Algebraic neural network theory

Kathlén Kohn
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

SFSG SWEDISH FOUNDATIONS' STARTING GRANT

Ragnar Söderbergs STIFTELSE

Göran Gustafssons Stiftelser

neural tangent kernel

NTK approach

$$\begin{tikzcd} \boxed{\text{target network}} \arrow[d, "\text{increase width}"] \\ \boxed{\infty\text{-width network}} \end{tikzcd}$$

linearized models of ∞ dimension

neural
tangent
kernel

NTK approach

$$\begin{array}{c} \boxed{\text{target network}} \\ \Big\downarrow {\scriptstyle \text{increase width}} \\ \boxed{\infty\text{-width network}} \end{array}$$

linearized models

of ∞ dimension

algebraic
geometry

AG approach

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

nonlinear models in

finite-dimensional ambient spaces

Stone-Weierstraß

continuous functions

Let X compact Hausdorff space & A subalgebra of $C(X,\mathbb{R})$ containing a nonzero constant function.

A is dense in $C(X,\mathbb{R})$ in supremum norm $\iff$ A separates points (i.e., $\forall x \neq y \in X \ \exists f \in A : f(x) \neq f(y)$)

Cor.: $X \subseteq \mathbb{R}^n$ compact, $f : X \to \mathbb{R}^m$ continuous, $\varepsilon > 0$.

$\Rightarrow \exists p : X \to \mathbb{R}^n$ polynomial function such that $\forall x \in X : \|f(x) - p(x)\| < \varepsilon$.

Example: MLPs

multilayer perceptrons

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

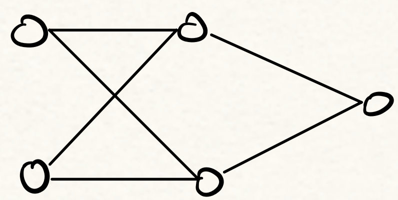
α_i = <u>learnable</u> affine linear functions

σ = nonlinear activation function, applied entrywise

we assume: σ is a univariate polynomial

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

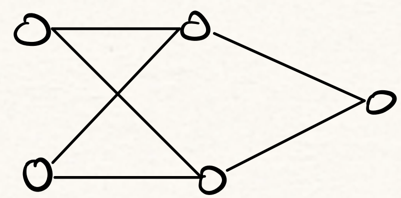
$$\begin{bmatrix} e & f \end{bmatrix} \sigma\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}\right)$$



Which functions does this MLP parametrize?

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

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



Which functions does this MLP parametrize?

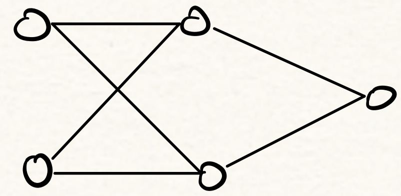
$$e(ax+by)^2 + f(cx+dy)^2$$
$$= \underbrace{(a^2e + c^2f)}_{A} x^2 + \underbrace{2(abe + cdf)}_{B} xy + \underbrace{(b^2e + d^2f)}_{C} y^2$$

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?

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

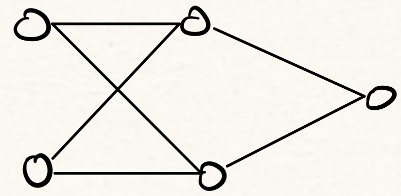
↖ homogeneous quadratic polynomials in x,y

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?

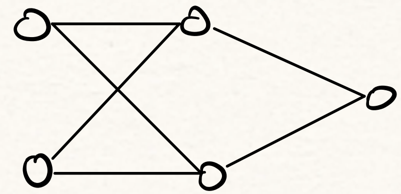
$$e(ax+by)^3 + f(cx+dy)^3$$
$$= \underbrace{(a^3e + c^3f)}_{A}x^3 + \underbrace{3(a^2be + c^2df)}_{B}x^2y + \underbrace{3(ab^2e + cd^2f)}_{C}xy^2 + \underbrace{(b^3e + d^3f)}_{D}y^3$$

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



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

Which functions does this MLP parametrize?

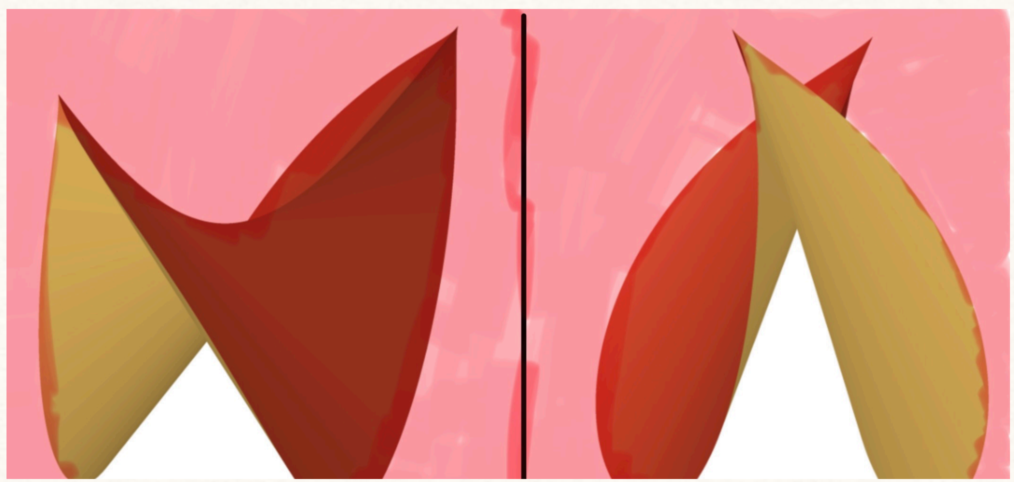
$$e(ax+by)^3 + f(cx+dy)^3$$
$$= \underbrace{(a^3e + c^3f)}_{A}x^3 + \underbrace{3(a^2be + c^2df)}_{B}x^2y + \underbrace{3(ab^2e + cd^2f)}_{C}xy^2 + \underbrace{(b^3e + d^3f)}_{D}y^3$$

