold $\mathcal{M}$ for $a = d = t = 2, d' = 1$.

It is singular along the orange curve, and has boundary points where the curve leaves/enters $\mathcal{M}$.

It is not a variety, but a semialgebraic set.

Henry, Marchetti, K.: **ICLR 2025**

a dictionary

machine learning	algebraic geometry
sample complexity & expressivity	dimension, degree, covering number
subnetworks & implicit bias	singularities
identifiability & hidden symmetries	fibers of the parametrization
optimization & gradient descent	critical point theory, discriminants, dynamical invariants

dimension, degree, covering number

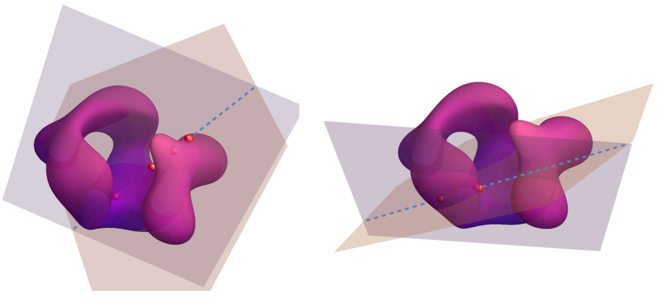
The dimension of the neuromanifold $\mathcal{M}$ measures how many functions can be exactly expressed by the network.

dimension, degree, covering number

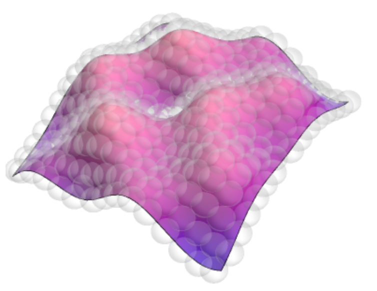
The dimension of the neuromanifold $\mathcal{M}$ measures how many functions can be exactly expressed by the network.

The degree of an algebraic variety is the number of intersections (over $\mathbb{C}$) with a generic linear space (of the correct dimension).

It measures how curvy/twisted the variety is.

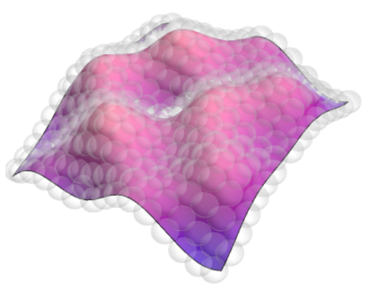


dimension, degree, covering number



covering number $\mathcal{N}_\varepsilon(\mathcal{M})$ = minimum number of metric balls of radius ε required to cover $\mathcal{M}$

dimension, degree, covering number

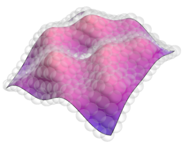


covering number $\mathcal{N}_\varepsilon(\mathcal{M})$ = minimum number of metric balls of radius ε required to cover $\mathcal{M}$

$$\log \mathcal{N}_\varepsilon(\mathcal{M}) = \mathcal{O}\left(\dim(\mathcal{M}) \cdot \log \frac{\mathrm{degree}(\mathcal{M})}{\varepsilon} + C\right)$$

(cf. Weyl's Tube Formula)

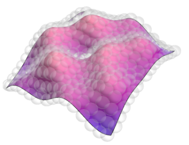
dimension, degree, covering number



covering number $\mathcal{N}_\varepsilon(\mathcal{M})$ = minimum number of metric balls of radius ε required to cover $\mathcal{M}$

relation to sample complexity:
the number of data samples required to infer the function that best approximates the distribution of data (with high probability, and within a given generalization loss margin ε) scales logarithmically in $\mathcal{N}_\varepsilon(\mathcal{M})$.

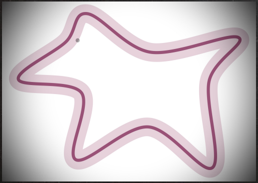
dimension, degree, covering number



covering number $\mathcal{N}_\varepsilon(\mathcal{M})$ = minimum number of metric balls of radius ε required to cover $\mathcal{M}$

relation to sample complexity:
the number of data samples required to infer the function that best approximates the distribution of data (with high probability, and within a given generalization loss margin ε) scales logarithmically in $\mathcal{N}_\varepsilon(\mathcal{M})$.

relation to approximative expressivity:



the volume of the ε-tube around $\mathcal{M}$ measures how many functions can be approximated within an error of ε.

it is $\leq \mathcal{N}_\varepsilon(\mathcal{M}) \cdot \mathrm{vol}(\text{ball of radius } 2\varepsilon)$

dimension, degree, covering number

Takeaway

Dimension and degree are the most fundamental invariants of an algebraic neuromanifold.

