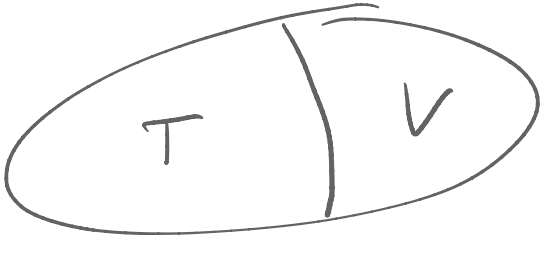
Lecture 2

$$h_1^* \in \operatorname*{argmin}_{h \in H_1} L_T(h)$$

$$h_2^* \in \operatorname*{argmin}_{h \in H_2} L_T(h)$$

$$H_1 \subset H_2$$

$$L_T(h_1^*) \geq L_T(h_2^*)$$



$$L_V(h_1^*) \lessgtr L_V(h_2^*)$$

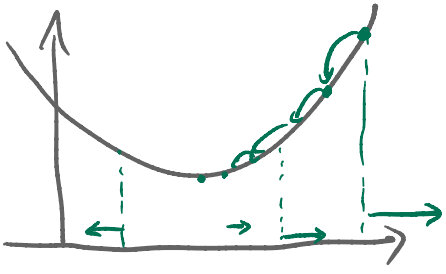
→ PICK h_1^*

h_2^* IS OVERFITTING

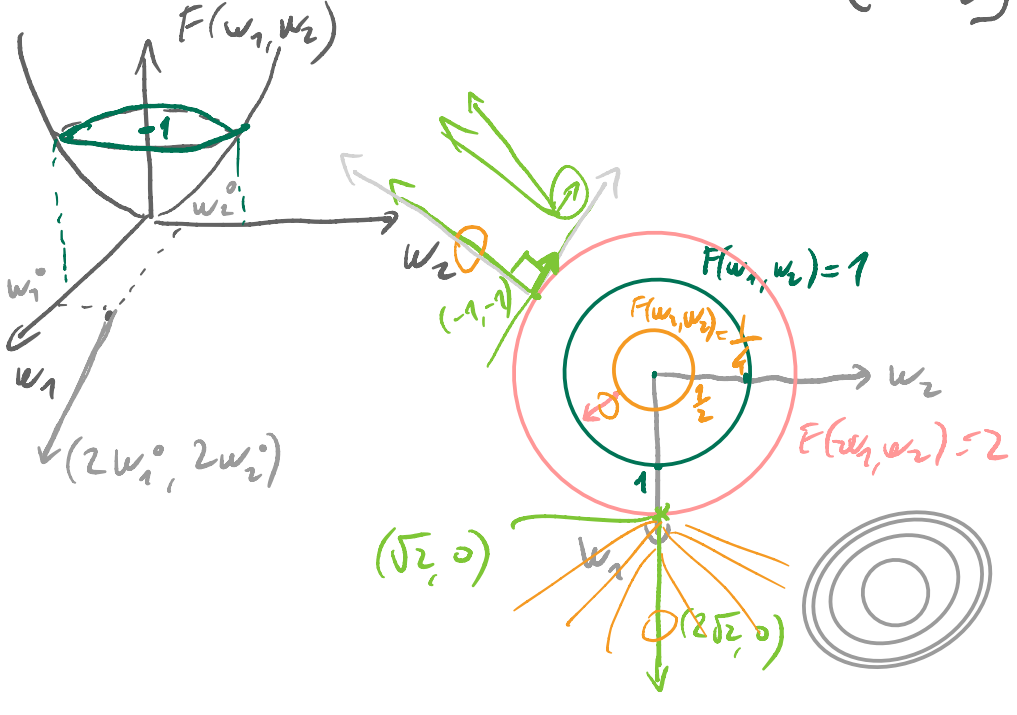
→ PICK h_2^*

h_1^* IS UNDERFITTING

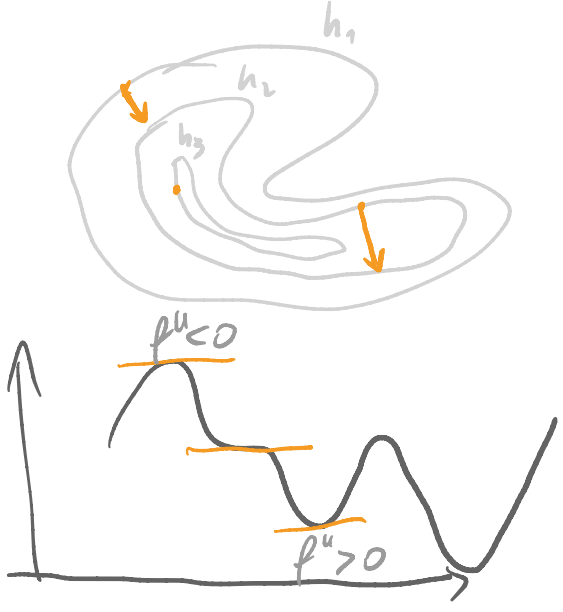
$$\nabla F(w) = \begin{bmatrix} \frac{\partial F}{\partial w_1} \\ \vdots \\ \frac{\partial F}{\partial w_n} \end{bmatrix}$$