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.
$A = 1$
$B = 0$
$C = -1$
$D = 0$



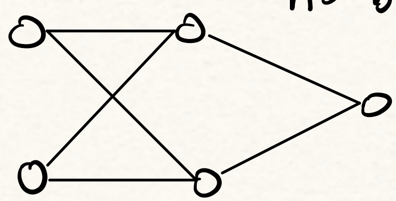
Neuromanifolds

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

parameters inputs outputs

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

Examples:



no bias

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

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

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

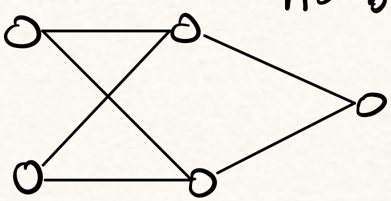
Neuromanifolds

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

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

Examples:

no bias



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

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

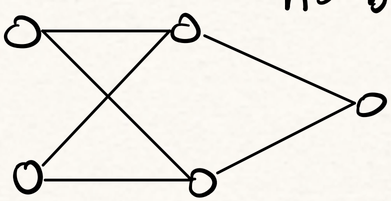
$\sigma(x) = x \quad \Rightarrow \quad \mathcal{M} = \mathbb{R}^{1 \times 2}$

Neuromanifolds

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

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

Examples:

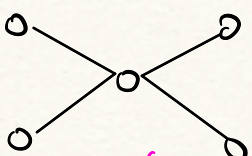


no bias

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

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

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



$\sigma(x) = x$

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

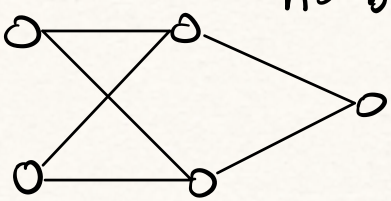
Neuromanifolds

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

parameters inputs outputs

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

Examples:

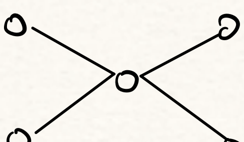


no bias

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

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

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



$\sigma(x) = x$

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

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

$\Rightarrow \mathcal{M} = ?$

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

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

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

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

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

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

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

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

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

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

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!

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

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

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

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

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

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

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

nonlinear models in finite-dimensional ambient spaces

Network training = 'distance' minimization

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

neuromanifold

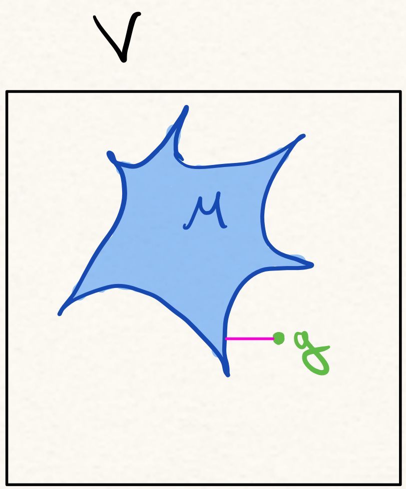
$S \subseteq \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ finite dataset,

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 $f \in \mathcal{M}$.



Why?

Network training = 'distance' minimization

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

neuromanifold

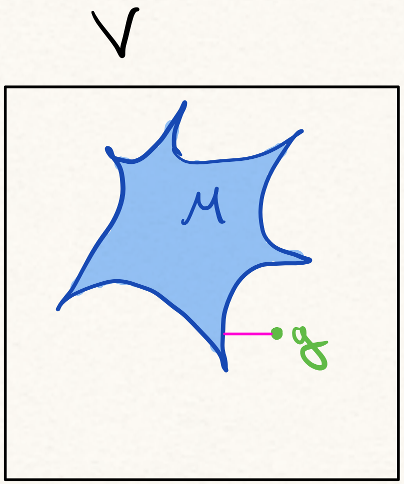
$S \subseteq \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ finite dataset,

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 $f \in \mathcal{M}$.



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 $f(x) = \nu_D(x)\cdot c_f$,

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,

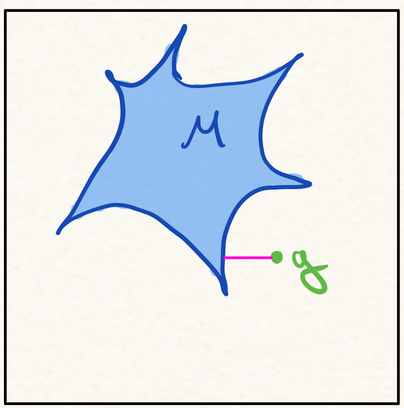
mean squared error

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



Assume: $d_L = 1$

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

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,

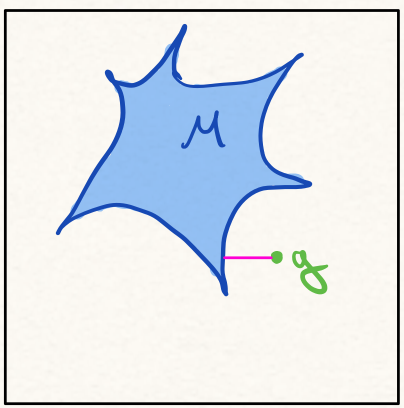
mean squared error

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

[$\mathrm{dist}(f,g) = 0$ 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



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 $f(x) = \nu_D(x)\cdot c_f$,

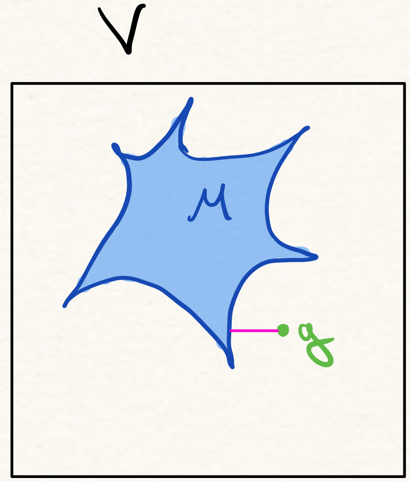
A & B matrices whose rows are $\nu_D(a)$ & b, resp., over all $(a,b)\in S$

pseudoinverse

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

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

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



Observations $(d_L = 1)$:

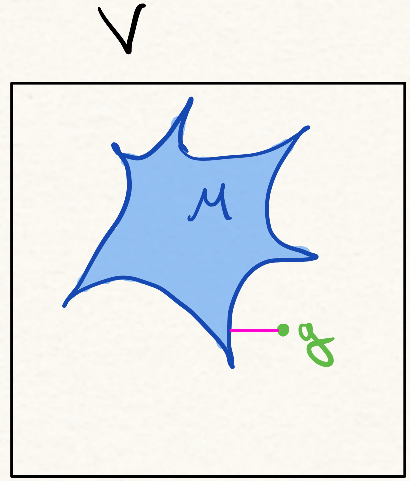
① A^TA depends only on input data,
A^+B on both input & output

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

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

③

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



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

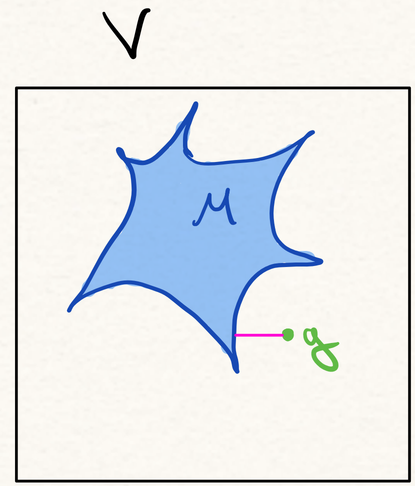
③ even when $|S| \gg \dim V$, A^TA is <u>not</u> an arbitrary symmetric PD matrix,
while A^+B yields all vectors $\in \mathbb{R}^{\dim V}$

Why?

Which matrices can be obtained?

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

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



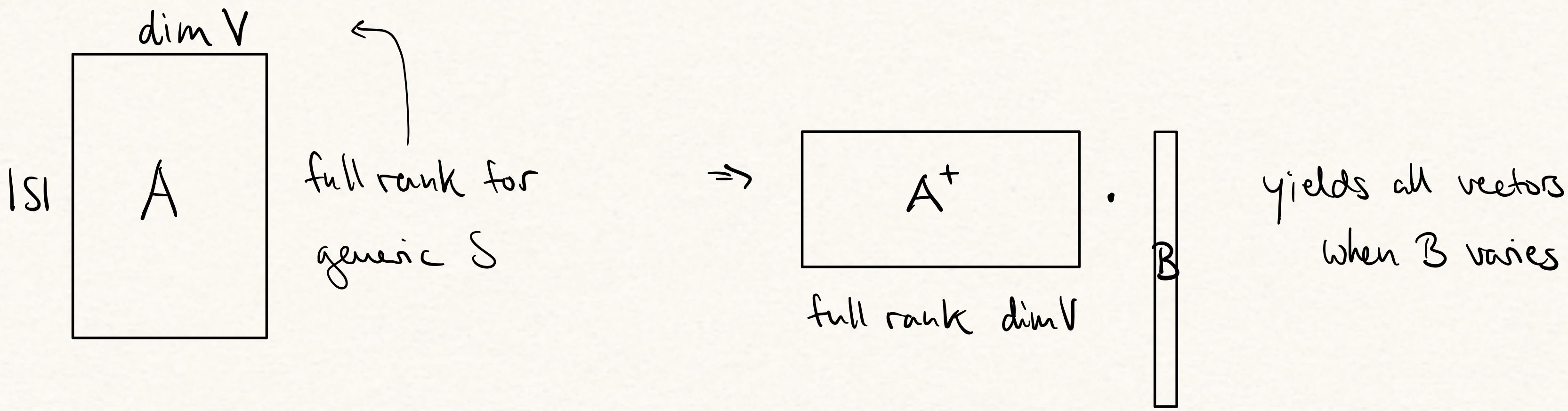
Observations ($d_L = 1$):

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

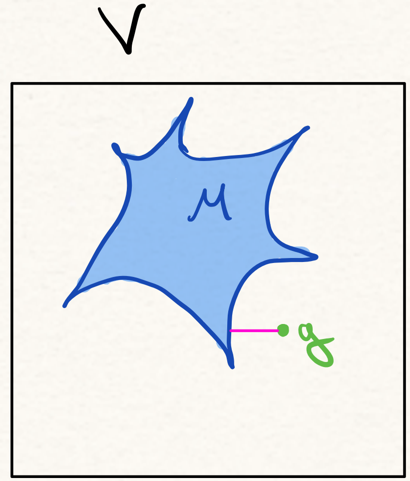
(LLMs: $|S| < \dim \mathcal{M}$)

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 $\in \mathbb{R}^{\dim V}$



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



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 $\in \mathbb{R}^{\dim V}$

$$A^TA = \begin{array}{c} \\ \\ i \rightarrow \end{array} \begin{bmatrix} | & & | \\ v(a_1) & \cdots & v(a_{|S|}) \\ | & & | \end{bmatrix} \begin{array}{c} \downarrow j \\ \begin{bmatrix} \text{—}\ v(a_1)\ \text{—} \\ \vdots \\ \text{—}\ v(a_{|S|})\ \text{—} \end{bmatrix} \end{array} \quad \text{has } (i,j) \text{ entry } \sum_{(a,b)\in S} \underbrace{v_i(a)\, v_j(a)}_{\substack{\text{monomial of degree} \le 2D \\ \text{that can be factored in several ways}}}$$

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.: $d_0 = 1$

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

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

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

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

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

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

Network training = 'distance' minimization

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

neuromanifold

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

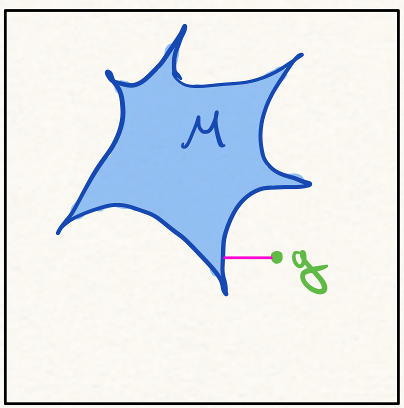
mean squared error

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 $\mathrm{dist}: V \times V \to \mathbb{R}_{\ge 0}$ 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



$d_L > 1$

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

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

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

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

Loss Landscape

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

params input output

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

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

Loss Landscape

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

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

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

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:

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

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

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

Loss Landscape

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

can be studied in a decoupled way:

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

$\rightsquigarrow$ 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?

