→ **Deep Learning:** Learning Theorey ✓ Finite hypothesis

Infinite hypothesis → VC-dimension

✓ $$m \geq \frac{1}{\epsilon}\left(\ln |H| + \ln\left(\frac{1}{\delta}\right)\right)$$

for finite hypothesis cs train error = 0

✓ $$m \geq \frac{1}{2\epsilon^2}\left(\ln |H| + \ln\left(\frac{1}{\delta}\right)\right)$$

finite hypothesis cs trainError ≠ 0

$|H| \to \infty$ Model Complexity is Computed Using VC dimension of the hypothesis class.

VC-dimension: The maximum number of samples that can be $\boxed{\text{Shattered}}$ by a hypothesis class is called VC-dimension.

Vapnik-Chervonenkis

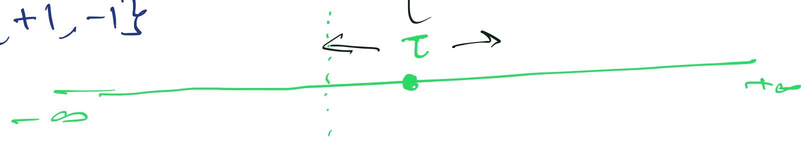
**SHATTER:** For all possible labellings of the samples, if there exist a hypothesis (model) 'h' with zero error (correctly classify all samples) in the hypothesis class 'H' is called shattering.

**Example:**

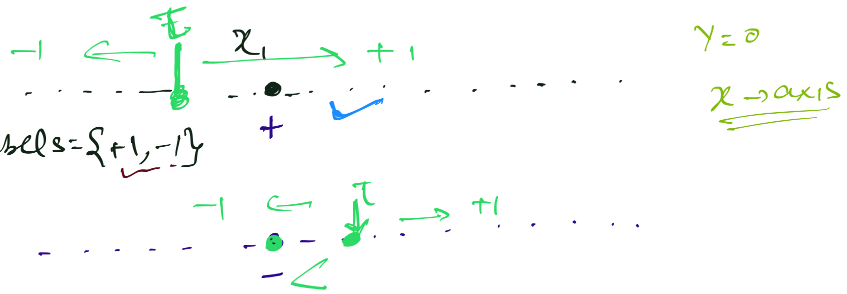
Let Threshold: $\tau$ $\begin{cases} +1 & \text{if } \tau \geq 0 \\ -1 & \text{otherwise} \end{cases}$

Class: $\{+1, -1\}$

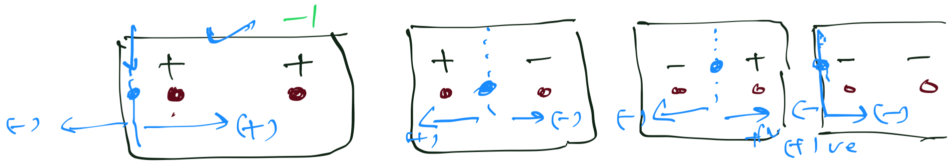
Samples:



$x_1$




Class labels = {+1, -1}

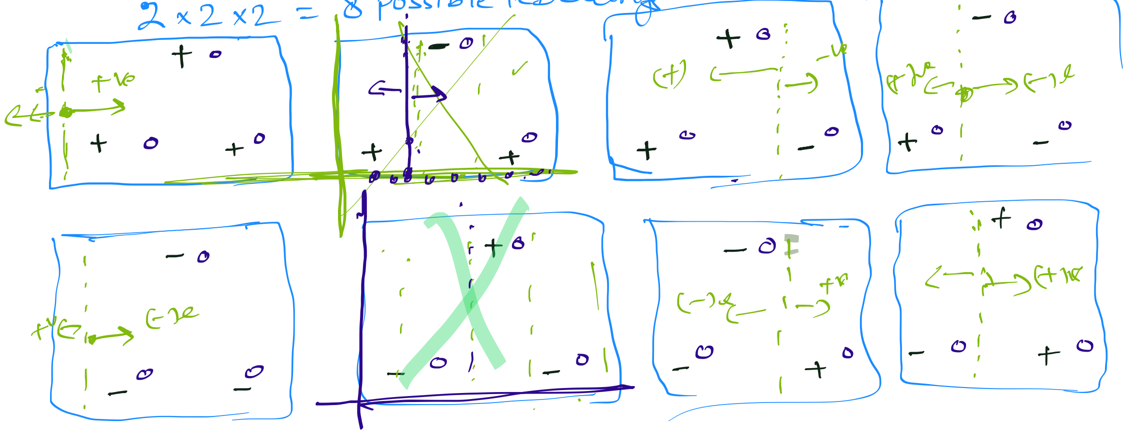



2 - Samples, 2 Classes :

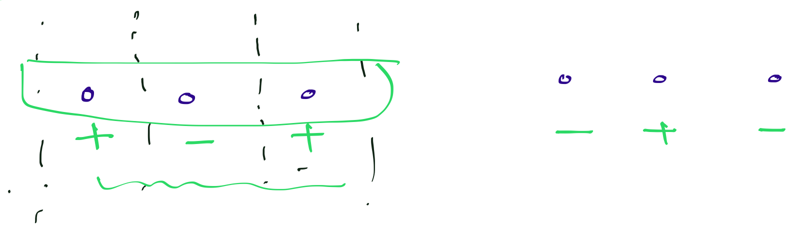
$2 \times 2 = 4$

3 - Sample, 2 Class

$2 \times 2 \times 2 = 8$ possible lebellings



threshold does not exist which can shatter.



VC-dimension for threshold hypothesis class is 2. It can shatter a maximum of 2 samples.

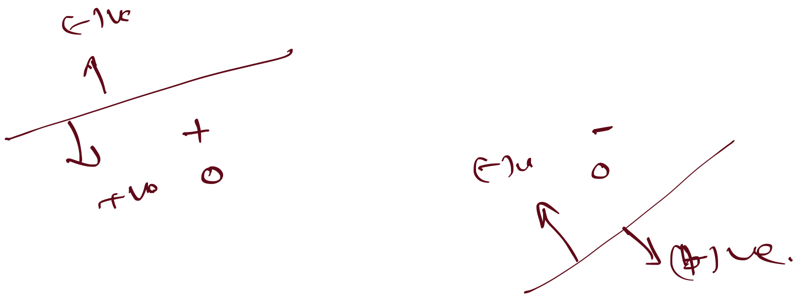
Example 2.

H: Linear models | Linear hyperplane.

Let example be H: linear model in 2D (line)

