Algebraic Neural Network Theory

Kathlén Kohn

KTH
Digital Futures

deep learning in a nutshell

parameter space

function space

Goal:
find θ that minimize

θ
learnable weights

$\longmapsto$

f_θ
network function

$\longmapsto$

$L(f_\theta)$
loss value

deep learning in a nutshell

neuromanifold (Amari et. al 2001)

extremely hard to understand!

Idea:
study polynomial networks

parameter space

function space

Goal:
find θ that minimize

θ learnable weights	$\longmapsto$	f_θ network function	$\longmapsto$	$L(f_\theta)$ loss value

Example: MLPs

multilayer perceptrons

$$\alpha_L \circ \sigma \circ \dots \circ \sigma \circ \alpha_2 \circ \sigma \circ \alpha_1$$

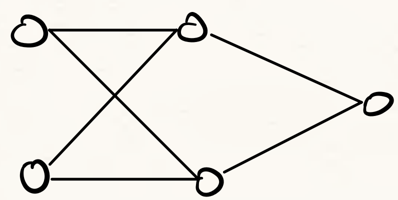
α_i = learnable affine linear functions

σ = nonlinear activation function, applied entrywise

we assume: σ is a univariate polynomial

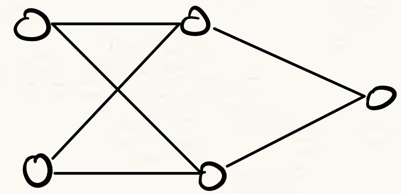
Ex: $\sigma(x) = x^2$

$$\begin{bmatrix} e & f \end{bmatrix} \sigma\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right)$$



Which functions does this MLP parametrize?

Ex.: $\sigma(x) = x^2$



$$[e\ \ f]\ \sigma\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}\right)$$

Which functions does this MLP parametrize?

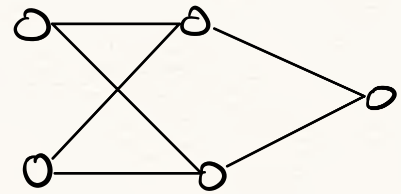
$$e(ax+by)^2 + f(cx+dy)^2$$
$$= \underbrace{(a^2e + c^2f)}_{A}x^2 + \underbrace{2(abe + cdf)}_{B}xy + \underbrace{(b^2e + d^2f)}_{C}y^2$$

Can you obtain **all** of $\mathbb{R}[x,y]_2$?

↖ homogeneous quadratic polynomials in x,y

i.e., are all values for A, B, C possible?

Ex.: $\sigma(x) = x^2$



$$\begin{bmatrix} e & f \end{bmatrix} \sigma\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right)$$

Which functions does this MLP parametrize?

$$e(ax+by)^2 + f(cx+dy)^2$$
$$= \underbrace{(a^2e + c^2f)}_{A} x^2 + \underbrace{2(abe + cdf)}_{B} xy + \underbrace{(b^2e + d^2f)}_{C} y^2$$

Can you obtain <u>all</u> of $\mathbb{R}[x,y]_2$?

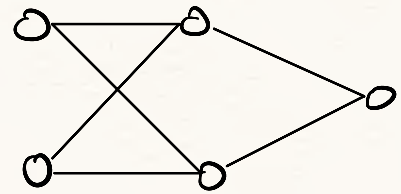
↖ homogeneous quadratic polynomials in x,y

i.e., are all values for A, B, C possible?

YES

What about $\sigma(x) = x^3$?

Ex.: $\sigma(x) = x^3$



$$\begin{bmatrix} e & f \end{bmatrix} \sigma\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}\right)$$

Which functions does this MLP parametrize?

$$e(ax+by)^3 + f(cx+dy)^3$$
$$= \underbrace{(a^3e+c^3f)}_{A}x^3 + \underbrace{3(a^2be+c^2df)}_{B}x^2y + \underbrace{3(ab^2e+cd^2f)}_{C}xy^2 + \underbrace{(b^3e+d^3f)}_{D}y^3$$

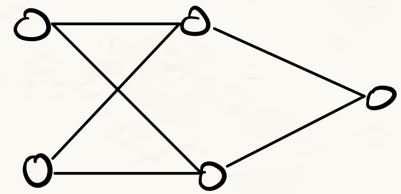
Can you obtain <u>all</u> of $\mathbb{R}[x,y]_3$?

↖ homogeneous cubic polynomials in x,y

i.e., are all values for A,B,C,D possible?

Ex.: $\sigma(x) = x^3$

$$\begin{bmatrix} e & f \end{bmatrix} \sigma\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}\right)$$



Which functions does this MLP parametrize?

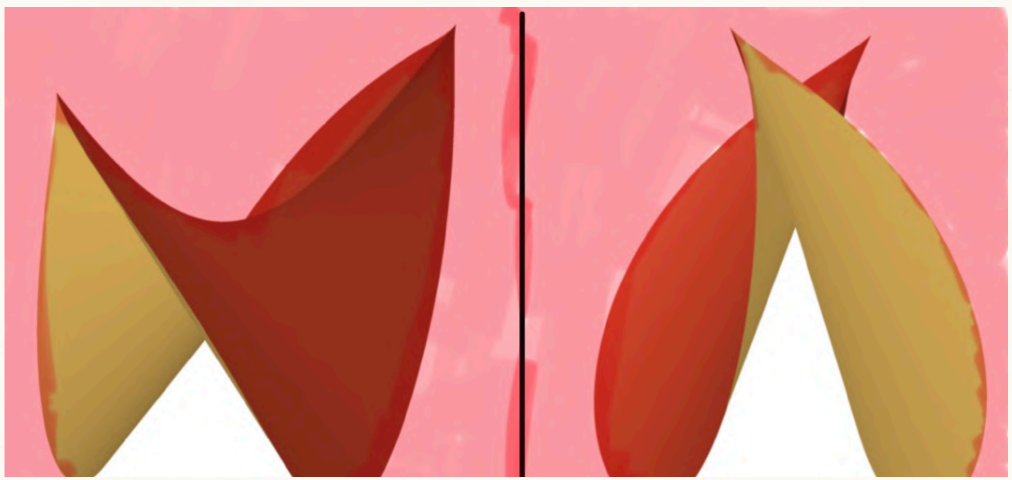
$$e(ax+by)^3 + f(cx+dy)^3$$
$$= \underbrace{(a^3e + c^3f)}_{A}x^3 + \underbrace{3(a^2be + c^2df)}_{B}x^2y + \underbrace{3(ab^2e + cd^2f)}_{C}xy^2 + \underbrace{(b^3e + d^3f)}_{D}y^3$$

Can you obtain <u>all</u> of $\mathbb{R}[x,y]_3$?

↖ homogeneous cubic polynomials in x,y

i.e., are all values for A, B, C, D possible?

No, e.g. $A = 1$, $B = 0$, $C = -1$, $D = 0$



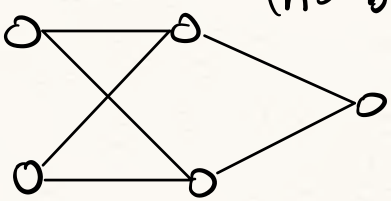
Neuromanifolds

A **parametric machine learning** model is a map $\mu: \Theta \times X \longrightarrow Y$.

parameters inputs outputs

Its **neuromanifold** is $\mathcal{M} := \{\mu(\theta,\cdot): X \to Y \mid \theta \in \Theta\}$.

Example MLPs:



(no bias vectors = linear layer functions)

$\sigma(x) = x^2 \quad \Rightarrow \quad \mathcal{M} = \mathbb{R}[x,y]_2$

$\sigma(x) = x^3 \quad \Rightarrow \quad \mathcal{M} \subsetneq \mathbb{R}[x,y]_3$

$\sigma(x) = x \quad \Rightarrow \quad$?

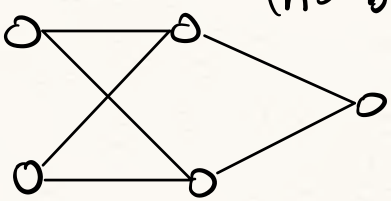
Neuromanifolds

A **parametric machine learning** model is a map $\mu : \Theta \times X \longrightarrow Y$.

parameters inputs outputs

Its **neuromanifold** is $\mathcal{M} := \{ \mu(\theta, \cdot) : X \to Y \mid \theta \in \Theta \}$.

Example MLPs:

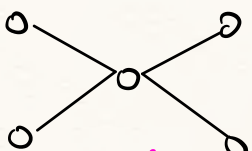


(no bias vectors = linear layer functions)

$\sigma(x) = x^2 \quad \Rightarrow \quad \mathcal{M} = \mathbb{R}[x,y]_2$

$\sigma(x) = x^3 \quad \Rightarrow \quad \mathcal{M} \subsetneq \mathbb{R}[x,y]_3$

$\sigma(x) = x \quad \Rightarrow \quad \mathcal{M} = \mathbb{R}^{1\times 2}$



$\sigma(x) = x$

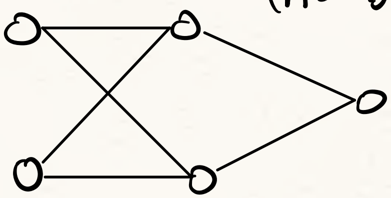
$\begin{bmatrix} c \\ d \end{bmatrix} \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad \Rightarrow \quad \mathcal{M} = ?$

Neuromanifolds

A **parametric machine learning** model is a map $\mu : \underbrace{\Theta}_{\text{parameters}} \times \underbrace{X}_{\text{inputs}} \longrightarrow \underbrace{Y}_{\text{outputs}}$.

Its **neuromanifold** is $\mathcal{M} := \{ \mu(\theta, \cdot) : X \to Y \mid \theta \in \Theta \}$.

