Singularities

Neuromanifolds $\mathcal{M}$ are **not** smooth manifolds!

Example: Linear MLPs

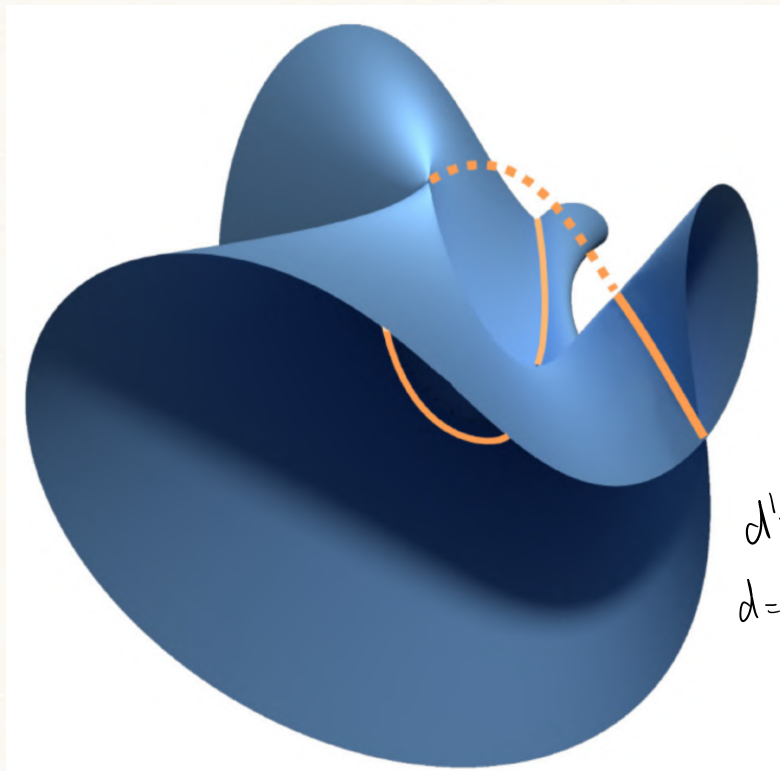
$\mathcal{M} = \{ W \in \mathbb{R}^{d_L \times d_0} \mid \mathrm{rk}(W) \leq r \}$ Zariski closed ($\rightsquigarrow \partial\mathcal{M} = \emptyset$)

$\mathrm{Sing}(\mathcal{M}) = \{ W \in \mathbb{R}^{d_L \times d_0} \mid \mathrm{rk}(W) \leq r-1 \}$ if $0 < r < \min\{d_0, d_L\}$

Example: unnormalized / linear / lightning **self-attention mechanism**

$$\begin{aligned} \mathbb{R}^{d \times t} &\longrightarrow \mathbb{R}^{d' \times t} \\ X &\longmapsto V X X^T K^T Q X \end{aligned}$$

learnable parameters $V \in \mathbb{R}^{d' \times d}$, $K, Q \in \mathbb{R}^{a \times d}$



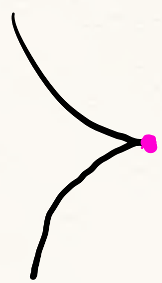
$d' = 1$

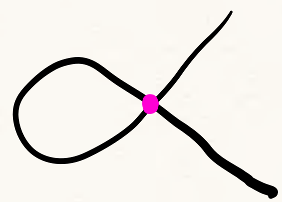
$d = t = a = 2$

Singularities

$X \subseteq \mathbb{R}^n$ irreducible variety.

Intuitive definition: $x \in X$ is singular if X doesn't look like $\mathbb{R}^{\dim X}$ locally around x.

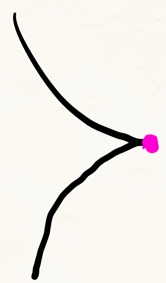


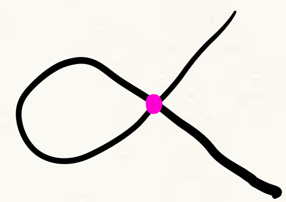


Singularities

$X \subseteq \mathbb{R}^n$ irreducible variety.

Intuitive definition: $x \in X$ is singular if X doesn't look like $\mathbb{R}^{\dim X}$ locally around x.





Formal definition: • vanishing ideal $I_X := \{ f \in \mathbb{R}[x_1, \dots, x_n] \mid \forall x \in X : f(x) = 0 \}$

Singularities

$X \subseteq \mathbb{R}^n$ irreducible variety.

Intuitive definition: $x \in X$ is singular if X doesn't look like $\mathbb{R}^{\dim X}$ locally around x.

Formal definition: • vanishing ideal $I_X := \{ f \in \mathbb{R}[x_1, \dots, x_n] \mid \forall x \in X: f(x) = 0 \}$

• generating set $\{f_1, \dots, f_s\}$, i.e., $I_X = \{ g_1 f_1 + \dots + g_s f_s \mid g_i \in \mathbb{R}[x_1, \dots, x_n] \}$ (exists by Hilbert's basis theorem)

Singularities

$X \subseteq \mathbb{R}^n$ irreducible variety.

Intuitive definition: $x \in X$ is **singular** if X doesn't look like $\mathbb{R}^{\dim X}$ locally around x.