$$F(w_1, w_2) = w_1^2 + w_2^2 \qquad \nabla F = \begin{bmatrix} 2w_1 \\ 2w_2 \end{bmatrix}$$



$h_3 > h_2 > h_1$



$f' \longrightarrow \nabla F$

$f'' \longrightarrow$ Hessian of $F \equiv H_F$

IS A MATRIX

If $F: \mathbb{R}^n \to \mathbb{R}$ $\quad H_F$ is a $n \times n$ MATRIX

$$(H_F)_{ij} = \frac{\partial}{\partial w_i} \frac{\partial}{\partial w_j} F$$

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

$$\begin{bmatrix} \frac{\partial^2 F}{\partial w_1^2} & \frac{\partial}{\partial w_1}\frac{\partial F}{\partial w_2} \\ \frac{\partial}{\partial w_2}\frac{\partial}{\partial w_1}F & \frac{\partial^2}{\partial w_2^2}F \end{bmatrix} =$$

$$= \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

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

$f'' > 0 \longleftrightarrow H_F$ is a positive definite matrix.

$f'' \geq 0 \longleftrightarrow H_F$ is positive semidefinite matrix.

A is positive semidefinite if $\forall x \in \mathbb{R}^n \quad x^T A x \geq 0$

$A \in \mathbb{R}^{1\times 1} \quad x^T A x = x^2 A \geq 0 \Rightarrow A \geq 0$

A positive definite

$\forall x \in \mathbb{R}^n - \{\underline{0}\} \quad x^T A x > 0$

if $\bar{w}$: $\nabla F(\bar{w}) = 0$ & $H_F(\bar{w})$ is POSITIVE DEFINITE

$\Rightarrow$ $\bar{w}$ IS A LOCAL MINIMUM

H_F $\quad F''(w)$ $\quad \nabla^2 F$

A is diagonalizable, if $\exists$ V (invertible) and Λ diagonal, such that

$A = V \Lambda V^{-1}$

Λ is diagonal

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \lambda_n \end{pmatrix}$$

λ_i eigenvalues

IF $\forall_i \lambda_i \geq 0 \Rightarrow A$ is positive semidefinite

IF $\forall_i \lambda_i > 0 \Rightarrow A$ is positive definite

---

$A = A^T$ (A is symmetric)

$\Rightarrow A$ is diagonalizable

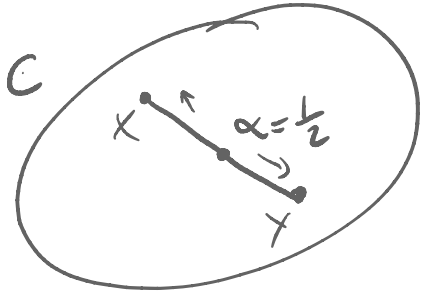
and $V^{-1} = V^T$ V is orthonormal

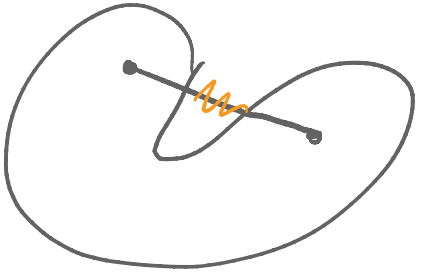
$$A = V \Lambda V^{-1} = V \Lambda V^T$$

