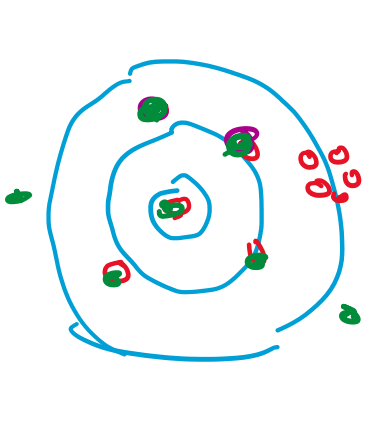
Lecture 17

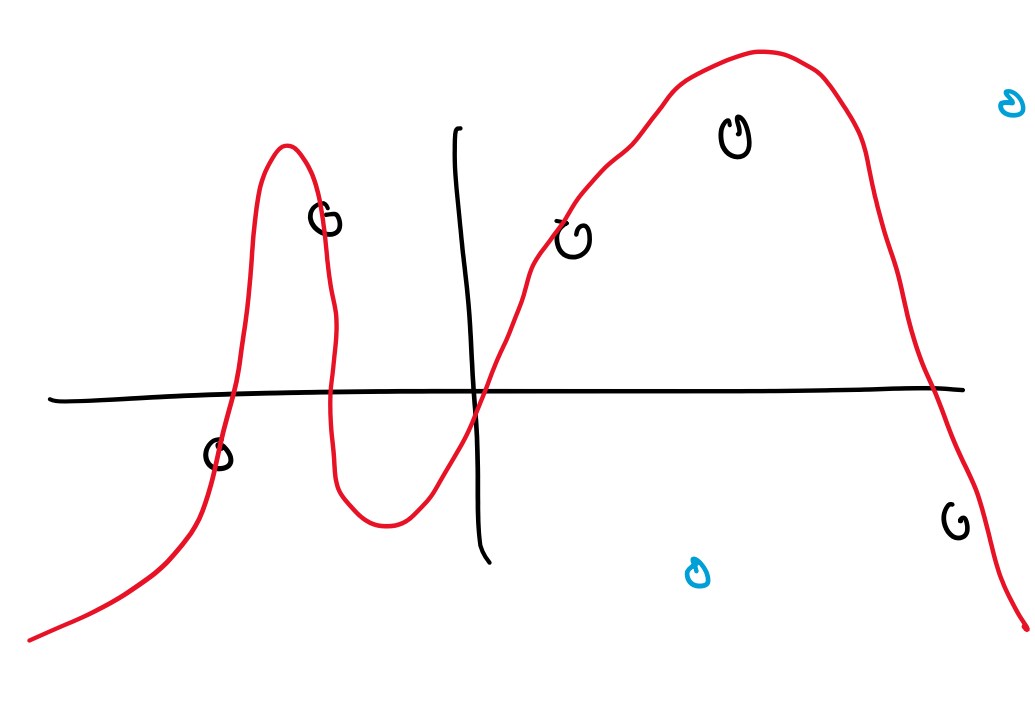
Thursday, November 7, 2024 10:02 AM

Regularization

Regression/ classification

$\hat{y} = f_{\theta}(\vec{x})$ $\theta \equiv$ parameters





Complex models ⇒ high variance

Simple model ⇒ high bias

$$\min_{\theta} \underbrace{\sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i)}_{\text{MSE or CE}} + \alpha \underbrace{\|\theta\|^2}_{\text{penalty}}$$

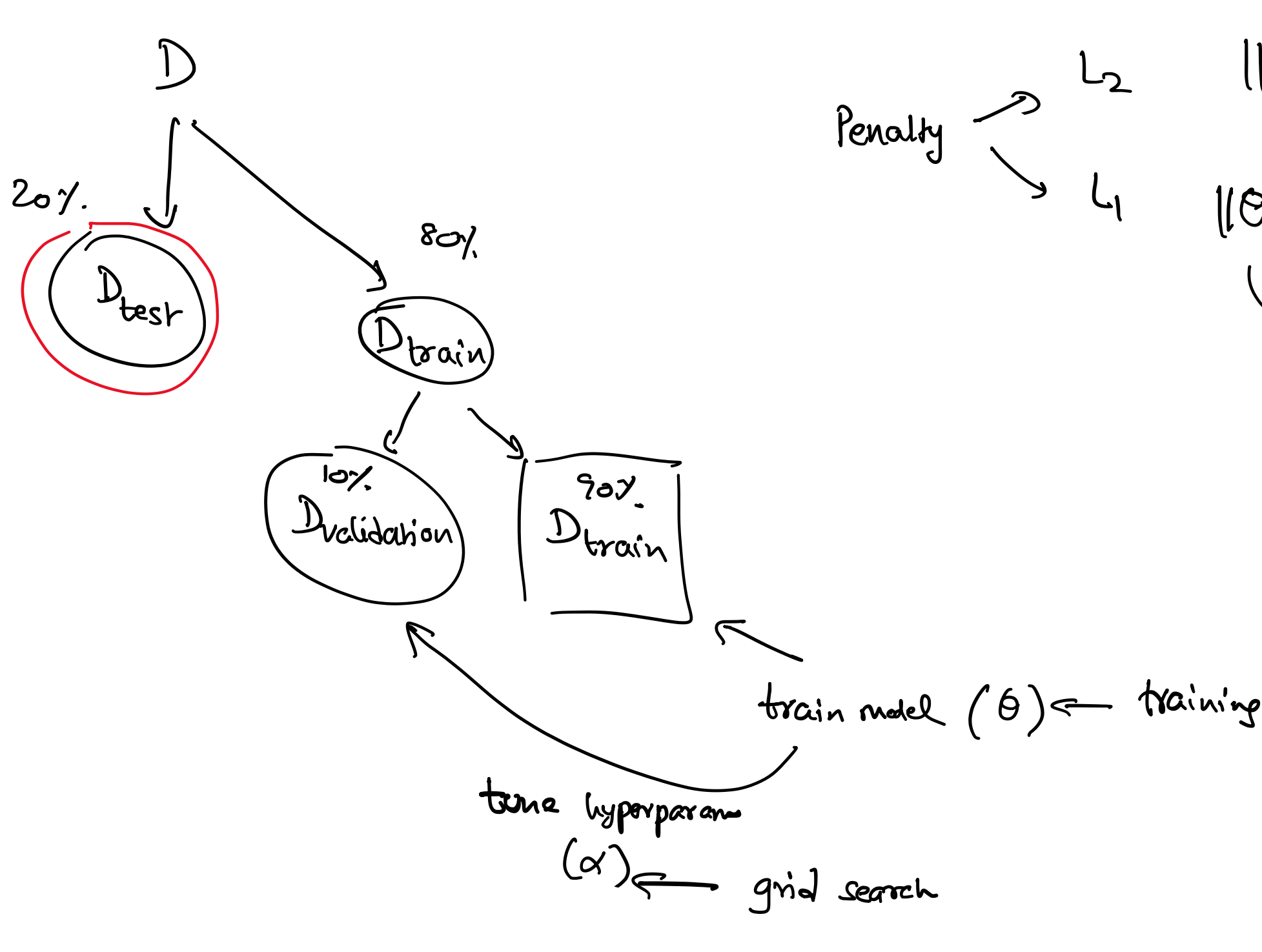
Regularization constant (user chosen – hyperparam)

