Module 5 : 4 Lectures 7/9 – 7/18 $\boxed{\text{HW5}}$

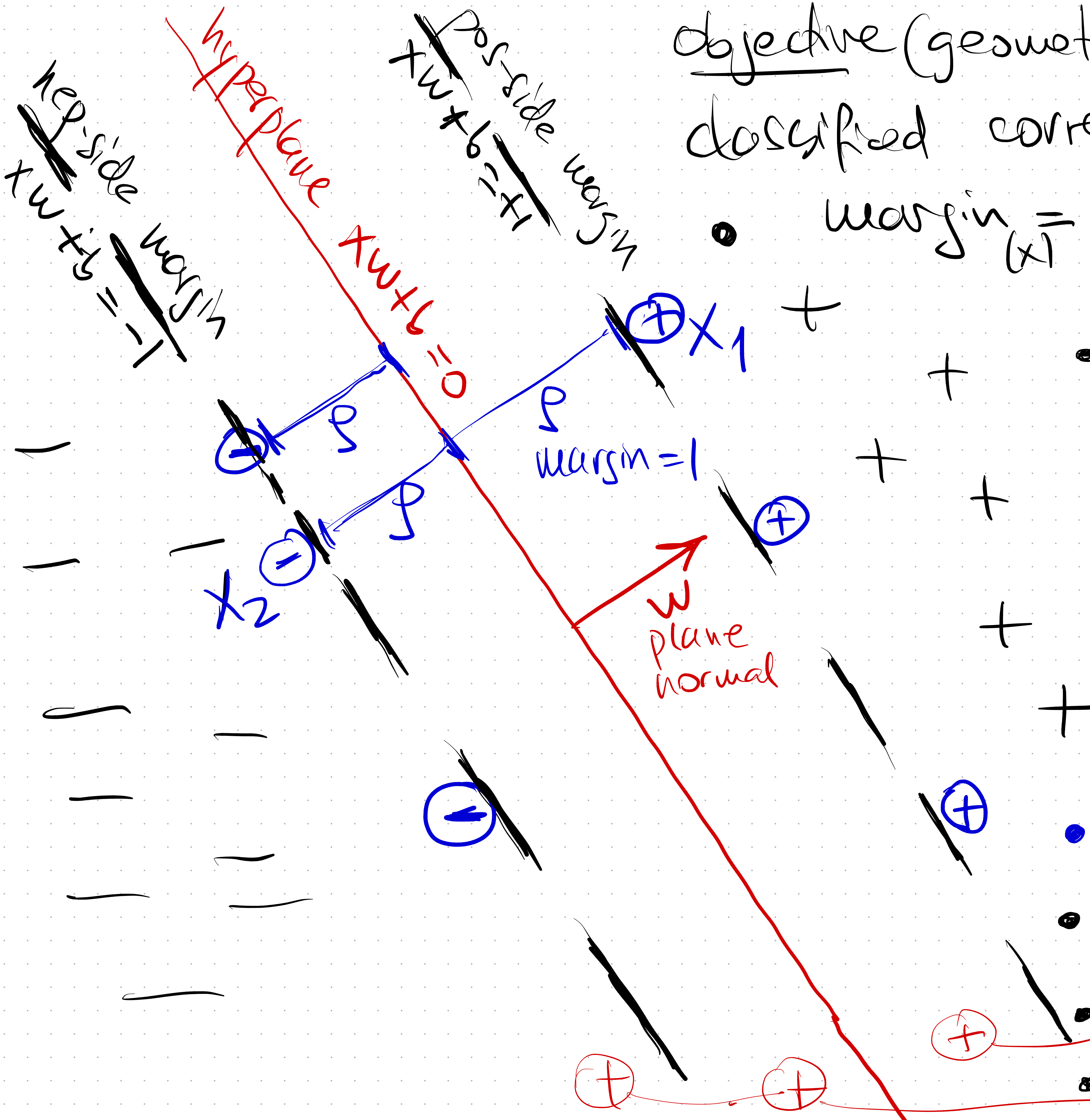
- Kernels
- Kernelization / dualization of ML algorithms
- $\boxed{\text{SVM}}$, Sequential Minim Optimization
- active learning

