OPTIMAL LECTURE 2, 26/1/2021

$$\underset{h \in H}{\text{minimize}} \; R_S(h)$$

ERM
Empirical
Risk
Minimization

In most cases h is parameterized by a vector of parameters $w \in \mathbb{R}^d$

ex. $h(x) = \sum_{i=1}^{d} w_i \cdot x_i$

$$h(w, x) = \sum_{i=1}^{d} w_i x_i$$

$$R_S(h) = \sum_{i=1}^{|S|} \ell(h, (x_i, y_i))$$

$$\text{ex} = \underbrace{\sum_{i=1}^{|S|} (h(w, x_i) - y_i)^2}_{} = \sum_{i=1}^{|S|} f(w, i)$$

$$F(w) = R_S(h)$$

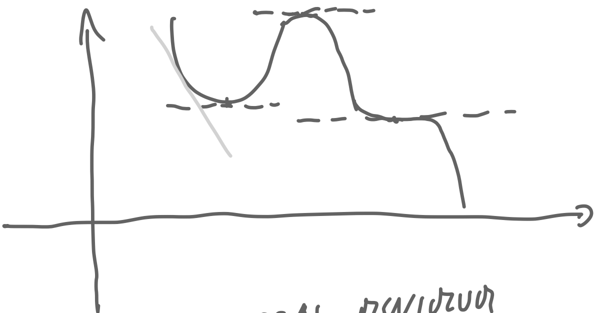
$$\boxed{\underset{w \in \mathbb{R}^d}{\text{minimize}} \; F(w)}$$

$$\underset{w \in W \subset \mathbb{R}^d}{\text{minimize}} \; F(w)$$

$$\underset{w \in \mathbb{R}^d}{\text{minimize}} \quad F(w)$$

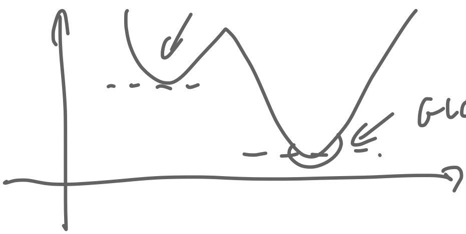
$d = 1$ $\quad F'(w^*) = 0$

w^* CRITICAL POINT



$F''(w^*) > 0$

$\Rightarrow w^*$ IS A MINIMUM



$d > 1$

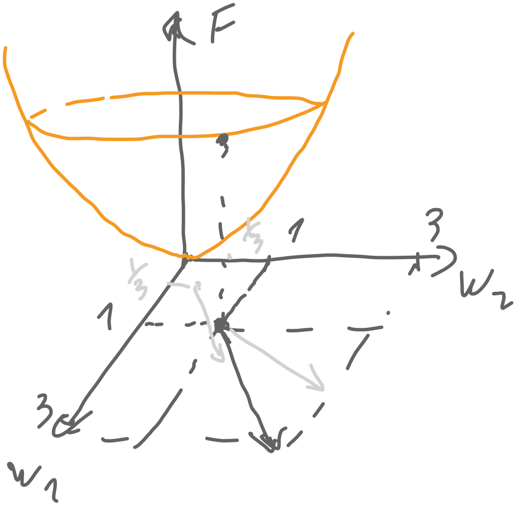
$$\frac{\partial F}{\partial w_1}$$

$$\nabla F(w) = \begin{bmatrix} \frac{\partial F}{\partial w_1} \\ \frac{\partial F}{\partial w_2} \\ \vdots \\ \frac{\partial F}{\partial w_d} \end{bmatrix}$$

$$F(w) = w_1^2 + w_2^2$$

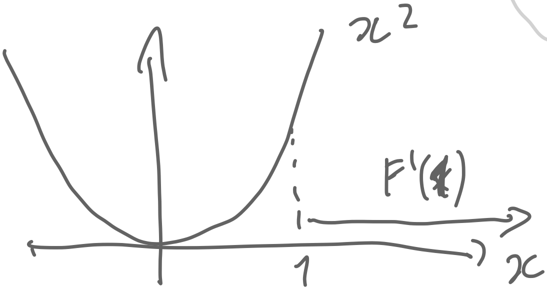
$$\frac{\partial F}{\partial w_1} = 2w_1$$