Example MLPs:

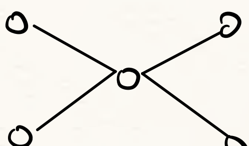


(no bias vectors = linear layer functions)

$\sigma(x) = x^2 \quad \Rightarrow \quad \mathcal{M} = \mathbb{R}[x,y]_2$

$\sigma(x) = x^3 \quad \Rightarrow \quad \mathcal{M} \subsetneq \mathbb{R}[x,y]_3$

$\sigma(x) = x \quad \Rightarrow \quad \mathcal{M} = \mathbb{R}^{1 \times 2}$



$\sigma(x) = x$

$\begin{bmatrix} c \\ d \end{bmatrix} \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad \Rightarrow \quad \mathcal{M} = \{ W \in \mathbb{R}^{2\times 2} \mid \operatorname{rk}(W) \leq 1 \}$

Linear MLPs: $\quad \alpha_L \circ \dots \circ \alpha_2 \circ \alpha_1$, where

$\alpha_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ linear

$$\Rightarrow \mathcal{M} = \{ W \in \mathbb{R}^{d_L \times d_0} \mid \operatorname{rk}(W) \leq \min\{d_0, d_1, \dots, d_L\} \}$$

Linear MLPs: $\alpha_L \circ \dots \circ \alpha_2 \circ \alpha_1$, where

$\alpha_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ linear

$$\Rightarrow \mathcal{M} = \{ W \in \mathbb{R}^{d_L \times d_0} \mid \mathrm{rk}(W) \leq \min\{d_0, d_1, \dots, d_L\} \}$$

Polynomial MLPs: $\alpha_L \circ \sigma \circ \dots \circ \sigma \circ \alpha_2 \circ \sigma \circ \alpha_1$, where

$\alpha_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ affine linear

$\sigma \in \mathbb{R}[x]_{\leq \delta}$

$\Rightarrow \mathcal{M}$ lives in a <u>finite-dimensional</u> vector space, namely

$$\left(\mathbb{R}[x_1, \dots, x_{d_0}]_{\leq \delta^{L-1}} \right)^{d_L}$$

Linear MLPs: $\alpha_L \circ \dots \circ \alpha_2 \circ \alpha_1$, where

$\alpha_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ linear

$$\Rightarrow \mathcal{M} = \{ W \in \mathbb{R}^{d_L \times d_0} \mid \mathrm{rk}(W) \leq \min\{d_0, d_1, \dots, d_L\} \}$$

Polynomial MLPs: $\alpha_L \circ \sigma \circ \dots \circ \sigma \circ \alpha_2 \circ \sigma \circ \alpha_1$, where

$\alpha_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ affine linear

$\sigma \in \mathbb{R}[x]_{\leq \delta}$

$\Rightarrow \mathcal{M}$ lives in a finite-dimensional vector space, namely

$$\left(\mathbb{R}[x_1, \dots, x_{d_0}]_{\leq \delta^{L-1}} \right)^{d_L}$$

Polynomial MLPs are the only ones with that property!

Universal Approximation Theorem

Leshno, Lin, Pinkus, Schocken: Multilayer feedforward networks with a non-polynomial activation function can approximate any function.
Neural Networks 6, 1993:

Theorem 1:

Let $\sigma \in M$. Set

$$\Sigma_n = \text{span}\,\{\sigma(\mathbf{w}\cdot\mathbf{x}+\theta) : \mathbf{w} \in R^n, \theta \in R\}.$$

Then Σ_n is dense in $C(R^n)$ if and only if σ is not an algebraic polynomial (a.e.).

Universal Approximation Theorem

Leshno, Lin, Pinkus, Schocken: Multilayer feedforward networks with a non-polynomial activation function can approximate any function.
Neural Networks 6, 1993:

Theorem 1:

Let $\sigma \in M$. Set

$$\Sigma_n = \operatorname{span}\{\sigma(\mathbf{w}\cdot\mathbf{x}+\theta) : \mathbf{w}\in R^n, \theta \in R\}.$$

Then Σ_n is dense in $C(R^n)$ if and only if σ is not an algebraic polynomial (a.e.).

polynomials are the choice to approximate networks with finite-dimensional models

AG approach

$$\begin{array}{c} \boxed{\text{algebraic network}} \\ \big\downarrow \text{Stone-Weierstraß} \\ \boxed{\text{target network}} \end{array}$$

nonlinear models in finite-dimensional ambient spaces

NTK approach

neural tangent kernel

$\boxed{\text{target network}}$

$\Big\downarrow$ increase width

$\boxed{\infty\text{-width network}}$

linearized models of ∞ dimension

AG approach

algebraic geometry

$\boxed{\text{algebraic network}}$ nonlinear models in finite-dimensional ambient spaces

Stone-Weierstraß $\Big\downarrow$

$\boxed{\text{target network}}$

Neural Tangent Kernel: Convergence and Generalization in Neural Networks

Arthur Jacot, Franck Gabriel, Clement Hongler

Advances in Neural Information Processing Systems 31 (NeurIPS 2018)

> 4000 citations

Network training = 'distance' minimization

Let $\mathcal{M} \subseteq V := \left(\mathbb{R}[x_1, \ldots, x_{d_0}]_{\leq D}\right)^{d_L}$,

neuromanifold

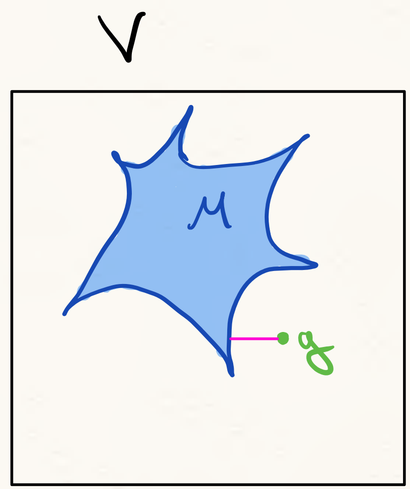
$S \subseteq \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ finite dataset,

mean squared error

MSE loss : $\mathcal{L}(f) := \sum_{(a,b) \in S} \| f(a) - b \|^2$

[dist$(f,g) = 0$ possible for $f \neq g$]

Proposition: There is a pseudometric dist $: V \times V \to \mathbb{R}_{\geq 0}$ and some $g \in V$ such that minimizing $\mathcal{L}(f)$ over $f \in \mathcal{M}$ is equivalent to minimizing dist(f,g) over $f \in \mathcal{M}$.



Why ?

Network training = 'distance' minimization

Let $\mathcal{M} \subseteq V := \left(\mathbb{R}[x_1, \ldots, x_{d_0}]_{\leq D}\right)^{d_L}$,

neuromanifold

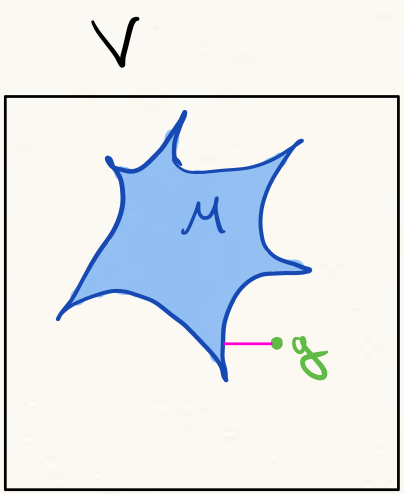
$S \subseteq \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ finite dataset,

mean squared error

MSE loss: $\mathcal{L}(f) := \sum_{(a,b) \in S} \|f(a) - b\|^2$

[$\mathrm{dist}(f,g) = 0$ possible for $f \neq g$]

Proposition: There is a pseudometric $\mathrm{dist}: V \times V \to \mathbb{R}_{\geq 0}$ and some $g \in V$ such that minimizing $\mathcal{L}(f)$ over $f \in \mathcal{M}$ is equivalent to minimizing $\mathrm{dist}(f,g)$ over $f \in \mathcal{M}$.



Assume: $d_L = 1$

Veronese embedding

Let $\nu_D: (x_1, \ldots, x_{d_0}) \longmapsto$ (all monomials in $x_1, \ldots, x_{d_0}$ of degree $\leq D$),

c_f be coefficient vector of $f \in V$ such that $f(x) = \nu_D(x) \cdot c_f$,

Network training = 'distance' minimization

Let $\mathcal{M} \subseteq V := \left(\mathbb{R}[x_1, \ldots, x_{d_0}]_{\leq D}\right)^{d_L}$,

neuromanifold

$S \subseteq \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ finite dataset,

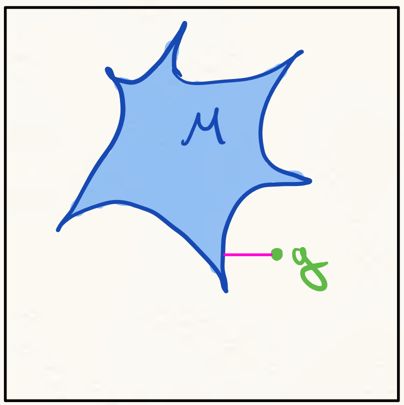
mean squared error

MSE loss: $\mathcal{L}(f) := \sum_{(a,b) \in S} \| f(a) - b \|^2$

[dist(f,g) = 0 possible for $f \neq g$]

Proposition: There is a pseudometric dist : $V \times V \to \mathbb{R}_{\geq 0}$ and some $g \in V$ such that minimizing $\mathcal{L}(f)$ over $f \in \mathcal{M}$ is equivalent to minimizing dist(f, g) over $f \in \mathcal{M}$.

V



Assume: $d_L = 1$

Veronese embedding

Let $\nu_D : (x_1, \ldots, x_{d_0}) \longmapsto$ (all monomials in $x_1, \ldots, x_{d_0}$ of degree $\leq D$),

c_f be coefficient vector of $f \in V$ such that $f(x) = \nu_D(x) \cdot c_f$,

A & B matrices whose rows are $\nu_D(a)$ & b, resp., over all $(a,b) \in S$

$\Rightarrow \mathcal{L}(f) = \| A c_f - B \|^2$

Network training = 'distance' minimization

Let $\mathcal{M} \subseteq V := \left(\mathbb{R}[x_1, \dots, x_{d_0}]_{\leq D}\right)^{d_L}$,

neuromanifold

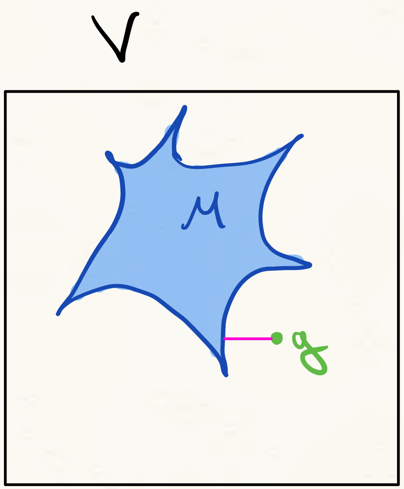
$S \subseteq \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ finite dataset,

mean squared error

MSE loss: $\mathcal{L}(f) := \sum_{(a,b)\in S} \| f(a) - b\|^2$

[dist(f,g) = 0 possible for $f \neq g$]

Proposition: There is a pseudometric dist : $V \times V \to \mathbb{R}_{\geq 0}$ and some $g \in V$ such that minimizing $\mathcal{L}(f)$ over $f \in \mathcal{M}$ is equivalent to minimizing dist(f,g) over $f \in \mathcal{M}$.