They control metric quantities such as covering numbers,
which in turn measure approximate expressivity and sample complexity.

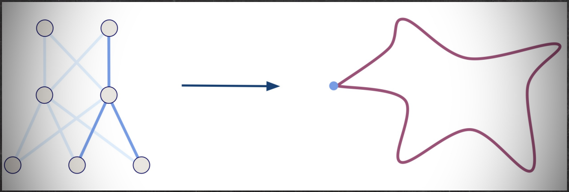
singularities

Singularities of a variety are points where the variety does not look locally like a smooth manifold.

singularities

Singularities of a variety are points where the variety does not look locally like a smooth manifold.

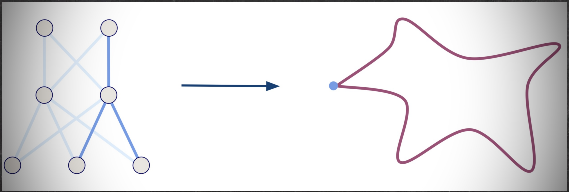
Conjecture: The singularities of neuromanifolds correspond to subnetworks.
(known for convolutional & fully-connected networks with polynomial activation)



singularities

Singularities of a variety are points where the variety does not look locally like a smooth manifold.

Conjecture: The singularities of neuromanifolds correspond to subnetworks.
(known for convolutional & fully-connected networks with polynomial activation)



Potential explanation for *lottery ticket hypothesis*: the tendency of deep networks to discard weights during learning.

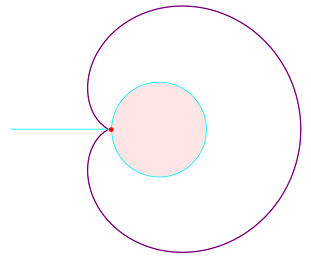
singularities

A singularity might, depending on its type, attract a large portion of the ambient space during training – explaining implicit bias.

singularities

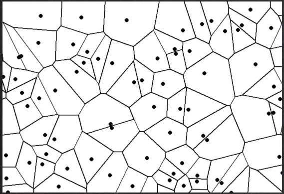
A singularity might, depending on its type, attract a large portion of the ambient space during training – explaining implicit bias.

This is captured by the Voronoi cell of the singularity:



voronoi cells

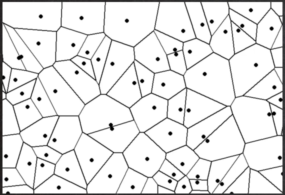
Given a set $\mathcal{M} \subseteq \mathbb{R}^n$, the Voronoi cell of $x \in \mathcal{M}$ consists of all $u \in \mathbb{R}^n$ such that x is "closest" among all points in $\mathcal{M}$.



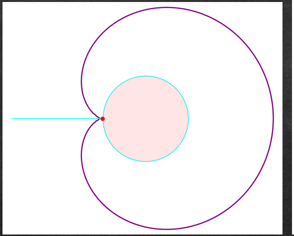
$\mathcal{M}$ might be finite

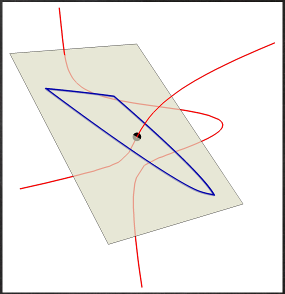
voronoi cells

Given a set $\mathcal{M} \subseteq \mathbb{R}^n$, the Voronoi cell of $x \in \mathcal{M}$ consists of all $u \in \mathbb{R}^n$ such that x is “closest” among all points in $\mathcal{M}$.



$\mathcal{M}$ might be finite



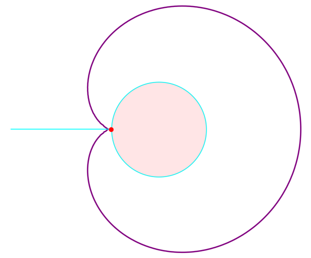


or a manifold, variety, semi-algebraic set, etc.

singularities

A singularity might, depending on its type, attract a large portion of the ambient space during training – explaining implicit bias.

This is captured by the Voronoi cell of the singularity:



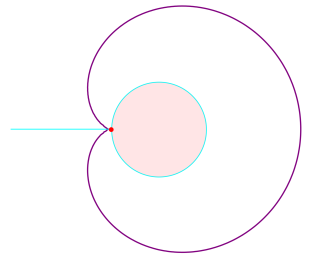
$\mathcal{M} \subseteq \mathbb{R}^2$ is the purple curve

loss $=$ Euclidean distance

singularities

A singularity might, depending on its type, attract a large portion of the ambient space during training – explaining implicit bias.