Support Vector Machine : linear classifier

binary labels $y \in \{\pm 1\}$

$$h(x) = xw + \underbrace{b}_{\text{bias}}$$

$\boxed{\text{objective}}$: increase geometrical margin to closest points



objective (geometry): want all points classified correctly

- $\text{margin}_{(x)} = (xw+b)\cdot \boxed{y} > 0$ (label: y)

CONFIDENCE

- margin far from plane

$\text{margin}(x) \geq 1$

(both sides)

blue: points closest to margin $=1$

points type

- support $(xw+b)y = 1$
- loose (correct) $(xw+b)y > 1$
- tight (correct) $1 > (xw+b)y > 0$
- mistakes $(xw+b)y < 0$

x_2: $x_2 w + b = \boxed{-1}$
x_1: $x_1 w + b = \boxed{+1}$

fix scale

$\Rightarrow (x_1 - x_2) w = 1 - (-1) = 2$

geometry: $2\rho = \|x_1 - x_2\|$

$\Rightarrow \|(x_1 - x_2) w\| = 2$

$2\rho \cdot \|w\| = 2$

$\rho = \frac{1}{\|w\|}$

ISOLATE geom for support vectors

want margin ρ (to support vectors) as high as possible and equal on both sides.

PRIMAL (w)

SVM problem: find separation params (w, b) OBJ: Quad

(constrained optimization)

$\max \rho \iff \min \frac{1}{2}\|w\| \iff \boxed{\min \frac{1}{2}\|w\|^2}$

subject to: constraints = linear

$\boxed{(x_i w + b) y_i \geq 1}$ for all points $(x_i, \boxed{y_i})$

+1 or −1

all points have margin ≥ 1

Support vectors = points with margin ≤ 1

$(xw + b) = 1$

Solution part 1 : Constrained optimization $\Rightarrow$ Lagrangian Multiplier

$$L(w,b,\alpha) = \underbrace{\frac{1}{2}\|w\|^2}_{\text{OBJ}} - \sum_{i=1}^{N} \boxed{\alpha_i} \underbrace{\left[y_i(x_i w + b) - 1\right]}_{\text{constraint}}$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N} \alpha_i y_i x_i^T \overset{\text{want}}{=} 0$$

$$\Rightarrow \underset{\text{primal}}{\boxed{w}} = \sum_{i=1}^{N} \underset{\text{dual}}{\boxed{\alpha_i}} y_i x_i^T$$

α

= linear combination of datapoints with coef $(\alpha_i)_{i=1:N}$

DUALITY

$$\frac{\partial L}{\partial b} = 0 - \sum_{i=1}^{N} \alpha_i y_i \overset{\text{want}}{=} 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \Rightarrow \underset{\text{+ force pushing}}{\boxed{\sum_{y_i=+1} \alpha_i}} = \underset{\text{− force pushing}}{\boxed{\sum_{y_j=-1} \alpha_j}}$$

Rewrite $L(w)$ with $w = \sum \alpha_i y_i x_i^T$

$$L = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i (y_i(x_i w + b) - 1)$$

$$= \frac{1}{2} w^T w - \sum \alpha_i y_i x_i \boxed{w} - \sum \alpha_i y_i b + \sum \alpha_i$$

$$= \boxed{\frac{1}{2}} \underbrace{\left(\sum_i x_i y_i x_i^T\right)^T \left(\sum_j x_j y_j x_j^T\right)}_{\text{same}} - \underbrace{\left(\sum \alpha_i y_i x_i\right)\left(\sum \alpha_j y_j x_j\right)}_{\text{same}} + \sum \alpha_i \quad - \sum \alpha_i y_i b$$

$$= -\frac{1}{2}\left(\sum x_i y_i x_i^T\right)^T \left(\sum_j x_j y_j x_j^T\right) - b \boxed{\sum \alpha_i y_i} + \sum \alpha_i$$