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 $\mathcal{M}$?

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 $\theta \in \Theta$ & $x \in X$)

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

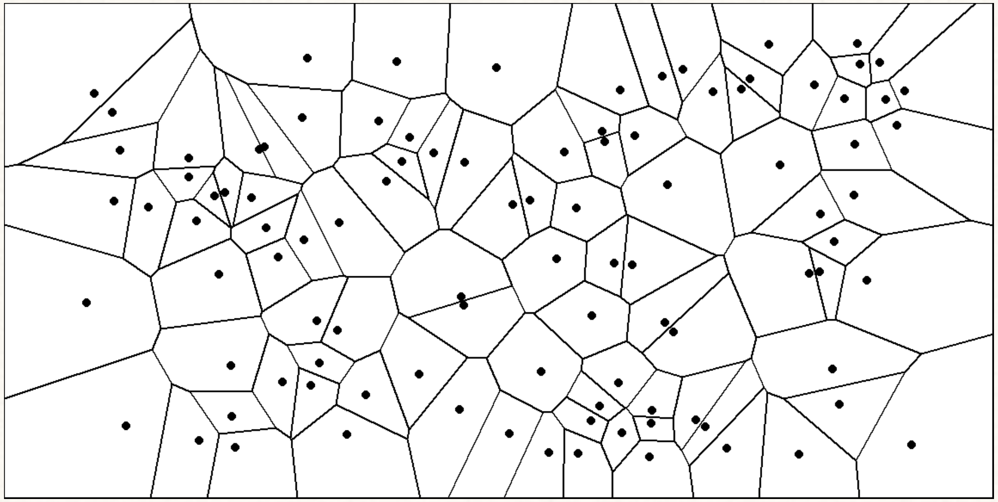
What kind of object is $\mathcal{M}$?

A semialgebraic set!

describable by polynomial equations & inequalities

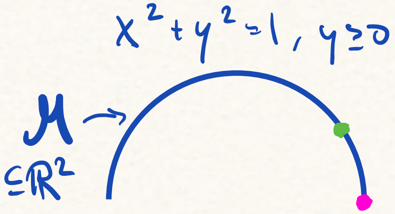
Euclidean distance minimization can be implicitly biased to singularities & boundaries of $\mathcal{M}$

Voronoi cells



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

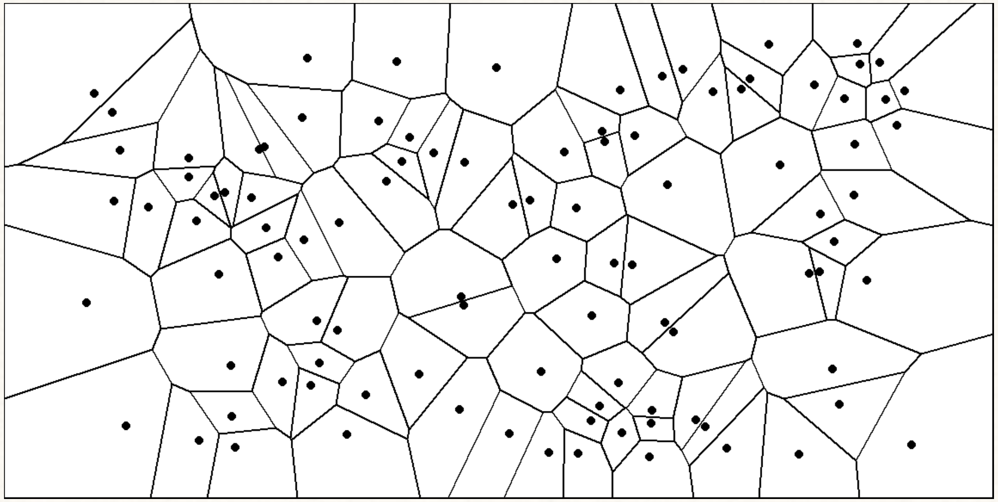
$$Vor_S(p) := \{u \in \mathbb{R}^n \mid \forall q \in S,\ q \neq p : \|p-u\|_2 < \|q-u\|_2\}$$



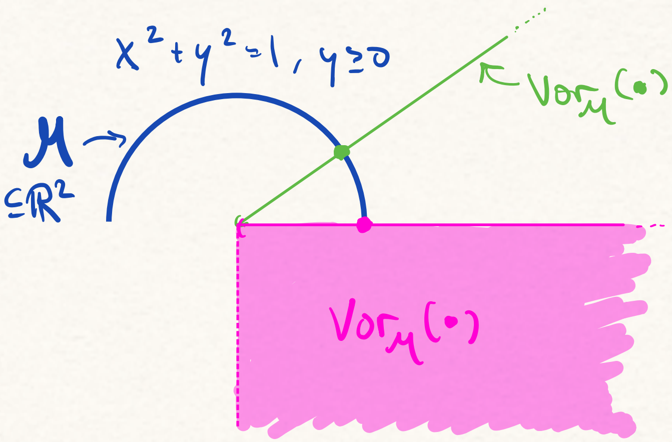
What is the Voronoi cell at • ?

What is the Voronoi cell at • ?

Voronoi cells



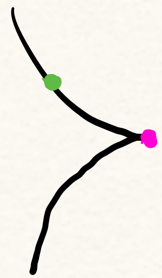
For $S \subseteq \mathbb{R}^n$, the **Voronoi cell** at $p \in S$ is
$$\mathrm{Vor}_S(p) := \{ u \in \mathbb{R}^n \mid \forall q \in S,\ q \neq p : \|p-u\|_2 < \|q-u\|_2 \}$$



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

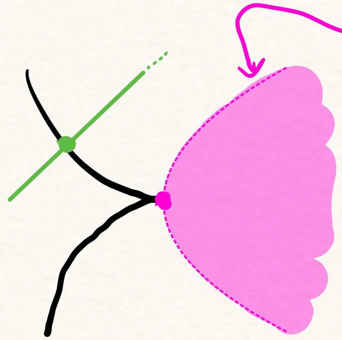


What are the Voronoi cells at and ?

singularities

$y^2+x^3=0$

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



Challenge: Compute this curve!

