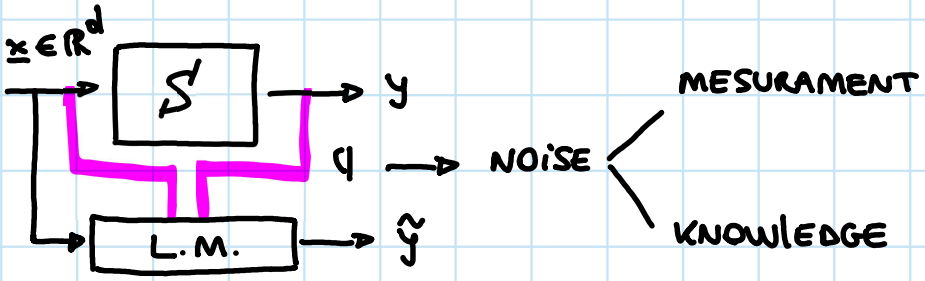
ML AND DATA ANALYSIS	GRADIENT DESCENT	Oct. 31. 2023



$D = \{(x_1, y_1), \dots, (x_N, y_N)\}$ i.i.d. HYPOTHESIS

- MORE DATA I HAVE, MORE INFORMATION ABOUT THE SYSTEM
- SYSTEM DOESN'T CHANGE IN TIME.

$f(x) = \tilde{y} \in F$

- PARAMETERS $w x$
- HYPERPARAMETERS $w \phi(x)$

HERE WE SEARCH FOR A GOOD QUALITY OF THE SOLUTION.

LOSS FUNCTION: $\ell(f(x), y)$ (HYPERPARAMETER)

COMPLEXITY MEASURE: $C(f)$ (HYPERPARAMETER) HERE WE SEARCH FOR THE SEMPLICITY OF THE SOLUTION

WE WANT THE MODEL TO BE GOOD ON PREVIOUSLY UNSEEN DATA, SO WE WANT TO MINIMIZE THE RISK OF OUR FUNCTION.

$$R(f) = \mathbb{E}_{(x,y)} \ell(f(x), y)$$

I JUST CARE TO BE MORE GOOD ON THE DATA THAT IS MORE FREQUENT (HYPER PARAMETER)

WE CANNOT MEASURE THIS RISK, I ONLY HAVE THE DATA, SO I CAN MEASURE THE RISK ON MY DATA.

$$\hat{R}(f) = \frac{1}{m} \sum_{i=1}^{m} \ell(f(x_i), y_i)$$

(EMPIRICAL ERROR)

USING STATISTICALLY LEARNING THEORY: WHEN F IS INDP. FROM D_m

TRADE OFF: COMPLEX FUNC → REDUCE RISK ON MY DATA, SIMPLER FUNC → MORE RISK ON MY DATA

$$R(f) \underset{1-\delta}{\leq} \hat{R}(f) + \lambda C(f) + \Phi(m,\delta)$$

- $\lambda C(f)$: CHOICE RISK (DUE TO THE LEARNING PHASE)
- $\Phi(m,\delta)$: DATA RISK

↳ I NEED TO BALANCE THIS TWO NEEDS.

I SELECT AS BEST FUNCTION:

$$f^* = \underset{f \in F}{\text{ARG min}}\ \hat{R}(f) + \lambda C(f)$$

↳ HOW TO SOLVE THIS MINIMA?

IF:
$$\begin{cases} f(x) = \underline{w}\underline{x} \\ \ell = (y-f)^2 \\ C = \|\underline{w}\|^2 \end{cases}$$

→ THE SOLUTION IS $O(d^2)$ OF $A\underline{w} = b$

→ THE SOLUTION IS $O(n^2)$ OF $A\underline{\alpha} = b$

] OPTIMAL SOLUTION

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

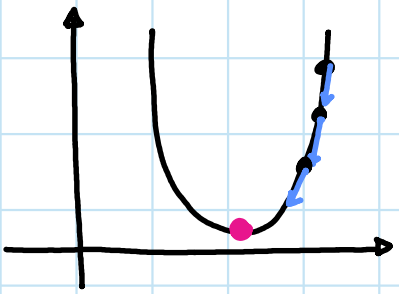
THIS FORMULATION IS GOOD BECAUSE I DON'T HAVE CONSTRAINTS. IT'S THE MINIMA OF A PARABOLOID.