$$\frac{\partial F}{\partial w_2} = 2w_2$$



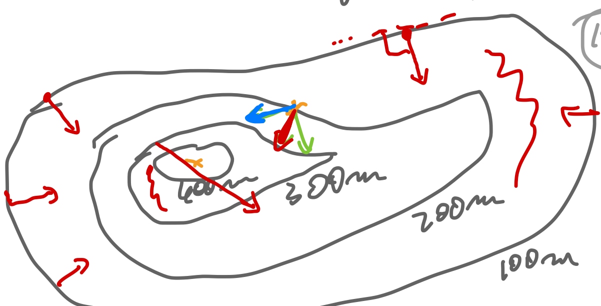
$$F(w) = w_1^2 + w_2^2$$

$$\nabla F(w) = \begin{bmatrix} 2w_1 \\ 2w_2 \end{bmatrix}$$



∇F → points in the direction where the function is growing the most

→ $\|\nabla F\|$ tells us how fast is growing in that direction



ISO LEVEL CURVES

THIS FUNCT. SHOULD BE INCREASING

POSSIBLE?

∇F gives the direction of maximum increase of F

$\Downarrow$

it is orthogonal to the isolevel curves of the function

$$w^* \in \arg\min F(w) \Rightarrow \nabla F(w^*) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\frac{\partial F}{\partial w_1} = 0 \qquad \frac{\partial F}{\partial w_2} = 0$$

HESSIAN OF F

$F(w_1, w_2)$

$$\frac{\partial^2}{\partial w_2 \partial w_1} F(w_1, w_2) =$$

$$= \frac{\partial}{\partial w_2}\left(\frac{\partial}{\partial w_1} F(w_1, w_2)\right)$$

$$EX = \frac{\partial}{\partial w_2}\left(\frac{\partial}{\partial w_1} w_1^2 + w_2^2\right) =$$

$$= \frac{\partial}{\partial w_2}(2w_1) = 0$$

$H_F(w)$ IS A MATRIX

$$[H_F(w)]_{i,j} = \frac{\partial^2}{\partial w_i \, \partial w_j} F(w)$$

ex $$H_F(w) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\frac{\partial}{\partial w_1} \frac{\partial}{\partial w_1} F(w) = 2$$

$$F''(w) \iff H_F(w)$$

$$F''(w) \geq 0 \overset{?}{\Longrightarrow} \cancel{\det(H_F(w)) > 0}$$

$H_F(w)$ IS SEMIDEFINITE POSITIVE

$$\forall x \in \mathbb{R}^d, \quad x^T H_F(w) x \geq 0$$

$$F''(w) > 0 \Rightarrow H_F(w)$$ IS DEFINITE POSITIVE

$$\forall x \in \mathbb{R}^d - \{0\}, \; x^T H_F(w) \, x > 0$$

$H_F(w)$ IS POSITIVE SEMIDEFINITE

$\forall x \in \mathbb{R}^d \qquad x^T H_F(w)\, x \geq 0$

$d = 1 \qquad H_F(w) = \left[\frac{\partial^2 H_F}{\partial^2 w_1}\right] = H_F''(w)$

$$x\, H_F''(w)\, x = x^2 \times H_F''(w) \geq 0$$

$$x^2 \geq 0 \Rightarrow H_F''(w) \geq 0$$

$H_F(w)$ COMPUTE SINGULAR VALUES

A IS DIAGONALIZABLE

$$\boxed{A = U D U^T} \qquad U^T U = I$$

U IS ORTHOGONAL MATRIX

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & & 0 \\ 0 & 0 & \ddots & \lambda_d \end{pmatrix}$$

$\{\lambda_i\}$ EIGENVALUES

$A = U D U^T \qquad AU = U D U^T U =$

$AU = UD$

$A u_i = \lambda_i \cdot u_i$

U IS THE MATRIX ARE EIGEN VECTORS

$A = A^T \Rightarrow A$ IS DIAGONALIZABLE

$H_F = H_F^T$ $\quad H_F$ IS "

