The Geometry of Attention Networks and Polynomial Networks

Kathlén Kohn

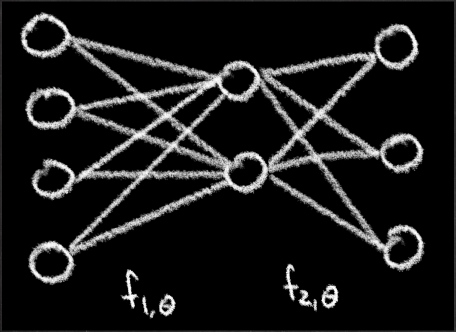
based on joint works with

Nathan Henry
Univ. of Toronto

Giovanni Marchetti
KTH

Vahid Shahverdi
KTH

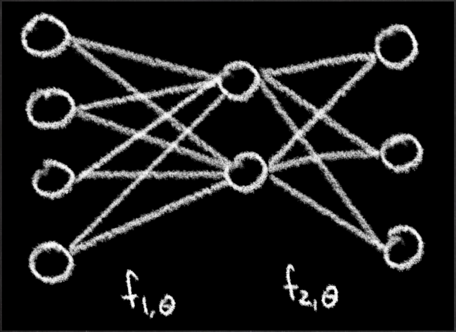
feedforward neural networks



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

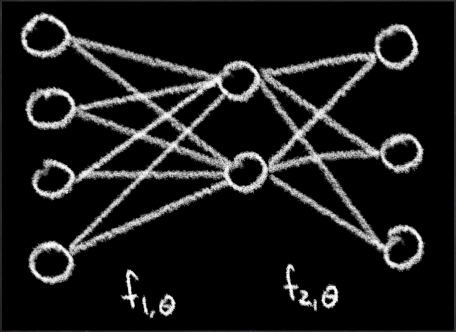


are parametrized families of functions

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

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

feedforward neural networks

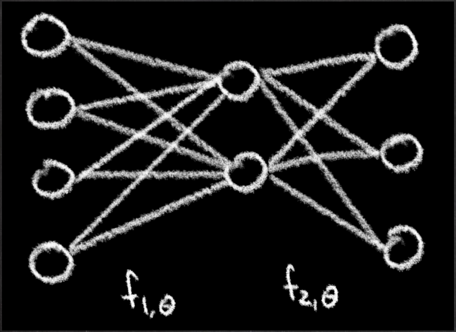


are parametrized families of functions

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

$L = \#$ 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

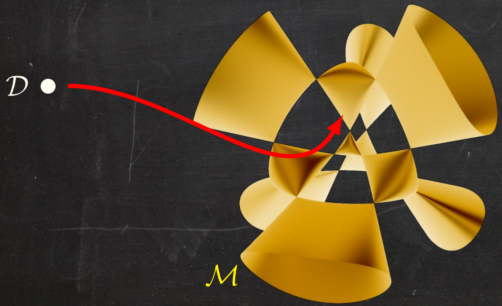
$$\mu : \mathbb{R}^N \longrightarrow \mathcal{M},$$
$$\theta \longmapsto f_{L,\theta} \circ \ldots \circ f_{1,\theta}$$

$L = \#$ 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

training a network

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

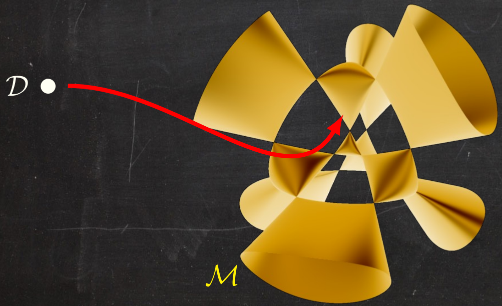
$$\mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}.$$



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

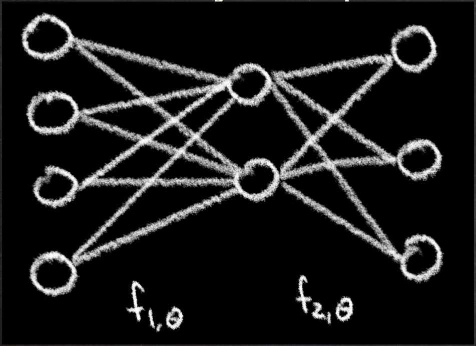
For piecewise algebraic activation, the neuromanifold is a semi-algebraic set (defined by polynomial equalities and inequalities).

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.

