Algebra & Geometry of Neural Networks

neural tangent kernel

NTK approach

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

linearized models of ∞ dimension

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

algebraic geometry

AG approach

$$\begin{tikzcd} \boxed{\text{algebraic network}} \arrow[d, "\text{Stone-Weierstraß}"'] \\ \boxed{\text{target network}} \end{tikzcd}$$

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

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

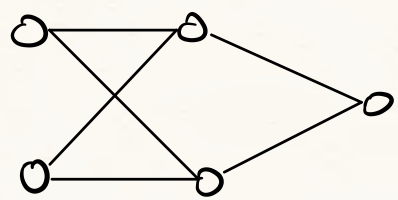
α_i = <u>learnable</u> affine linear functions

σ = nonlinear activation function, applied entrywise

we assume: σ is a univariate polynomial

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

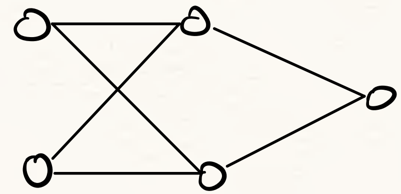
$$\begin{bmatrix} e & f \end{bmatrix} \sigma\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right)$$



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?

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

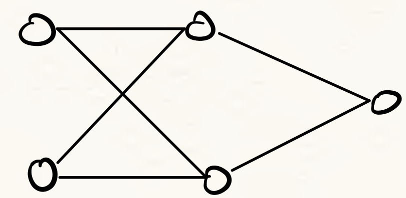
Can you obtain all of $\mathbb{R}[x,y]_2$?

↖ 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 $\mathbb{R}[x,y]_2$?

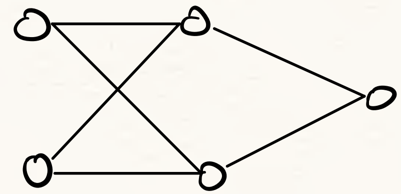
↖ homogeneous quadratic polynomials in x,y

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

YES

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

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



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

Which functions does this MLP parametrize?

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

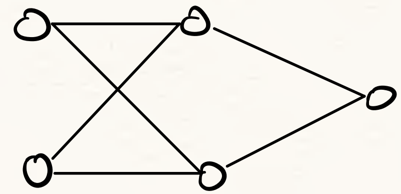
Can you obtain *all* of $\mathbb{R}[x,y]_3$?

↖ homogeneous cubic polynomials in x,y

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

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

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



Which functions does this MLP parametrize?

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

Can you obtain all 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. $A = 1$
$B = 0$
$C = -1$
$D = 0$

Macaulay2

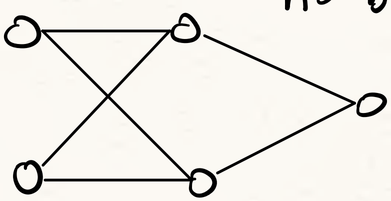
Neuromanifolds

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

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

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

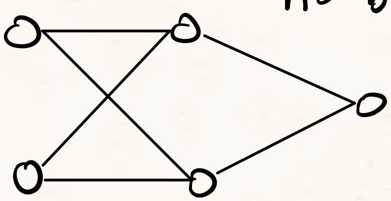
$\sigma(x) = x \quad \Rightarrow \quad$?

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

Examples:



no bias

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

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

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

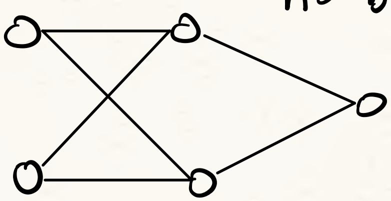
Neuromanifolds

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

parameters inputs outputs

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

Examples:

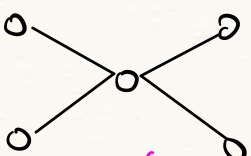


no bias

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

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

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



$\sigma(x) = x$

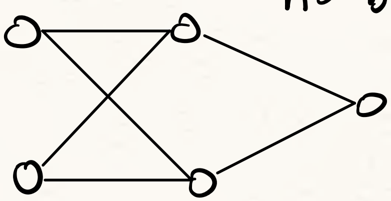
$\begin{bmatrix} c \\ d \end{bmatrix} \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad \Rightarrow \quad \mathcal{M} = ?$

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

Examples:

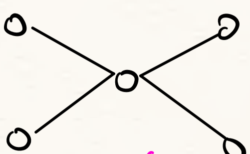


no bias

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

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

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



$\sigma(x) = x$

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

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

$\Rightarrow \mathcal{M} = ?$

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

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

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

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

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

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

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

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

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

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

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

Polynomial MLPs are the <u>only</u> 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:

<u>**Theorem 1:**</u>

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.).

