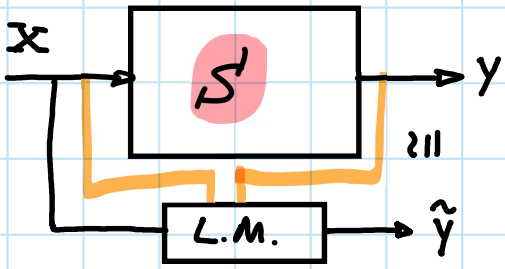
ML AND DATA ANALYSIS	RLS AND KERNELS	October 5 2023



DEFINITIONS	MEANING
$D = \{(x_1, y_1), \dots, (x_N, y_N)\}$	DATASET, THE ONLY KNOWLEDGE THAT WE HAVE ABOUT THE SYSTEM (I.I.D)
$f(x) = \sum_{i=0}^{p} a_i x^i \in F$	CLASS OF FUNCTIONS THAT WE USE FOR APPROSIMATING THE INPUT-OUTPUT RELATION
$\ell(f(x), y) = (y - f(x))^2$	MEASURE OF QUALITY FOR F. HOW MUCH THE THINGS THE f PREDICTS ARE CLOSE TO THE THINGS I MEASURE. LOSS FUNCTION
$R(f) = \mathbb{E}_{xy}(\ell(f(x), y))$	RISK OF A FUNCTION. QUANTITY THAT WE WANT TO MINIMIZE. IT'A CHOICE!
$\hat{R}(f) = \frac{1}{m} \sum_{i=1}^{m} \ell(f(x), y)$	EMPIRICAL ERROR ON my DATA. I CAN MEASURE.

IF D I.I.D. AND $\ell \in [a,b]$ (BOUNDED) THEN:

- IF $|F| = 1$ AND INDEPENDENT FROM D $\Rightarrow$ $|R(f) - \hat{R}(f)|_{1-\delta} \leq \sqrt{\frac{1}{m\delta}}$
- IF $|F| = m_f$ AND INDEPENDENT FROM D $\Rightarrow$ $|R(f) - \hat{R}(f)|_{1-\delta} \leq \sqrt{\frac{1}{m\delta}} + \sqrt{\frac{m_f - 1}{m\delta}}$

$\min_{f \in F} R(f) \rightarrow \min_{f \in F} \hat{R}(f) + \lambda C(f)$

TRADE OFF BETW. ACCURACY AND COMPLEXITY.

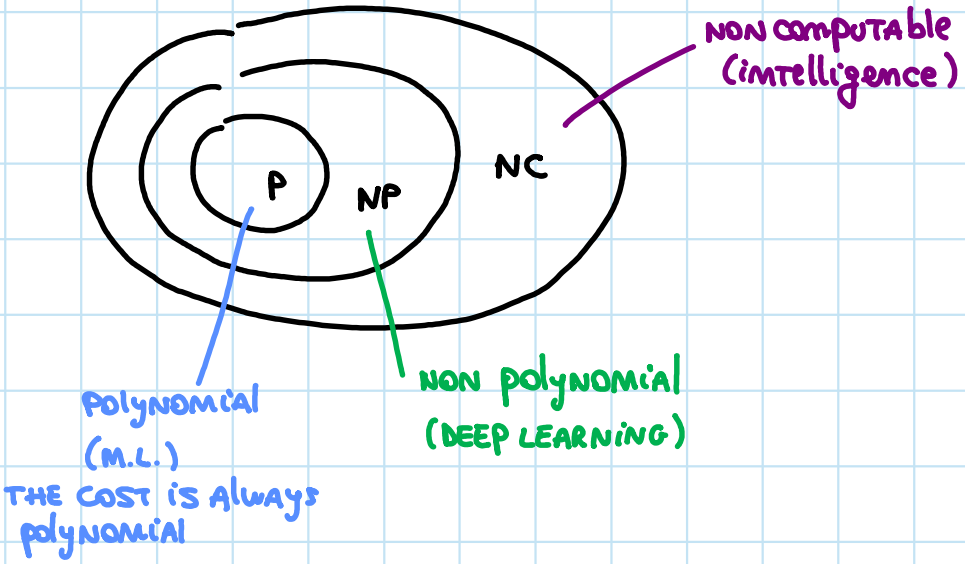
THE SQUARE LOSS IS:
- DIFFERENTIABLE
- CONVEX (good computational properties)

UNKNOWN, ERROR ON THE POPULATION
GENERALITATION ERROR

NO LEARNING, ONLY A STATISTICAL RISK (LIMITED)