$\boxed{\sum \alpha_i y_i} = 0$

all pairs

$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j x_i x_j^T$$

dual $L(\,)$

$(\alpha_i)_{i=1:N}$

Dual SUM problem

Max $\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \boxed{x_i x_j^T}$ (K_{ij})

subject to $\alpha_i \geq 0$; $\boxed{\sum_{i=1}^{N} \alpha_i y_i = 0}$ linear constraint

KKT conditions for duality ≡ primal problem Sol = saddle point

1) $\min L(w) \equiv \max L(\alpha)$ with approp. constraints

2) Lag Multipliers $\alpha \geq 0$

3) at solution (saddle point) $\frac{\partial L}{\partial w} = 0$

force > 0
⇒ support vectors

4) constraints tight (equality $y_i(xw+b) - 1 = 0 \Rightarrow \boxed{\alpha_i > 0}$

5) constraints not tight ($y_i(xw+b) > 1$) $\Rightarrow \boxed{\alpha_i = 0}$ no force

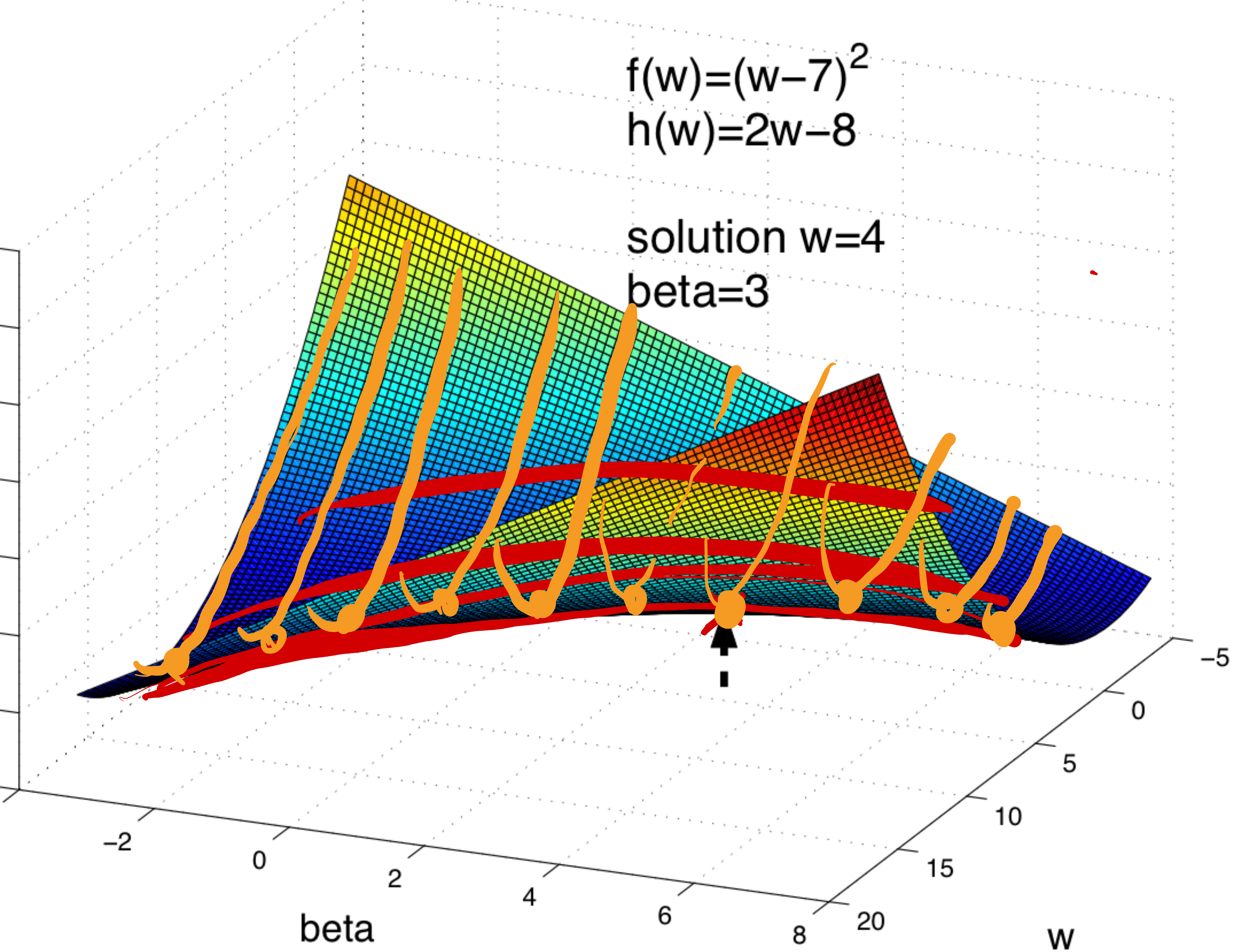


Figure 13: Saddle Point

= min (max (sections Red)) = max (min (sections Orange))

<u>TO DO</u>: 1) deal with points inside margin (mistakes)
2) solve dual pb in α
3) pipeline for train / prediction