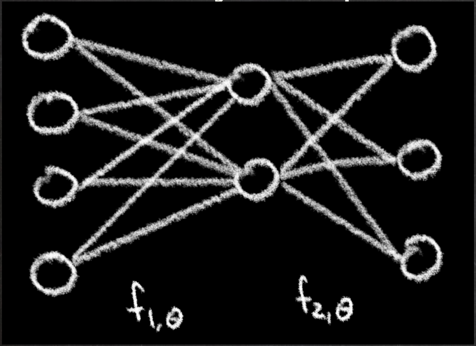
baby example: linear dense networks



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

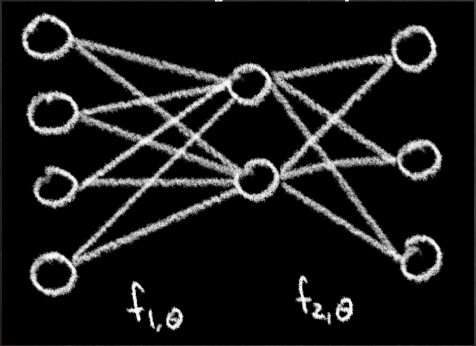


In this example:

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

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

baby example: linear dense networks



In this example:

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

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

In general:

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

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

example: attention networks

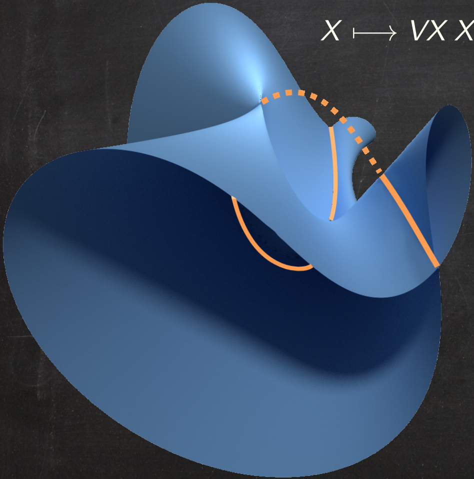
A single-layer lightning self-attention network with weights $Q, K \in \mathbb{R}^{a\times d}$ and $V \in \mathbb{R}^{d'\times d}$ is

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

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

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



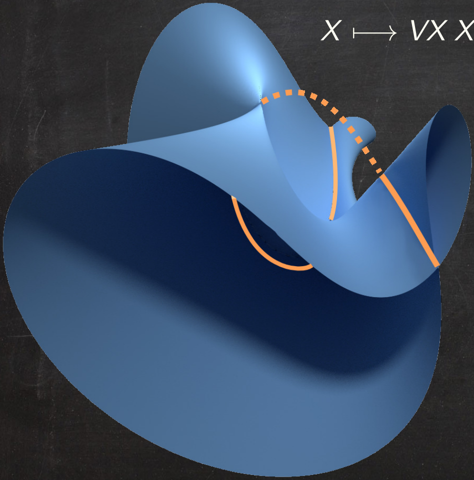
A slice of the 5-dimensional neuromanifold $\mathcal{M}$ for $a = d = t = 2, d' = 1$.

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

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.

a dictionary

machine learning	algebraic geometry
sample complexity	dimension
identifiability	fibers
expressivity	degree
subnetworks & hidden bias	singularities
learning dynamics	algebraic critical point theory

dimension and fibers

fundamental theorem:
The dimension of the neuromanifold $\mathcal{M}$ scales linearly with the sample complexity of learnability (in the PAC sense).

dimension and fibers

fundamental theorem:
The dimension of the neuromanifold $\mathcal{M}$ scales linearly with the sample complexity of learnability (in the PAC sense).

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

dimension and fibers

fundamental theorem:
The dimension of the neuromanifold $\mathcal{M}$ scales linearly with the sample complexity of learnability (in the PAC sense).

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 $\mu^{-1}(f)$?

dimension and fibers

fundamental theorem:
The dimension of the neuromanifold $\mathcal{M}$ scales linearly with the sample complexity of learnability (in the PAC sense).

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/image theorem:
The dimension of the image of an algebraic map equals the co-dimension of its generic fiber.

degree

The degree of an affine/projective algebraic variety is the number of intersections with a linear space (of the correct dimension).

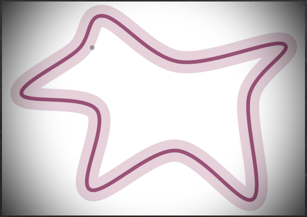
It measures how twisted the variety is,

degree

The degree of an affine/projective algebraic variety is the number of intersections with a linear space (of the correct dimension).