$\alpha = 0 \Rightarrow$ no regularization

$\alpha > 0$

Penalty → L_2 $\|\theta\|^2$

Penalty → L_1 $\|\theta\|_1$ → $\sum_{i=1}^{d+1} |\theta_i|$



Linear regression

$\hat{y}_i = \tilde{w}^T \tilde{x}_i$ ← augmented data & $\vec{w}$ vector

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \boxed{\sum_{i=1}^{n} (y_i - \tilde{w}^T \tilde{x}_i)^2}$$

$$= \|\underset{\text{response}}{Y} - \underset{\text{Predicted}}{\hat{Y}}\|^2 = \|\underset{n\times 1}{Y} - \underbrace{\underset{n\times(d+1)}{\tilde{D}}\ \underset{(d+1)\times 1}{\tilde{w}}}_{n\times 1}\|^2$$

n-dim

$$\min_{\tilde{w}} J = \underbrace{\|Y - \tilde{D}\tilde{w}\|^2}_{L} + \alpha \underbrace{\|\tilde{w}\|^2}_{\text{regularization penalty}}$$

$\alpha \geq 0$ $\|\tilde{w}\|^2 = \tilde{w}^T\tilde{w}$

$$\frac{\partial J}{\partial \tilde{w}} = -2\tilde{D}^T Y + 2(\tilde{D}^T\tilde{D})\tilde{w} + 2\alpha\tilde{w} = 0$$

$$(\tilde{D}^T\tilde{D} + \alpha I)\tilde{w} = \tilde{D}^T Y$$ $(\lambda_k + \alpha)$

$$\boxed{\tilde{w} = (\tilde{D}^T\tilde{D} + \alpha I)^{-1} \tilde{D}^T Y}$$

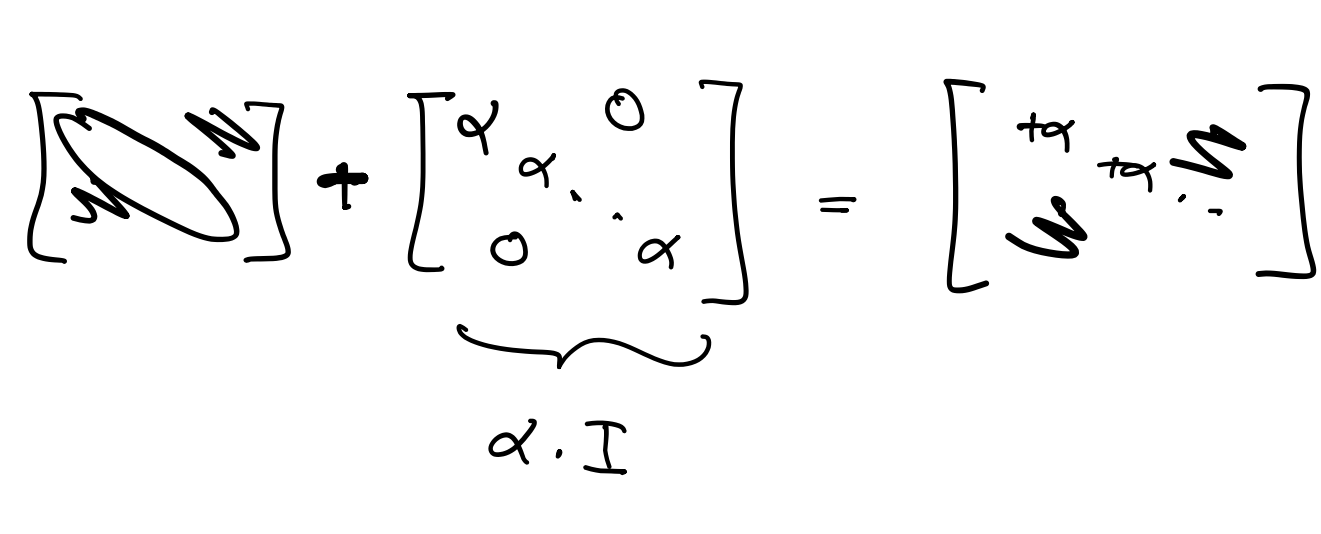
symmetric PSD $(d+1)\times(d+1)$

uncentered scatter matrix

$\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots \geq \lambda_{d+1}$

$\det = \lambda_1 \times \lambda_2 \times \cdots \times \lambda_{d+1}$

$= \prod \lambda_i$



$\alpha = 0$

$(\tilde{D}^T\tilde{D})^{-1}\tilde{D}^T Y$

inverse may not exist

$\lambda_k = 0$

ridge regression

→ regularization by $\alpha\|\tilde{w}\|^2$

Logistic regression is classification

$y_i \in \{C_1, C_2 \ldots C_k\}$ categorical response

real → squared error

prob → cross entropy

Support vector machines

↳ new objective

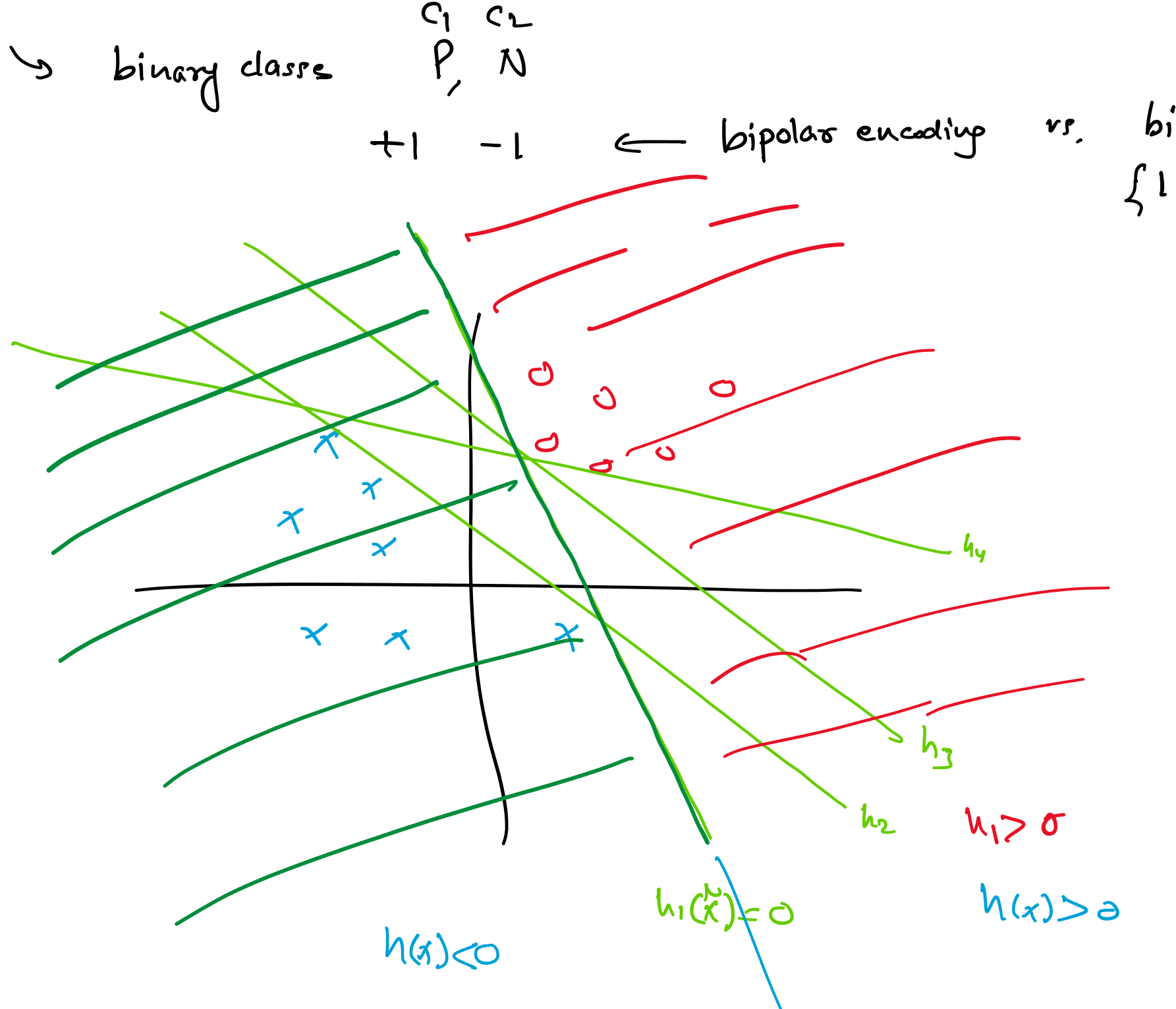
(max margin)