This is captured by the Voronoi cell of the singularity:



$\mathcal{M} \subseteq \mathbb{R}^2$ is the purple curve

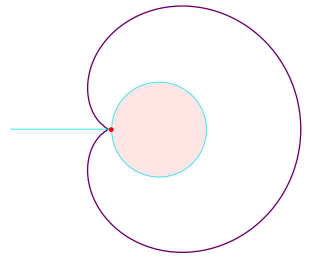
loss $=$ Euclidean distance

at all smooth points $x \in \mathcal{M}$, the Voronoi cell is a line segment

singularities

A singularity might, depending on its type, attract a large portion of the ambient space during training – explaining implicit bias.

This is captured by the Voronoi cell of the singularity:



- $\mathcal{M} \subseteq \mathbb{R}^2$ is the purple curve
- loss $=$ Euclidean distance
- at all smooth points $x \in \mathcal{M}$, the Voronoi cell is a line segment
- the Voronoi cell at the singularity is 2-dimensional, i.e., that point is the closest with positive probability

singularities

Takeaway

Singularities of the neuromanifold can introduce implicit biases in the learning process.

They often correspond to subnetworks, favoring the selection of simpler models.

fibers of the parametrization

Recall: The neuromanifold is the image of parametrization map

$$\mu : \mathbb{R}^N \longrightarrow \mathcal{M}, \quad \theta \longmapsto f_{L,\theta} \circ \ldots \circ f_{1,\theta}.$$

Identifiability / hidden symmetries:
Which network parameters give rise to the same function?

fibers of the parametrization

Recall: The neuromanifold is the image of parametrization map

$$\mu : \mathbb{R}^N \longrightarrow \mathcal{M},$$
$$\theta \longmapsto f_{L,\theta} \circ \ldots \circ f_{1,\theta}.$$

Identifiability / hidden symmetries:
Which network parameters give rise to the same function?

In algebraic geometry terms:
Given $f \in \mathcal{M}$, which parameters θ are in the fiber $\mu^{-1}(f)$?

fibers of the parametrization

Recall: The neuromanifold is the image of parametrization map

$$\mu : \mathbb{R}^N \longrightarrow \mathcal{M},$$
$$\theta \longmapsto f_{L,\theta} \circ \ldots \circ f_{1,\theta}.$$

Identifiability / hidden symmetries:
Which network parameters give rise to the same function?

In algebraic geometry terms:
Given $f \in \mathcal{M}$, which parameters θ are in the fiber $\mu^{-1}(f)$?

fiber-dimension theorem:
The dimension of the image of an algebraic map equals the co-dimension of its generic fiber. (nonlinear version of rank-nullity theorem)

fibers of the parametrization

Recall: The neuromanifold is the image of parametrization map

$$\mu : \mathbb{R}^N \longrightarrow \mathcal{M}, \quad \theta \longmapsto f_{L,\theta} \circ \ldots \circ f_{1,\theta}.$$

More generally: All geometric features of the neuromanifold are caused by μ.

fibers of the parametrization

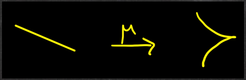
Recall: The neuromanifold is the image of parametrization map

$$\mu : \mathbb{R}^N \longrightarrow \mathcal{M},$$
$$\theta \longmapsto f_{L,\theta} \circ \ldots \circ f_{1,\theta}.$$

More generally: All geometric features of the neuromanifold are caused by μ.

For instance, singularities on $\mathcal{M}$ can arise in 2 ways:

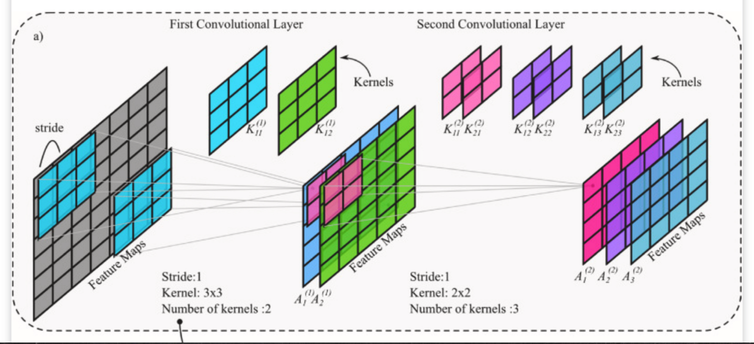
- from critical points of μ



- from special (i.e., non-generic) fibers of μ

example: polynomial convolutional networks

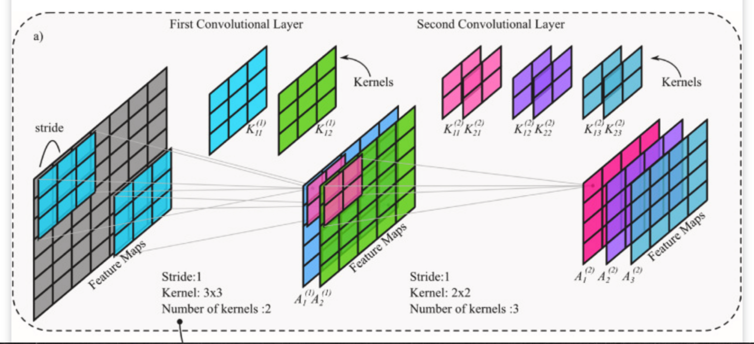
We now consider convolutional networks



where the activation function is a monomial: $\sigma(x) = x^r$.

example: polynomial convolutional networks

We now consider convolutional networks



where the activation function is a monomial: $\sigma(x) = x^r$.

Weierstrass Approximation Theorem:

Any activation function can be approximated by polynomial ones.

Any CNN neuromanifold can be approximated by polynomial ones.

example: polynomial convolutional networks

$$\sigma(x) = x^r$$

Theorem: Let $r > 1$.
The neuromanifold is an algebraic variety (i.e., described by polynomial equations) and closed in Euclidean topology.

example: polynomial convolutional networks

$$\sigma(x) = x^r$$

Theorem: Let $r > 1$.
The neuromanifold is an algebraic variety (i.e., described by polynomial equations) and closed in Euclidean topology.

For a generic function $f \in \mathcal{M}$, the only symmetries in the fiber $\mu^{-1}(f)$ are rescalings of the layers.

Shahverdi, Marchetti, K.: **AISTATS 2025**

example: polynomial convolutional networks

$$\sigma(x) = x^r$$

Theorem: Let $r > 1$.
The neuromanifold is an algebraic variety (i.e., described by polynomial equations) and closed in Euclidean topology.

For a generic function $f \in \mathcal{M}$, the only symmetries in the fiber $\mu^{-1}(f)$ are rescalings of the layers.

After modding out the layer scaling, the network parametrization map becomes

- an isomorphism almost everywhere

example: polynomial convolutional networks

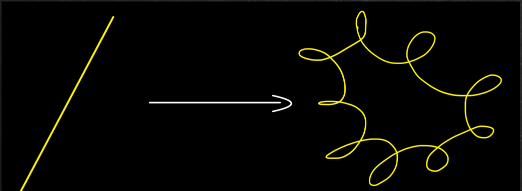
$$\sigma(x) = x^r$$

Theorem: Let $r > 1$.
The neuromanifold is an algebraic variety (i.e., described by polynomial equations) and closed in Euclidean topology.

For a generic function $f \in \mathcal{M}$, the only symmetries in the fiber $\mu^{-1}(f)$ are rescalings of the layers.

After modding out the layer scaling, the network parametrization map becomes

- an isomorphism almost everywhere
- that has finite fibers ($\Leftrightarrow$ singularities)
- and is regular (constant-rank Jacobian)



example: polynomial convolutional networks

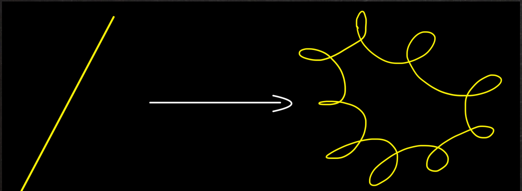
$$\sigma(x) = x^r$$

Theorem: Let $r > 1$.
The neuromanifold is an algebraic variety (i.e., described by polynomial equations) and closed in Euclidean topology.

For a generic function $f \in \mathcal{M}$, the only symmetries in the fiber $\mu^{-1}(f)$ are rescalings of the layers.