$H_F(w) = U D U^T$

H_F IS POSITIVE SEMIDEFINITE

IFF $\lambda_i \geq 0 \quad \forall i$

H_F IS POSITIVE DEFINITE

IFF $\lambda_i > 0 \quad \forall i$

Imagine $\lambda_j < 0$ I pick $x = u_j$

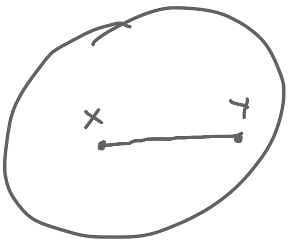
$$u_j^T H_F u_j = u_j^T \lambda_j u_j = \underbrace{\|u_j\|^2}_{>0} \lambda_j < 0$$

$\Rightarrow H_F$ IS NOT POSITIVE SEMIDEFINITE

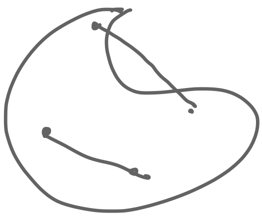
$H_F(w) \longrightarrow \nabla^2 F(w)$

CONVEXITY

Convex set



A

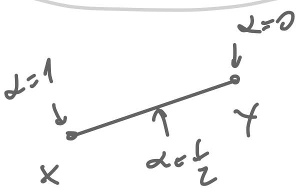


B NOT CONVEX

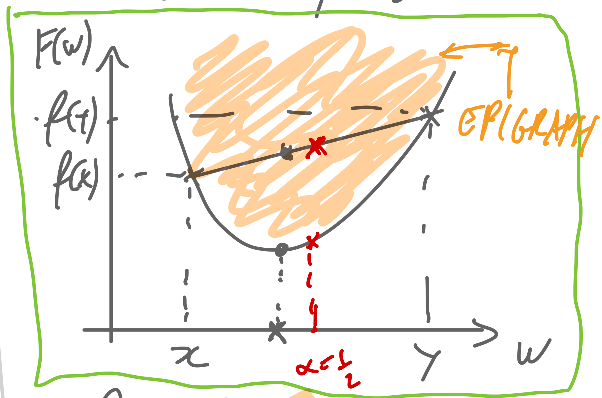
A is convex if

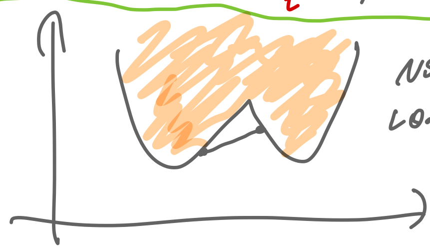
$\forall\, x, y \in A,\ \forall\, \alpha \in [0,1]$

$$\alpha x + (1-\alpha) y \in A$$



Convex function





NOT CONVEX

F IS CONVEX IF

$\forall\, x, y \quad \forall\, \alpha \in [0,1]$

$$f(\alpha x + (1-\alpha) y) \boxed{\leq} \alpha f(x) + (1-\alpha) f(y)$$

WHY DO WE CARE ABOUT CONVEXITY

CONVEX OPT. PROBLEM

$$\underset{w \in W}{\text{minimize}} \ F(w)$$

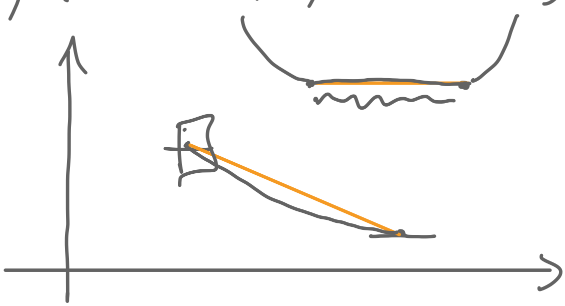
W is a convex set

F is a convex function

WHY CONVEX OPT. PROBLEMS ARE NICE?