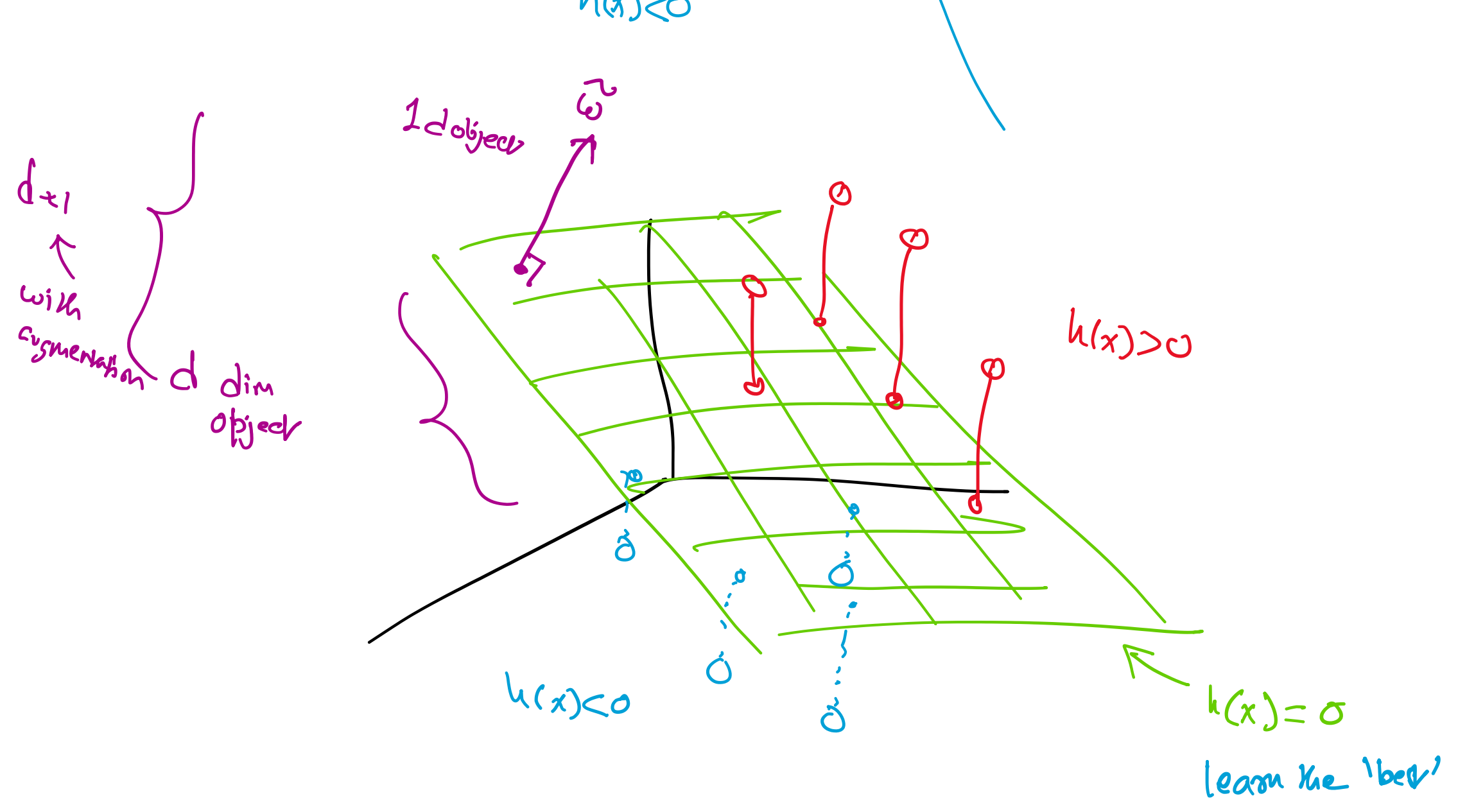
structural

↳ $\tilde{w}$ ← learn a direction.