After modding out the layer scaling, the network parametrization map becomes

- an isomorphism almost everywhere
- that has finite fibers ($\Leftrightarrow$ singularities)
- and is regular (constant-rank Jacobian)



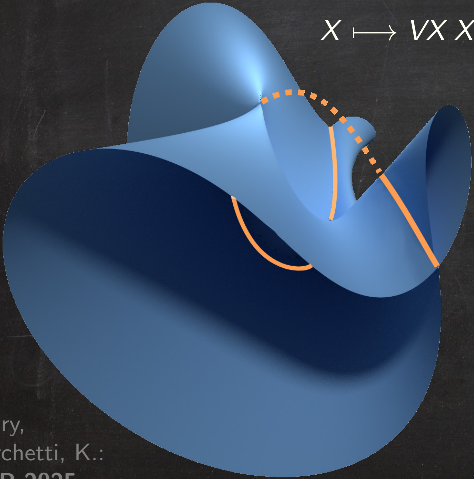
The singularities correspond to subnetworks.

Shahverdi, Marchetti, K.: **AISTATS 2025**

comparison: lightning self-attention

A single-layer lightning self-attention network with weights $Q, K \in \mathbb{R}^{a\times d}$ and $V \in \mathbb{R}^{d'\times d}$ is

$$\mathbb{R}^{d\times t} \longrightarrow \mathbb{R}^{d'\times t},$$
$$X \longmapsto VX\, X^{\top}K^{\top}QX.$$



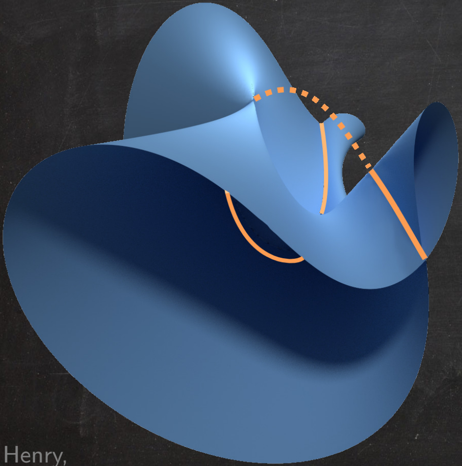
The neuromanifold is semialgebraic but not a variety (polynomial inequalities needed!)

It has both nodal and cuspidal singularities.

Henry, Marchetti, K.: **ICLR 2025**

comparison: lightning self-attention

$$VXX^\top K^\top QX$$



cusps 

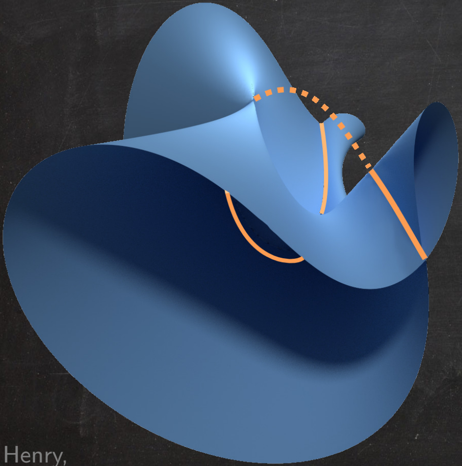
$\Leftrightarrow$ boundary points

$\Leftrightarrow$ Jacobian rank drops

Henry, Marchetti, K.: **ICLR 2025**

comparison: lightning self-attention

$$VXX^\top K^\top QX$$



cusps 

$\Leftrightarrow$ boundary points

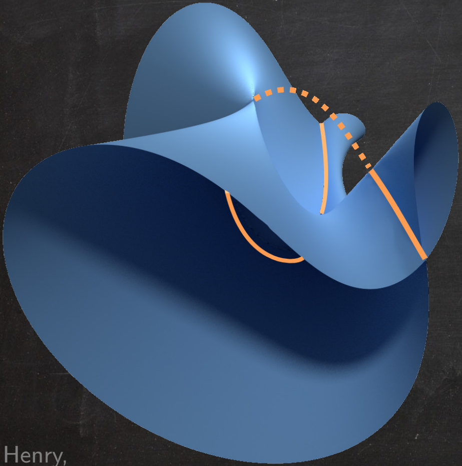
$\Leftrightarrow$ Jacobian rank drops

Theorem: For generic $f \in \mathcal{M}$, the only symmetries in the fiber $\mu^{-1}(f)$ are the "obvious" ones:

Henry, Marchetti, K.: **ICLR 2025**

comparison: lightning self-attention

$$VXX^\top K^\top QX$$



cusps 
$\Leftrightarrow$ boundary points
$\Leftrightarrow$ Jacobian rank drops

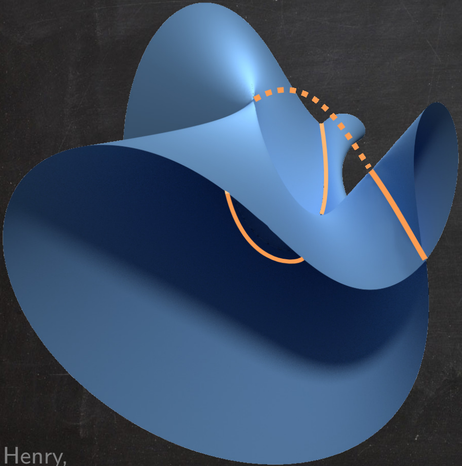
Theorem: For generic $f \in \mathcal{M}$, the only symmetries in the fiber $\mu^{-1}(f)$ are the “obvious” ones:

- layer rescalings

Henry, Marchetti, K.: **ICLR 2025**

comparison: lightning self-attention

$VXX^\top K^\top QX$



cusps 

$\Leftrightarrow$ boundary points

$\Leftrightarrow$ Jacobian rank drops

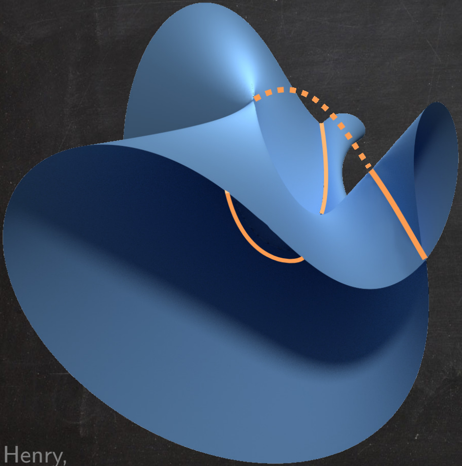
Theorem: For generic $f \in \mathcal{M}$, the only symmetries in the fiber $\mu^{-1}(f)$ are the "obvious" ones:

- layer rescalings
- $\mathrm{GL}(a)$-symmetries of K and Q in each layer

Henry, Marchetti, K.: **ICLR 2025**

comparison: lightning self-attention

$$VXX^\top K^\top QX$$



cusps 

$\Leftrightarrow$ boundary points

$\Leftrightarrow$ Jacobian rank drops

Theorem: For generic $f \in \mathcal{M}$, the only symmetries in the fiber $\mu^{-1}(f)$ are the "obvious" ones:

- layer rescalings
- $\mathrm{GL}(a)$-symmetries of K and Q in each layer
- $\mathrm{GL}(d)$-symmetries of V and $K^\top Q$ of neighboring layers

Henry, Marchetti, K.: **ICLR 2025**

fibers of the parametrization

Takeaway

Fibers of the parameterization control the dimension and symmetries of the neuromanifold.

Together with the parameterization's critical points, they explain the singularities of the neuromanifold.

critical point theory & discriminants

Goal: minimize the loss

$$\mathcal{L}_{\mathcal{D}} : \mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}.$$

critical point theory & discriminants

Goal: minimize the loss

$$\mathcal{L}_{\mathcal{D}} : \mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}.$$

Critical points of $\mathcal{L}_{\mathcal{D}}$ arise in various ways:

1. they can be caused by the parametrization μ (i.e., $\theta \in \mathrm{Crit}(\mu)$) such that $\theta \in \mathrm{Crit}(\mathcal{L}_{\mathcal{D}})$ but $\mu(\theta) \notin \mathrm{Crit}(\ell_{\mathcal{D}})$)

critical point theory & discriminants

Goal: minimize the loss

$$\mathcal{L}_{\mathcal{D}} : \mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}.$$

Critical points of $\mathcal{L}_{\mathcal{D}}$ arise in various ways:

1. they can be caused by the parametrization μ
   (i.e., $\theta \in \mathrm{Crit}(\mu)$ such that $\theta \in \mathrm{Crit}(\mathcal{L}_{\mathcal{D}})$ but $\mu(\theta) \notin \mathrm{Crit}(\ell_{\mathcal{D}})$)
   spurios critical points

critical point theory & discriminants

Goal: minimize the loss

$$\mathcal{L}_{\mathcal{D}} : \mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}.$$

Critical points of $\mathcal{L}_{\mathcal{D}}$ arise in various ways:

1. they can be caused by the parametrization μ
   (i.e., $\theta \in \mathrm{Crit}(\mu)$) such that $\theta \in \mathrm{Crit}(\mathcal{L}_{\mathcal{D}})$ but $\mu(\theta) \notin \mathrm{Crit}(\ell_{\mathcal{D}})$)
   spurios critical points

   e.g. appear as local minima in polynomial MLPs with positive probability but not in polynomial CNNs

critical point theory & discriminants

Goal: minimize the loss

$$\mathcal{L}_{\mathcal{D}} : \mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}.$$