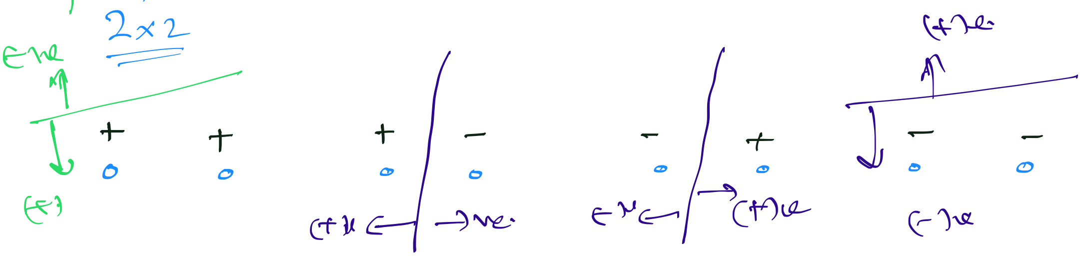
Class = {+1, -1}

1. Sample:




2 Samples, 2 Classes. $2^2 = 4$

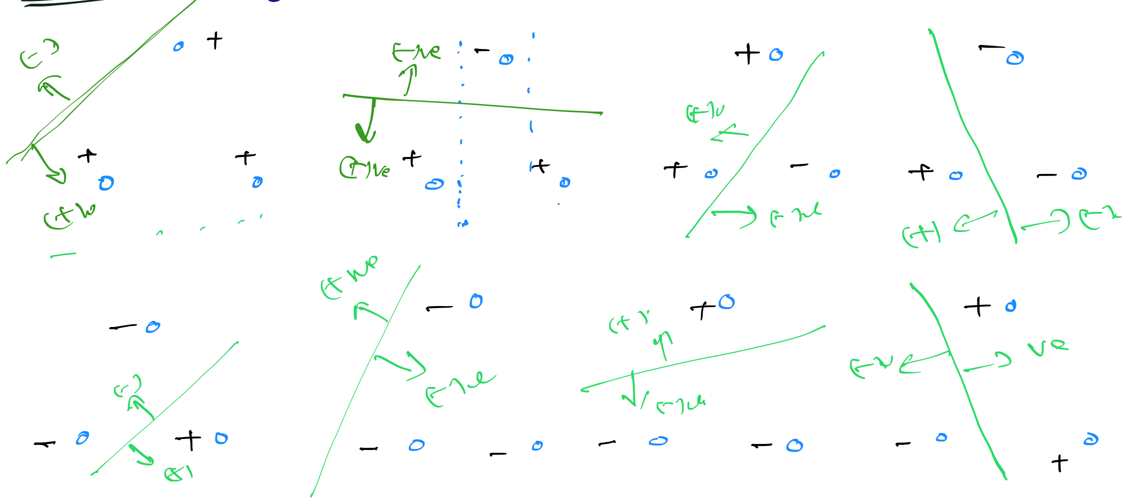
2×2

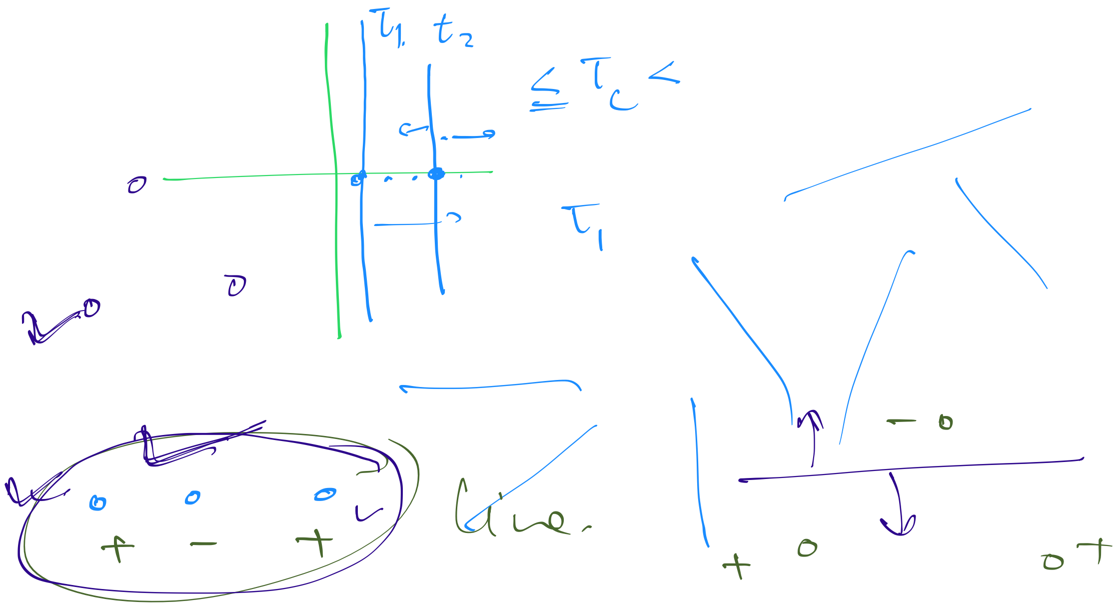



i.e we can shatter 2 points.