IN A CONVEX OPT PROBLEM

1) IF $\nabla F(w^*) = 0 \Rightarrow w^*$ IS A LOCAL MINIMUM



2) IF w^* IS A LOCAL MINIMUM IT IS ALSO A GLOBAL MINIMUM

$\Downarrow$

$$\boxed{\text{THERE ARE EFFICIENT ALGORITHMS TO SOLVE CONVEX OPTIMIZATION PROBLEMS}}$$

CLASSIFICATION PROBLEM

$$\ell(h,(x_i,y_i)) = \mathbb{1}_{h(x_i)\neq y_i}$$ NATURAL CHOICE

COMMON CHOICE

CROSS-ENTROPY LOSS

$$\ell(h,(x,y)) = -y \log h(x) - (1-y)\log(1-h(x))$$

$$h(x) = w^T x$$

THIS LOSS IS DIFFERENTIABLE

THIS LOSS IS CONVEX IN w

$\boxed{\text{SURROGATE LOSS}}$

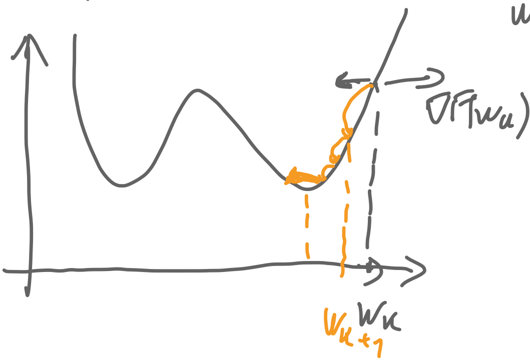
NATURAL CHOICE FOR $\ell_{NAT}(h,(x,y))$

YOU PICK $\ell_{SUR}(h,(x,y))$

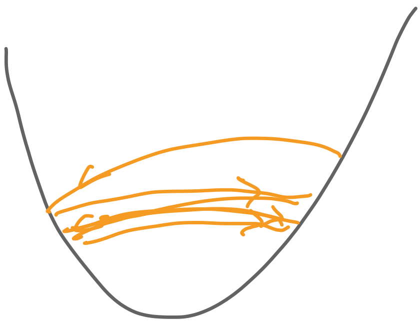
1) $\ell_{NAT}(h,(x,y)) \leq \ell_{SUR}(h,(x,y))$

2) $\ell_{SUR}(h,(x,y))$ IS CONVEX

$$\underset{w \in \mathbb{R}^d}{\text{minimize}} \; F(w)$$



$$w_{k+1} = w_k - \underbrace{\alpha_k}_{\text{LEARNING RATE}} \nabla F(w_k)$$



α CONST

$$\nabla F(w_k) \simeq 0$$

$$F(w) = \sum_{i=1}^{|S|} f(w, i)$$

$$w_{k+1} = w_k - \alpha \nabla F(w_k)$$

FULL-BATCH GRADIENT METHOD

$$\nabla F(w_k) = \nabla \left(\sum_i f(w, i) \right) = \sum_i \nabla f(w, i)$$

BATCH = DATASET

AT EVERY ITERATION YOU TAKE ONLY ONE POINT IN THE DATASET UNIFORMLY AT RANDOM ξ

$$w_{k+1} = w_k - \eta \nabla f(w, \xi)$$

STOCHASTIC GRADIENT METHOD
DESCENT

YOU CAN PICK A RANDOM SUBSET ξ_k OF SAMPLES FROM THE DATASET

$$\boxed{\frac{1}{n_k} \sum_{i=1}^{n_k} \nabla f(w, \xi_{k,i})}$$

$$\xi_k = \{ \xi_{k,1}, \xi_{k,2}, \dots \xi_{k,n_k} \}$$

MINI-BATCH GRADIENT METHOD

$n_k = 1 \equiv$ SGD

$n_k = |S| \equiv$ FULL-BATCH

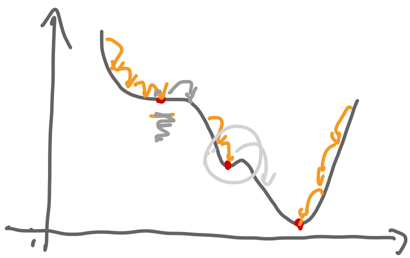
minimise $F(w) = \sum_{i=1}^{|S|} f(w, i)$

WHY NOT ALWAYS FULL-BATCH?