↳ binary classes $\overset{C_1}{P}$, $\overset{C_2}{N}$

+1 −1 ← bipolar encoding vs. binary {1, 0}

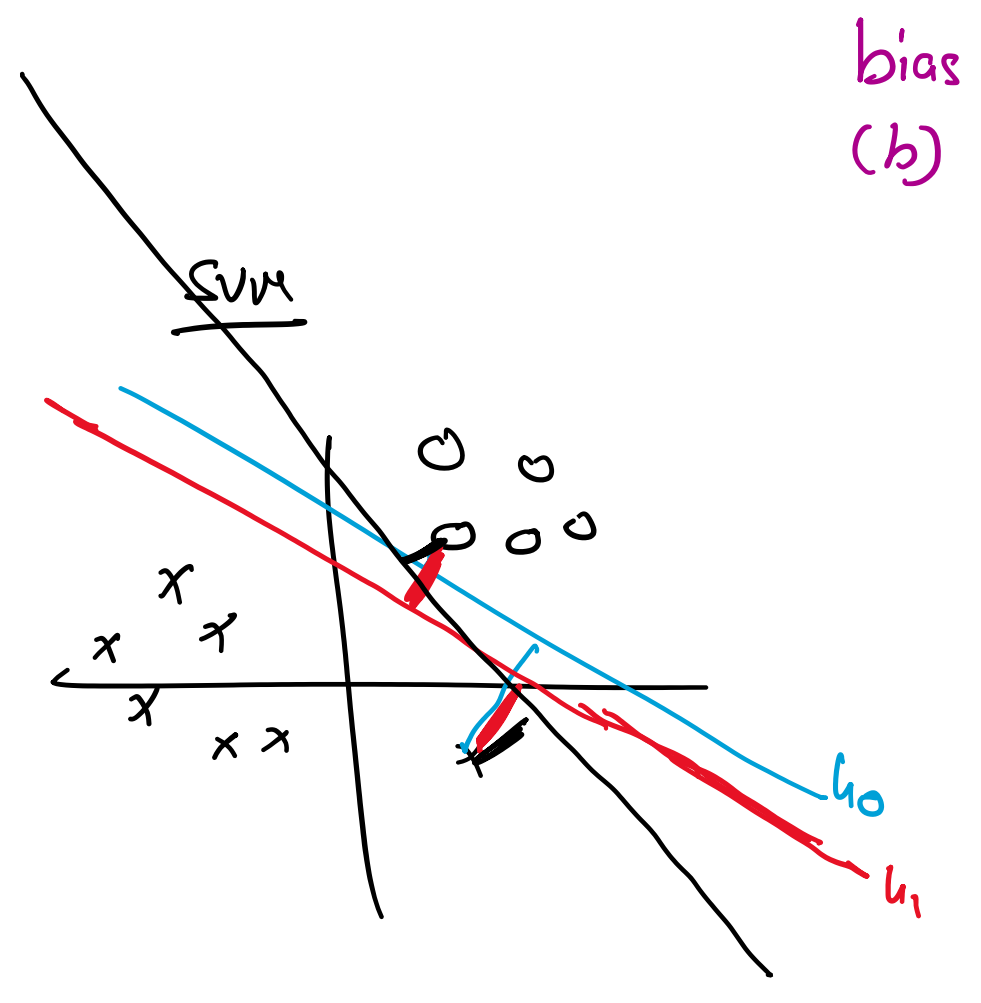




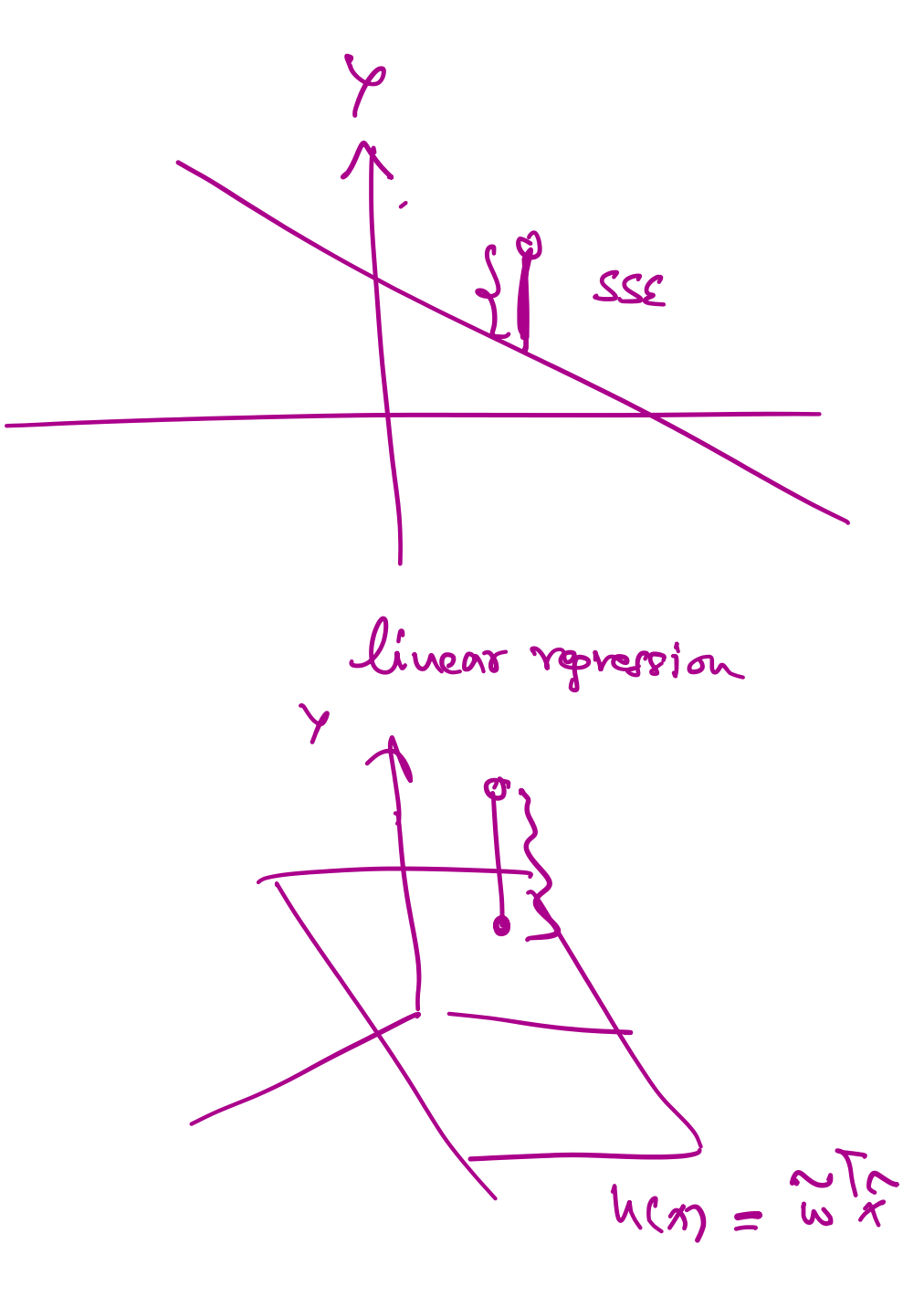
$\tilde{w}$ = normal vector

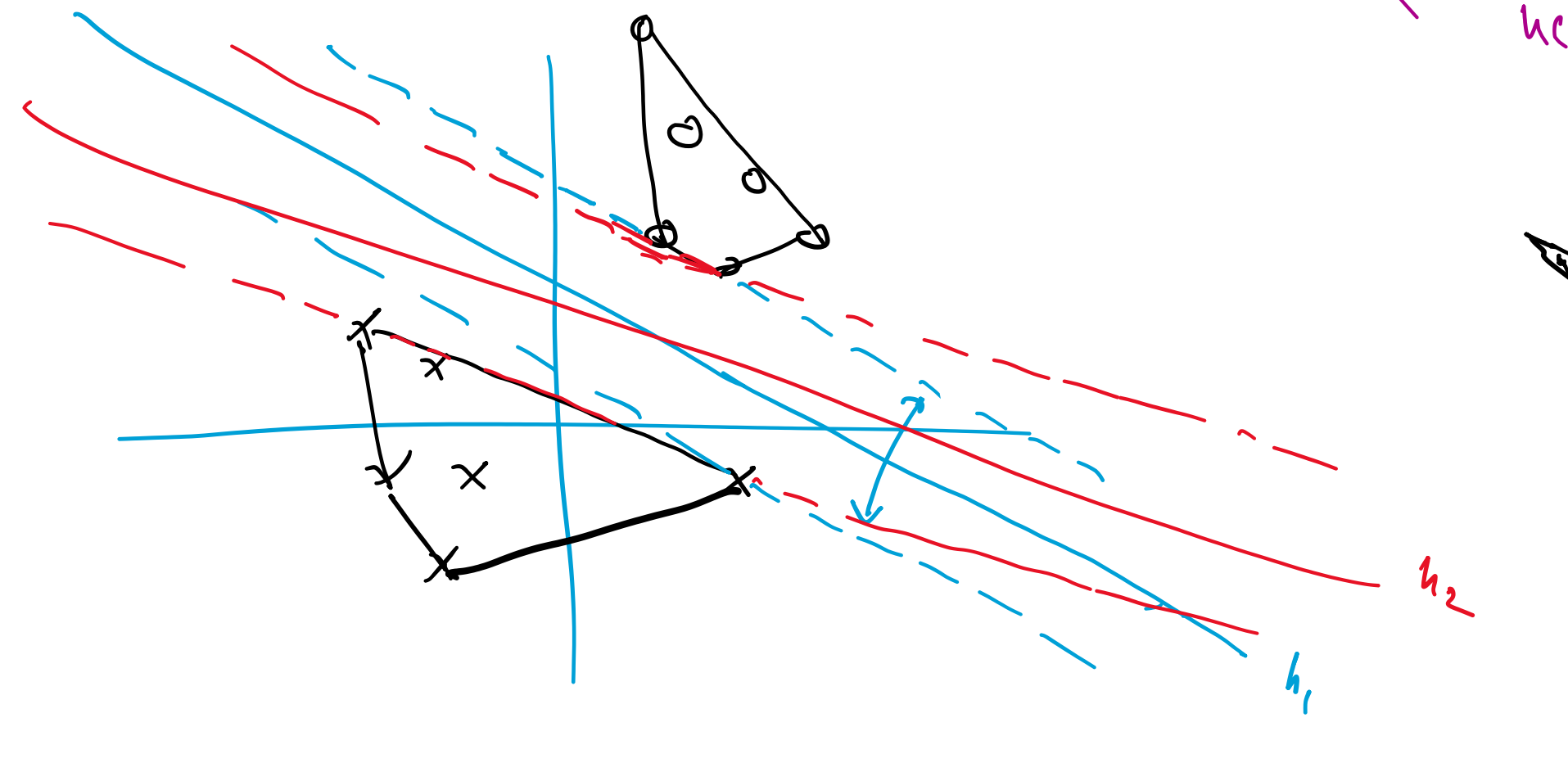
↳ ⊥ ≡ perpendicular

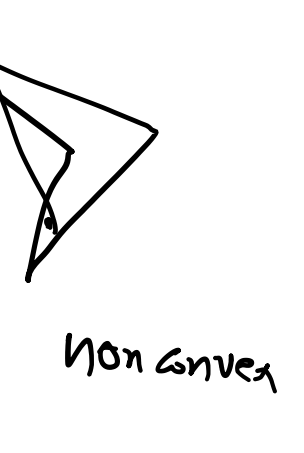
$$h(\tilde{x}) = \tilde{w}^T\tilde{x} = \underset{\text{bias (b)}}{w_0} + \underbrace{w_1 x_1 + w_2 x_2 + \cdots + w_d x_d}$$



Max margin







max margin objective: maximize the distance of a point to the hyperplane (from either class)

margin = distance of the closest point to h(x)

$$\delta^* = \min_{x} \underbrace{d(x, h)}_{\text{distance of } x \text{ to hyperplane } h}$$