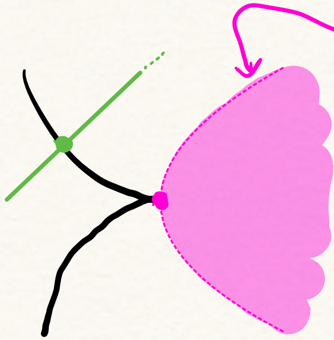
$\rightsquigarrow$ implicit bias towards $\mathrm{Sing}(M)$

What are the Voronoi cells at • and • ?

singularities

$y^2 + x^3 = 0$

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



Challenge: Compute this curve!

$\leadsto$ implicit bias towards Sing(M)

What are the Voronoi cells at • and • ?

Tradeoff

learning close to singularity

$\leadsto$ slow & numerical instability

[Amari et al]

singular solution generalizes better:

① stable global minimum when perturbing data

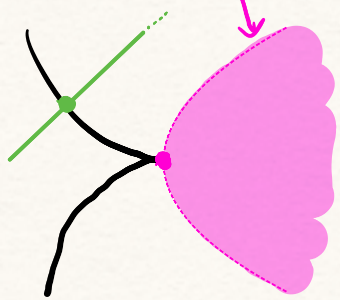
② **Conjecture**: singularities of neuromanifolds are sparse subnetworks

[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 $\mathrm{Sing}(\mathcal{M})$

What are the Voronoi cells at • and • ?

Tradeoff

☹

learning close to singularity

$\rightsquigarrow$ slow & numerical instability

[Amari et al]

☺

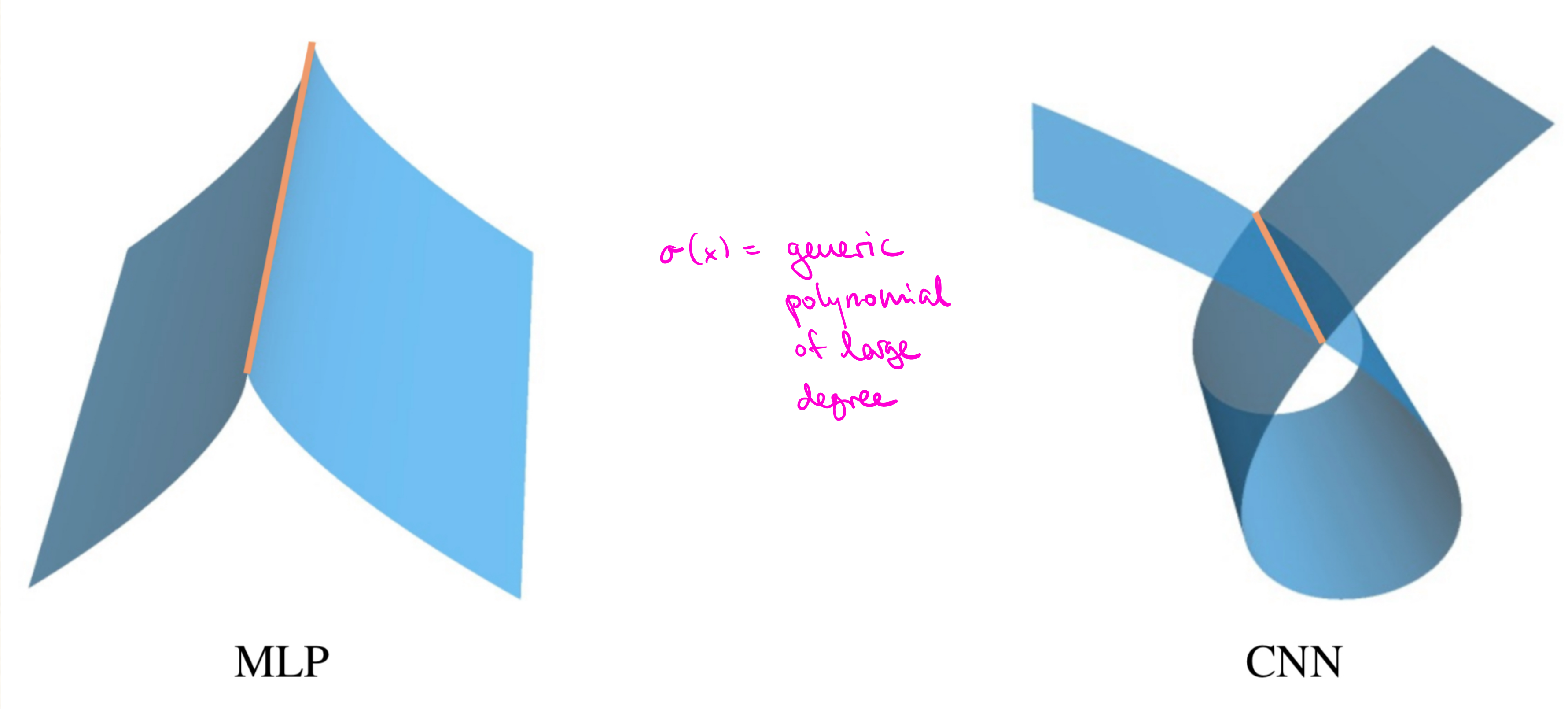
singular solution generalizes better:

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



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

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)

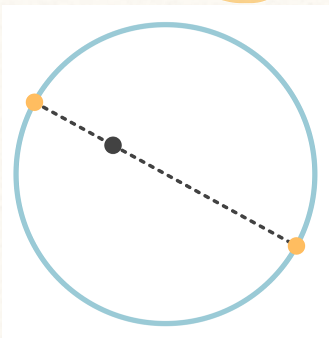
Q symmetric PD $n \times n$ matrix

Fact: For almost all $u \in \mathbb{R}^n$, the number of <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 $\mathrm{EDD}_{\begin{bmatrix}1&0\\0&1\end{bmatrix}}(\bigcirc)$?