It measures how twisted the variety is, and its approximation capabilities:

Weyl Tube Formula:
The volume of the ε-tube around an algebraic variety of dimension n, co-dimension m, and degree d increases as $O(nd\varepsilon)^m$.



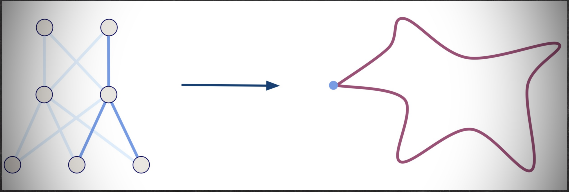
singularities

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.

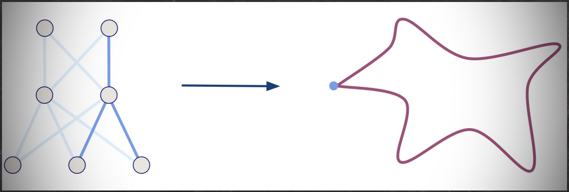
Conjecture: The singularities of neuromanifolds correspond to subnetworks. (known for convolutional & fully-connected networks with polynomial activation)



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.

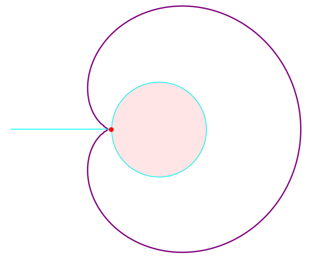
singularities

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

singularities

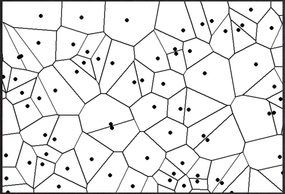
A singularity might, depending on its type, attract a large portion of the ambient space during training – explaining implicit bias.

This is captured by the Voronoi cell of the singularity:



voronoi cells

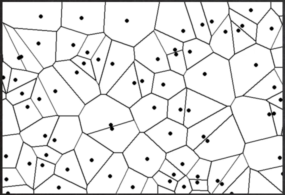
Given a set $\mathcal{M} \subseteq \mathbb{R}^n$, the Voronoi cell of $x \in \mathcal{M}$ consists of all $u \in \mathbb{R}^n$ such that x is "closest" among all points in $\mathcal{M}$.



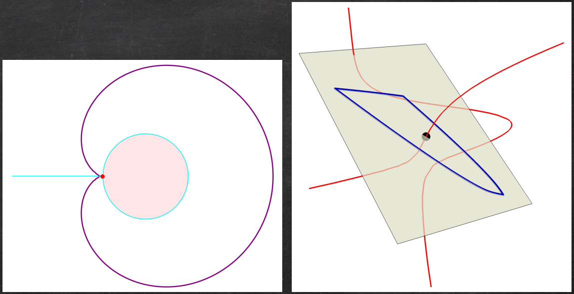
$\mathcal{M}$ might be finite

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

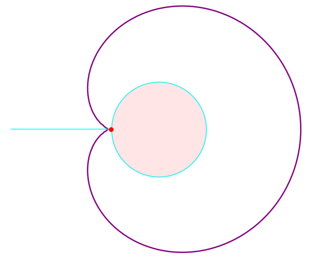


or a manifold, variety, semi-algebraic set, etc.

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:



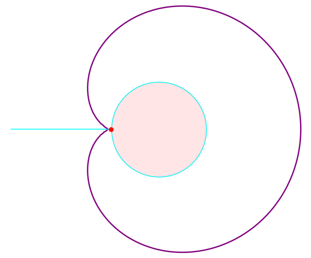
$\mathcal{M} \subseteq \mathbb{R}^2$ is the purple curve

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

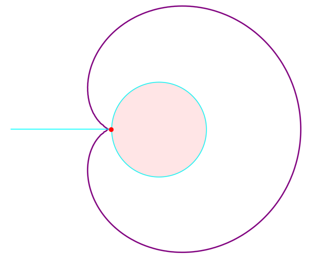
loss $=$ Euclidean distance

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

algebraic critical point theory can ...

- $\diamond$ distinguish pure from spurious critical points that only come from the network parametrization μ

(recall: $\mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}$.)

algebraic critical point theory can . . .

◇ distinguish pure from spurious critical points that only come from the network parametrization μ

(recall: $\mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}$.)