Objective: $$\max_{\tilde{w}} \underbrace{\left\{ \min_{x'} d(x, h) \right\}}_{\text{margin for } \tilde{w}}$$

$\vec{x}_i \to y_i$

{+1 −1}

2 classes

$$\hat{y}_i = \begin{cases} +1 & \text{if } h(x_i) \geq 0 \\ -1 & \text{if } h(x_i) < 0 \end{cases}$$

$$h(\vec{x}_i) = \underset{\text{bias (b)}}{w_0} + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$$

$$\tilde{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} \qquad \tilde{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix}$$

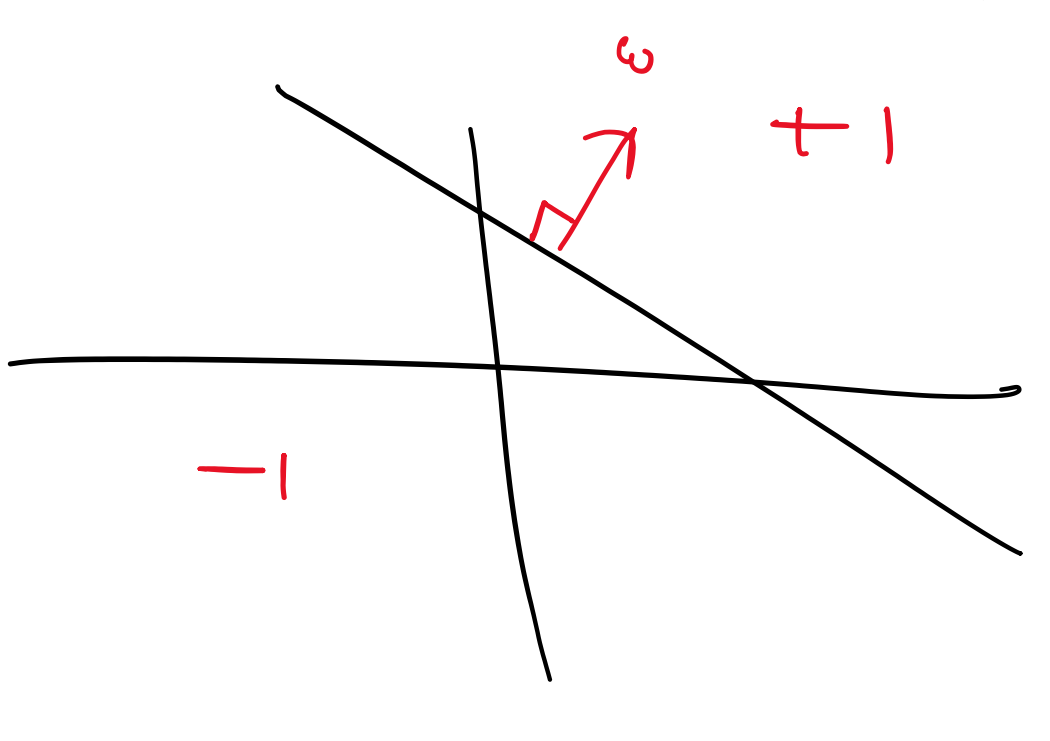
$\tilde{D}$

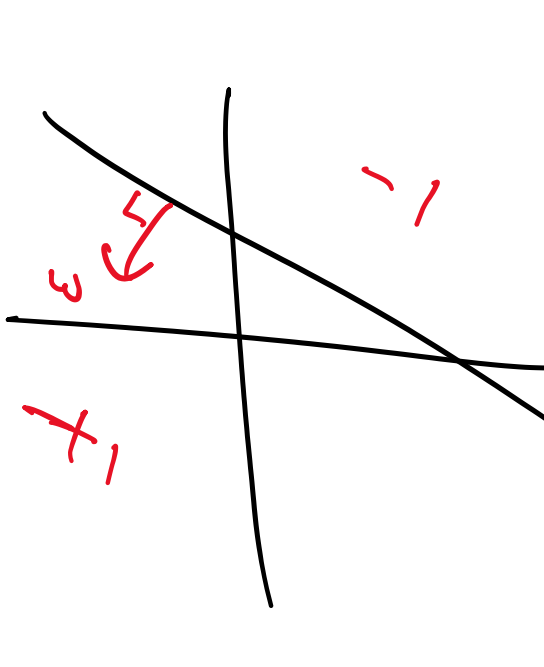
$h(\tilde{x}_i) = \tilde{w}^T\tilde{x}_i$