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)

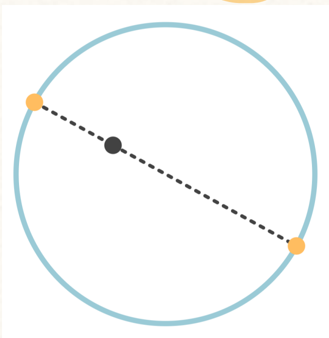
Q symmetric PD $n \times n$ matrix

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

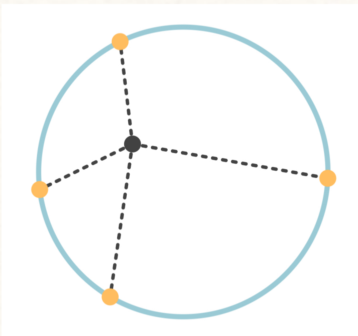
$$\min_{x \in \mathcal{M} \setminus \mathrm{Sing}(\mathcal{M})} \|x-u\|_Q^2$$

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

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



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)

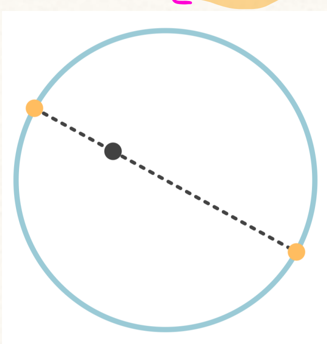
Q symmetric PD $n \times n$ matrix

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

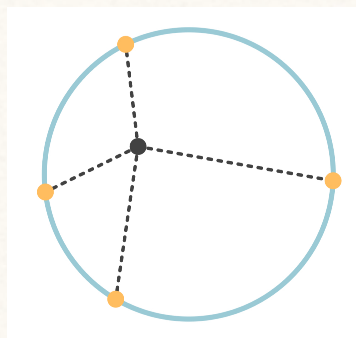
$$\min_{x \in \mathcal{M} \setminus \mathrm{Sing}(\mathcal{M})} \|x-u\|_Q^2$$

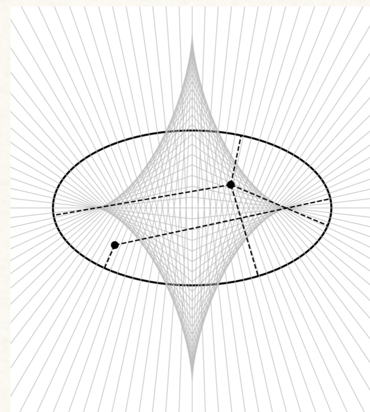
is the same, called the **Euclidean Distance Degree**: $\mathrm{EDD}_Q(\mathcal{M})$.

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



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



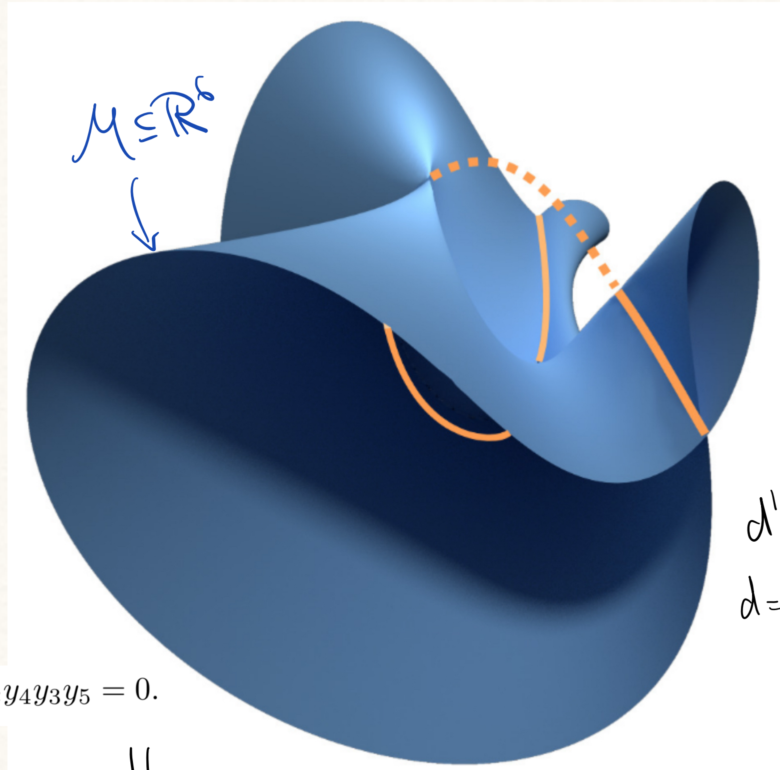


Lightning Self-Attention (single head, single layer)

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

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,

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

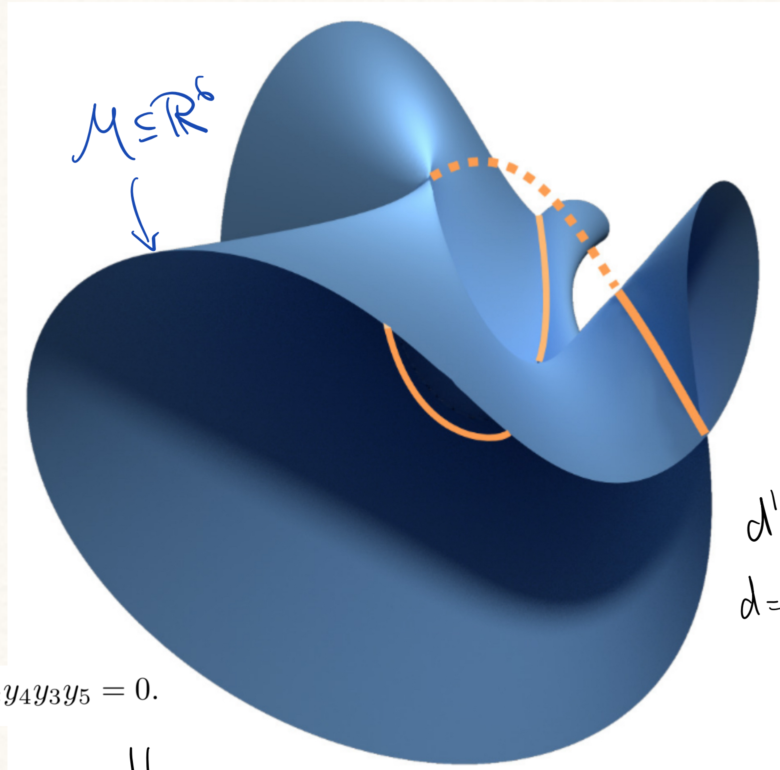
What happens if Q becomes degenerate?