Critical points of $\mathcal{L}_{\mathcal{D}}$ arise in various ways:

1. they can be caused by the parametrization μ
   (i.e., $\theta \in \mathrm{Crit}(\mu)$ such that $\theta \in \mathrm{Crit}(\mathcal{L}_{\mathcal{D}})$ but $\mu(\theta) \notin \mathrm{Crit}(\ell_{\mathcal{D}})$)
   spurios critical points

   e.g. appear as local minima in polynomial MLPs with positive probability but not in polynomial CNNs

2. they correspond to critical points of the loss in function space
   (i.e., $\theta \in \mathrm{Crit}(\mathcal{L}_{\mathcal{D}})$ and $\mu(\theta) \in \mathrm{Crit}(\ell_{\mathcal{D}})$)

critical point theory & discriminants

Goal: minimize the loss

$$\mathcal{L}_\mathcal{D} : \mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_\mathcal{D}} \mathbb{R}.$$

Critical points of $\mathcal{L}_\mathcal{D}$ arise in various ways:

1. they can be caused by the parametrization μ
   (i.e., $\theta \in \mathrm{Crit}(\mu)$ such that $\theta \in \mathrm{Crit}(\mathcal{L}_\mathcal{D})$ but $\mu(\theta) \notin \mathrm{Crit}(\ell_\mathcal{D})$)
   spurios critical points

   e.g. appear as local minima in polynomial MLPs with positive probability but not in polynomial CNNs

2. they correspond to critical points of the loss in function space
   (i.e., $\theta \in \mathrm{Crit}(\mathcal{L}_\mathcal{D})$ and $\mu(\theta) \in \mathrm{Crit}(\ell_\mathcal{D})$)
   the function $\mu(\theta)$ can be either a
   a) singular point on $\mathcal{M}$ or
   b) in the smooth locus of $\mathcal{M}$

critical point theory & discriminants

Goal: minimize the loss

$$\mathcal{L}_{\mathcal{D}} : \mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}.$$

Critical points of $\mathcal{L}_{\mathcal{D}}$ arise in various ways:

1. they can be caused by the parametrization μ
   (i.e., $\theta \in \mathrm{Crit}(\mu)$ such that $\theta \in \mathrm{Crit}(\mathcal{L}_{\mathcal{D}})$ but $\mu(\theta) \notin \mathrm{Crit}(\ell_{\mathcal{D}})$)
   spurios critical points

   e.g. appear as local minima in polynomial MLPs with positive probability but not in polynomial CNNs

2. they correspond to critical points of the loss in function space
   (i.e., $\theta \in \mathrm{Crit}(\mathcal{L}_{\mathcal{D}})$ and $\mu(\theta) \in \mathrm{Crit}(\ell_{\mathcal{D}})$)
   the function $\mu(\theta)$ can be either a
   a) singular point on $\mathcal{M}$ or
   b) in the smooth locus of $\mathcal{M}$ Morse theory

critical point theory & discriminants

for algebraic optimization problems (e.g. mean squared error or cross entropy loss), the number of complex critical points of $\mathcal{L}_{\mathcal{D}}$ is constant for generic $\mathcal{D}$

critical point theory & discriminants

for algebraic optimization problems (e.g. mean squared error or cross entropy loss), the number of complex critical points of $\mathcal{L}_{\mathcal{D}}$ is constant for generic $\mathcal{D}$
$\rightsquigarrow$ measures intrinsic optimization degree

over $\mathbb{R}$, the number or type (local / global minima, strict / non-strict saddle, etc.) of the critical points changes when $\mathcal{D}$ crosses an algebraic discriminant hypersurface

critical point theory & discriminants

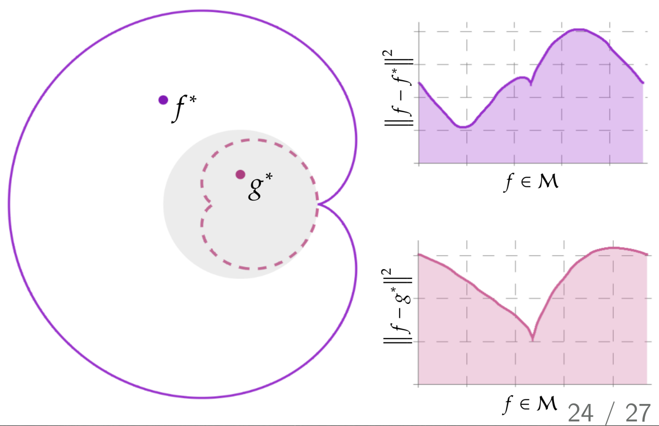
for algebraic optimization problems (e.g. mean squared error or cross entropy loss), the number of complex critical points of $\mathcal{L}_{\mathcal{D}}$ is constant for generic $\mathcal{D}$ $\rightsquigarrow$ measures intrinsic optimization degree