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.).

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

AG approach

$$\boxed{\text{algebraic network}} \xrightarrow{\text{Stone-Weierstraß}} \boxed{\text{target network}}$$

nonlinear models in finite-dimensional ambient spaces

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,

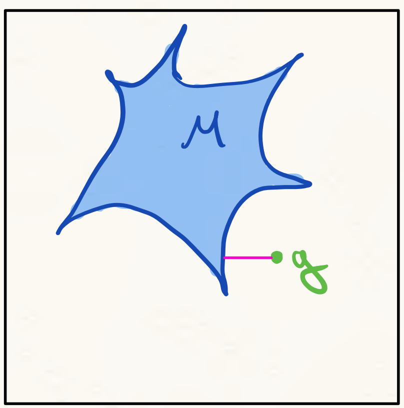
mean squared error

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



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

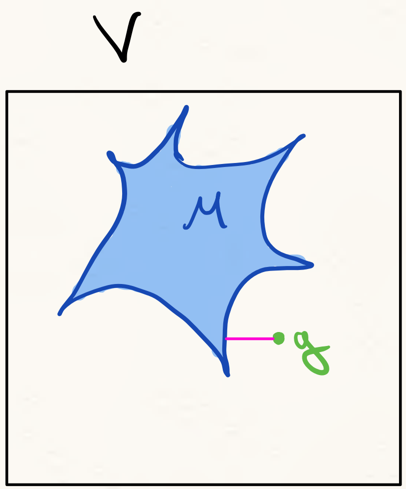
$S \subseteq \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ finite dataset,

mean squared error

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

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

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,

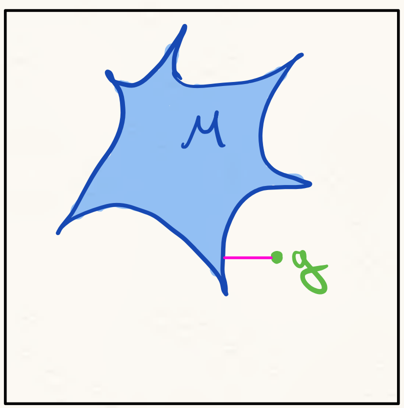
mean squared error

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

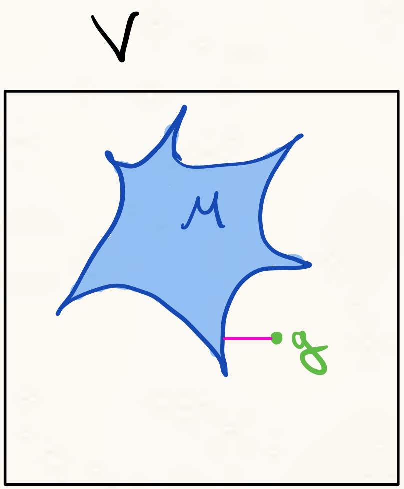
$S \subseteq \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ finite dataset,

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 $f \in \mathcal{M}$.



Veronese embedding

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 $\leq D$),

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

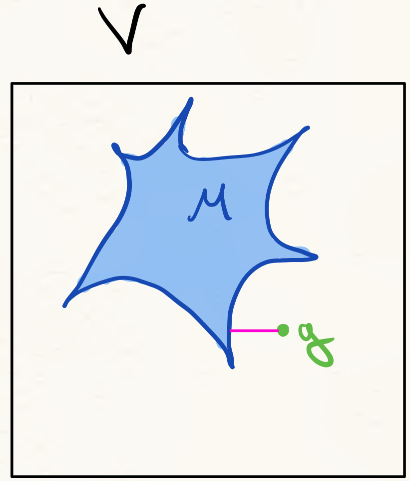
A & B matrices whose rows are $\nu_D(a)$ & b, resp., over all $(a,b) \in S$

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 ($d_L = 1$):