## 3-Samples, 2 classes:

2 2 2 = 8

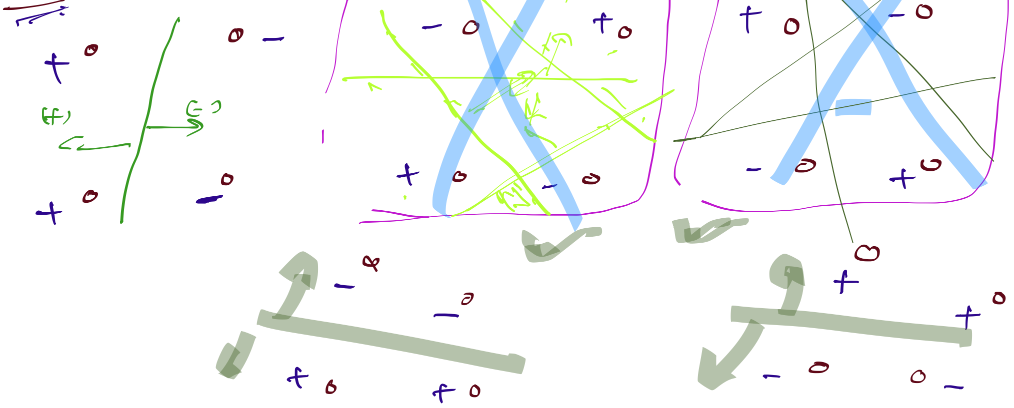


Line Can shatter 3 Samples.

4 Samples, 2 classes : $\underline{2} \times \underline{2} \times \underline{2} \times \underline{2} = 16$ Possible labelling.

FEW:



Line Can not Shatter 4 - points

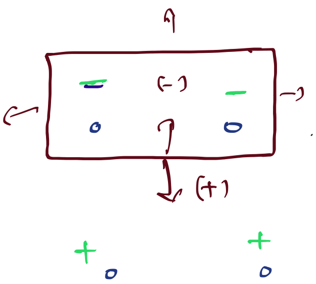
Line can shatter ~~at~~ maximum of 3-points.

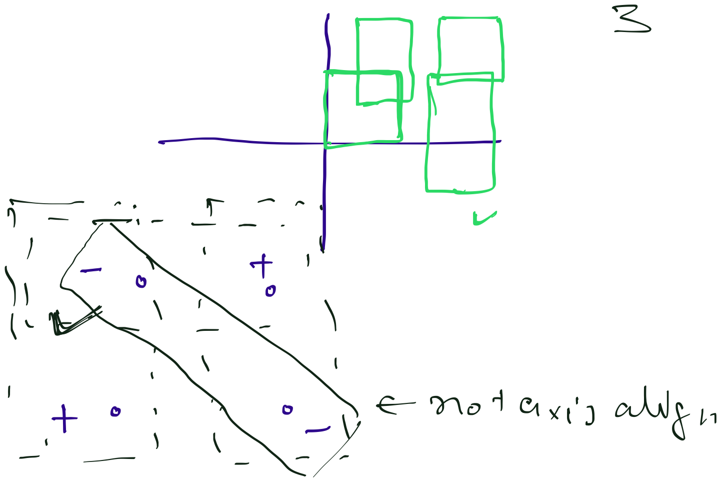
VC-dimension (Line (or linear hypothesis in 2D) is 3

VC-dimension of Linear hyperplane in d-dimension is $d+1$

Hypothesis is Axis-aligned Rectangle (2D)

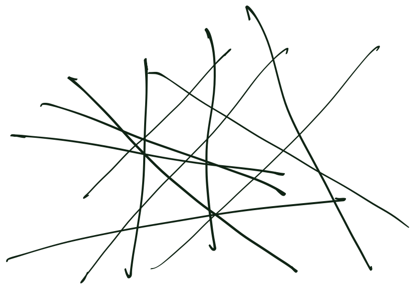
4-Sample class = 2



Rectangle in 2D → VC-dim = (4)