◇ count critical points

algebraic critical point theory can . . .

◇ distinguish pure from spurious critical points that only come from the network parametrization μ

(recall: $\mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}$.)

◇ count critical points

◇ determine the critical points' type (local / global minimal, strict / non-strict saddle points, etc.) and location (e.g., on singular locus)

algebraic critical point theory can . . .

◇ distinguish pure from spurious critical points that only come from the network parametrization μ

(recall: $\mathbb{R}^N \xrightarrow{\mu} \mathcal{M} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}$.)

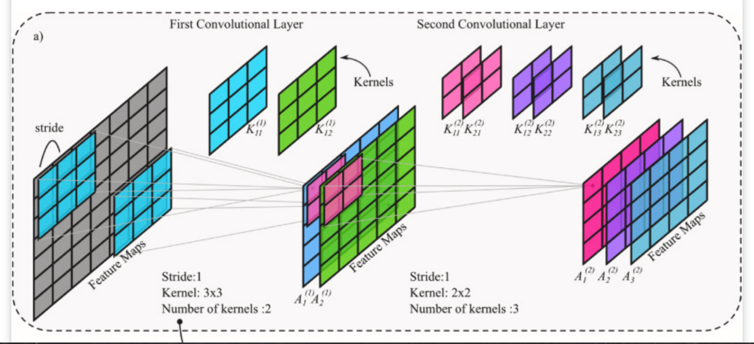
◇ count critical points

◇ determine the critical points' type (local / global minimal, strict / non-strict saddle points, etc.) and location (e.g., on singular locus)

◇ identify particularly areas on the neuromanifold that are particularly exposed (implicit bias) or have many critical points

example: polynomial convolutional networks

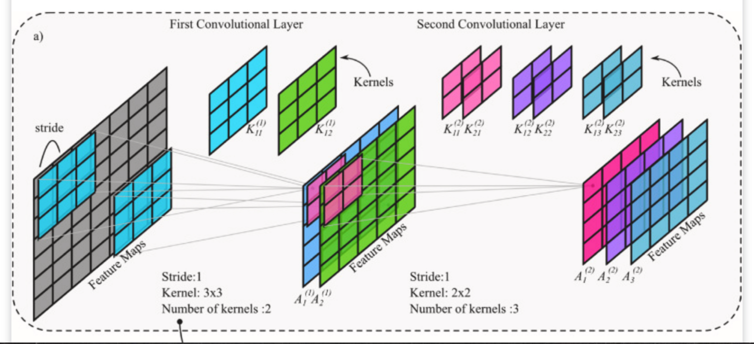
We now consider convolutional networks



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.

Its dimension ($\sim$ sample complexity) is linear in the depth.

Its degree ($\sim$ expressivity) is super exponential in the depth.

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.

Its dimension ($\sim$ sample complexity) is linear in the depth.
$\dim(\mathcal{M}) = L(k-1)+1$ for $L = \#$layers, $k =$ filter size

Its degree ($\sim$ expressivity) is super exponential in the depth.
$\mathrm{degree}(\mathcal{M}) = (L(k-1))!\,\frac{r^{L(L-1)(k-1)/2}}{(k-1)!^L}$

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.

Its dimension ($\sim$ sample complexity) is linear in the depth.
$\dim(\mathcal{M}) = L(k-1)+1$ for $L = \#$layers, $k =$ filter size

Its degree ($\sim$ expressivity) is super exponential in the depth.
$\mathrm{degree}(\mathcal{M}) = (L(k-1))! \, \frac{r^{L(L-1)(k-1)/2}}{(k-1)!^L}$

} explains why depth is important!

example: polynomial convolutional networks

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

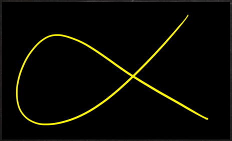
Theorem: Let $r > 1$.
The neuromanifold is an algebraic variety (i.e., described by polynomial equations) and closed in Euclidean topology.

Its dimension ($\sim$ sample complexity) is linear in the depth.
$\dim(\mathcal{M}) = L(k-1) + 1$ for $L = \#$layers, $k =$ filter size

Its degree ($\sim$ expressivity) is super exponential in the depth.
$\text{degree}(\mathcal{M}) = (L(k-1))! \, \frac{r^{L(L-1)(k-1)/2}}{(k-1)!^L}$

} explains why depth is important!