$$\begin{cases} \min_{\underline{x}} M(\underline{x}) \\ \text{S.T.} \\ I(\underline{x}) \leq 0,\ E(\underline{x}) = 0 \end{cases}$$

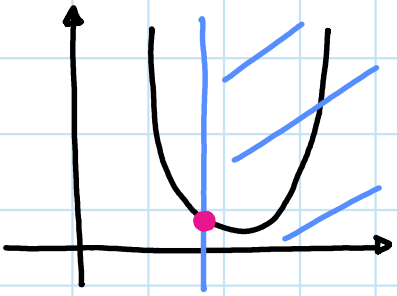
↑ INEQUALITY AND EQUALITY CONSTRAINTS

I WOULD LIKE TO DEAL ALWAYS WITH NON CONSTRAINED OPTIMIZATION PROBLEMS.

$$\min_{\underline{x}} M(\underline{x})$$



UNCONSTRAINED PROBLEM



CONSTRAINED PROBLEM

GRADIENT DESCENT:

$$\begin{cases} \underline{x}_0 = \underline{0} \\ \underline{x}_{i+1} = \underline{x}_i - \eta \nabla_{\underline{x}} M(\underline{x})\Big|_{x_i} \end{cases}$$

NEWTON METHOD

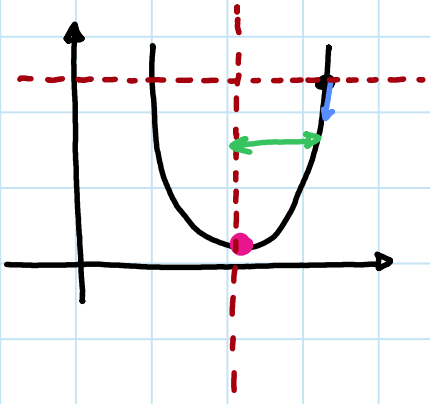
η: LEARNING RATE (HYPERPARAMETER)

WHY USE GRADIENT DESCENT AND NOT THE METHOD ABOVE, SOLVING THE LINEAR SYSTEM?

IF MY FUNCTION $M(x)$ IS L SMOOTH, SO IF $\forall x$ IT'S BOUNDED BY A PARABOLOID

$$M(\underline{x}') \leq M(\underline{x}) + \nabla_{\underline{x}} M(x) \cdot (\underline{x}' - \underline{x}) + \frac{L}{2} \|\underline{x}' - \underline{x}\|^2$$

THERE'S A PARABOLA SMALL ENOUGH THAT CAN BE INSERTED IN EACH POINT.



I KNOW WHICH IS THE MAX VALUE OF η SO THAT I DON'T START TO GO IN THE OPPOSITE DESIDERATE DIRECTION.

$$\eta_{MAX} = \frac{1}{L}$$

AND WE CAN PROVE THAT:

$$M(\underline{x}_t) - M(\underline{x}^*) \leq \frac{L}{t} \|\underline{x}_0 - \underline{x}^*\|^2$$

REACHED POINT t — OPTIMAL POINT

THE DISTANCE BETWEEN THE POINT THAT I AM AND THE POINT THAT I WANT TO REACH, GOES DOWN LINEARLY IN THE NUMBER OF ITERATIONS.

↳ IF I WANT TO BE 2 TIMES CLOSER, I HAVE TO DOUBLE THE NUMBER OF ITERATIONS.

LINEAR SYSTEM

$O(m^2)$

$A\underline{x} = b$

↳ $x'x$ OR xx'

$$X = \begin{bmatrix} \text{---} & x_1' & \text{---} \\ & \vdots & \\ \text{---} & x_N' & \text{---} \end{bmatrix} \in \mathbb{R}^{n \times d}$$

I NEED TO HAVE ALL THE DATA ON MY P.C.

A ALSO NEED TO BE SEMIDEFINITE POSITIVE

↓

NOT VERY PARALLELIZABLE

VERY TIME COMSUMING!