H = line on 2D →



$y = mx + c$

$$\boxed{w^T x_i + b = 0}$$

- b: bias (Distance of the plane from origin of the Vector space)
- w: weight (Vector perpendicular to the plane having $x_i$)

In a d-dimension — $i^{th}$ sample $x_i$

VC(H) ∝ # parameter

$$x_i = \begin{bmatrix} x_i^1 \\ x_i^2 \\ \vdots \\ x_i^d \end{bmatrix}_{d\times 1} ; \quad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} ; \quad b \in \mathbb{R}$$

$h_{w,b}(x_i) = w^T x_i + b$

$$h(x_i) = \boxed{w^T x_i + b = 0}$$

$x_i \in \mathbb{R}^d$

$w \in \mathbb{R}^d$

$VC(H) = d+1$

Linear model in d-dimension.

$$m \geq O(VC(H))$$

$$m \geq O(VC(H) \lg VC(H))$$

$$m \geq O(VC(H)^2)$$

Structure

m: # samples

$\epsilon$: ( ~~generalization~~ Error | Tolerance

$$m \geq O\left(\frac{1}{\delta}\right)$$

$$\boxed{\simeq m \geq O((1-\delta)^{?})}$$

$$m \geq O\left(\frac{1}{\epsilon}\right)$$

Zero True

$$m \geq O\left(\frac{1}{\epsilon^2}\right)$$

$$m \geq O(\ln |H|)$$

$$\boxed{\text{True Error / Generalization Error} = \text{Train Error} + \epsilon}$$

Structural Error

→ Finite Hyp

(i) $m \geq \frac{1}{\epsilon}\left(\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right)$ # Train Error = 0

(ii) $m \geq \frac{1}{2\epsilon^2}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right)$ # Train Err ≠ 0

(iii) $m \geq \frac{1}{\epsilon^2}\left(C_1 . VC(H) + C_2 . \ln\left(\frac{C_3}{\delta}\right)\right)$

$\underbrace{C_1, C_2, C_3}$ # Constant

(i) $\epsilon \leq \frac{\ln(|H|) + \ln\left(\frac{1}{\delta}\right)}{m}$

$\epsilon \propto (1-\delta)$

Prob Conf

(ii) $\epsilon = \sqrt{\frac{\ln(|H|) + \ln\left(\frac{1}{\delta}\right)}{2m}}$

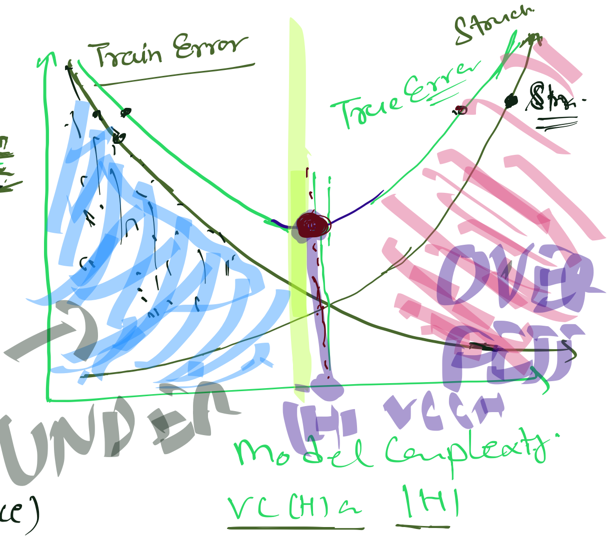
(iii) $\epsilon \simeq \sqrt{\frac{C_1 . VC(H) + C_2 \ln(C_3/\delta)}{m}}$

$\frac{m}{\delta}$

Real time Error — empirical error — Structure err $\epsilon$

$\boxed{\text{True Error} \leq \text{Train Error} + \epsilon}$

True is Error is high
Train Error is low
overfitting (Variance)




Train Error → high : Underfitting. (bias problem)

Good hypothesis : Low overfitting Low Under fit.

high Variance → overfitting.
high bias → Underfitting

Bias - Variance Trade Off

Model : low bias & low variance is a good