Assume: $d_L = 1$

Veronese embedding

Let $\nu_D : (x_1, \dots, x_{d_0}) \longmapsto$ (all monomials in $x_1, \dots, x_{d_0}$ of degree $\leq D$),

c_f be coefficient vector of $f \in V$ such that $f(x) = \nu_D(x) \cdot c_f$,

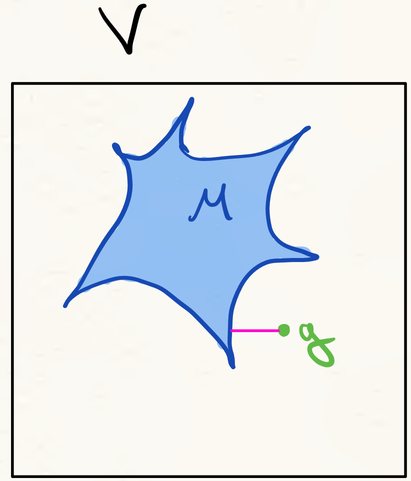
A & B matrices whose rows are $\nu_D(a)$ & b, resp., over all $(a,b) \in S$

pseudoinverse

$$\Rightarrow \mathcal{L}(f) = \|A c_f - B\|^2 = \|c_f - A^+ B\|^2_{A^T A} + \text{const.}$$

$\|c\|_Q := c^T Q c$

$$\underset{f\in\mathcal{M}}{\operatorname{argmin}}\ \mathcal{L}(f) = \underset{f\in\mathcal{M}}{\operatorname{argmin}}\ \| c_f - A^+B \|^2_{A^TA}$$



Observations $(d_L = 1)$:

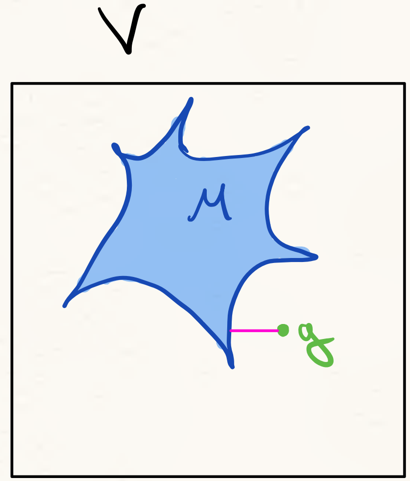
① A^TA depends only on input data,
A^+B on both input & output

(LLMs: $|S| < \dim \mathcal{M}$)

② $A^TA \in \mathbb{R}^{\dim V \times \dim V}$ is rank-deficient whenever $|S| < \dim V$ $\rightsquigarrow$ <u>pseudo</u>metric

③

$$\underset{f\in\mathcal{M}}{\operatorname{argmin}}\ \mathcal{L}(f) = \underset{f\in\mathcal{M}}{\operatorname{argmin}}\ \| c_f - A^+B\|^2_{A^TA}$$



Observations ($d_L=1$):

1. A^TA depends only on input data, A^+B on both input & output

2. $A^TA \in \mathbb{R}^{\dim V \times \dim V}$ is rank-deficient whenever $|S| < \dim V$ (LLMs: $|S| < \dim \mathcal{M}$) $\rightsquigarrow$ <u>pseudo</u>metric

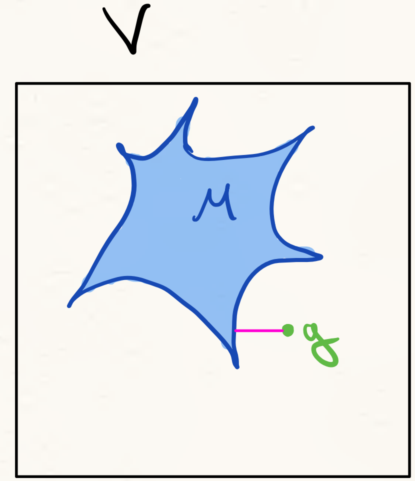
3. even when $|S| \gg \dim V$, A^TA is <u>not</u> an arbitrary symmetric PD matrix, while A^+B yields all vectors $\in \mathbb{R}^{\dim V}$

Why?

Which matrices can be obtained?

(try for $d_0 = 1$: $v(x) = (1, x, x^2, \dots, x^D)$)

$$\underset{f \in \mathcal{M}}{\operatorname{argmin}} \; \mathcal{L}(f) = \underset{f \in \mathcal{M}}{\operatorname{argmin}} \; \| c_f - A^+ B \|^2_{A^T A}$$



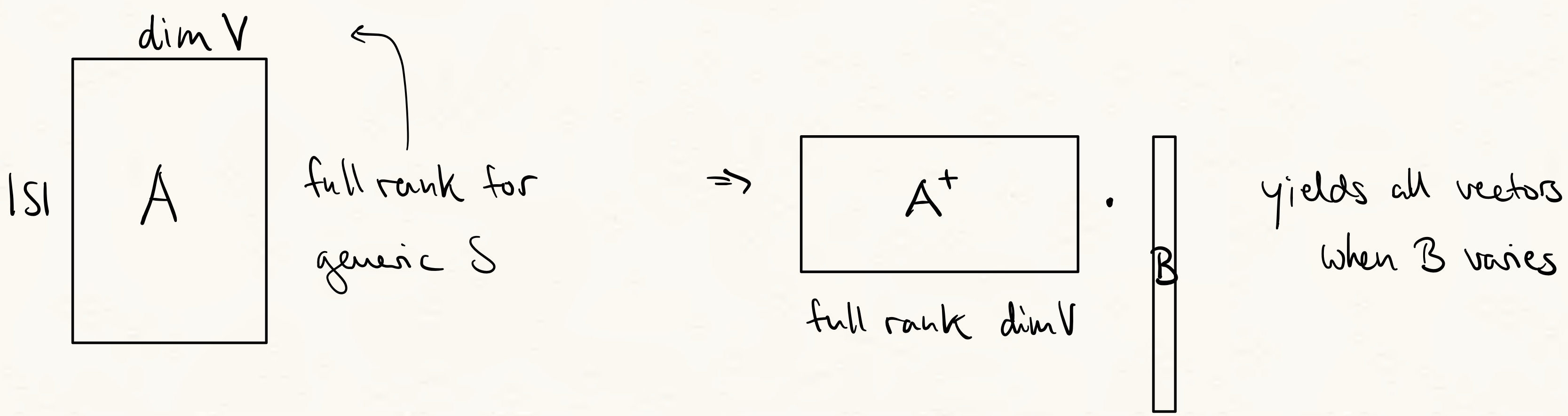
Observations $(d_L = 1)$:

① $A^T A$ depends only on input data,
$A^+ B$ on both input & output

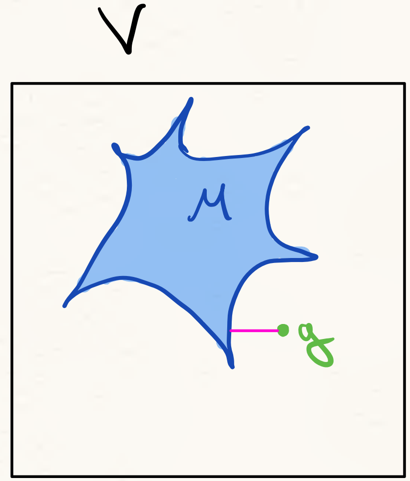
② $A^T A \in \mathbb{R}^{\dim V \times \dim V}$ is rank-deficient whenever $|S| < \dim V$ $\rightsquigarrow$ pseudometric

(LLMs: $|S| < \dim \mathcal{M}$)

③ even when $|S| \gg \dim V$, $A^T A$ is not an arbitrary symmetric PD matrix,
while $A^+ B$ yields all vectors $\in \mathbb{R}^{\dim V}$



$$\operatorname*{argmin}_{f\in\mathcal{M}} \mathcal{L}(f) = \operatorname*{argmin}_{f\in\mathcal{M}} \|c_f - A^+B\|^2_{A^TA}$$



Observations ($d_L = 1$):

1. A^TA depends only on input data, A^+B on both input & output

(LLMs: $|S| < \dim \mathcal{M}$)

2. $A^TA \in \mathbb{R}^{\dim V \times \dim V}$ is rank-deficient whenever $|S| < \dim V$ ⇝ <u>pseudo</u>metric

3. even when $|S| \gg \dim V$, A^TA is <u>not</u> an arbitrary symmetric PD matrix, while A^+B yields all vectors $\in \mathbb{R}^{\dim V}$

$$A^TA = \begin{array}{c} \\ i \rightarrow \end{array}\begin{bmatrix} | & & | \\ v(a_1) & \cdots & v(a_{|S|}) \\ | & & | \end{bmatrix} \begin{array}{c} \downarrow j \\ \begin{bmatrix} — \; v(a_1) \; — \\ \vdots \\ — \; v(a_{|S|}) \; — \end{bmatrix} \end{array}$$

has (i,j) entry $\displaystyle\sum_{(a,b)\in S} \underbrace{v_i(a)\, v_j(a)}$

monomial of degree $\le 2D$
that can be factored in several ways

Ex.: $d_0 = 1$

$$\Rightarrow v(x) = (1, x, x^2, \dots, x^D)$$

$$\Rightarrow A = \begin{bmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^D \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_{|S|} & a_{|S|}^2 & \cdots & a_{|S|}^D \end{bmatrix}$$ Vandermonde matrix

$$\Rightarrow A^TA = \begin{bmatrix} |S| & \sum a_k & \sum a_k^2 & \cdots & \sum a_k^D \\ \sum a_k & \sum a_k^2 & \sum a_k^3 & \cdots & \sum a_k^{D+1} \\ \sum a_k^2 & \sum a_k^3 & \sum a_k^4 & \cdots & \sum a_k^{D+2} \\ \vdots & \vdots & \vdots & & \\ \sum a_k^D & \sum a_k^{D+1} & \sum a_k^{D+2} & \cdots & \sum a_k^{2D} \end{bmatrix}$$ Hankel matrix

Ex.: $d_0 = 1$

$$\Rightarrow v(x) = (1, x, x^2, \dots, x^D)$$

$$\Rightarrow A = \begin{bmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^D \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_{|S|} & a_{|S|}^2 & \cdots & a_{|S|}^D \end{bmatrix} \quad \text{Vandermonde matrix}$$

$$\Rightarrow A^TA = \begin{bmatrix} |S| & \sum a_k & \sum a_k^2 & \cdots & \sum a_k^D \\ \sum a_k & \sum a_k^2 & \sum a_k^3 & \cdots & \sum a_k^{D+1} \\ \sum a_k^2 & \sum a_k^3 & \sum a_k^4 & \cdots & \sum a_k^{D+2} \\ \vdots & \vdots & \vdots & & \\ \sum a_k^D & \sum a_k^{D+1} & \sum a_k^{D+2} & \cdots & \sum a_k^{2D} \end{bmatrix} \quad \text{Hankel matrix}$$

Ex.: $d_0 = 2, \ D = 2$

$$\Rightarrow v(x,y) = (1, x, y, x^2, xy, y^2)$$

$$\Rightarrow A^TA = \sum_{\substack{(a,b)\in S \\ a=(x,y)}} \begin{array}{cc} \begin{array}{cccccc} 1 & x & y & x^2 & xy & y^2 \end{array} & \\ \begin{bmatrix} 1 & x & y & x^2 & xy & y^2 \\ x & x^2 & xy & x^3 & x^2y & xy^2 \\ y & xy & y^2 & x^2y & xy^2 & y^3 \\ x^2 & x^3 & x^2y & x^4 & x^3y & x^2y^2 \\ xy & x^2y & xy^2 & x^3y & x^2y^2 & xy^3 \\ y^2 & xy^2 & y^3 & x^2y^2 & x^{3} & y^4 \end{bmatrix} & \begin{array}{c} 1 \\ x \\ y \\ x^2 \\ xy \\ y^2 \end{array} \end{array}$$

Network training = 'distance' minimization

Let $\mathcal{M} \subseteq V := \left(\mathbb{R}[x_1,\dots,x_{d_0}]_{\leq D}\right)^{d_L}$,

neuromanifold

$S \subseteq \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ finite dataset,

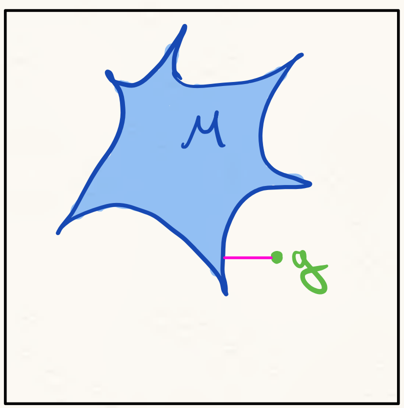
mean squared error

MSE loss: $\mathcal{L}(f) := \sum_{(a,b)\in S} \|f(a)-b\|^2$

[$\mathrm{dist}(f,g)=0$ possible for $f \neq g$]

Proposition: There is a pseudometric $\mathrm{dist}: V\times V \to \mathbb{R}_{\geq 0}$ and some $g \in V$ such that minimizing $\mathcal{L}(f)$ over $f\in\mathcal{M}$ is equivalent to minimizing $\mathrm{dist}(f,g)$ over $f \in \mathcal{M}$.