Formal definition: • **vanishing ideal** $I_X := \{ f \in \mathbb{R}[x_1, \dots, x_n] \mid \forall x \in X : f(x) = 0 \}$

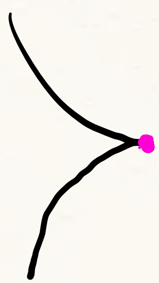
- generating set $\{f_1, \dots, f_s\}$, i.e., $I_X = \{ g_1 f_1 + \dots + g_s f_s \mid g_i \in \mathbb{R}[x_1, \dots, x_n] \}$ (exists by **Hilbert's basis theorem**)
- Jacobian $s \times n$-matrix: $J = \left[\frac{\partial f_i}{\partial x_j} \right]_{i,j}$

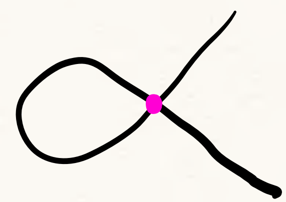
Fact: There is a subvariety $\Delta \subsetneq X$ such that for all $x \in X \setminus \Delta$, $\operatorname{rk}(J(x)) = n - \dim X$.

Singularities

$X \subseteq \mathbb{R}^n$ irreducible variety.

Intuitive definition: $x \in X$ is singular if X doesn't look like $\mathbb{R}^{\dim X}$ locally around x.





Formal definition:
- vanishing ideal $I_X := \{ f \in \mathbb{R}[x_1, \dots, x_n] \mid \forall x \in X : f(x) = 0 \}$
- generating set $\{f_1, \dots, f_s\}$, i.e., $I_X = \{ g_1 f_1 + \dots + g_s f_s \mid g_i \in \mathbb{R}[x_1, \dots, x_n] \}$ (exists by Hilbert's basis theorem)
- Jacobian $s \times n$-matrix: $J = \left[\frac{\partial f_i}{\partial x_j} \right]_{i,j}$

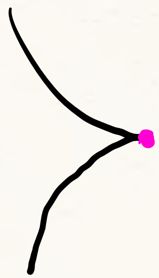
Fact: There is a subvariety $\Delta \subsetneq X$ such that for all $x \in X \setminus \Delta$, $\mathrm{rk}(J(x)) = n - \dim X$.
Such x are called smooth / regular (note: $\ker J(x) = T_x X \cong \mathbb{R}^{\dim X}$). Why?
x with $\mathrm{rk}(J(x)) < n - \dim X$ are singular; they form a subvariety $\mathrm{Sing}(X)$ of X.

Singularities

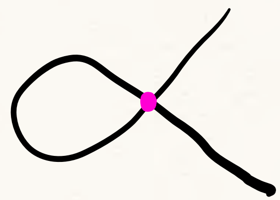
$X \subseteq \mathbb{R}^n$ irreducible variety.

Intuitive definition: $x \in X$ is singular if X doesn't look like $\mathbb{R}^{\dim X}$ locally around x.

$y^2 + x^3 = 0$



Compute Sing(X)!

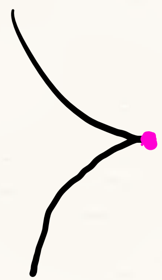


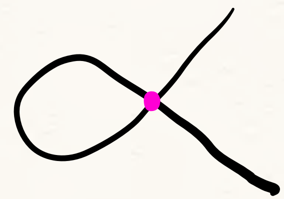
$y^2 - x^2(x+1) = 0$

Formal definition:
- vanishing ideal $I_X := \{ f \in \mathbb{R}[x_1, \dots, x_n] \mid \forall x \in X: f(x) = 0 \}$
- generating set $\{f_1, \dots, f_s\}$, i.e., $I_X = \{ g_1 f_1 + \dots + g_s f_s \mid g_i \in \mathbb{R}[x_1, \dots, x_n] \}$ (exists by Hilbert's basis theorem)
- Jacobian $s \times n$-matrix: $J = \left[\frac{\partial f_i}{\partial x_j} \right]_{i,j}$

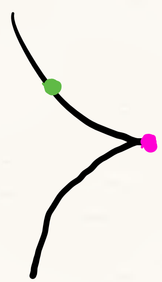
Fact: There is a subvariety $\Delta \subsetneq X$ such that for all $x \in X \setminus \Delta$, $\mathrm{rk}(J(x)) = n - \dim X$. Such x are called smooth / regular (note: $\ker J(x) = T_x X \cong \mathbb{R}^{\dim X}$). Why?
x with $\mathrm{rk}(J(x)) < n - \dim X$ are singular; they form a subvariety $\mathrm{Sing}(X)$ of X.

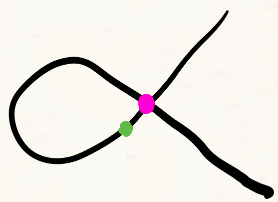
Why do we care about singularities?





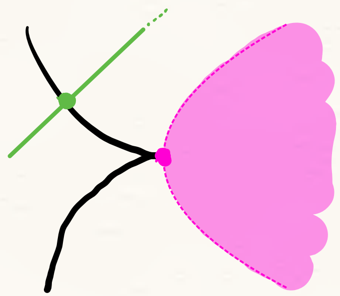
Why do we care about singularities?

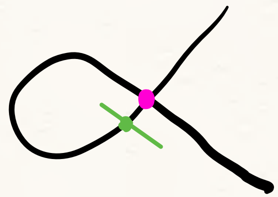




What are the Voronoi cells at • and • ?

Why do we care about singularities?

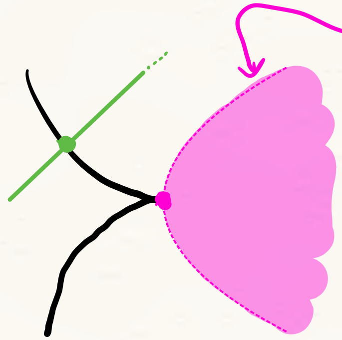




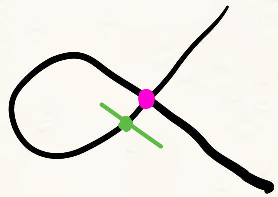
What are the Voronoi cells at • and • ?