over $\mathbb{R}$, the number or type (local / global minima, strict / non-strict saddle, etc.) of the critical points changes when $\mathcal{D}$ crosses an algebraic discriminant hypersurface

over $\mathbb{C}$: always 4 critical points

over $\mathbb{R}$: 4 or 2 critical points

discriminant = dashed



critical point theory, discriminants, dynamical invariants

Takeaway

The critical points of the loss arise from the geometry of the neuromanifold and its parametrization.

Their number and type can change suddenly as data crosses discriminants.

Moreover, algebraic invariants of gradient flow govern the training dynamics...

many future questions

- Describe all singularities of attention neuromanifolds explicitly, and compute their Voronoi cells. ($\rightsquigarrow$ implicit bias?)
- Compare the type of critical points and more generally the loss landscape of
  - attention networks
  - polynomial convolutional networks
  - polynomial dense networks
- How do skip connections and inhomogeneous activations regularize μ (i.e., less spurious critical points) and smoothen out singularities?
- What happens to the neuromanifold when imposing group equivariance?
- What about ReLU networks, or more generally piecewise rational activation?
- Beyond algebraic geometry: tame geometry of o-minimal structures

thanks for your attention!

machine learning	algebraic geometry
sample complexity & expressivity	dimension, degree, covering number
subnetworks & implicit bias	singularities
identifiability & hidden symmetries	fibers of the parametrization
optimization & gradient descent	critical point theory, discriminants, dynamical invariants

An Invitation to Neuroalgebraic Geometry

Giovanni Luca Marchetti [*1] Vahid Shahverdi [*1] Stefano Mereta [*1] Matthew Trager [*2] Kathlén Kohn [*1]

Abstract

In this expository work, we promote the study of function spaces parameterized by machine learning models through the lens of algebraic geometry. To this end, we focus on algebraic models, such as neural networks with polynomial activations, whose associated function spaces are semi-algebraic varieties. We outline a dictionary between algebro-geometric invariants of these varieties, such as dimension, degree, and singularities, and fundamental aspects of machine learning, such as sample complexity, expressivity, training dynamics, and implicit bias.

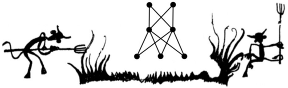


Figure 1. A neural variation of a celebrated doodle from the algebraic geometry literature (Grothendieck, 1968).

Neuroalgebraic Geometry

algebraic NNs

↓ polynomial approximation

target NN

studies nonlinear models
in finite-dimensional ambient space

aims to draw conclusions
in the limit

NTK approach

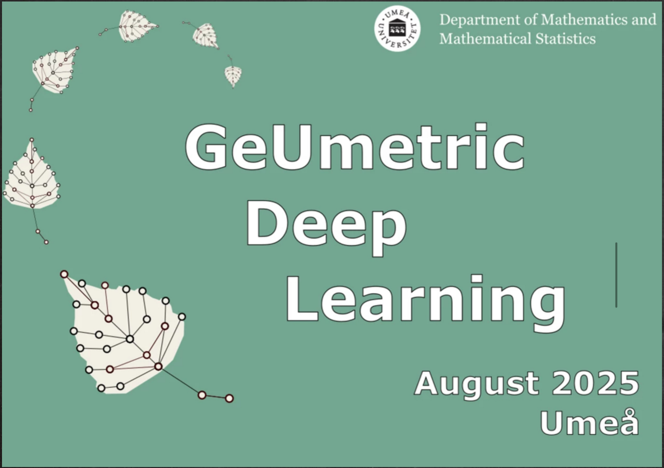
target NN

↓ increase width

∞-width NN

studies linearized models
in ∞-dimensional ambient space

aims to draw conclusions
from the limit



invited speakers:

Elisenda Grigsby
Boston College

Emanuele Rodolá
University of Rome

emphasis:
presentations by young researchers

August 19-21

Axel Flinth (Umeå Univ), Jan E. Gerken (Chalmers), Kathlén Kohn (KTH), Giovanni Marchetti (KTH), Stefano Mereta (KTH), Fredrik Ohlsson (Umeå Univ)