equation

hyperplane $\equiv \{\tilde{x}_i \mid h(\tilde{x}_i) = 0\}$

$\Rightarrow \tilde{w}^T\tilde{x}_i = 0 \quad \forall x_i \in$ hyperplane

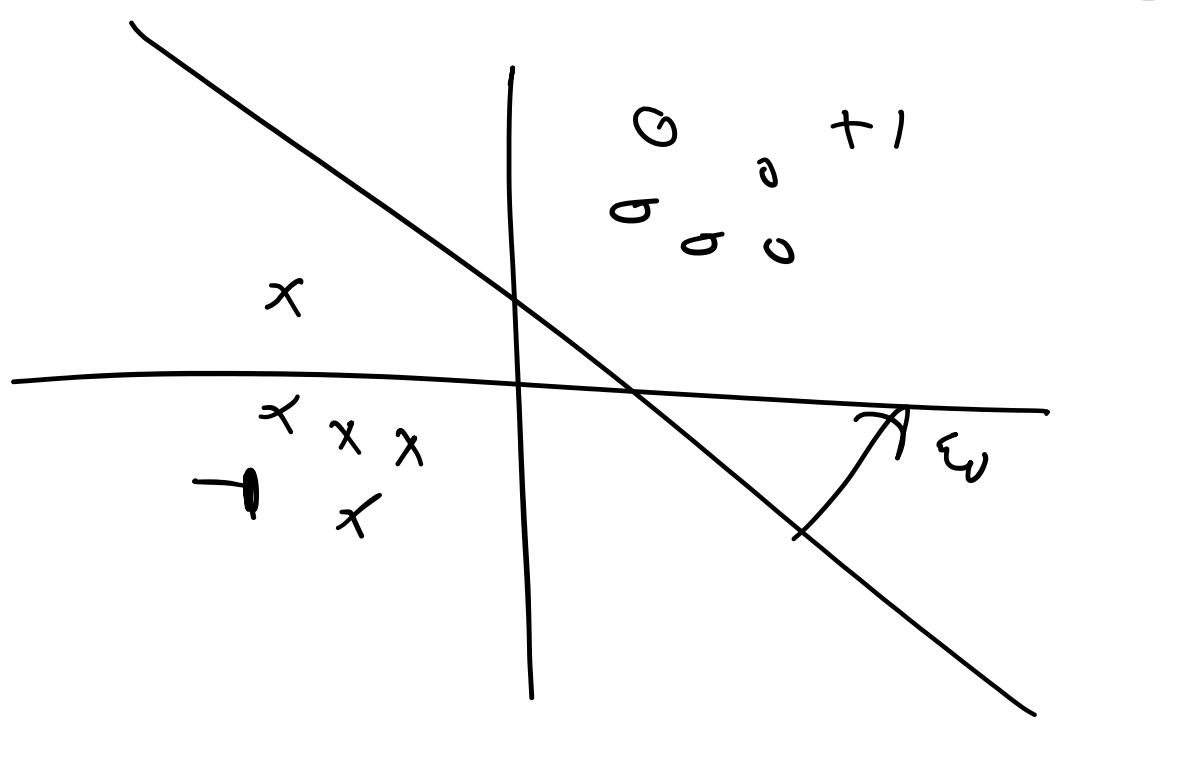


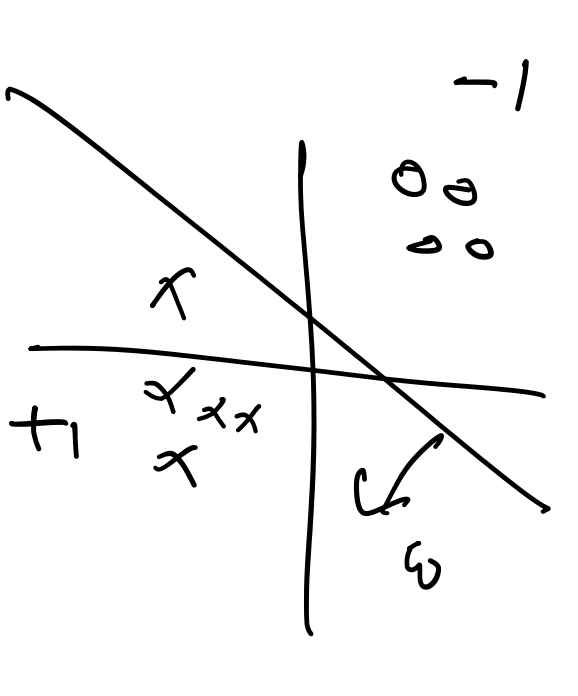


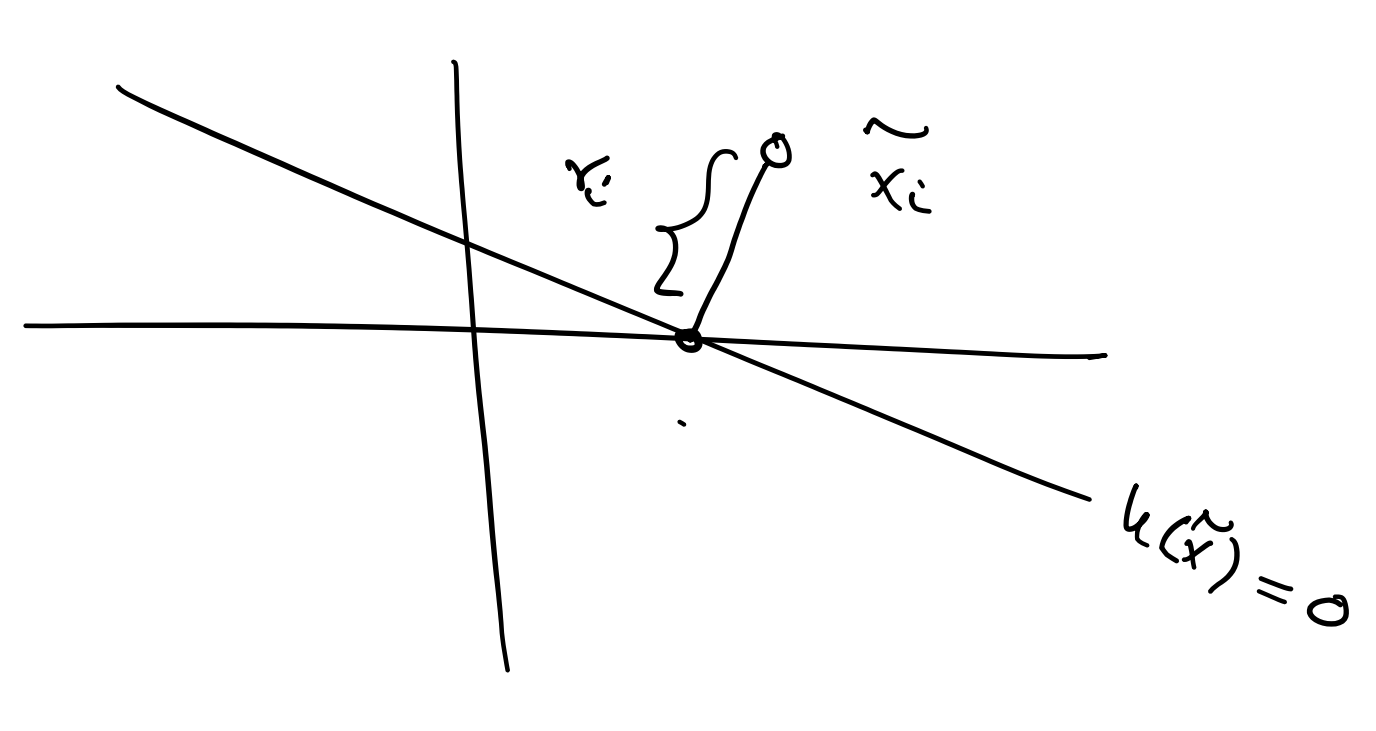
h should respect the class labels

$h(\tilde{x}) > 0$ when $y = +1$

$h(\tilde{x}) < 0$ when $y = -1$







r_i = 'signed' distance $= \dfrac{h(\tilde{x}_i)}{\|\tilde{w}\|} = \dfrac{\tilde{w}^T\tilde{x}_i}{\|\tilde{w}\|}$