Why do we care about singularities?

$y^2 + x^3 = 0$



Challenge: Compute this curve!



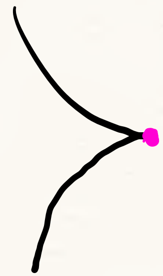
$y^2 - x^2(x+1) = 0$

What are the Voronoi cells at • and • ?

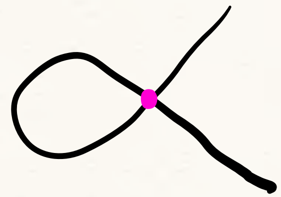
How do singularities on $\mathcal{M}$ arise?

How do singularities on $\mathcal{M}$ arise?

$\varphi : t \longmapsto (-t^2, t^3)$



$\mathrm{Br}(\varphi) = \{\bullet\}$

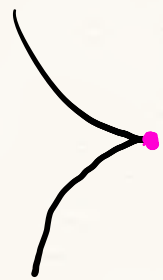


$\varphi : t \longmapsto (t^2-1,\ t(t^2-1))$

$|\varphi^{-1}(\bullet)| = 2$ (while $|\varphi^{-1}(\cdot)| = 1$ for all other points)

How do singularities on $\mathcal{M}$ arise?

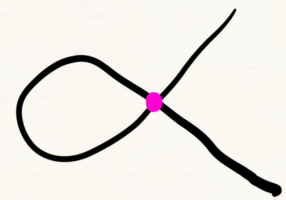
$\varphi : t \mapsto (-t^2, t^3)$



$\mathrm{Br}(\varphi) = \{\bullet\}$

φ injective

$\varphi : t \mapsto (t^2-1,\ t(t^2-1))$

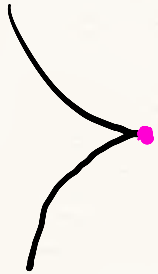


$|\varphi^{-1}(\bullet)| = 2$ (while $|\varphi^{-1}(\cdot)| = \underline{1}$ for all other points)

$\mathrm{Br}(\varphi) = \emptyset$

How do singularities on $\mathcal{M}$ arise?

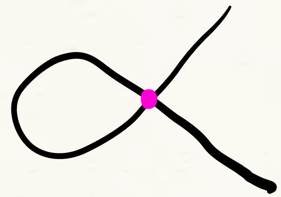
$\varphi : t \mapsto (-t^2, t^3)$



$\mathrm{Br}(\varphi) = \{{}_2\bullet\}$

φ injective

$\varphi : t \mapsto (t^2-1,\ t(t^2-1))$



$|\varphi^{-1}(\bullet)| = 2$ (while $|\varphi^{-1}(\cdot)| = 1$ for all other points)

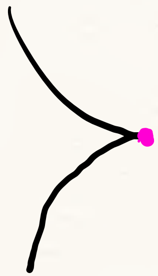
$\mathrm{Br}(\varphi) = \emptyset$

Singularities on $\mathcal{M}$ can be caused by the branch locus or the special $\underline{\text{fibers}}$ of $\varphi : \Theta \longrightarrow \mathcal{M}$.

$\varphi^{-1}(f)$

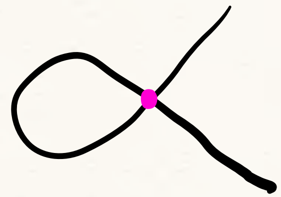
How do singularities on $\mathcal{M}$ arise?

$\varphi : t \mapsto (-t^2, t^3)$



$\mathrm{Br}(\varphi) = \{\bullet\}$

φ injective



$\varphi : t \mapsto (t^2-1, t(t^2-1))$

$|\varphi^{-1}(\bullet)| = 2$ (while $|\varphi^{-1}(\cdot)| = 1$ for all other points)

$\mathrm{Br}(\varphi) = \emptyset$

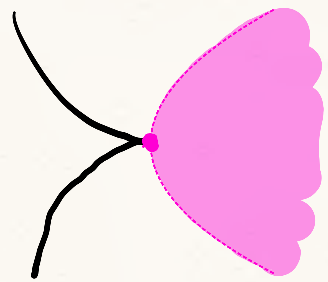
Singularities on $\mathcal{M}$ can be caused by the branch locus or the special <u>fibers</u> of $\varphi : \Theta \longrightarrow \mathcal{M}$.

$\varphi^{-1}(f)$

$\varphi : t \mapsto t^3$ has smooth image although $\mathrm{Br}(\varphi) = \{0\}$

Tradeoff: Good generalization vs. efficient optimization

$\varphi : t \mapsto (-t^2, t^3)$

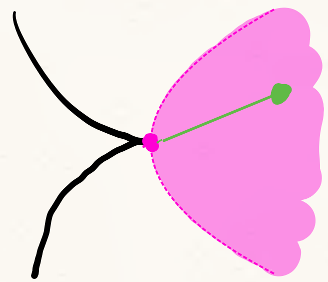


Jacobian rank drop at $t=0$:

numerical instability close to $t=0$

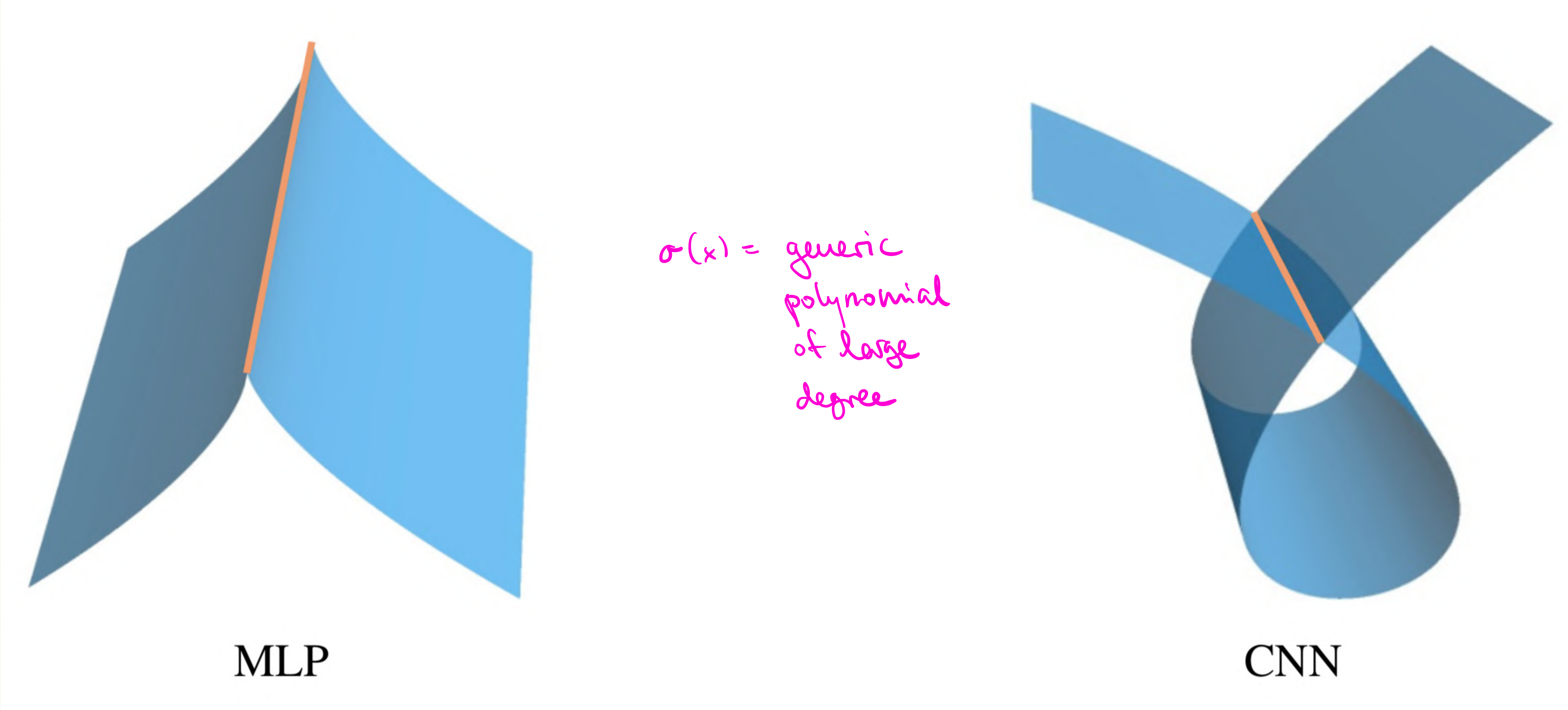
Tradeoff: Good generalization vs. efficient optimization

$\varphi : t \mapsto (-t^2, t^3)$



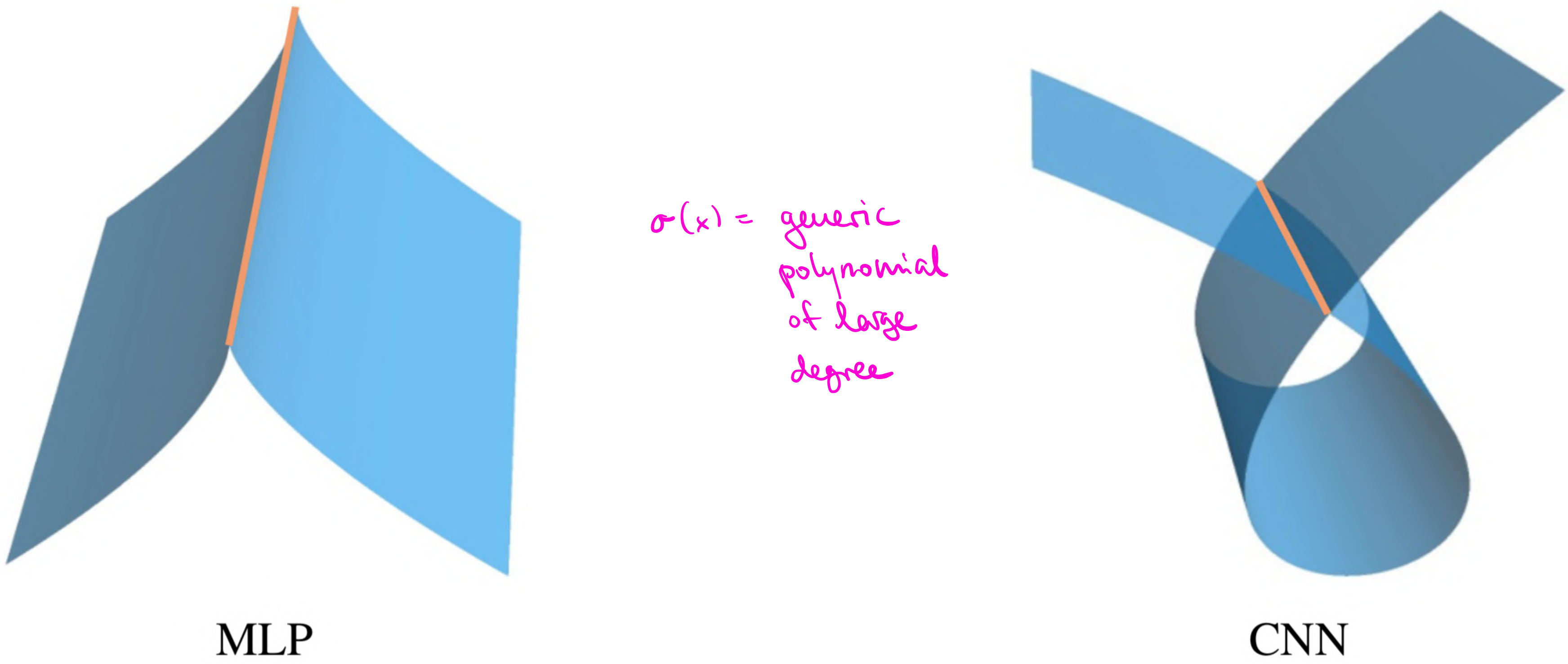
Jacobian rank drop at $t=0$:
numerical instability close to $t=0$