- I MAY NOT HAVE ENOUGH MEMORY

$\tilde{S} \subset S$ minimise $\sum_{(x_i, y_i) \in \tilde{S}} \ell(h, (x_i, y_i))$

- NOISE CAN BE USEFUL ~~OVERFITTING~~

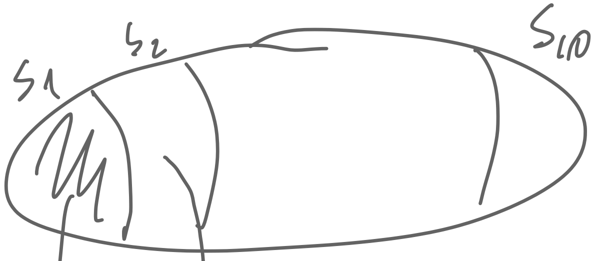


FB could converge to any of them

THE NOISE CAN HELP ME AVOIDING TO GET STUCK IN BAD CRITICAL POINTS.

IT DOESN'T HAPPEN FOR CONVEX PROBLEM.

- GO BACK MEMORY



RAM $\to \dfrac{1}{|S_1|} \sum_{i \in S_1} \nabla f(w, i) = g_1$

RAM $\to \dfrac{1}{|S_2|} \sum_{i \in S_2} \nabla f(w, i) = g_2$

$$\frac{g_1 + g_2 + \dots g_{10}}{10} = \frac{1}{|S|} \sum_{i \in S} \nabla f(w, i)$$

$$\equiv FB$$

FB

$$W_2 = W_1 - \alpha \frac{1}{|S|} \sum_{i \in S} \nabla f(w_1)$$

MB

10

$$W_2 = W_1 - \alpha \frac{1}{|S_1|} \sum_{i \in S_1} \nabla f(w_1)$$

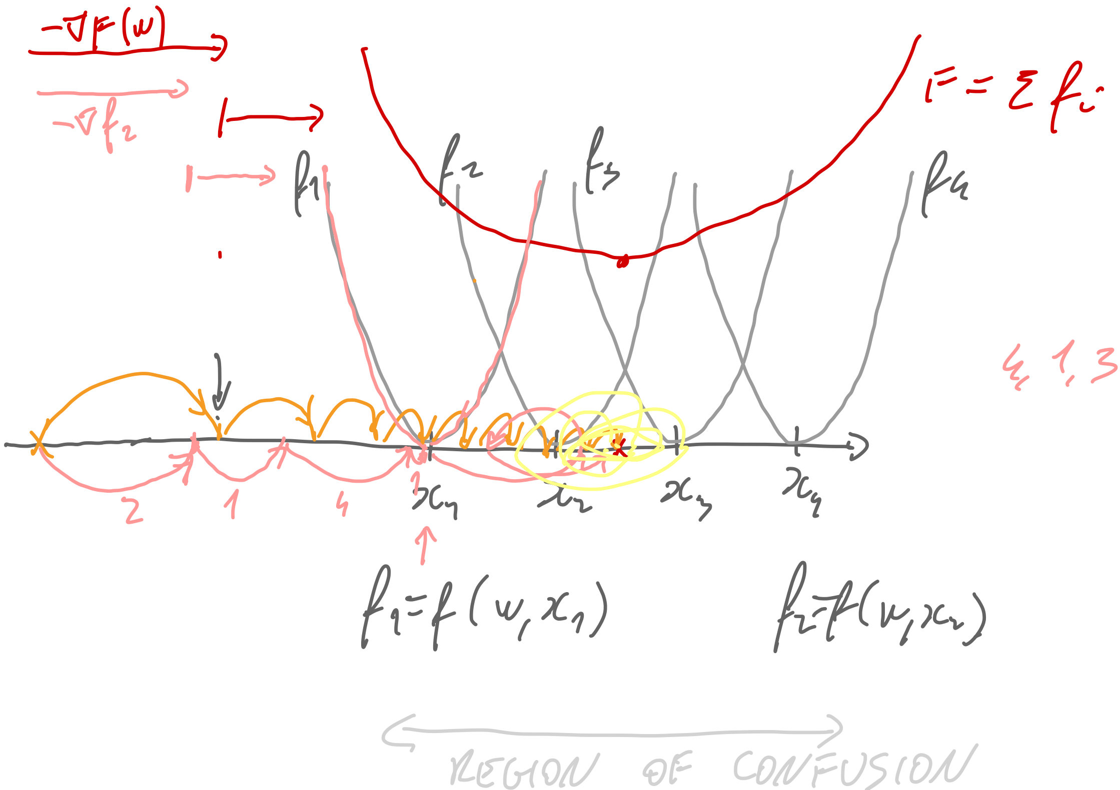
$$W_3 = W_2 - \alpha \frac{1}{|S_2|} \sum_{i \in S_2} \nabla f(w_2)$$