<u>PRIMAL PB</u>

$$\min \frac{1}{2}\|w\|^2 + \boxed{C \sum_{i=1}^{N} \xi_i}$$

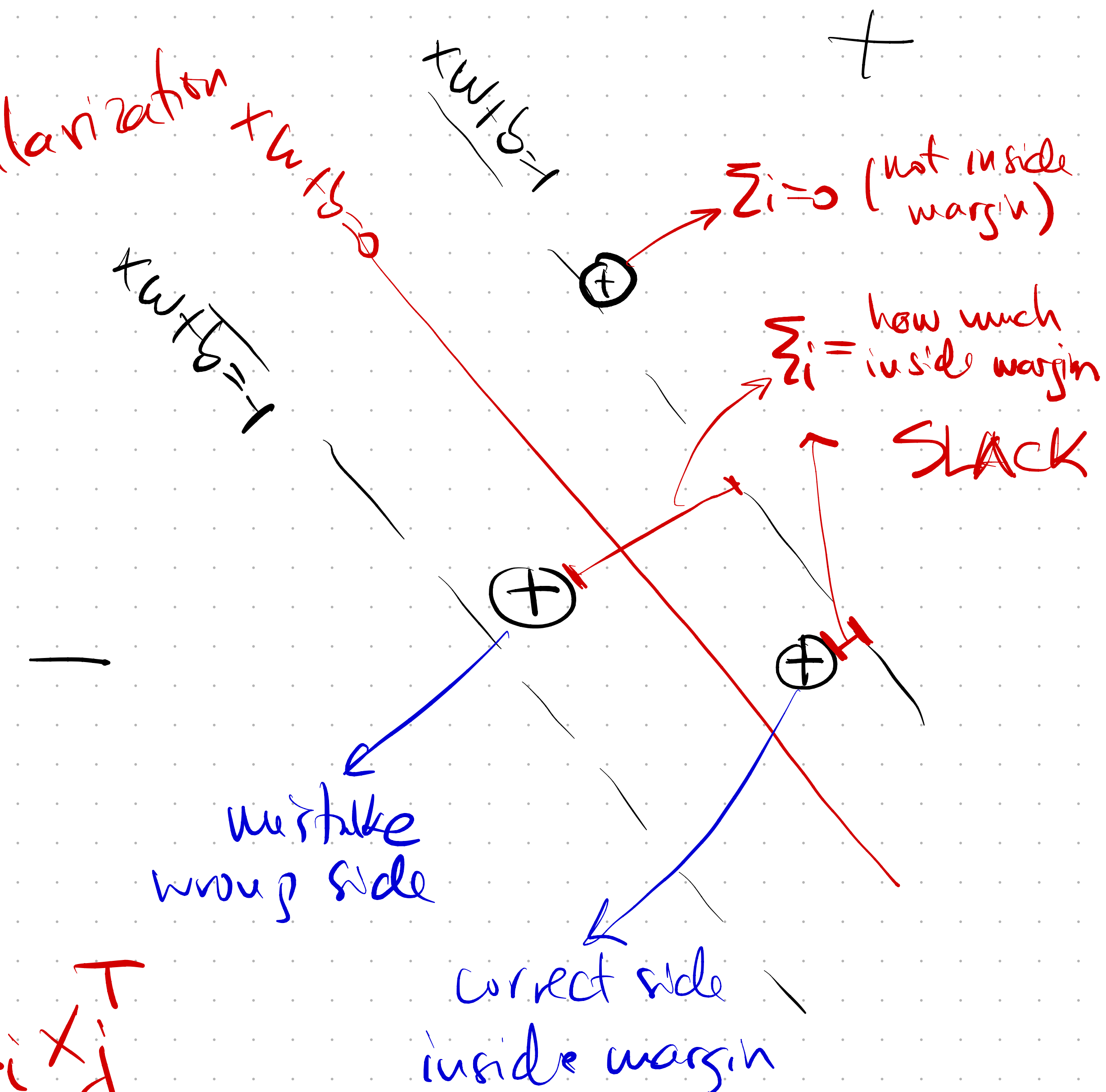
~ L1 regularization

subject to

$$y_i(xw+b) \geq 1 - \xi_i$$

allowance inside margin

$$\xi_i \geq 0$$



Dual SVM Problem with SLACK ξ_i

$$\underline{\max} \sum \alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j x_i x_j^T$$