• is a stable solution for every • $\in \mathrm{Vor}_{\mathcal{M}}(\bullet)$,
i.e., • stays minimizer when perturbing •



Some singularities known
(work in progress to determine all)

All singularities known..



$\sigma(x) =$ generic polynomial of large degree

Some singularities known
(work in progress to determine all)

All singularities known.

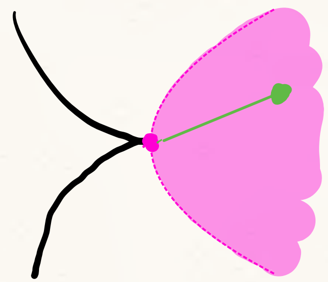
true in both cases :)

[Shahverdi, Marchetti, K. 2025]

Conjecture. The singularities of neuromanifolds $\mathcal{M}$ correspond to (certain) subnetworks, i.e., networks of a smaller architecture that can be embedded into the given architecture.

Tradeoff: Good generalization vs. efficient optimization

$\varphi: t \mapsto (-t^2, t^3)$



Jacobian rank drop at $t=0$:
numerical instability close to $t=0$

• is a stable solution for every • $\in \mathrm{Vor}_u($•$)$,
i.e., • stays minimizer when perturbing •

and • is a sparse solution (conjecturally)

Fibers

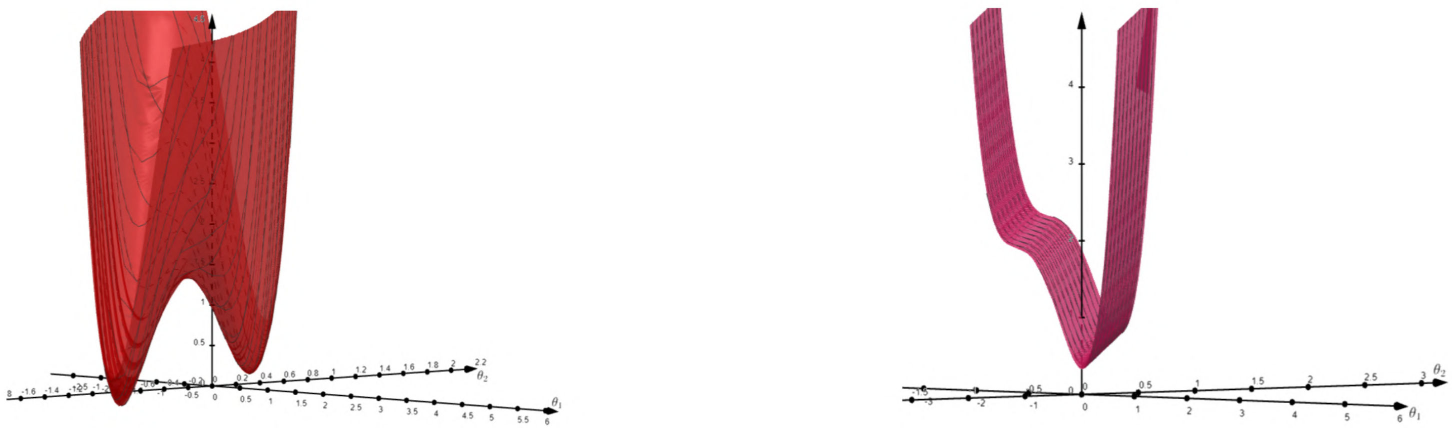
$$
\begin{array}{ccccc}
\Theta & \xrightarrow{\ \varphi\ } & \mathcal{M} & \xrightarrow{\ \mathcal{L}\ } & \mathbb{R} \\
\theta & \longmapsto & f_\theta & &
\end{array}
$$

each minimizer $\hat{f} \in \mathcal{M}$ can give many minimizers $\varphi^{-1}(\hat{f})$ in Θ $\Longleftarrow$ potentially many global minimizers $\hat{f}$ (in large model regime)

Fibers

$$\Theta \xrightarrow{\ \varphi\ } \mathcal{M} \xrightarrow{\ \mathcal{L}\ } \mathbb{R}$$
$$\theta \longmapsto f_\theta$$

each minimizer $\hat{f} \in \mathcal{M}$ can give many minimizers $\varphi^{-1}(\hat{f})$ in Θ $\Longleftarrow$ potentially many global minimizers $\hat{f}$ (in large model regime)



(a) $\mu_1 : \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix} \mapsto \begin{bmatrix}\frac{5}{4}\theta_1^2+5\theta_1\theta_2+5\theta_2^2+\theta_1+2\theta_2\\ \frac{1}{4}\theta_1^2+\theta_1\theta_2+\theta_2^2+\frac{3}{2}\theta_1+2\theta_2\end{bmatrix}$ with generic fiber $\mu_1^{-1}(\mu_1(\theta))$ being 2 points.