δ_i = distance (non-negative) $= \dfrac{y_i h(\tilde{x}_i)}{\|\tilde{w}\|}$ ← $h(\tilde{x}_i) > 0$ if $y_i = +1$; $h(\tilde{x}_i) < 0$ if $y_i = -1$

margin for $h(\tilde{x})$ $\tilde{w}$ $\equiv \delta^* = \min_{\tilde{x}_i} \{\delta_i\}$

$$\delta^* = \min_{\tilde{x}_i} \left\{ \frac{y_i h(\tilde{x}_i)}{\|\tilde{w}\|} \right\}$$

$$\boxed{\max_{\tilde{w}} \text{ margin} = \max_{\tilde{w}} \left\{ \min_{\tilde{x}_i} \left\{ \frac{y_i h(\tilde{x}_i)}{\|\tilde{w}\|} \right\} \right\}}$$

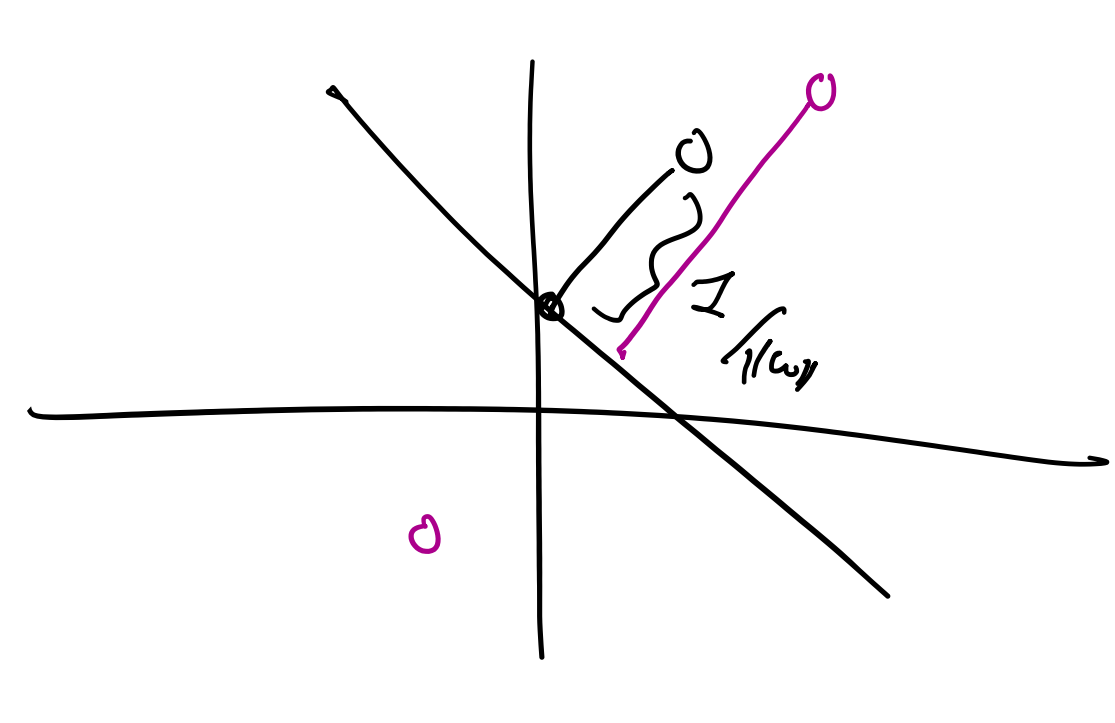
$$\delta_i = \frac{y_i \tilde{w}^T\tilde{x}_i}{\|\tilde{w}\|}$$ ← just make bigger by increasing the scale of $\tilde{w}$