V



$d_L > 1$

$f = (f_1,\dots,f_{d_L})$, $\quad C_f := \begin{bmatrix} | & & | \\ c_{f_1} & \cdots & c_{f_{d_L}} \\ | & & | \end{bmatrix}$

$\Rightarrow f(x) = \nu_D(x)\cdot C_f$

$\|C\|_Q^2 := \mathrm{tr}(C^T Q C)$

$\Rightarrow \mathcal{L}(f) = \|A C_f - B\|^2_{\mathrm{Frob}} = \|C_f - A^{+}B\|^2_{A^TA} + \text{const.}$

Loss Landscape

$$= \{(\theta, \mathcal{L}(f_\theta)) \mid \theta \in \Theta\}$$

$$\underset{\text{params}}{\Theta} \times \underset{\text{input}}{X} \to \underset{\text{output}}{Y}$$

$$\mu : \Theta \times X \to Y$$

$$\mathcal{M} = \{\underbrace{\mu(\theta, \cdot)}_{= f_\theta} \mid \theta \in \Theta\}$$

Loss Landscape

$$= \{ (\theta, \mathcal{L}(f_\theta)) \mid \theta \in \Theta \}$$

$$\mu : \overset{\text{params}}{\Theta} \times \overset{\text{input}}{X} \longrightarrow \overset{\text{output}}{Y}$$

$$\mathcal{M} = \{ \underbrace{\mu(\theta, \cdot)}_{= f_\theta} \mid \theta \in \Theta \}$$

can be studied in a decoupled way:

$$\begin{tikzcd}
\Theta \arrow[r] & \mathcal{M} \arrow[r, "\mathcal{L}"] & \mathbb{R} \\
\theta \arrow[r, mapsto] & f_\theta &
\end{tikzcd}$$

loss landscape in function space:

$$= \{ (f, \mathcal{L}(f)) \mid f \in \mathcal{M} \} \subseteq V \times \mathbb{R}$$

Loss Landscape

$$= \{(\theta, \mathcal{L}(f_\theta)) \mid \theta \in \Theta\}$$

params, input, output

$$\mu : \Theta \times X \to Y$$

$$\mathcal{M} = \{\underbrace{\mu(\theta, \cdot)}_{= f_\theta} \mid \theta \in \Theta\}$$

can be studied in a decoupled way:

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

loss landscape in function space:

$$= \{(f, \mathcal{L}(f)) \mid f \in \mathcal{M}\} \subseteq V \times \mathbb{R}$$

Geometry of $\mathcal{M}$ affects loss landscape!

How?

Which geometric properties does $\mathcal{M}$ have?

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]

Machine Learning	Algebraic Geometry
sample complexity and expressivity	dimension, degree, and covering number
subnetworks and implicit bias	singularities
identifiability and invariance	fibers of the parameterization
optimization and gradient descent	critical point theory, discriminants, and dynamical invariants

Identifiability

$$\begin{array}{ccc} \Theta & \xrightarrow{\mu} & \mathcal{M} \\ \theta & \longmapsto & f_\theta \end{array}$$

Given (generic) $f \in \mathcal{M}$, what is $\mu^{-1}(f)$?

Identifiability

$$\begin{array}{ccc} \Theta & \xrightarrow{\mu} & \mathcal{M} \\ \theta & \longmapsto & f_\theta \end{array}$$

Given (generic) $f \in \mathcal{M}$, what is $\mu^{-1}(f)$?

For monomial MLP with $\sigma(x) = x^r$, $r >> 0$:

$$\mu : \mathbb{R}^{d_L \times d_{L-1}} \times \dots \times \mathbb{R}^{d_1 \times d_0} \longrightarrow \mathcal{M}$$

$$(W_L, \dots, W_1) \longmapsto W_L \circ \sigma \circ \dots \circ \sigma \circ W_2 \circ \sigma \circ W_1$$

What is the generic fiber?

Identifiability

$$\Theta \xrightarrow{\mu} \mathcal{M}$$
$$\theta \longmapsto f_\theta$$

Given (generic) $f \in \mathcal{M}$, what is $\mu^{-1}(f)$?

For monomial MLP with $\sigma(x) = x^r$, $r >> 0$:

$$\mu : \mathbb{R}^{d_L \times d_{L-1}} \times \dots \times \mathbb{R}^{d_1 \times d_0} \longrightarrow \mathcal{M}$$
$$(W_L, \dots, W_1) \longmapsto W_L \circ \sigma \circ \dots \circ \sigma \circ W_2 \circ \sigma \circ W_1$$

2019

On the Expressive Power of Deep Polynomial Neural Networks

Joe Kileel* Princeton University
Matthew Trager* New York University
Joan Bruna New York University

Observation: $D_i \in GL(d_i)$ diagonal
$P_i \in GL(d_i)$ permutation matrix

$$\Rightarrow \mu(W_L D_{L-1}^{-r} P_{L-1}^T, \dots, P_2 D_2 W_2 D_1^{-r} P_1^T, P_1 D_1 W_1)$$
$$= \mu(W_L, \dots, W_2, W_1)$$

Identifiability

$$\begin{array}{ccc} \Theta & \xrightarrow{\mu} & \mathcal{M} \\ \theta & \longmapsto & f_\theta \end{array}$$

Given (generic) $f \in \mathcal{M}$, what is $\mu^{-1}(f)$?

For **monomial MLP** with $\sigma(x) = x^r$, $r >> 0$:

Activation degree thresholds and expressiveness of polynomial neural networks 2024

Bella Finkel[*], Jose Israel Rodriguez[†], Chenxi Wu, Thomas Yahl

Proven that those parameter symmetries are the generic fiber (implicitly described all fibers!)

follow-ups (2025):

THE ALEXANDER-HIRSCHOWITZ THEOREM FOR NEUROVARIETIES

ALEX MASSARENTI AND MASSIMILIANO MELLA

Identifiability of Deep Polynomial Neural Networks

Konstantin Usevich[*], Ricardo Borsoi, Clara Dérand, Marianne Clausel[†]
Université de Lorraine, CNRS, CRAN

Identifiability

$$\begin{array}{ccc} \Theta & \xrightarrow{\mu} & \mathcal{M} \\ \theta & \longmapsto & f_\theta \end{array}$$

Given (generic) $f \in \mathcal{M}$, what is $\mu^{-1}(f)$?

For monomial MLP with $\sigma(x) = x^r$, $r >> 0$:

Activation degree thresholds and expressiveness of polynomial neural networks 2024

Bella Finkel,[*] Jose Israel Rodriguez,[†] Chenxi Wu, Thomas Yahl

Proposition 16. Let $\mathbb{K}$ be a subfield of $\mathbb{C}$. Given integers d, k, there exists an integer $\tilde{r} = \tilde{r}(k)$ with the following property. If $r > \tilde{r}(k)$ and $p_1, \ldots, p_k \in \mathbb{K}[x_1, \ldots, x_d]$ are pairwise non-proportional, then $p_1^r, \ldots, p_k^r$ are linearly independent (over $\mathbb{K}$). Moreover, $\tilde{r}(k) = 6(k-1)^2 - 6(k-1) + 1$ has the desired property.

Identifiability

$$\begin{array}{ccc} \Theta & \xrightarrow{\quad\mu\quad} & \mathcal{M} \\ \theta & \longmapsto & f_\theta \end{array}$$

Given (generic) $f \in \mathcal{M}$, what is $\mu^{-1}(f)$?

For polynomial MLP with $\sigma(x) \in \mathbb{R}[x]_{\leq r}$ generic, $r >> 0$:

Conjecture: Generic fiber $\mu^{-1}(f)$ consists only of permutations.

Conjecture: Let $d, k \in \mathbb{Z}_{>0}$.

There is $\tilde{r} \in \mathbb{Z}_{>0}$ such that all $r > \tilde{r}$ satisfy:

There is $\mathcal{U} \subseteq \mathbb{R}[x]_{\leq r}$ Zariski open such that:

For all $\sigma \in \mathcal{U}$ and all $p_1, \ldots, p_k \in \mathbb{R}[x_1, \ldots, x_d]$ non-constant & pairwise distinct:

$\sigma(p_1), \ldots, \sigma(p_k)$ are linearly independent.

Geometry of Neuromanifolds

$\mu : \Theta \times X \longrightarrow Y$ polynomial (in both $\theta \in \Theta$ & $x \in X$)

$$\begin{aligned} \Theta &\longrightarrow \mathcal{M} \\ \theta &\longmapsto \mu(\theta, \cdot) \end{aligned}$$

What kind of object is $\mathcal{M}$?

A semialgebraic set!

describable by polynomial equations & inequalities

Geometry of Neuromanifolds

$\mu : \Theta \times X \longrightarrow Y$ polynomial (in both $\theta \in \Theta$ & $x \in X$)

$$\begin{array}{ccc} \Theta & \longrightarrow & \mathcal{M} \\ \theta & \longmapsto & \mu(\theta, \cdot) \end{array}$$

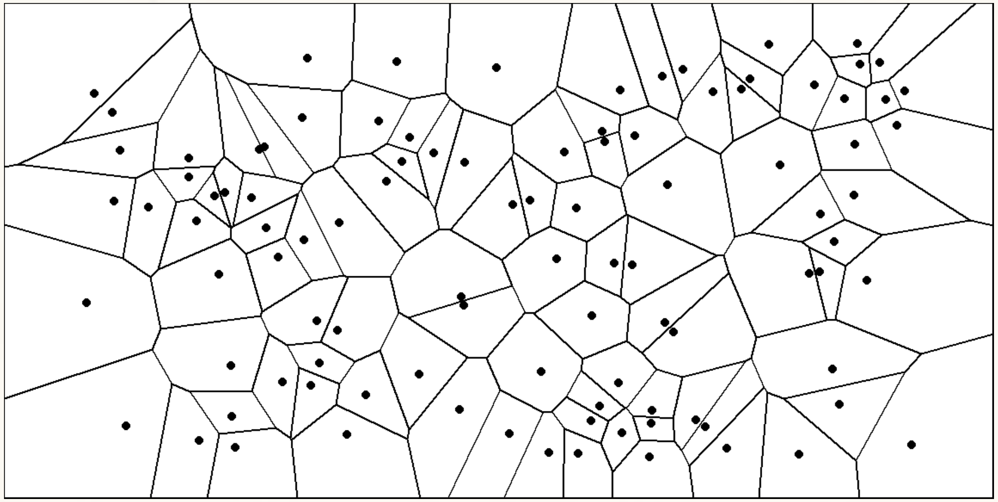
What kind of object is $\mathcal{M}$?