<u>subject to</u>

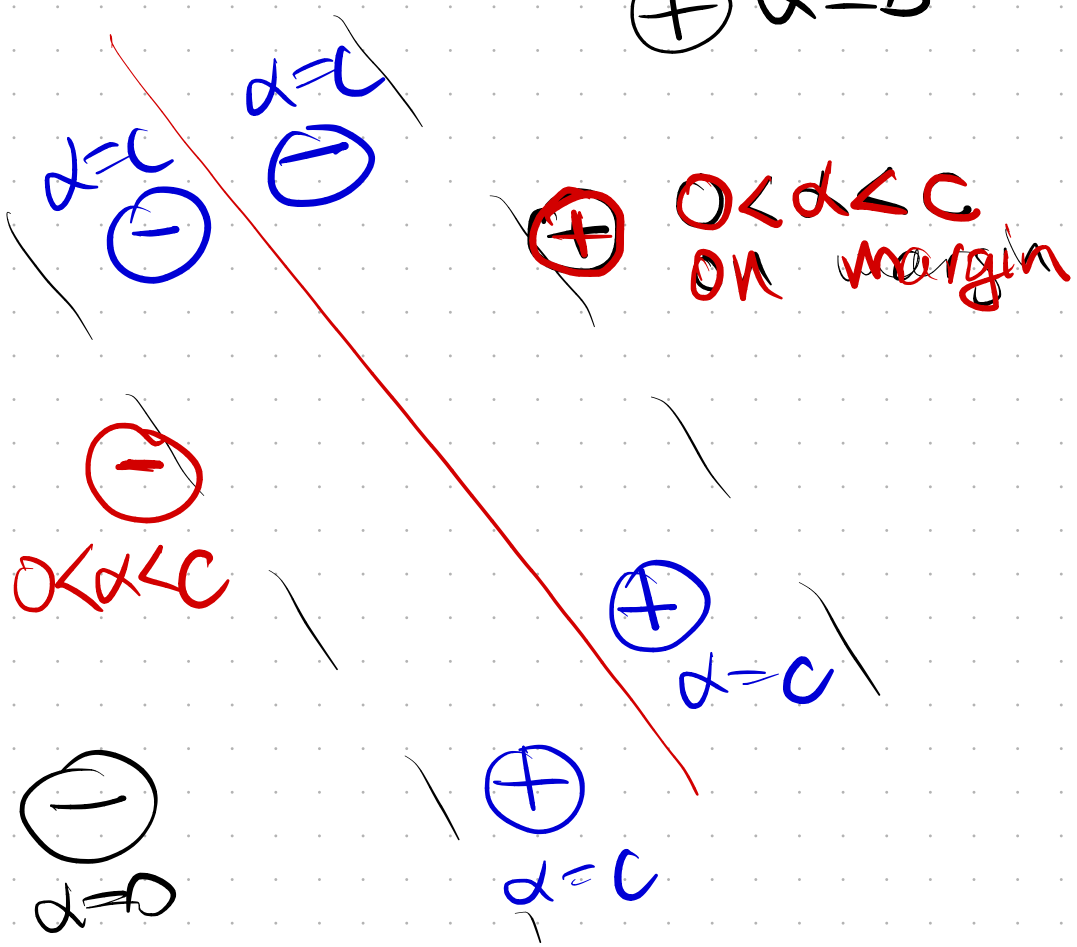
$$0 \leq \boxed{\alpha_i \leq C}$$

due to slack constraints

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

from KKT: far from margin

$\oplus$ $\alpha=0$



three types of points

- far from margin (correct) $\alpha=0$
- on margin $0<\alpha<C$
- inside margin (possibly incorrect) $\alpha=C$

How to solve the dual pb in α variables? $(\alpha_i)_{i=1:N}$ N variables

$\underline{\max}$ $\sum \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j x_i x_j^T$

subject to: $0 \leq \alpha_i \leq C$; $\sum \alpha_i y_i = 0$

Quadratic Solver? possible, efficient alg - math difficult - not scalable

Empirical optimization: $\boxed{S}$equential $\boxed{M}$inimal $\boxed{O}$ptimization

much easier, faster, not nec. optimal

- optimize 2 "α" at a time (α_i, α_j)
- repeat process until convergence?
- pick at each round (α_i, α_j) pair by bigger-diff criteria

random (α_i, α_j) possible, much slower

- $\underline{\text{why not}}$ one α at a time?

select $\alpha_i \Rightarrow$ change its value while other $\alpha_{j \neq i}$ fixed

not possible due to constraint $\sum \alpha_i y_i = 0$