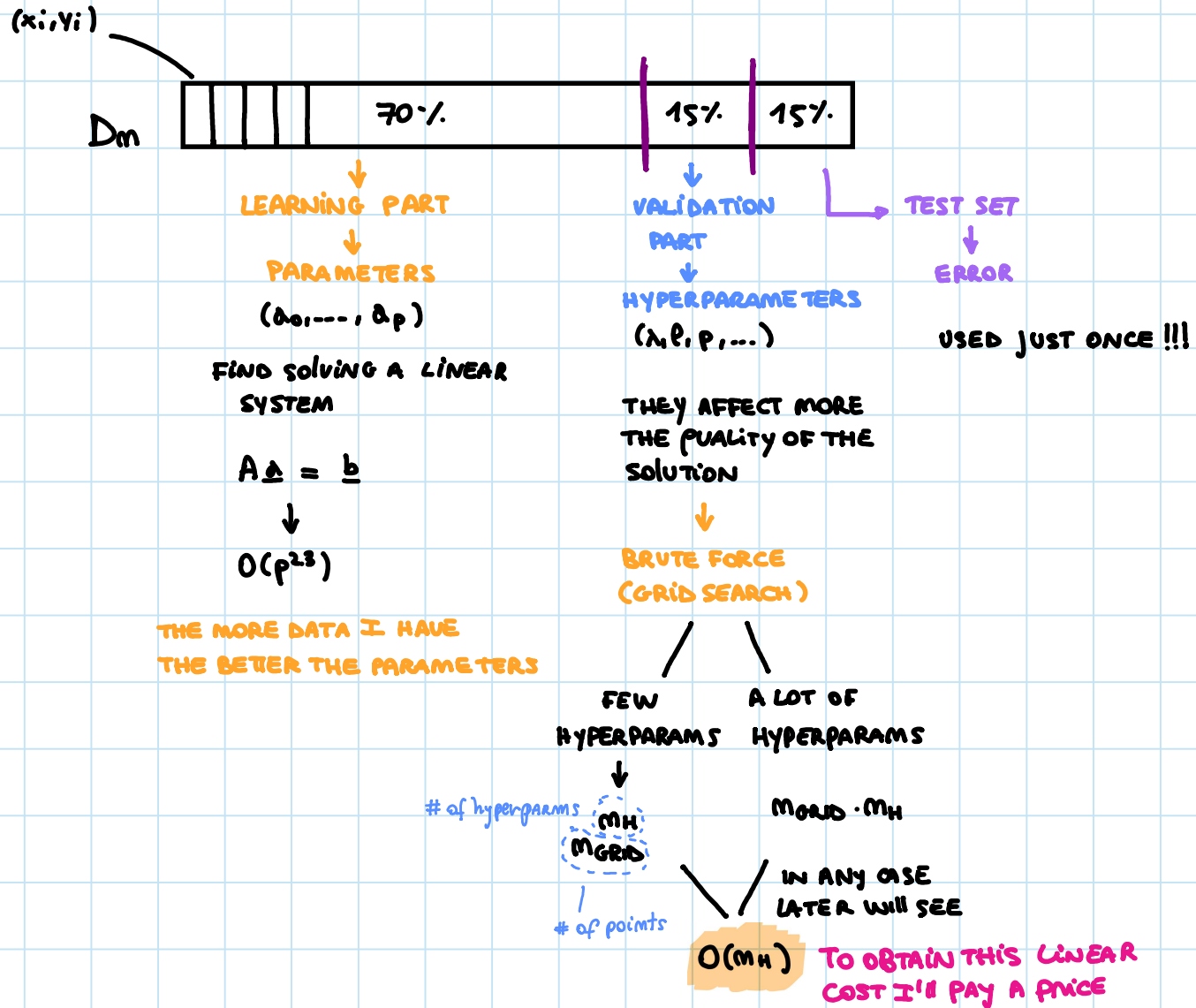
THE MORE FUNC. I USE TO CHECK IF A FUNC. IS GOOD ON THE DATA, THE MORE THE RISK OF FINDING A FUNC. THAT RANDOMLY FITS my DATA INCREASES $\Rightarrow$ I DON'T HAVE ANY KNOWLEDGE OF THE DISTRIBUTION.

TYPES OF PROBLEMS:



WE ARE TRYING TO CONSTRUCT AN INTELLIGENT MACHINE WITH A POLYNOMIAL COST.

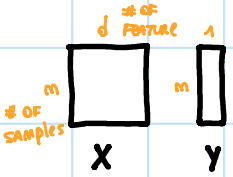
WE CANNOT GET THE CHOICES A PRIORI BUT WE NEED TO TEST THEM ON THE DATA.



IN THE LEARNING PART THE PROBLEM IS CONSTRAINED, BUT IN THE VALIDATION THE PROBLEM BORNS AS NON POLYNOMIAL. TO MAKE IT LINEAR I'LL HAVE TO PAY A PRICE.