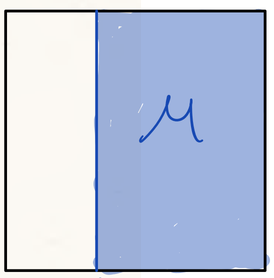
(b) $\mu_2 : \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix} \mapsto \begin{bmatrix}\frac{5}{4}\theta_1^2+5\theta_1\theta_2+5\theta_2^2+\theta_1+2\theta_2\\ \frac{1}{4}\theta_1^2+\theta_1\theta_2+\theta_2^2+\frac{3}{2}\theta_1+3\theta_2\end{bmatrix}$ with fibers $\mu_2^{-1}(\mu_2(\theta))$ being lines.

Figure 1. Loss landscapes of $\|\mu_i(\theta) - \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]\|_2^2$.

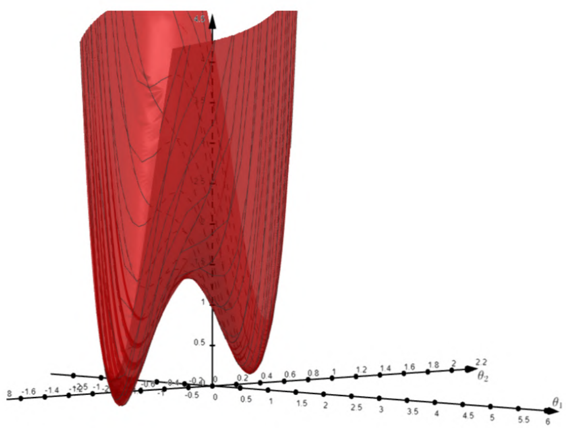
Fibers

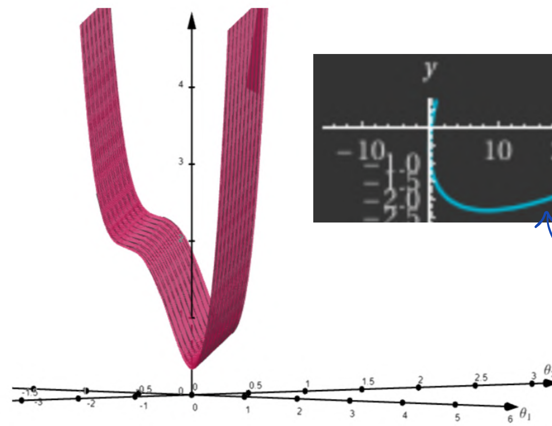
$$\begin{array}{ccccc} \Theta & \xrightarrow{\quad\varphi\quad} & \mathcal{M} & \xrightarrow{\quad\mathcal{L}\quad} & \mathbb{R} \\ \theta & \longmapsto & f_\theta & & \end{array}$$

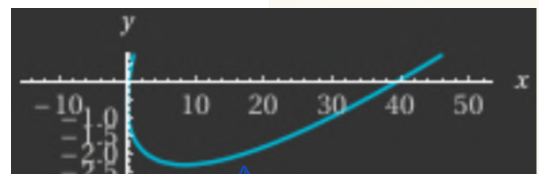
each minimizer $\hat{f} \in \mathcal{M}$ can give many minimizers $\varphi^{-1}(\hat{f})$ in Θ $\Longleftarrow$ $\underbrace{\qquad\qquad\qquad}$ potentially many global minimizers $\hat{f}$ (in large model regime)



$x = -\frac{1}{5}$







$\mathcal{M}$

$x^2 - 10xy + 25y^2 - 39x + 26y = 0$

(a) $\mu_1 : \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix} \mapsto \begin{bmatrix}\frac{5}{4}\theta_1^2+5\theta_1\theta_2+5\theta_2^2+\theta_1+2\theta_2 \\ \frac{1}{4}\theta_1^2+\theta_1\theta_2+\theta_2^2+\frac{3}{2}\theta_1+2\theta_2\end{bmatrix}$ with generic fiber $\mu_1^{-1}(\mu_1(\theta))$ being 2 points.

(b) $\mu_2 : \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix} \mapsto \begin{bmatrix}\frac{5}{4}\theta_1^2+5\theta_1\theta_2+5\theta_2^2+\theta_1+2\theta_2 \\ \frac{1}{4}\theta_1^2+\theta_1\theta_2+\theta_2^2+\frac{3}{2}\theta_1+3\theta_2\end{bmatrix}$ with fibers $\mu_2^{-1}(\mu_2(\theta))$ being lines.

Figure 1. Loss landscapes of $\|\mu_i(\theta) - \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]\|_2^2$.

Fiber-Dimension Theorem

Let $\varphi : X \longrightarrow \mathbb{R}^m$ morphism between irreducible varieties.

Zariski closure

$$\Rightarrow \; Y := \overline{\varphi(X)} \text{ is irreducible}$$

Thm.: There is a subvariety $\Delta \subsetneq X$ such that for all $x \in X \setminus \Delta$,

$$\dim X = \dim Y + \dim \varphi^{-1}(\varphi(x)).$$

Does this ring a bell from linear algebra?