Constrain $\tilde{w}$ such that the closest point is at 'distance' 1

$s(y_i \tilde{w}^T\tilde{x}_i) = 1$

$s = \dfrac{1}{y_i \tilde{w}^T x_i}$

then margin $= \dfrac{1}{\|\tilde{w}\|}$



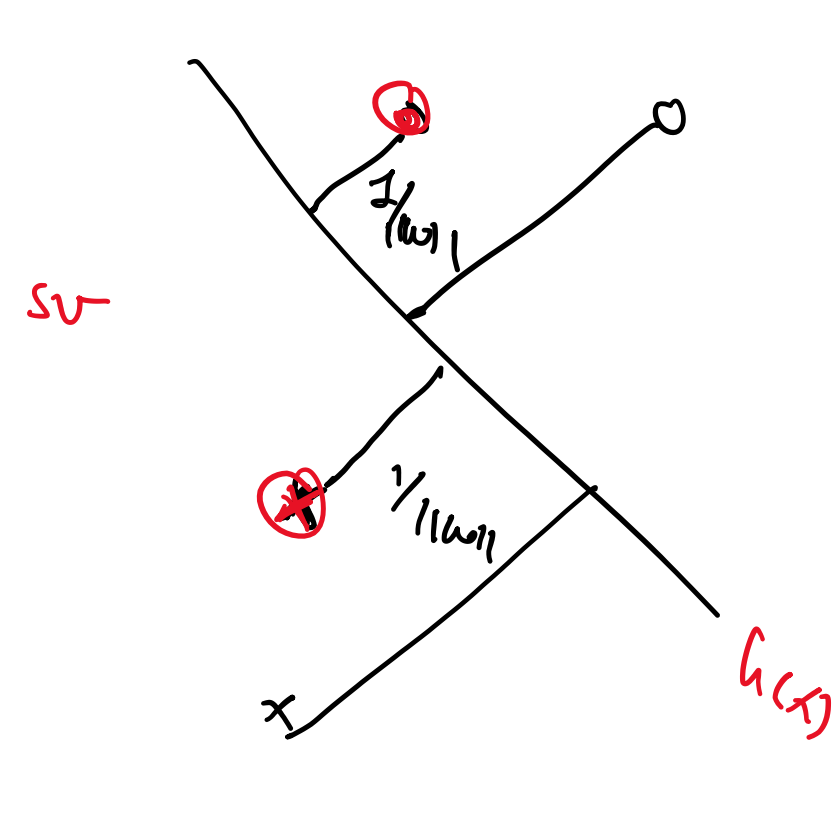
but introduce n new constraints

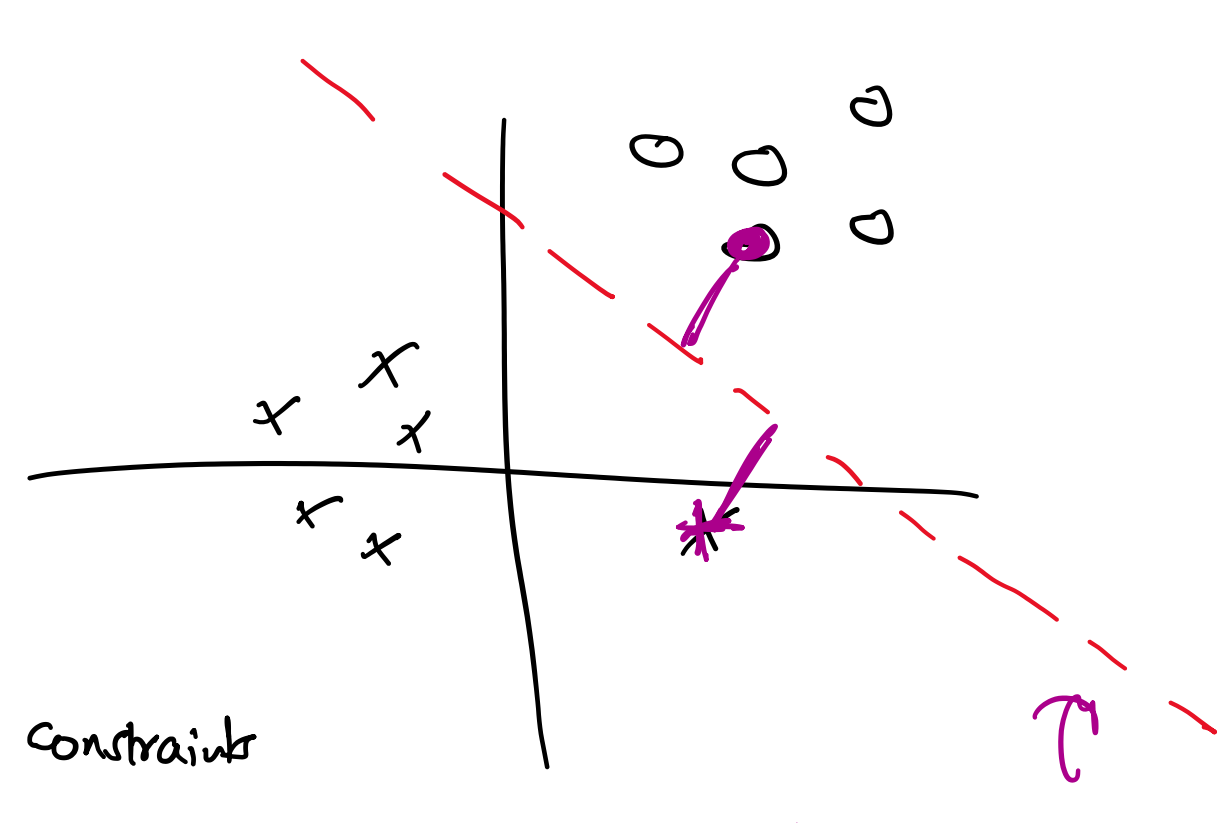
$y_i h(x_i) \geq 1 \quad \forall x_i$

(Since we scale the closest point to be 'distance' 1)

$y_i h(x_i) = 1$ for the closest point

support vectors





h(x) is defined by the support vectors

$$\boxed{\max_{\tilde{w}} \left\{ \frac{1}{\|\tilde{w}\|} \right\}}$$ margin

subject to n constraints

$y_i h(\tilde{x}_i) \geq 1 \quad \forall \tilde{x}_i$

$\Rightarrow \boxed{y_i \tilde{w}^T\tilde{x}_i \geq 1 \quad \forall x_i}$

linear constraints

⇓

$\min_{\tilde{w}} \frac{1}{2}\|\tilde{w}\|^2$

$$\frac{1}{\|\tilde{w}\|} = \frac{1}{\sqrt{\tilde{w}^T\tilde{w}}}$$

SVM objective: $\min_{\tilde{w}} \frac{1}{2}\|\tilde{w}\|^2$ s.t. $\boxed{y_i \tilde{w}^T\tilde{x}_i \geq 1} \ \forall \tilde{x}_i \in D$

introduce Lagrange multiplier $\alpha_i \geq 0$ for each constraint

$\underbrace{y_i \tilde{w}^T x_i - 1 \geq 0}_{\alpha_i}$

unconstrained new objective

$$\min_{\tilde{w}} J = \underbrace{\frac{1}{2}\|\tilde{w}\|^2}_{\text{margin}} - \sum_{i=1}^{n} \alpha_i \underbrace{(y_i \tilde{w}^T x_i - 1)}_{\text{from the constraints}}$$

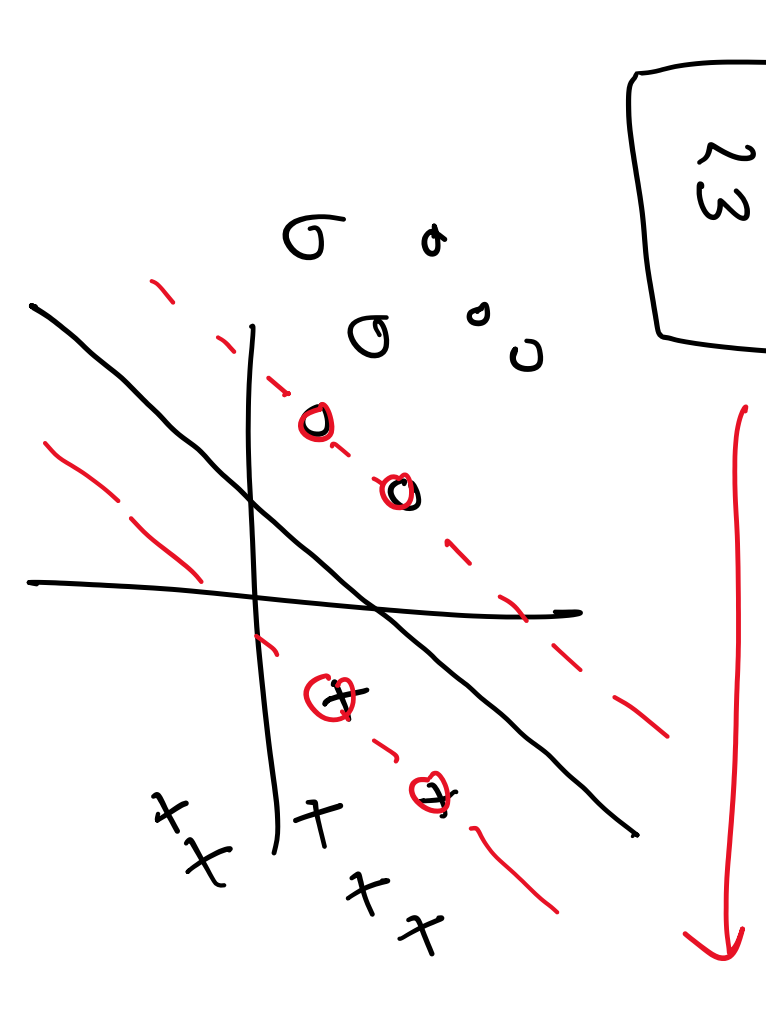
Solution: differentiate with respect to $\tilde{w}$

$$\frac{\partial J}{\partial \tilde{w}} = \frac{1}{\cancel{2}} \cancel{2}\tilde{w} - \sum_{i=1}^{n} \alpha_i y_i \tilde{x}_i = 0$$

$$\boxed{\tilde{w} = \sum_{i=1}^{n} \alpha_i y_i \tilde{x}_i}$$

$y_i \to +1, -1$

$\alpha_i \geq 0$



linear combination of the training points!

$\alpha_i = 0$ for all non SV

$\alpha_i > 0$ only for the support vectors SV

$$\tilde{w} = \sum_{\text{S.V.}} \alpha_i y_i \tilde{x}_i$$

$(\alpha_i > 0)$ ⇐ S.V.

$\tilde{w}$ unknown ⇒ α_i unknown

solve for α_i instead of $\tilde{w}$

$$\boxed{\max_{\alpha_i} J_{dual} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \tilde{x}_i^T \tilde{x}_j}$$

α_i

$\vec{\alpha} = \underbrace{(\alpha_1, \alpha_2, \ldots \alpha_n)}_{\text{n-dim, one per point}}$

Quadratic equation, Convex] quadratic form

⇓

Unique/global optimal solution

$\nabla_{\alpha_i} = \dfrac{\partial J}{\partial \alpha_i}$ gradient ascent!