ALL OF THIS WAS FOR JUST ONE OUTPUT!

$x \in \mathbb{R} \rightarrow x \in \mathbb{R}^d$



LET'S DO THE CHOICES NOW:

FIRST APPROACH DOESN'T WORK

$f(\underline{x}) = a_0$ FLAT MODEL $O(1)$

$a_{11} x_1 + a_{12} x_2 + \dots + a_{1d} x_d$ (First element x vector) LINEAR MODEL $O(d)$

$a_{11} x_1^2 + \dots + a_{1d} x_d^2 + a_{2d+1} x_1 x_2 + \dots$ QUADRATIC $O(d + \binom{d}{2}) \simeq O(d^2)$

$\vdots$

DEGREE p: $O\binom{d}{p} \simeq O(d^p)$ MC. LAURIN SERIES IN DIMENSION p

$d = 10^{12}$, $p = 3 \Rightarrow (10^{12})^3 = (2^{36})^3 \simeq 2^{100}$ JUST FOR STORING MY PARAMS

LET'S DO A LINEAR MODEL (# OF PARAMS SCALES LINEARLY WITH # FEATURES)

$f(\underline{x}) = \underline{w}\,\underline{x}$ EASIEST THING THAT I CAN DO

- CONSTRAINT: $f(\underline{x})$ PASSES IN $\underline{0} \Rightarrow b = 0$

$\ell(f(\underline{x}), y) = (y - f(\underline{x}))^2$

$C(f) = ?$ BEFORE WAS $\underline{w}' M \underline{w}$ $\Rightarrow C(f) = \|\underline{w}\|^2$

$\underline{w}' I \underline{w} = \|w\|^2$

MY MODEL

$$\min_{f \in F} \hat{R}(f) + \lambda C(f) = \min_{f \in F} \frac{1}{m} \sum_{i=1}^{m} (\underline{w}\underline{x}_i - y_i)^2 + \lambda \|\underline{w}\|^2$$

MATRICAL FORMAT:

$$X = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_m' \end{bmatrix} \in \mathbb{R}^{m \times d} \qquad \underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m$$

$$\min_{\underline{w}} \|X\underline{w} - \underline{y}\|^2 + \lambda \|\underline{w}\|^2$$

RIDGE REGRESSION (RLS)

PARABOLOID → CONVEX problem → 1 SOLUTION

SOLUTION:

$$\nabla_{\underline{w}} \left(\underline{w}' X' X \underline{w} - 2\underline{w}' X' \underline{y} + \underline{y}'\underline{y} + \lambda \underline{w}' I \underline{w} \right) = \underline{0}$$

$$2X'X\underline{w} - 2X'\underline{y} + 2\lambda I \underline{w} = \underline{0}$$

$$\underbrace{(X'X + \lambda I)}_{A \in \mathbb{R}^{d \times d}} \underline{w} = \underbrace{X'\underline{y}}_{\underline{b} \in \mathbb{R}^d}$$

LINEAR SYSTEM $O(d^2)$ CAN BE GOOD OR BAD DEPENDING ON THE SITUATION

CAN BE ALSO REWRITED IN: $\underline{w} = (X'X + \lambda I)^+ X'\underline{y}$ $O(d^{2.3})$

MERCE DISCOVERED KERNELS:

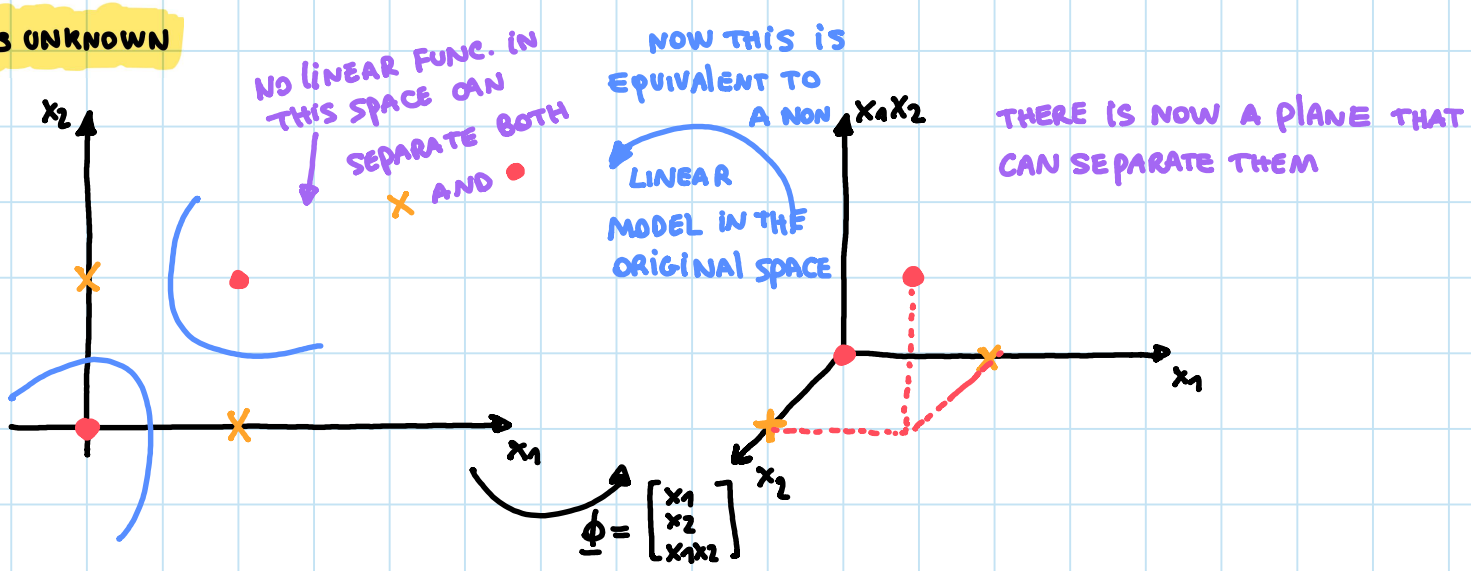
$$k(\underline{x}_i, \underline{x}_j) = e^{-\gamma \|\underline{x}_i - \underline{x}_j\|^2}$$

WORKS IN THE ORIGINAL SPACE BUT EQUIVALENT TO THIS:

$$k(\underline{x}_i, \underline{x}_j) = \underline{\phi}'(\underline{x}_i)\,\underline{\phi}(\underline{x}_j)$$

PROJECT MY DATA IN A NEW SPACE INDUCED BY ϕ AND COMPUTING THE SCALAR PRODUCT IN THIS SPACE

$\underline{\phi}$ IS UNKNOWN



WITH KERNELS I CAN DO SPECIFIC COMPUTATION WITHOUT COMPUTING ϕ.

$e^{-\gamma \|x_i - x_j\|^2}$ WHAT IS $|\phi| = ?$ IN HOW MANY DIMENSIONS THIS KERNEL IS PROJECTING MY DATA

CASE d=1

$$e^{-\gamma\|u-v\|^2} = e^{-\gamma u^2} e^{-\gamma v^2} e^{2\gamma uv} = e^{-\gamma u^2} e^{-\gamma v^2} \sum_{i=0}^{\infty} \frac{(2\gamma uv)^i}{i!} =$$

$$e^x = \sum_{i=0}^{\infty} \frac{1}{i!} \frac{d^i e^x}{dx^i}\bigg|_{x=0} x^i = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

MCLAURIN SERIES

$$= e^{-\gamma u^2} e^{-\gamma v^2} \begin{bmatrix} 1 \\ \sqrt{2\gamma}\, u \\ \frac{2\gamma}{\sqrt{2}} u^2 \\ \frac{\sqrt{(2\gamma)^3}}{\sqrt{3}} u^3 \\ \vdots \end{bmatrix}^T \begin{bmatrix} 1 \\ \sqrt{2\gamma}\, v \\ \frac{2\gamma}{\sqrt{2}} v^2 \\ \frac{\sqrt{(2\gamma)^3}}{\sqrt{3}} v^3 \\ \vdots \end{bmatrix}$$

$\underline{\phi}^T(u)$ $\underline{\phi}(v)$

SO $|\phi| = \infty \Rightarrow$ I CAN ALWAYS LINEARIZE

IN RLS:

$$\min_{\underline{w}} \|X\underline{w} - \underline{y}\|^2 + \lambda\|\underline{w}\|^2$$

$$\underline{w} = (X'X + \lambda I)^{+} X'\underline{y} \quad O(d^2)$$

WHAT IS THE SHAPE OF THIS SOLUTION?

I SUPPOSE THAT THE SOLUTION IS LINEAR IN THE DATA THAT I HAVE (THE SOLUTION DEPENDS ON THE DATA I COLLECTED)

$$\underset{d\times 1}{\underline{w}} = \sum_{i=1}^{m} a_i \underline{x}_i = \underset{d\times m}{X'} \underset{m\times 1}{\underline{a}}$$

$\Rightarrow$ THE SOLUTION LEAVES IN THE SPACE INDUCED BY THE MATRIX X

FOR THE POINTS THAT ARE PERPENDICULAR TO THE SPACE GENERATED BY X

$X\underline{\alpha} = \underline{0}$

↓

my HYPOTHESIS IS THAT my SOLUTION is PARALLEL TO THE SPACE INDUCED BY THE VECTORS INSIDE X