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

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?
(i.e., Q is symmetric positive <u>semi</u>definite)

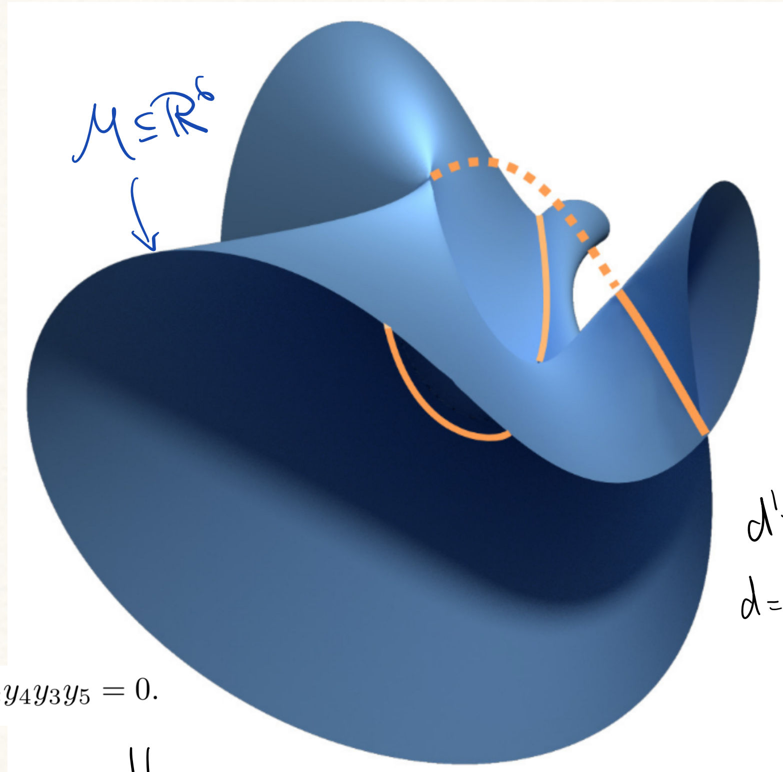
k	complex critical point set
0	14 points
1	14 points
2	4 points + a curve
3	a surface
4	a 3-dimensional subvariety
5	a 4-dimensional subvariety

$k := \dim\ker Q$

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



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

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

$\Downarrow$

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

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

$k := \dim \ker Q$

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

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

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

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

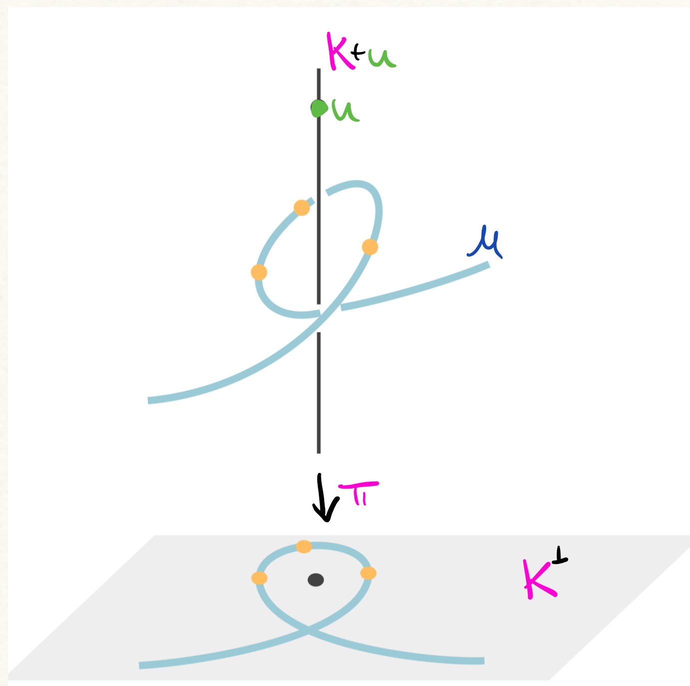
Case 1: Let $k < n - d$.

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

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

$$\Big\updownarrow \; 1:1$$

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



In general

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

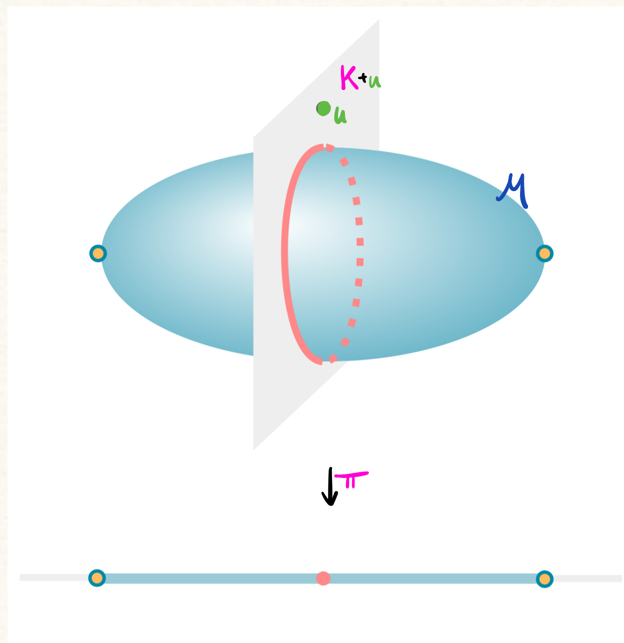
Q symmetric positive semidefinite $n \times n$ matrix
$K := \ker Q$