A semialgebraic set!

describable by polynomial equations & inequalities

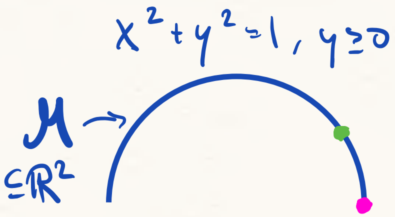
Euclidean distance minimization can be implicitly biased to singularities & boundaries of $\mathcal{M}$

Voronoi cells



For $S \subseteq \mathbb{R}^n$, the **Voronoi cell** at $p \in S$ is

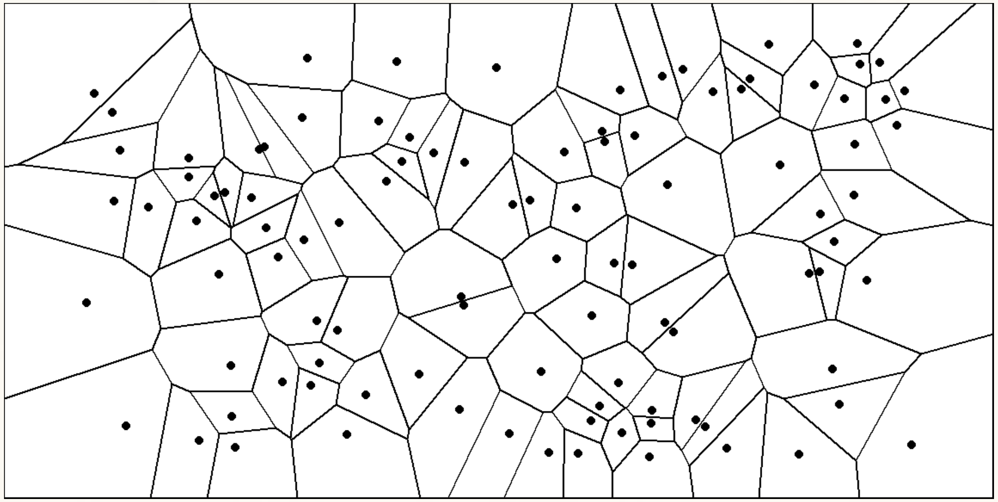
$$\mathrm{Vor}_S(p) := \{ u \in \mathbb{R}^n \mid \forall q \in S,\ q \neq p: \|p-u\|_2 < \|q-u\|_2 \}$$



What is the Voronoi cell at • ?

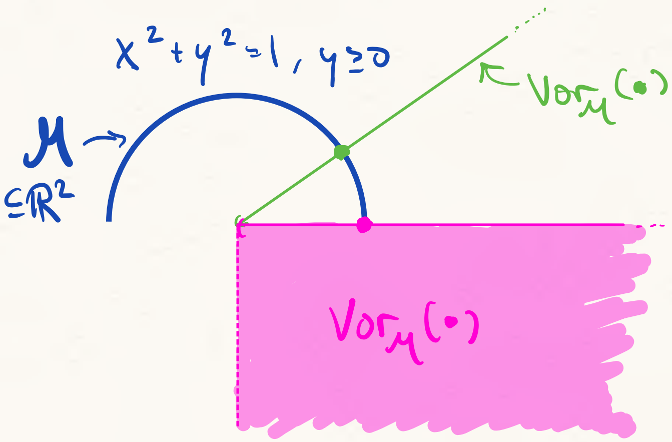
What is the Voronoi cell at • ?

Voronoi cells



For $S \subseteq \mathbb{R}^n$, the **Voronoi cell** at $p \in S$ is

$$\mathrm{Vor}_S(p) := \{u \in \mathbb{R}^n \mid \forall q \in S,\ q \neq p: \\ \|p-u\|_2 < \|q-u\|_2\}$$

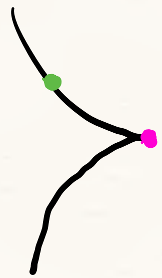


The 2 relative boundary points are the only points on $\mathcal{M}$ with full-dimensional Voronoi cells!

$\rightsquigarrow$ **implicit bias** towards $\partial\mathcal{M}$

points in $\partial\mathcal{M}$ are global minima with positive probability on data u

singularities

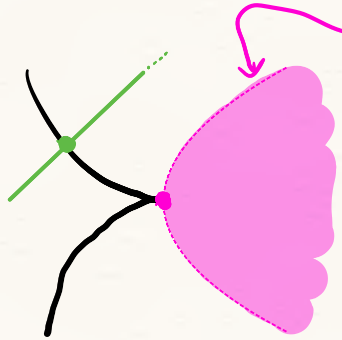


What are the Voronoi cells at • and • ?

singularities

$y^2 + x^3 = 0$

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



Challenge: Compute this curve!

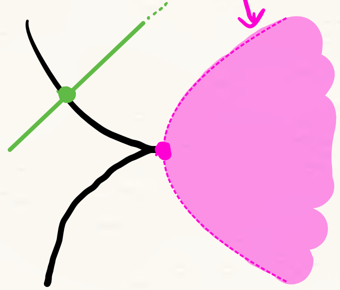
$\rightsquigarrow$ implicit bias towards $\mathrm{Sing}(M)$

What are the Voronoi cells at • and • ?

singularities

$y^2 + x^3 = 0$

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



Challenge: Compute this curve!

$\rightsquigarrow$ implicit bias towards Sing(M)

What are the Voronoi cells at • and • ?

Tradeoff

☹

learning close to singularity
$\rightsquigarrow$ slow & numerical instability
[Amari et al]

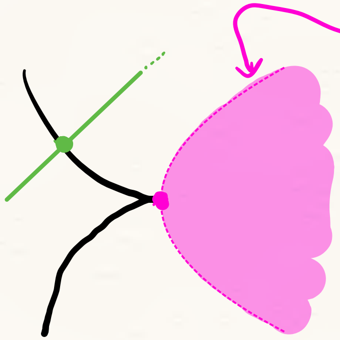
☺

singular solution generalizes better:
① stable global minimum when perturbing data
② **Conjecture**: singularities of neuromanifolds are sparse subnetworks
[We've proven this for MLPs & CNNs]

singularities

$y^2 + x^3 = 0$

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



Challenge: Compute this curve!

$\rightsquigarrow$ implicit bias towards Sing(M)

What are the Voronoi cells at • and • ?

Tradeoff

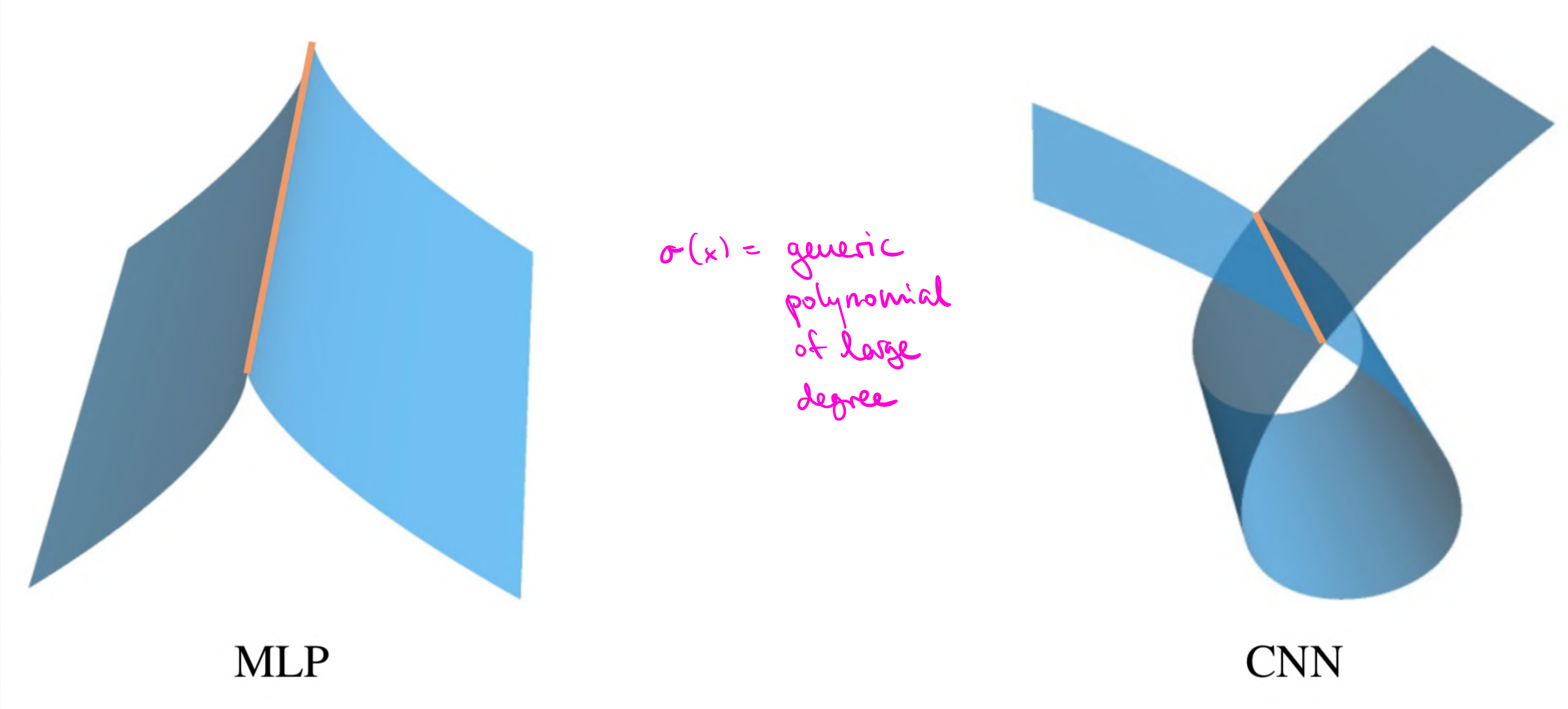
☹

learning close to singularity
$\rightsquigarrow$ slow & numerical instability
[Amari et al]

☺

singular solution generalizes better:
① stable global minimum when perturbing data
② **Conjecture:** singularities of neuromanifolds are sparse subnetworks
[We've proven this for MLPs & CNNs]

In general: depends on **type** of singularity



These singularities have that tradeoff, ... while these don't!

In both cases, they are sparse subnetworks ◡̈

What about smooth interior points?

$\mathcal{M} \subseteq \mathbb{R}^n$ algebraic variety (i.e. described by polynomial equations)

Q symmetric PD $n \times n$ matrix

Fact: For almost all $u \in \mathbb{R}^n$, the number of <u>complex</u> critical points of

$$\min_{x \in \mathcal{M} \setminus \mathrm{Sing}(\mathcal{M})} \|x-u\|_Q^2$$

is the <u>same</u>, called the **Euclidean Distance Degree**: $\mathrm{EDD}_Q(\mathcal{M})$.