GRADIENT DESCENT

$1/t$

$$x_{i+1} = x_i - \eta \nabla_x M(x)\Big|_{x_i}$$

↓

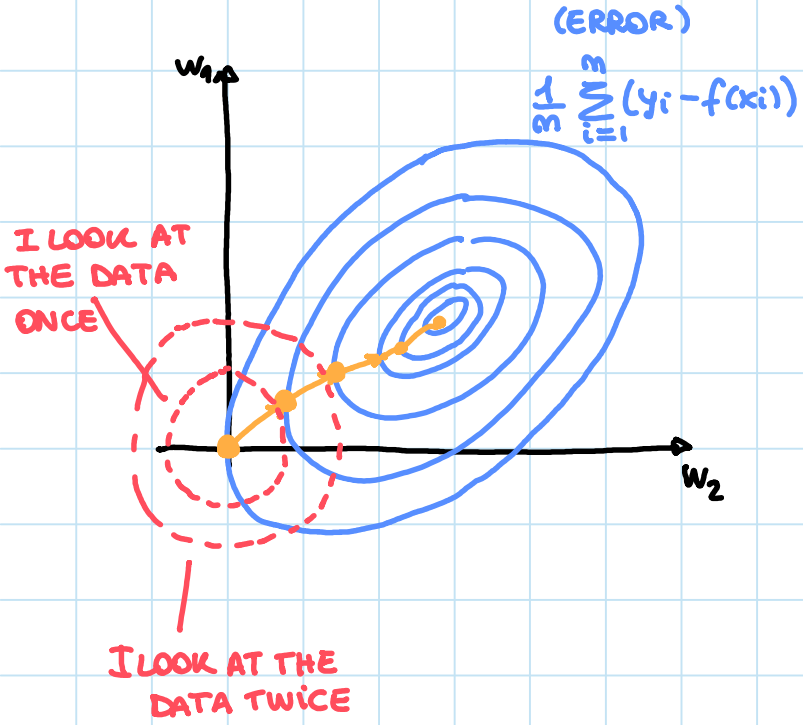
$$\min_{\underline{w}} \frac{1}{m} \sum_{i=1}^{m} (y_i - \underline{w}\underline{\phi}(\underline{x}_i))^2 + \lambda \|\underline{w}\|^2$$

I Apply HERE THE GRADIENT:

$$\frac{1}{m} \sum_{i=1}^{m} \nabla_{\underline{w}} (y_i - \underline{w}\underline{\phi}(\underline{x}_i))^2 + 2\lambda \underline{w}$$

THIS I PARALLELIZABLE!!

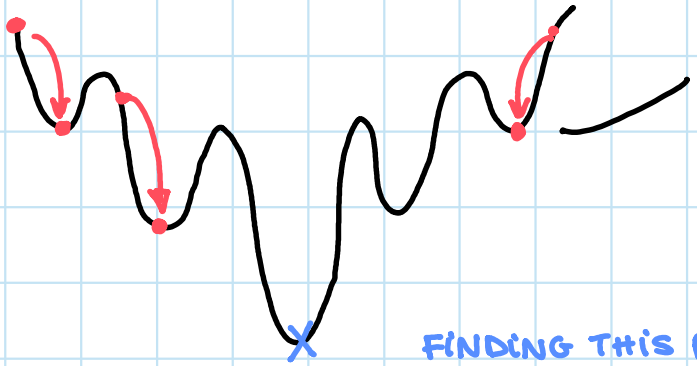
OSS: EVERY STEP THAT I DO I'M RE-LOOKING AT MY DATA → I'M OVERFITTING MY DATA



THE NUMBER OF STEPS ARE EQUIVALENT TO $\|w\|^2$ → IMPLICIT REGULARIZER

⇒ RIDGE REGRESSION IF I DON'T REACH THE MINIMA.

IF MY PROBLEM IS NOT CONVEX:



FINDING THIS POINT IS NON POLYNOMIAL

FROM THE OPTIMIZATION POINT OF VIEW THEY ARE WORSE, BUT FROM THE PRACTICAL THEY COULD BE BETTER (BEST TRADE OFF BETWEEN SEMPLICITY AND QUALITY)

MINIMA THAT ARE EASIER TO FIND ARE LESS PROBABLE TO OVERFIT.

NUMBER OF ITERATION IS A COMPLEXITY MEASURE IN GRADIENT DESCENT!