if all $\alpha_{j \neq i}$ fixed $\Rightarrow \alpha_i y_i = - \sum_{j \neq i} \alpha_j y_j \Rightarrow \alpha_i =$ fixed.

SMO Algorithm. (round) for many rounds

- choose pair (α_1, α_2) ?
- OBJ $\Rightarrow$ Quad equation (α_1, α_2)

$$\sum \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j x_i x_j^T = \dots$$

MAX

$$\dots = \alpha_1 + \alpha_2 - \frac{1}{2} K_{11} \alpha_1^2 - \frac{1}{2} K_{22} \alpha_2^2 - y_1 y_2 K_{12} \alpha_1 \alpha_2 - y_1 \alpha_1 \boxed{v_1}^{?} - y_2 \alpha_2 \boxed{v_2}^{?} + \text{const}$$

subject to $\alpha_1 y_1 + \alpha_2 y_2 = - \sum_{j \geq 3} \alpha_j y_j$ linear constraint

$\rightarrow$ Quad equation (α_1, α_2)
subject to linear constraint $\alpha_1 y_1 + \alpha_2 y_2 = \text{CONST}$

Train in dual form $\Rightarrow (\alpha_i)_{i=1:N}$

back to primal pb $\boxed{\underline{w} = \sum \alpha_i x_i y_i}$

$$b_+ = \min_{y_i=+1} x_i w \qquad b_- = \max_{y_j=-1} x_j w$$

avg $\boxed{b = \frac{b_+ + b_-}{2}}$

Z = test point

predict primal form: $\text{pred}(z) = Z \cdot w + b$

predict in dual form

$$\text{pred}(z) = \sum_{i=1}^{N} \alpha_i \boxed{K_{[x_i, z]}} + b$$

Kernel = similarity (x_i, z)