What is $\mathrm{EDD}_{\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]}(\bigcirc)$?

What about smooth interior points?

$\mathcal{M} \subseteq \mathbb{R}^n$ algebraic variety (i.e. described by polynomial equations)

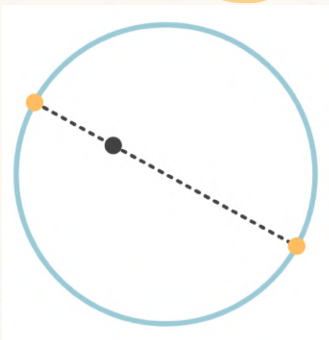
Q symmetric PD $n \times n$ matrix

Fact: For almost all $u \in \mathbb{R}^n$, the number of complex critical points of

$$\min_{x \in \mathcal{M} \setminus \mathrm{Sing}(\mathcal{M})} \|x-u\|_Q^2$$

is the same, called the **Euclidean Distance Degree**: $\mathrm{EDD}_Q(\mathcal{M})$.

What is $\mathrm{EDD}_{\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]}(\bigcirc)$?



What is $\mathrm{EDD}_{\left[\begin{smallmatrix}4&0\\0&1\end{smallmatrix}\right]}(\bigcirc)$?

What about smooth interior points?

$\mathcal{M} \subseteq \mathbb{R}^n$ algebraic variety (i.e. described by polynomial equations)

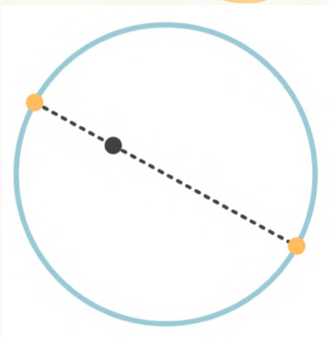
Q symmetric PD $n \times n$ matrix

Fact: For almost all $u \in \mathbb{R}^n$, the number of complex critical points of

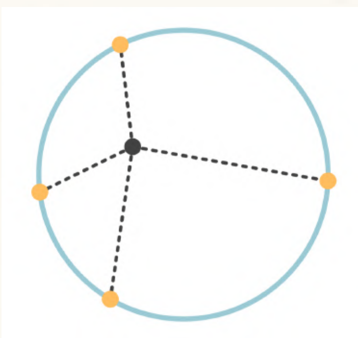
$$\min_{x \in \mathcal{M} \setminus \mathrm{Sing}(\mathcal{M})} \|x-u\|_Q^2$$

is the same, called the **Euclidean Distance Degree**: $\mathrm{EDD}_Q(\mathcal{M})$.

What is $\mathrm{EDD}_{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}(\bigcirc)$?



What is $\mathrm{EDD}_{\begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}}(\bigcirc)$?



What about smooth interior points?

$\mathcal{M} \subseteq \mathbb{R}^n$ algebraic variety (i.e. described by polynomial equations)

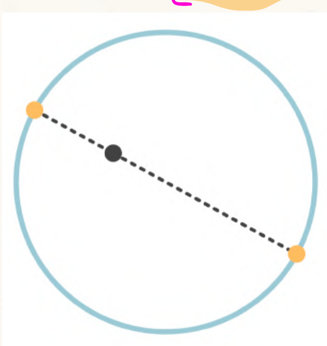
Q symmetric PD $n \times n$ matrix

Fact: For almost all $u \in \mathbb{R}^n$, the number of complex critical points of

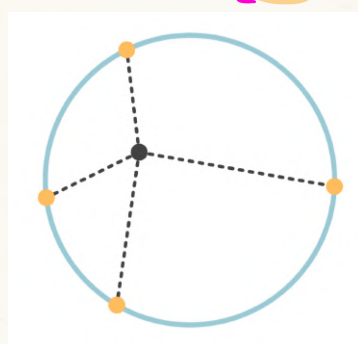
$$\min_{x \in \mathcal{M} \setminus \mathrm{Sing}(\mathcal{M})} \|x-u\|_Q^2$$

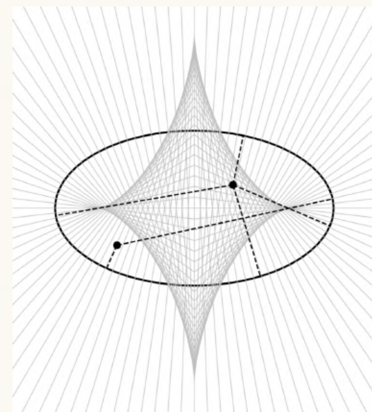
is the same, called the **Euclidean Distance Degree**: $EDD_Q(\mathcal{M})$.

What is $EDD_{\begin{bmatrix}1&0\\0&1\end{bmatrix}}(\bigcirc)$?



What is $EDD_{\begin{bmatrix}4&0\\0&1\end{bmatrix}}(\bigcirc)$?



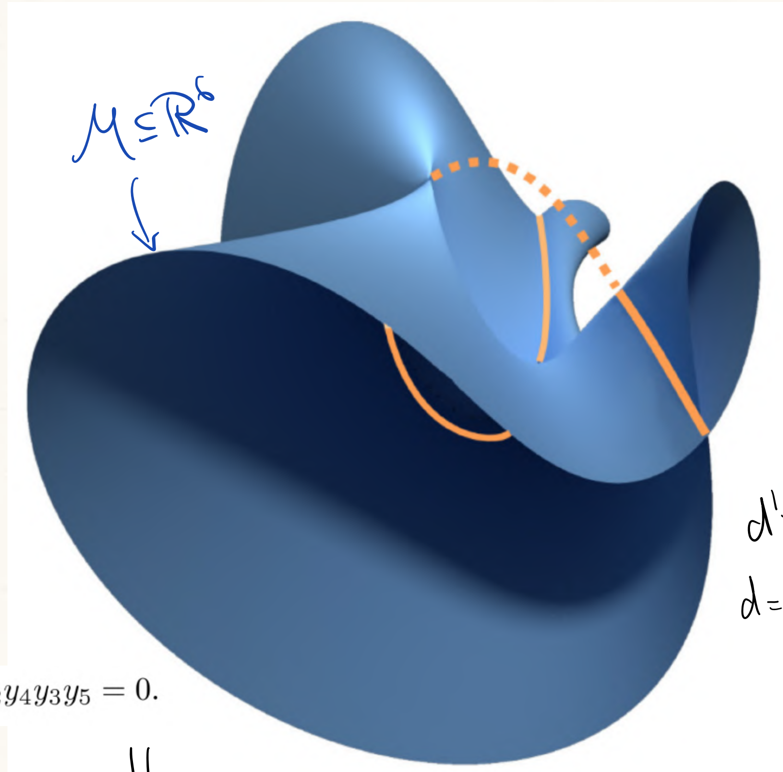


Lightning Self-Attention (single head, single layer)

$$\mathbb{R}^{d\times t} \longrightarrow \mathbb{R}^{d'\times t}$$
$$X \longmapsto VXX^TK^TQX$$

learnable parameters

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



$d'=1$

$d=t=a=2$

$$y_1^2y_6^2+y_4^2y_3^2+y_1y_3y_5^2+y_2^2y_4y_6-2y_1y_4y_3y_6-y_2y_1y_6y_5-y_2y_4y_3y_5=0.$$

$\Downarrow$

For almost all PD matrices Q,

$\mathrm{EDD}_Q(\mathcal{M})=14.$

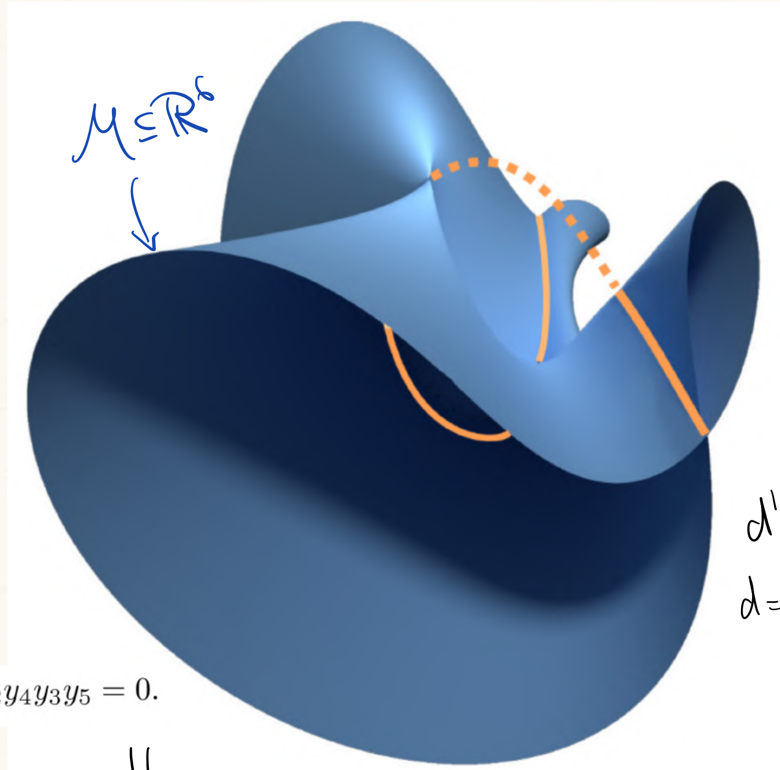
What happens if Q becomes degenerate?

(ie., Q is symmetric positive <u>semi</u>definite)

Lightning Self-Attention (single head, single layer)

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

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



$d' = 1$
$d = t = a = 2$

$$y_1^2 y_6^2 + y_4^2 y_3^2 + y_1 y_3 y_5^2 + y_2^2 y_4 y_6 - 2 y_1 y_4 y_3 y_6 - y_2 y_1 y_6 y_5 - y_2 y_4 y_3 y_5 = 0.$$

$\Downarrow$

For almost all PD matrices Q,
$\mathrm{EDD}_Q(\mathcal{M}) = 14$.