$$V^{-1} = V^T \Rightarrow V^T V = I$$

CONVEXITY

CONVEX SET





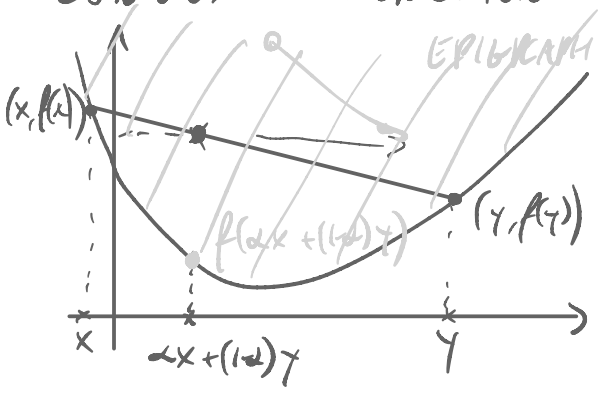
C is a convex set if

$\forall\ x, y \in C$

$\forall\ \alpha \in [0,1]$

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

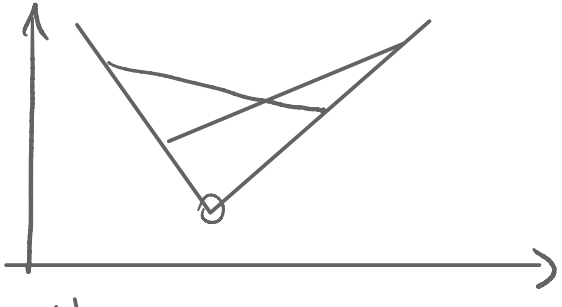
CONVEX FUNCTION

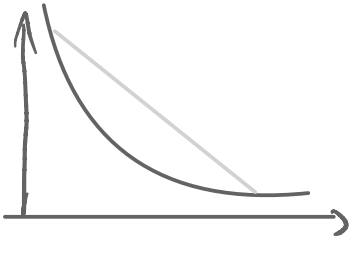


$f: B \to \mathbb{R}$ $\quad B \subset \mathbb{R}^K$ B convex

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

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



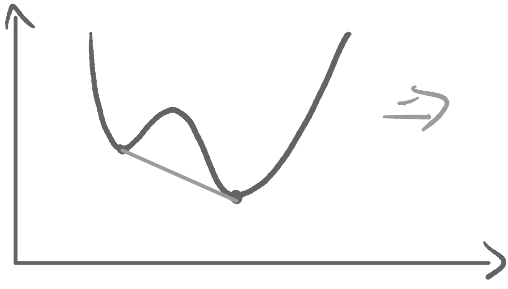


$\frac{1}{x}$ over $\mathbb{R}^+$

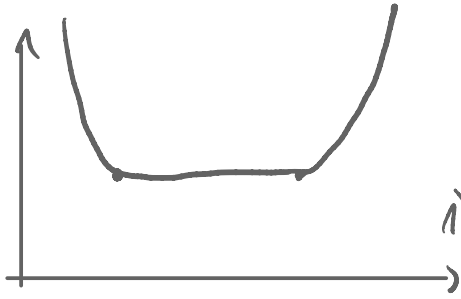
CONVEX OPTIMIZATION

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &x \in B \end{aligned}$$

B is convex f is convex

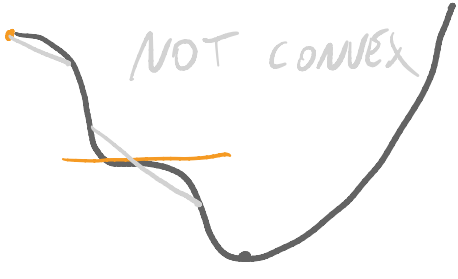


$\Rightarrow$ NO CONVEX

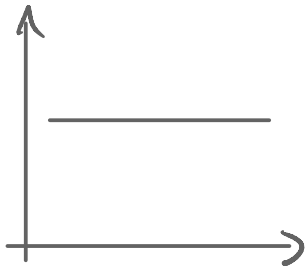


FOR CONVEX OPT PROBLEMS

1) A LOCAL MINIMIZER IS ALSO A GLOBAL ONE



2) A STATIONARY POINT ($\nabla f = 0$) IS A GLOBAL MINIMIZER



CLASSIFICATION PROBLEM

0-1 loss

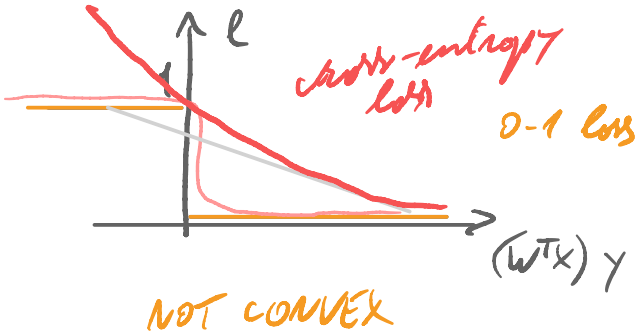
$$\ell(h, (x, y)) = \mathbb{1}_{h(x) \neq y} = \begin{cases} 1 & \text{if } h(x) \neq y \\ 0 & \text{otherwise} \end{cases}$$