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

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

Ⓐ $(K+u) \cap M$: 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
$K := \ker Q$

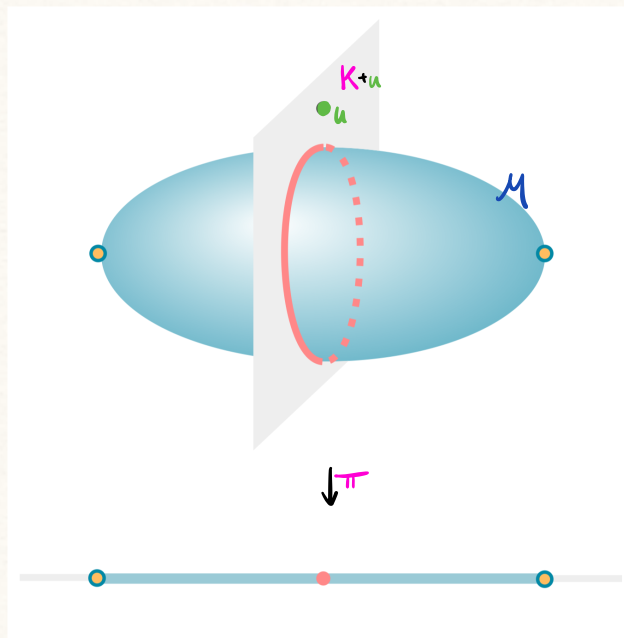
$\pi : \mathbb{R}^n \to K^\perp$
turns Q into nondegenerate quadric

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

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

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



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 $k \geq n-d$.

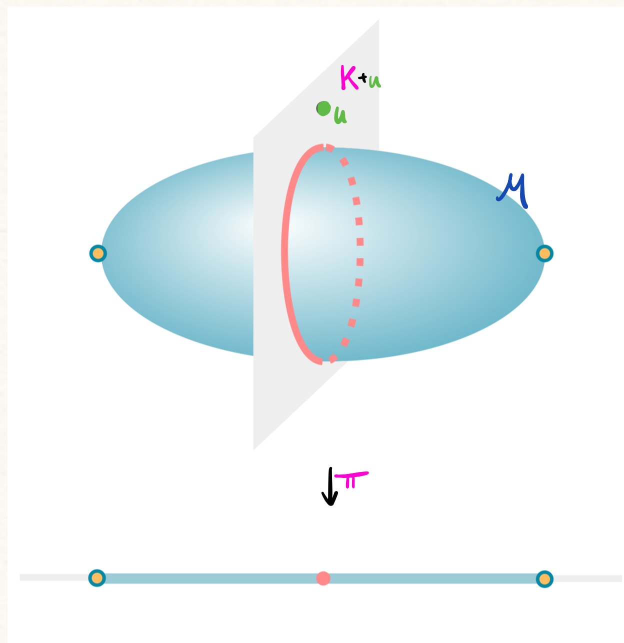
For almost all Q with $k = \dim K$ and almost all $u \in \mathbb{R}^n$, we have 2 types of critical points of $\min\limits_{x \in \mathcal{M} \setminus \mathrm{Sing}(\mathcal{M})} \|x-u\|_Q^2$:

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

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

$\updownarrow$ 1:1

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



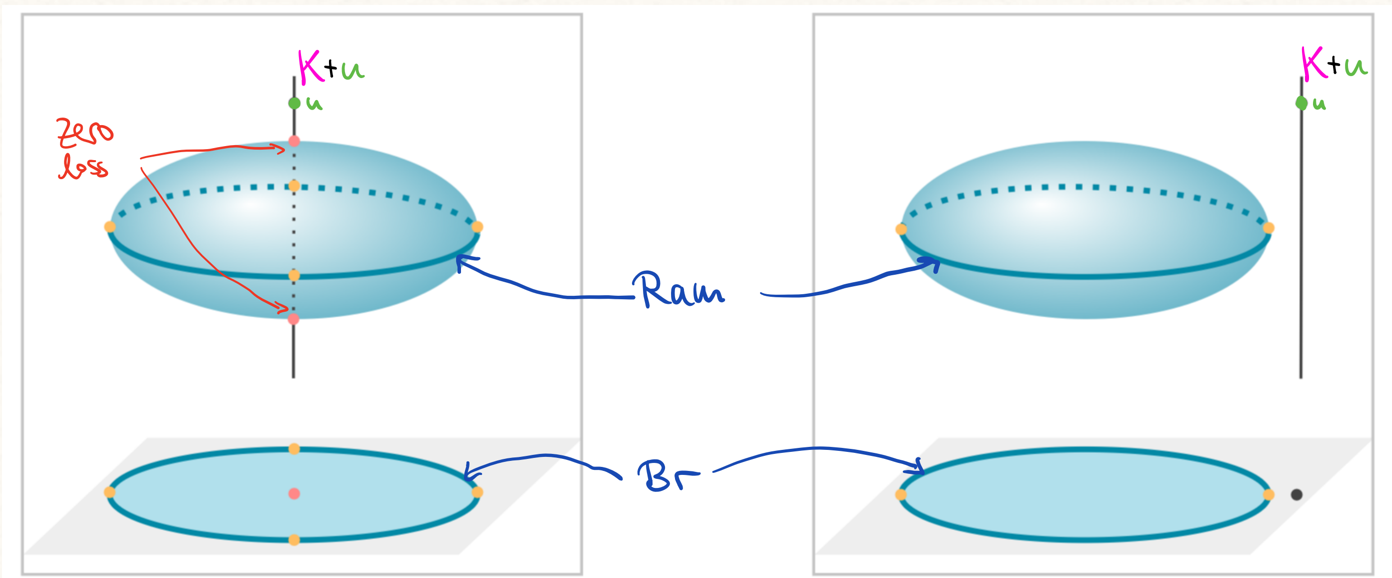
In general

$\mathcal{M} \subseteq \mathbb{R}^n$ 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 $k \geq n - d$.



Induced bias towards Ram!

depends only on K (not on Q) & not on u