Some known parameter symmetries

$$\Theta \xrightarrow{\varphi} \mathcal{M}$$

[Shahverdi, Marchetti, K. 2025]

Finkel, Rodriguez, Wu, Yahl 2024
Usevich, Dérand, Borsoi, Clausel 2025

MLPs: $\sigma(x) =$ generic polynomial of large degree: $\varphi^{-1}(\varphi(\Theta))$ finite generically

(almost proven: only permuting neurons)

$\sigma(x) =$ large degree monomial: $|\varphi^{-1}(\varphi(\Theta))| = \infty$, coming from permuting & scaling neurons

$\sigma(x) =$ sigmoid or tanh: $\varphi^{-1}(\varphi(\Theta))$ generically finite, coming from permutating neurons (and sign flips for tanh)

Fefferman 1994
Vlačić, Bölcskei 2022

Some known parameter symmetries

$$\Theta \xrightarrow[\varphi]{} \mathcal{M}$$

[Shahverdi, Marchetti, K. 2025]

Finkel, Rodriguez, Wu, Yahl 2024
Usevich, Dérand, Borsoi, Clausel 2025

MLPs: $\sigma(x)$ = generic polynomial of large degree: $\varphi^{-1}(\varphi(\Theta))$ finite generically
(almost proven: only permuting neurons)

$\sigma(x)$ = large degree monomial: $|\varphi^{-1}(\varphi(\Theta))| = \infty$, coming from permuting & scaling neurons

$\sigma(x)$ = sigmoid or tanh: $\varphi^{-1}(\varphi(\Theta))$ generically finite, coming from permutating neurons (and sign flips for tanh)

Fefferman 1994
Vlačić, Bölcskei 2022

CNNs: $\sigma(x)$ = generic polynomial of large degree: $\varphi^{-1}(\varphi(\Theta)) = \{\Theta\}$ for generic Θ

Some known parameter symmetries

$$\Theta \xrightarrow[\varphi]{} \mathcal{M}$$

Finkel, Rodriguez, Wu, Yahl 2024
Usevich, Dérand, Borsoi, Clausel 2025

[Shahverdi, Marchetti, K. 2025]

MLPs: $\sigma(x)$ = generic polynomial of large degree: $\varphi^{-1}(\varphi(\Theta))$ finite generically
(almost proven: only permuting neurons)

$\sigma(x)$ = large degree monomial: $|\varphi^{-1}(\varphi(\Theta))| = \infty$, coming from permuting & scaling neurons

$\sigma(x)$ = sigmoid or tanh: $\varphi^{-1}(\varphi(\Theta))$ generically finite, coming from permutating neurons (and sign flips for tanh)

Fefferman 1994
Vlačić, Bölcskei 2022

CNNs: $\sigma(x)$ = generic polynomial of large degree: $\varphi^{-1}(\varphi(\Theta)) = \{\Theta\}$ for generic Θ

Attention: 1 layer = $\mathbb{R}^{d \times t} \longrightarrow \mathbb{R}^{d' \times t}$

$$X \longmapsto V X \eta(X^T K^T Q X)$$

Can you see some parameter symmetries?

Some known parameter symmetries

$$\Theta \xrightarrow[\varphi]{} \mathcal{M}$$

[Shahverdi, Marchetti, K. 2025]

Finkel, Rodriguez, Wu, Yahl 2024
Usevich, Dérand, Borsoi, Clausel 2025

MLPs: $\sigma(x)$ = generic polynomial of large degree: $\varphi^{-1}(\varphi(\Theta))$ finite generically
(almost proven: only permuting neurons)

$\sigma(x)$ = large degree monomial: $|\varphi^{-1}(\varphi(\Theta))| = \infty$, coming from permuting & scaling neurons

$\sigma(x)$ = sigmoid or tanh: $\varphi^{-1}(\varphi(\Theta))$ generically finite, coming from permutating neurons (and sign flips for tanh)

Fefferman 1994
Vlačić, Bölcskei 2022

CNNs: $\sigma(x)$ = generic polynomial of large degree: $\varphi^{-1}(\varphi(\Theta)) = \{\Theta\}$ for generic Θ

Attention: 1 layer = $\mathbb{R}^{d \times t} \longrightarrow \mathbb{R}^{d' \times t}$

$$X \longmapsto V X \eta(X^T K^T Q X)$$

Can you see some parameter symmetries?

$\cdot\, G G^{-1} \cdot$

$V \mapsto CV$ in one layer &
$V' \mapsto V'C^{-1}$, $Q' \mapsto QC^{-1}$, $K' \mapsto KC^{-1}$ in next layer

Conjecture: These are all parameter symmetries generically (when η is the standard softmax).

Some known parameter symmetries

$$\Theta \xrightarrow[\varphi]{} \mathcal{M}$$

Finkel, Rodriguez, Wu, Yahl 2024
Usevich, Dérand, Borsoi, Clausel 2025

[Shahverdi, Marchetti, K. 2025]

MLPs: $\sigma(x)$ = generic polynomial of large degree: $\varphi^{-1}(\varphi(\Theta))$ finite generically
(almost proven: only permuting neurons)

$\sigma(x)$ = large degree monomial: $|\varphi^{-1}(\varphi(\Theta))| = \infty$, coming from permuting & scaling neurons

$\sigma(x)$ = sigmoid or tanh: $\varphi^{-1}(\varphi(\Theta))$ generically finite, coming from permutating neurons (and sign flips for tanh)

Fefferman 1994
Vlačić, Bölcskei 2022

CNNs: $\sigma(x)$ = generic polynomial of large degree: $\varphi^{-1}(\varphi(\Theta)) = \{\Theta\}$ for generic Θ

Attention: 1 layer = $\mathbb{R}^{d \times t} \longrightarrow \mathbb{R}^{d' \times t}$

$$X \longmapsto V X \eta(X^T K^T Q X)$$

Can you see some parameter symmetries?

$\cdot\, G G^{-1} \cdot$

$V \mapsto CV$ in one layer &
$V' \mapsto V'C^{-1}$, $Q' \mapsto QC^{-1}$, $K' \mapsto KC^{-1}$ in next layer

Conjecture: These are all parameter symmetries generically (when η is the standard softmax).