The singularities correspond to subnetworks and are nodal.



example: polynomial convolutional networks

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

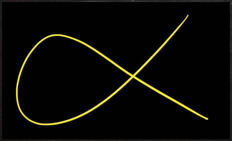
Theorem: Let $r > 1$.
The neuromanifold is an algebraic variety (i.e., described by polynomial equations) and closed in Euclidean topology.

Its dimension ($\sim$ sample complexity) is linear in the depth.
$\dim(\mathcal{M}) = L(k-1)+1$ for $L = \#$layers, $k =$ filter size

Its degree ($\sim$ expressivity) is super exponential in the depth.
$\text{degree}(\mathcal{M}) = (L(k-1))!\,\frac{r^{L(L-1)(k-1)/2}}{(k-1)!^L}$

} explains why depth is important!

The singularities correspond to subnetworks and are nodal.



These are typically not more exposed during training.

example: polynomial convolutional networks

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

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

example: polynomial convolutional networks

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

Theorem: Let $r > 1$.
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

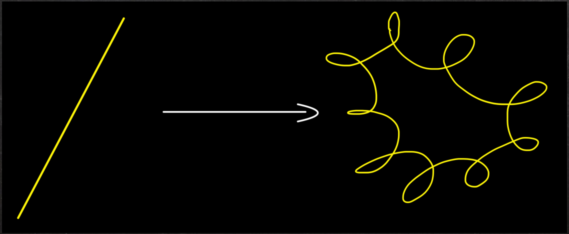
example: polynomial convolutional networks

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

Theorem: Let $r > 1$.
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)



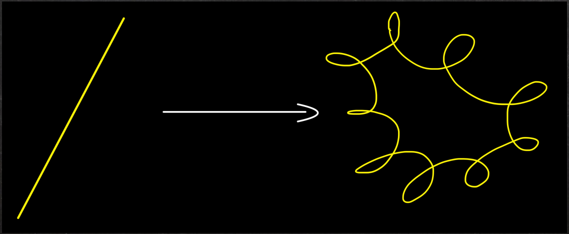
example: polynomial convolutional networks

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

Theorem: Let $r > 1$.
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

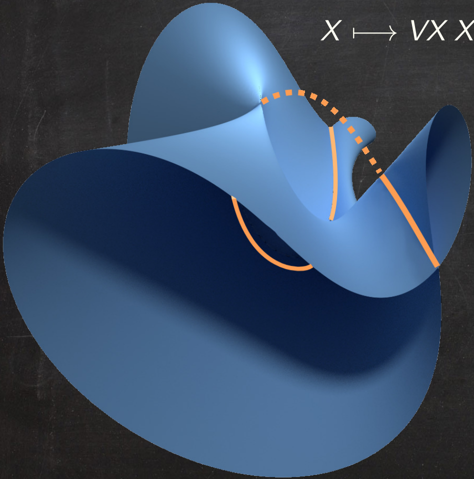
- an isomorphism almost everywhere
- that has finite fibers ($\Leftrightarrow$ singularities)
- and is regular (constant-rank Jacobian) $\Rightarrow$ no spurious critical points



comparison: lightning self-attention

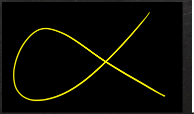
A single-layer lightning self-attention network with weights $Q, K \in \mathbb{R}^{a \times d}$ and $V \in \mathbb{R}^{d' \times d}$ is

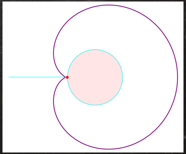
$$\mathbb{R}^{d \times t} \longrightarrow \mathbb{R}^{d' \times t},$$
$$X \longmapsto VX\,X^\top K^\top QX.$$



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

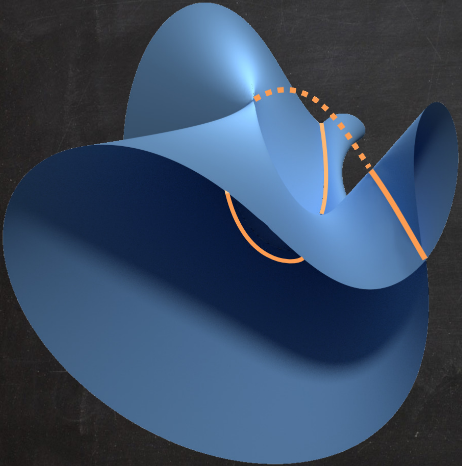
It has both nodal and cuspidal singularities.