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

$$\ell(w, (x, y)) = \mathbb{1}_{wx \neq y}$$

CROSS-ENTROPY LOSS

$$\ell(w, (x, y)) = -y \ln wx - (1-y) \ln(1-wx)$$



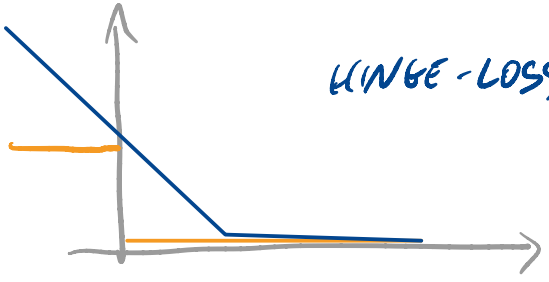
NOT CONVEX

class 1 $\quad y = +1$

class -1 $\quad y = -1$

$$h(x) = \operatorname{sign} w^T x$$

$w^T x > 0$	$y = 1$	$\ell = 0$
$w^T x < 0$	$y = -1$	$\ell = 0$
$w^T x > 0$	$y = -1$	$\ell = 1$
$w^T x < 0$	$y = 1$	$\ell = 1$



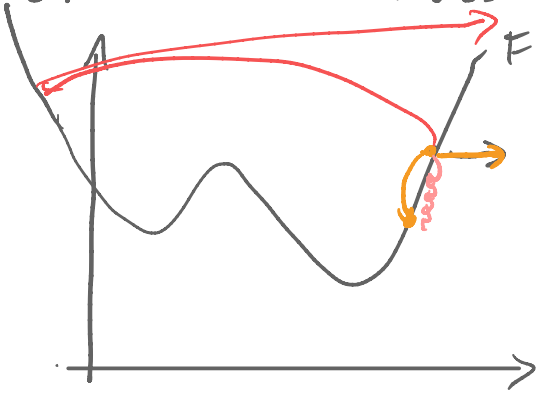
YOUR "NATURAL" LOSS ℓ IS NON-CONVEX

YOU CONSIDER AN ALTERNATIVE "SURROGATE" LOSS $\tilde{\ell}$

1) $\tilde{\ell} \geq \ell$

2) $\tilde{\ell}$ IS CONVEX

GRADIENT METHODS



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

α IS TOO LARGE

1) PICK AN INITIAL GUESS w_0

2) FOR $k = 1$ TO K

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

$$L_T(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, (x_i, y_i))$$

$$T = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$$

$$h : X \longrightarrow Y$$

EACH "h" WILL BE IDENTIFIED BY A VECTOR OF PARAMETERS $w \in \mathbb{R}^n$

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

$L_T(h)$

$$\underbrace{L_T(w)}_{F(w)} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell(h_w, (x_i, y_i))}_{f(w,i)}$$

$$\underset{w}{\text{minimize}} \quad F(w) = \frac{1}{n}\sum_{i=1}^{n} f(w,i)$$

w_0

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

$$= w_k - \alpha_k \nabla \left(\frac{1}{n}\sum_{i=1}^{n} f(w_k, i)\right) =$$

$$= w_k - \alpha_k \frac{1}{n}\sum_{i=1}^{n} \nabla f(w_k, i)$$

FULL-BATCH GRADIENT DESCENT

STOCHASTIC GRADIENT DESCENT

START FROM w_0

FOR $k = 0$ TO $K-1$

- PICK A RANDOM POINT ξ_k IN THE DATASET
- COMPUTE

$$w_{k+1} = w_k - \alpha_k \nabla f(w_k, \xi_k)$$

FB — SGD

Mini batch

PICK w_0

FOR $k = 0$ TO $K-1$

PICK m POINTS FROM T AT RANDOM $\xi_{k,1}\ \xi_{k,2}, \dots\ \xi_{k,m}$

$$w_{k+1} = w_k - \alpha_k \frac{1}{m} \sum_{i=1}^{m} \nabla f(w_k, \xi_{k,i})$$

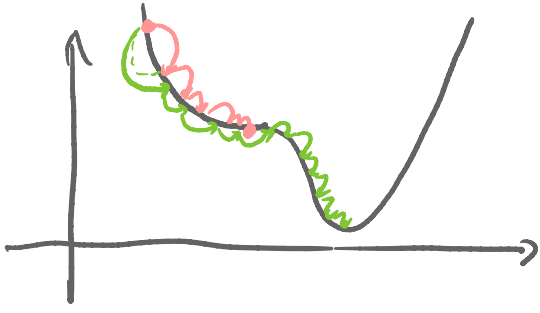
FB $n \times$ MORE EXPENSIVE THAN SGD

IF YOUR GPU IS ABLE TO PROCESS
B POINTS IN PARALLEL

THEN

FB IS ONLY $\frac{n}{B}$ × MORE EXPENSIVE THAN SGD

GPU WITH B=n



NOISE
CAN HELP
CONVERGENCE