Henry, Marchetti, K. 2025

Theorem: For lightning attention ($\eta = \mathrm{id}$), the only other generic parameter symmetry is scaling by a scalar within each layer.

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 $\mathcal{D}$

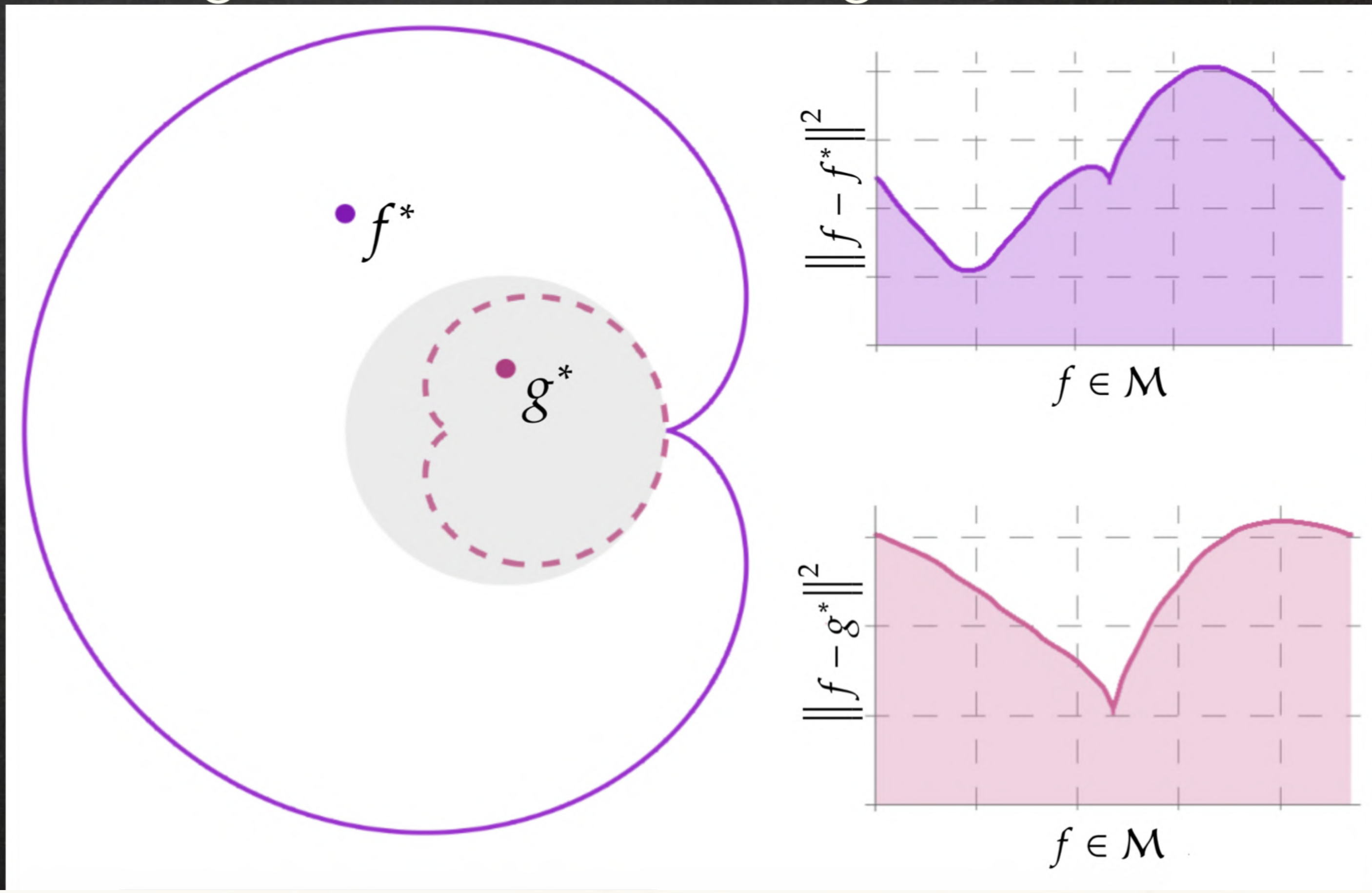
$\rightsquigarrow$ measures intrinsic optimization degree

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



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

Position: Algebra Unveils Deep Learning
An Invitation to Neuroalgebraic Geometry

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

Abstract

In this position paper, 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. Along the way, we review the literature and discuss ideas beyond the algebraic domain. This work lays the foundations of a research direction bridging algebraic geometry and deep learning, that we refer to as neuroalgebraic geometry.

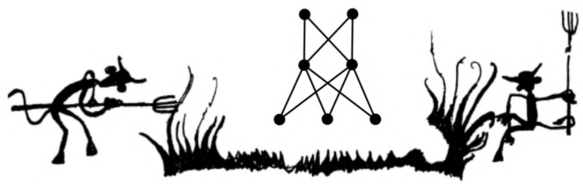


Figure 1. A neural variation of a celebrated doodle from the algebraic geometry literature (Grothendieck, 1968).

model towards an estimate of the ground-truth function. Consequently, geometric problems over neuromanifolds, such as nearest point problems, govern the training dynamics and provide insights into how neural networks learn.

Therefore, understanding the geometry of neuromanifolds offers a twofold potential. First, it serves as a powerful theoretical framework for analyzing and explaining empir-

Spotlight
Position
Paper
@ICML '25

thanks to

my amazing group

and co-authors:

Vahid Shahverdi

Giovanni Marchetti

Guido Montúfar

Stefano Mereta

Nathan Henry

Matthew Trager

Joan Bruna

Thomas Merkh

El Mehdi Achour

Holger Rauhut