What happens if Q becomes degenerate?
(i.e., Q is symmetric positive <u>semi</u>definite)

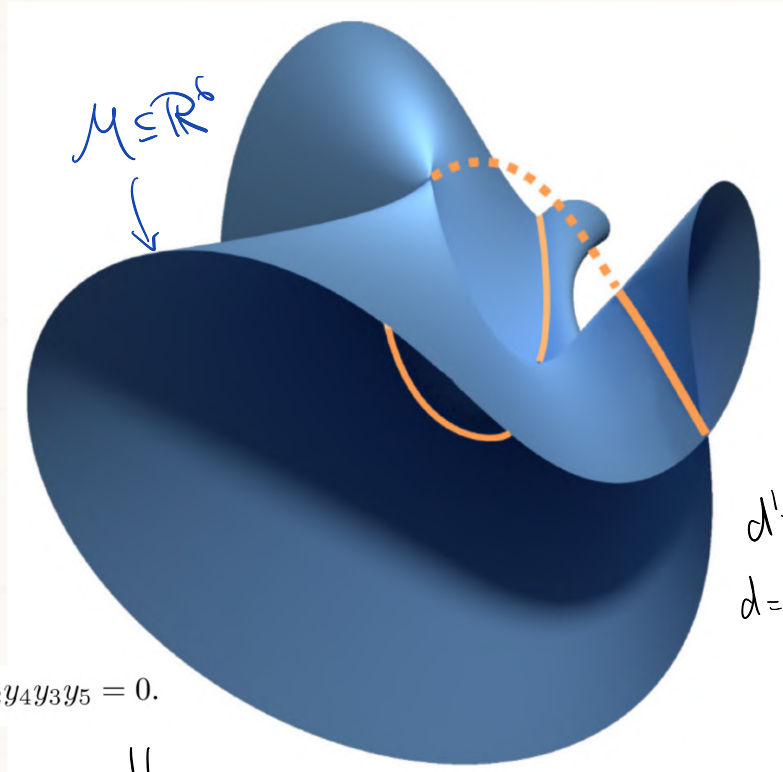
k	complex critical point set
0	14 points
1	14 points
2	4 points + a curve
3	a surface
4	a 3-dimensional subvariety
5	a 4-dimensional subvariety

$k := \dim \ker Q$

Lightning Self-Attention (single head, single layer)

$$\mathbb{R}^{d \times t} \longrightarrow \mathbb{R}^{d' \times t}$$
$$X \longmapsto VXX^TK^TQX$$

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



$d' = 1$
$d = t = a = 2$

$$y_1^2 y_6^2 + y_4^2 y_3^2 + y_1 y_3 y_5^2 + y_2^2 y_4 y_6 - 2 y_1 y_4 y_3 y_6 - y_2 y_1 y_6 y_5 - y_2 y_4 y_3 y_5 = 0.$$

$\Downarrow$

For almost all PD matrices Q,
$\mathrm{EDD}_Q(\mathcal{M}) = 14$.

What happens if Q becomes degenerate?
(i.e., Q is symmetric positive semidefinite)

k	complex critical point set
0	14 points
1	14 points
2	4 points + a curve
3	a surface
4	a 3-dimensional subvariety
5	a 4-dimensional subvariety

$k := \dim \ker Q$

$\mathcal{M} \cap (\ker(Q) + u)$
↳ zero loss solutions!

In general

$\mathcal{M} \subseteq \mathbb{R}^n$ algebraic variety, $d := \dim \mathcal{M}$.

Q symmetric positive semidefinite $n \times n$ matrix $K := \ker Q$	$\pi : \mathbb{R}^n \to K^\perp$ turns Q into nondegenerate quadric

In general

$M \subseteq \mathbb{R}^n$ algebraic variety, $d := \dim M$.

Q symmetric positive semidefinite $n \times n$ matrix
$K := \ker Q$

$\pi : \mathbb{R}^n \to K^\perp$
turns Q into nondegenerate quadric

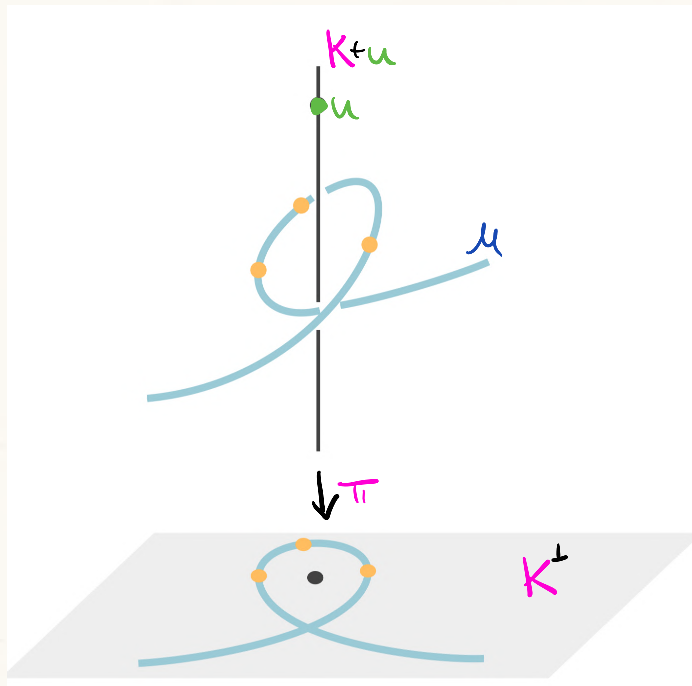
Case 1: Let $k < n - d$.

For almost all Q with $k = \dim K$ and almost all $u \in \mathbb{R}^n$,

$$\mathrm{EDD}_Q(M) \left\{ \text{critical points of } \min_{x \in M \setminus \mathrm{Sing}(M)} \|x - u\|_Q^2 \right.$$

$$\| \qquad \qquad \updownarrow 1:1$$

$$\mathrm{EDD}_{\pi(Q)}(\pi(M)) \left\{ \text{critical points of } \min_{x \in \pi(M) \setminus \mathrm{Sing}(\pi(M))} \|x - \pi(u)\|_{\pi(Q)}^2 \right.$$



In general

$\mathcal{M} \subseteq \mathbb{R}^n$ algebraic variety, $d := \dim \mathcal{M}$.

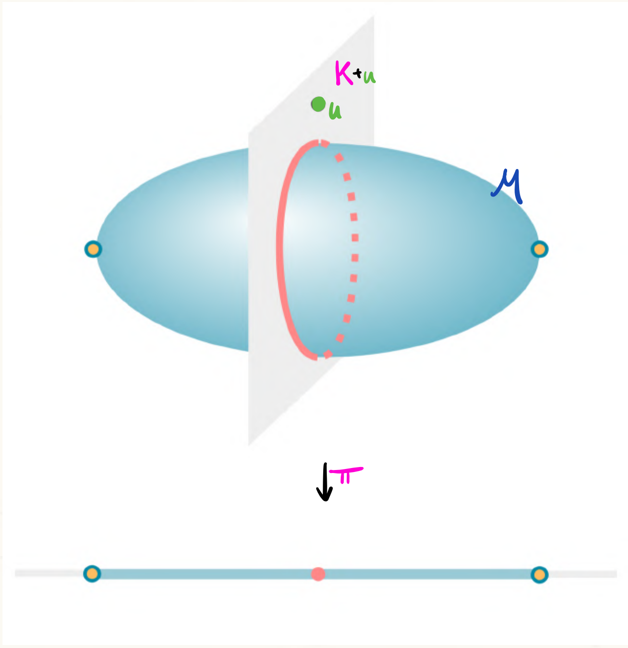
Q symmetric positive semidefinite $n \times n$ matrix
$K := \ker Q$

$\pi : \mathbb{R}^n \to K^{\perp}$
turns Q into nondegenerate quadric

Case 2: Let $k \geq n-d$.

For almost all Q with $k = \dim K$ and almost all $u \in \mathbb{R}^n$, we have 2 types of critical points of $\min_{x \in \mathcal{M} \setminus \mathrm{Sing}(\mathcal{M})} \|x-u\|_Q^2$:

(A) $(K+u) \cap \mathcal{M}$: zero loss solutions



In general

$\mathcal{M} \subseteq \mathbb{R}^n$ algebraic variety, $d := \dim \mathcal{M}$.

Q symmetric positive semi-definite $n \times n$ matrix
$K := \ker Q$

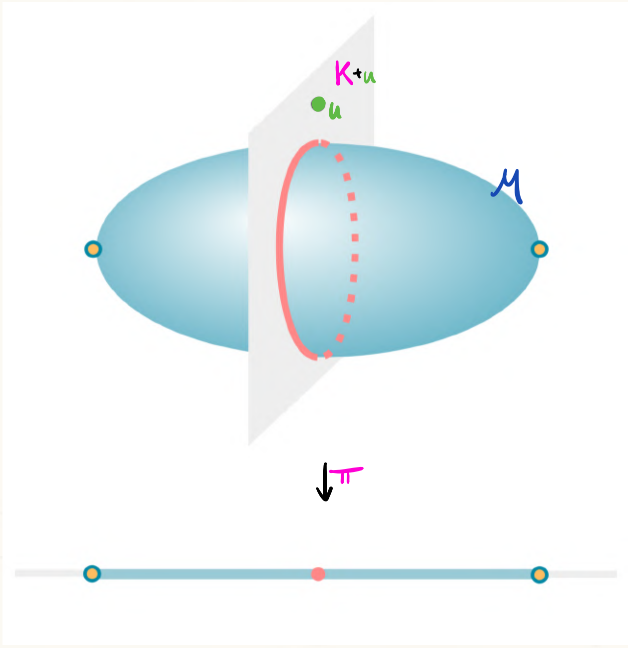
$\pi : \mathbb{R}^n \to K^\perp$
turns Q into nondegenerate quadric

Case 2: Let $k \geq n-d$.

For almost all Q with $k = \dim K$ and almost all $u \in \mathbb{R}^n$, we have 2 types of critical points of $\min_{x \in \mathcal{M} \setminus \mathrm{Sing}(\mathcal{M})} \|x-u\|_Q^2$:

Ⓐ $(K+u) \cap \mathcal{M}$: zero loss solutions

Ⓑ finitely many on the ramification locus $\mathrm{Ram}(\pi|_X)$
$:= \{\text{critical points of } \pi|_X\}$



In general

$\mathcal{M} \subseteq \mathbb{R}^n$ algebraic variety, $d := \dim \mathcal{M}$.

Q symmetric positive semi-definite $n \times n$ matrix
$K := \ker Q$

$\pi : \mathbb{R}^n \to K^{\perp}$
turns Q into nondegenerate quadric

Case 2: Let $k \geq n-d$.

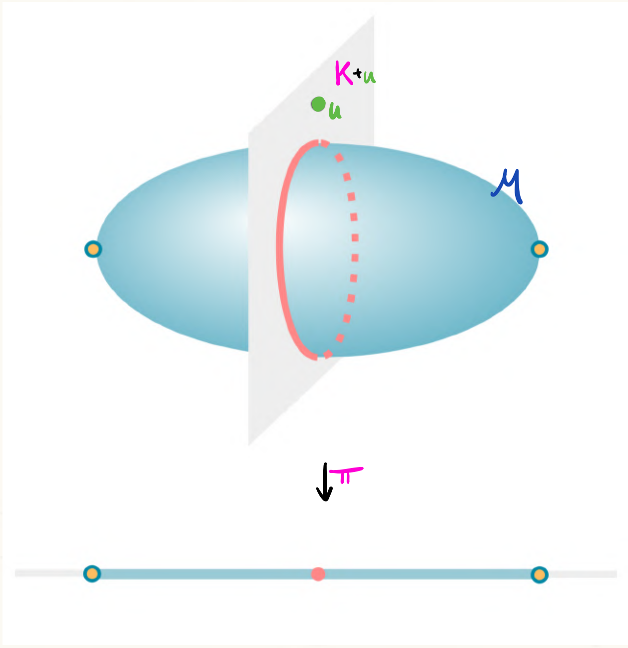
For almost all Q with $k = \dim K$ and almost all $u \in \mathbb{R}^n$, we have 2 types of critical points of $\min_{x \in \mathcal{M} \setminus \mathrm{Sing}(\mathcal{M})} \|x-u\|_Q^2$:

(A) $(K+u) \cap \mathcal{M}$: zero loss solutions

(B) finitely many on the ramification locus $\mathrm{Ram}(\pi|_X)$
$:= \{\text{critical points of } \pi|_X\}$

$\updownarrow$ 1:1

$\mathrm{EDD}_{\pi(Q)}(\mathrm{Br}) \leftarrow$ critical points of $\min_{x \in \mathrm{Br}(\pi|_X)} \|x-\pi(u)\|_{\pi(Q)}^2$ $\leftarrow$ Branch locus $\pi(\mathrm{Ram})$



In general

$\mathcal{M} \subseteq \mathbb{R}^n$ algebraic variety, $d := \dim \mathcal{M}$.