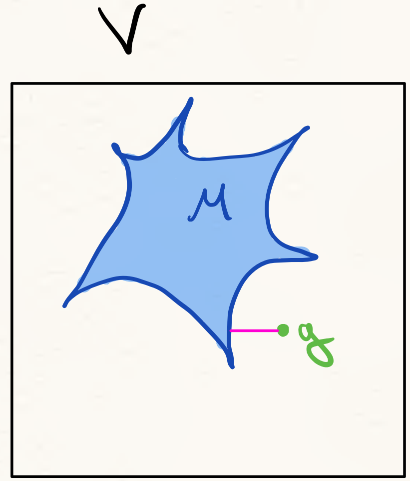
① A^TA depends only on input data, A^+B on both input & output

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

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

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 $|S| < \dim V$ $\rightsquigarrow$ <u>pseudo</u>metric

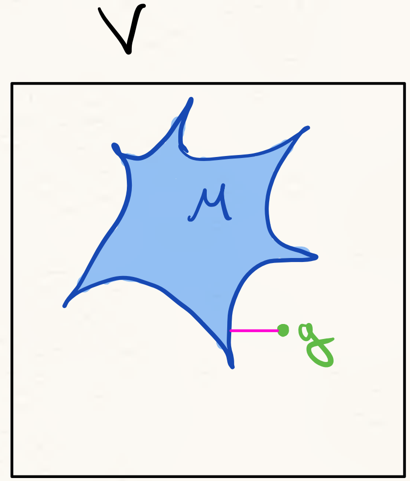
3. even when $|S| \gg \dim V$, A^TA is <u>not</u> an arbitrary symmetric PD matrix, while A^+B yields all vectors $\in \mathbb{R}^{\dim V}$

Why?

Which matrices can be obtained?

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

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



Observations $(d_L = 1)$:

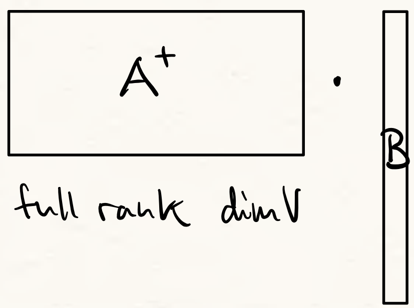
1. A^TA depends only on input data,
   A^+B on both input & output

2. $A^TA \in \mathbb{R}^{\dim V \times \dim V}$ is rank-deficient whenever $|S| < \dim V$ ⇝ pseudometric
   (LLMs: $|S| < \dim \mathcal{M}$)

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

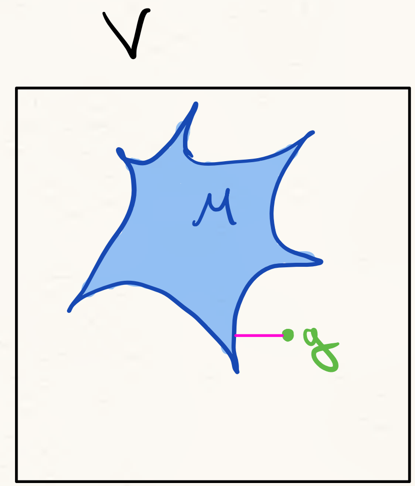
full rank for generic S

⇒



yields all vectors when B varies

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



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 $|S| < \dim V$ ⇝ <u>pseudo</u>metric
(LLMs: $|S| < \dim \mathcal{M}$)

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 $\sum_{(a,b)\in S} \underbrace{v_i(a)\, v_j(a)}_{}$

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

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

Ex.: $d_0 = 1$

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

$$\Rightarrow A = \begin{bmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^D \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_{|S|} & a_{|S|}^2 & \cdots & a_{|S|}^D \end{bmatrix}$$ 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.: $d_0 = 2, \; D = 2$

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

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

Network training = 'distance' minimization

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

neuromanifold

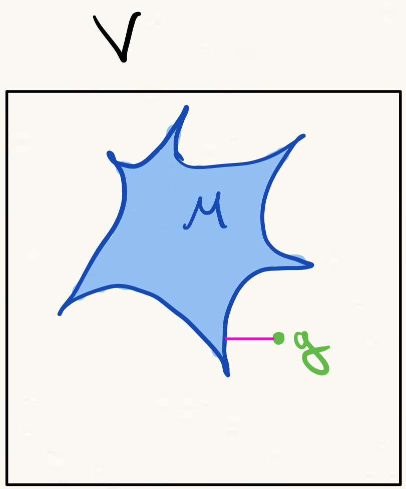
$S \subseteq \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ finite dataset,

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 $f \in \mathcal{M}$.



$d_L > 1$

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

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

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

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

Loss Landscape

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

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

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

Loss Landscape

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

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

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

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:

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

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:

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

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

How?

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