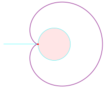




comparison: lightning self-attention

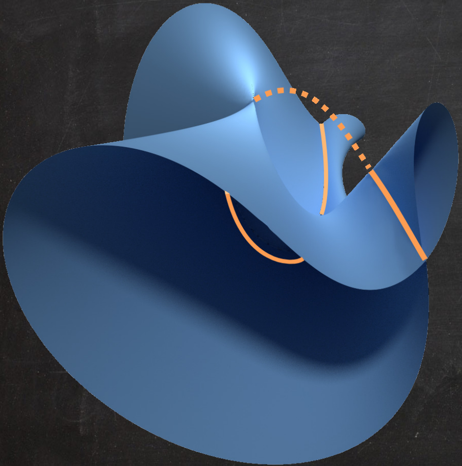
$$VXX^\top K^\top QX$$

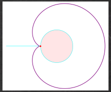


cusps 
⇔ boundary points
⇔ Jacobian rank drops

comparison: lightning self-attention

$VXX^\top K^\top QX$



cusps 

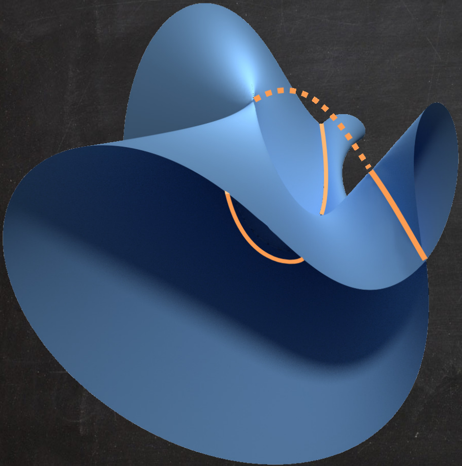
$\Leftrightarrow$ boundary points

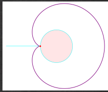
$\Leftrightarrow$ Jacobian rank drops

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

comparison: lightning self-attention

$VXX^\top K^\top QX$



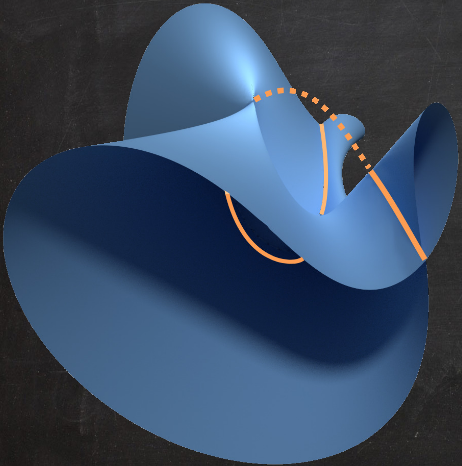
cusps 
$\Leftrightarrow$ boundary points
$\Leftrightarrow$ Jacobian rank drops

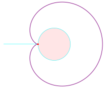
Theorem: For generic $f \in \mathcal{M}$, the only symmetries in the fiber $\mu^{-1}(f)$ are the "obvious" ones:

- layer rescalings

comparison: lightning self-attention

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



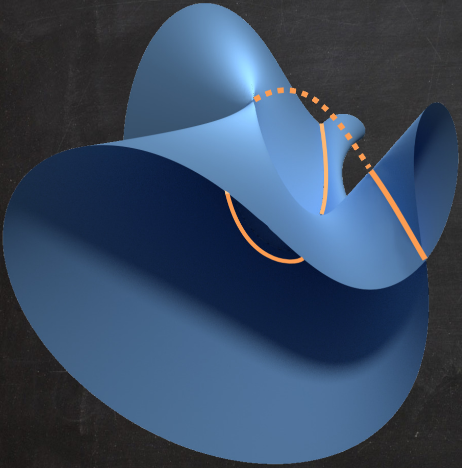
cusps 
⇔ boundary points
⇔ Jacobian rank drops

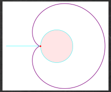
Theorem: For generic $f \in \mathcal{M}$, the only symmetries in the fiber $\mu^{-1}(f)$ are the "obvious" ones:

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

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

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
- Which properties carry over to the limit from polynomial networks to arbitrary networks?
- What happens to the neuromanifold when imposing group equivariance?
- What about ReLU networks, or more generally piecewise rational activation?

thanks for your attention!

machine learning	algebraic geometry
sample complexity	dimension
identifiability	fibers
expressivity	degree
subnetworks & hidden bias	singularities
learning dynamics	algebraic critical point theory