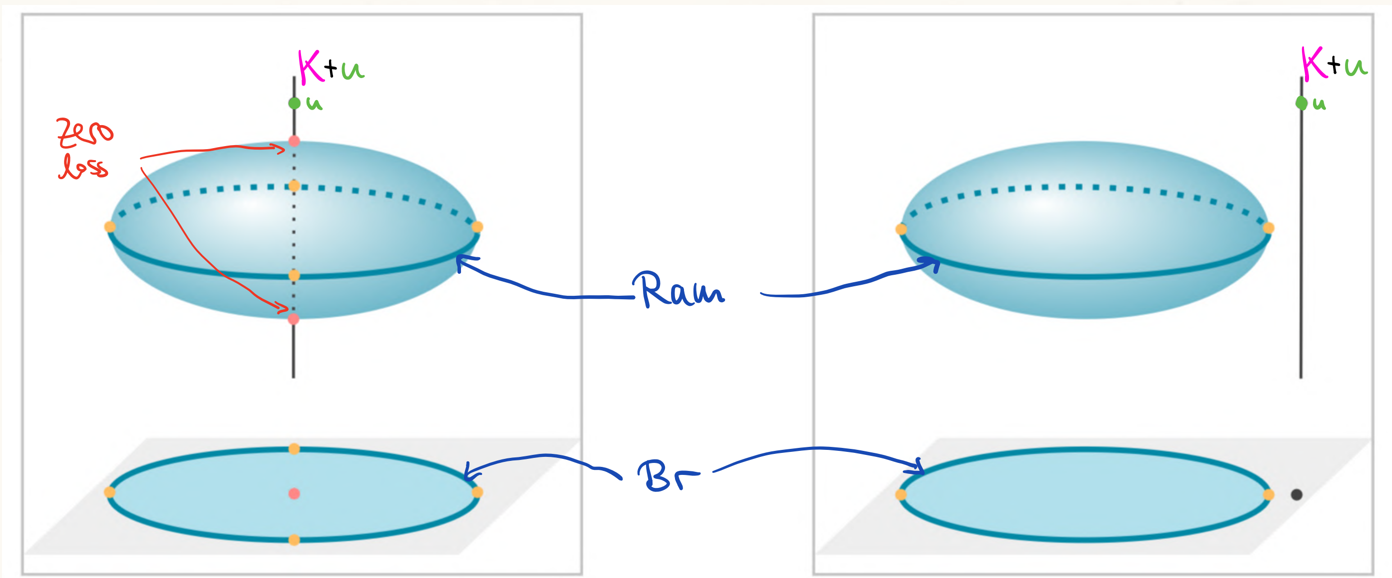
Q symmetric positive semidefinite $n \times n$ matrix

$K := \ker Q$

$\pi : \mathbb{R}^n \to K^\perp$

turns Q into nondegenerate quadric

Case 2: Let $k \geq n-d$.



Induced bias towards Ram!

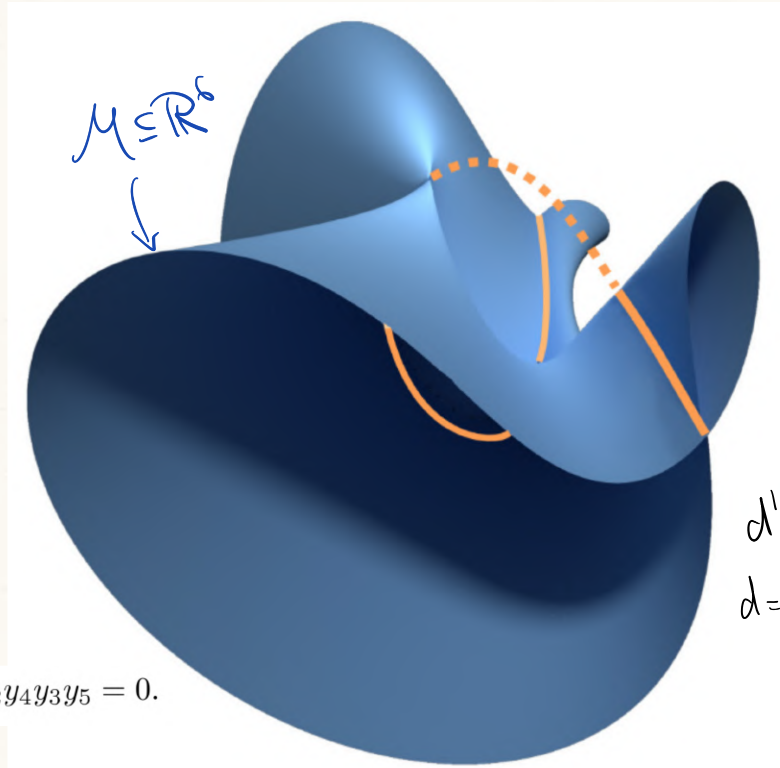
depends <u>only</u> on K (not on Q) & not on u

Lightning Self-Attention (single head, single layer)

$$\mathbb{R}^{d\times t} \longrightarrow \mathbb{R}^{d'\times t}$$
$$X \longmapsto VXX^TK^TQX$$

learnable parameters

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



$d' = 1$

$d = t = a = 2$

$$y_1^2y_6^2 + y_4^2y_3^2 + y_1y_3y_5^2 + y_2^2y_4y_6 - 2y_1y_4y_3y_6 - y_2y_1y_6y_5 - y_2y_4y_3y_5 = 0.$$

Q = from MSE loss with general dataset S

$\|S\|$	$k = \dim K$	complex critical point set
≥ 3	0	14 points
2	2	a curve and two lines
1	4	a 3-dimensional subvariety

Q = general symmetric positive semidefinite

k	complex critical point set
0	14 points
1	14 points
2	4 points + a curve
3	a surface
4	a 3-dimensional subvariety
5	a 4-dimensional subvariety

$k := \dim \ker Q$

$M \cap (\ker(Q) + u)$

↳ zero loss solutions!

algebraic neural network theory – an emerging field

Kileel, Trager, Bruna: On the expressive power of deep polynomial neural networks. **NeurIPS** 2019.

Trager, Kohn, Bruna: Pure and spurious critical points: a geometric study of linear networks. **ICLR** 2020.

Kohn, Merkh, Montúfar, Trager: Geometry of linear convolutional networks. **SIAM Journal on Applied Algebra & Geometry** 2022.

Kohn, Montúfar, Shahverdi, Trager: Function space & critical points of linear CNNs. **SIAM Journal on Applied Algebra & Geometry** 2024.

Kubjas, Li, Wiesmann: Geometry of polynomial neural networks. **Algebraic Statistics** 2024.

Marchetti, Shahverdi, Mereta, Trager, Kohn: Position: Algebra unveils deep learning – An invitation to Neuroalgebraic Geometry. **ICML** 2025.

Mody, Zubkov: Geometry of Rank Constraints in Shallow Polynomial Neural Networks. **ICML** 2025 Workshop.

Zhang, Kileel: Covering number of real algebraic varieties and beyond: Improved bounds and applications. **FoCM** 2025.

Finkel, Rodriguez, Wu, Yahl: Activation thresholds and expressiveness of polynomial neural networks. **Algebraic Statistics** 2025.

Arjevani, Bruna, Kileel, Polak, Trager: Geometry and optimization of shallow polynomial networks. arXiv:2501.06074.

Shahverdi, Marchetti, Kohn: On the geometry and optimization of polynomial convolutional networks. **AISTATS** 2025.

Henry, Marchetti, Kohn: Geometry of lightning self-attention: Identifiability and dimension. **ICLR** 2025.

Massarenti, Mella: The Alexander-Hirschowitz theorem for neurovarieties. arXiv:2511.19703.

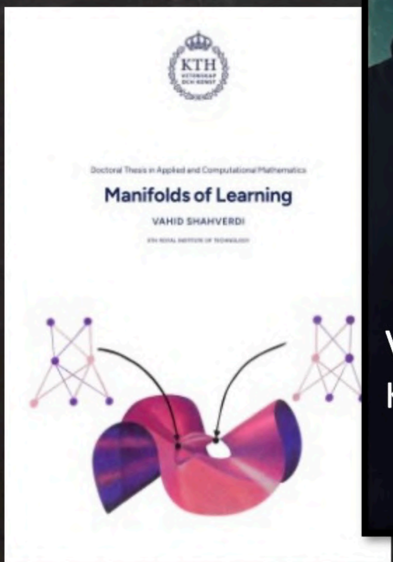
Grosdos, Robeva, Zubkov: Algebraic geometry of rational neural networks. arXiv:2509.11088.

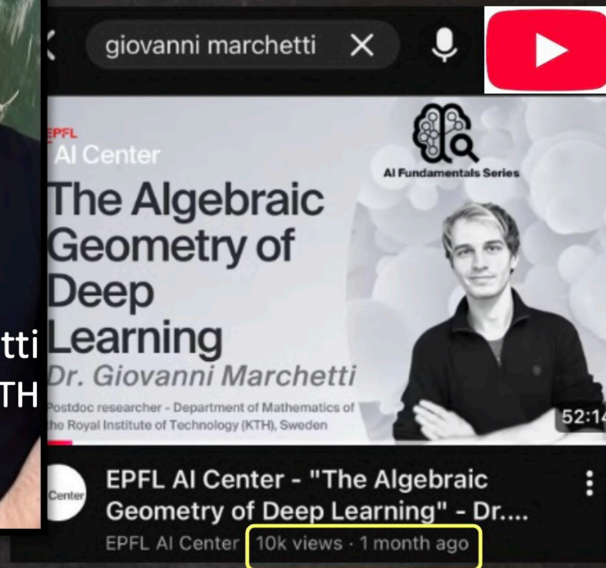
Shahverdi, Marchetti, Kohn: Learning on a razor's edge: the singularity bias of polynomial neural networks. ~~arXiv:2505.11846~~. ICLR 2026

Shahverdi: Algebraic complexity and neurovariety of linear convolutional networks. **Acta Univ. Sapientiae Math.** 2025.

Usevich, Dérand, Borsoi, Clausel: Identifiability of Deep Polynomial Neural Networks. **NeurIPS Oral** 2025.

special thanks to





Stefano Mereta
CUNEF Madrid

Matthew Trager
AWS AI Labs, NY

Joan Bruna
Courant Institute, NYU

Nathan Henry
UC Berkeley

Guido Montúfar
UCLA & MPI MiS Leipzig

Paul Breiding & Erin Connelly
Univ. Osnabrück