$$\cdots$$

$$W_{10} = W_9 - \alpha \frac{1}{|S_{10}|} \sum_{i \in S_{10}} \nabla f(w_9)$$

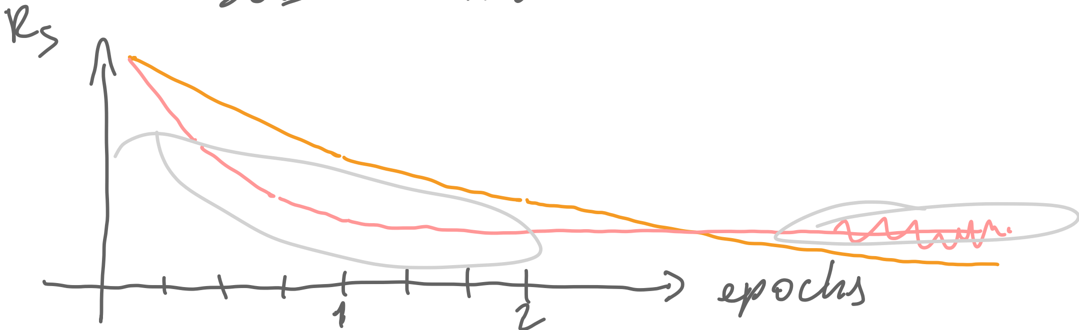
IS IT BETTER TO DO 1 PRECISE UPDATE (FB)

OR 10 NOT-SO PRECISE UPDATES (MB)?



1) WHEN WE ARE FAR FROM THE MINIMUM SGD GIVES US ≃ THE RIGHT DIRECTION

2) WHEN WE ARE CLOSE TO THE MINIMUM SGD CAN GET "CONFUSED"



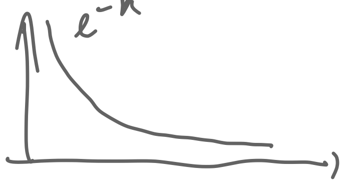
IDEA : START WITH SMALL n_k
AND PROGRESSIVELY INCREASE n_k

VARIANCE REDUCTION

Theoretical results

FB: $R_S(w_k) - R_S(w^*) \sim \rho^k \quad \rho < 1$

LINEAR CONVERGENCE



$\rho^k < \varepsilon \quad \Rightarrow \quad k_\varepsilon \in O\left(\ln \frac{1}{\varepsilon}\right)$

NUMBER OF ITERATIONS

SGD

$R_S(w_k) - R_S(w^*) \sim \frac{1}{k}$ SUBLINEAR

$k_\varepsilon \in O\left(\frac{1}{\varepsilon}\right)$

FB does $n = |S|$ calculations for iteration

SGD does 1 calculation for iteration

$$C_\varepsilon \begin{cases} \nearrow n \ln \frac{1}{\varepsilon} & \text{FB} \\ \searrow \frac{1}{\varepsilon} & \text{SGD} \end{cases}$$

$\uparrow$

operations to reach ε accuracy

If the dataset is huge
it can be $n \ln \frac{1}{\varepsilon} \gg \frac{1}{\varepsilon}$

T for one iteration

- Above $T \propto$ # operations
- $T = T_0 + \beta$ (# operations)

$\uparrow$

$\beta = 0$