Concept of Generalization in Learning: $E_{out} \approx 0$

$E_{in} \approx 0$ (Learning Algo) $\oplus$ $E_{in} \approx E_{out}$ (feasibility)

Hoeffding Inequality:

$$P\left[\,|E_{in}(g) - E_{out}(g)| > \epsilon\,\right] \leq 2\boxed{M}e^{-2\epsilon^2 N}$$

$\Rightarrow$? $\quad M \to$ # Hypothesis Set $\rightsquigarrow (\infty)$

$m_{\mathcal{H}}(N)$ = max dichotomies realized by N (over $\mathcal{H}$) $\leftarrow$ Growth Function

Break Point: $\boxed{m_{\mathcal{H}}(k) < 2^k}$ $\Rightarrow$ for all $i < k$, $m_{\mathcal{H}}(i) = 2^i$

▶ If k is finite $\Rightarrow$ $m_{\mathcal{H}}(N)$ is polynomial

Proof: $B(N, k) \leq B(N-1, k) + B(N-1, k-1)$

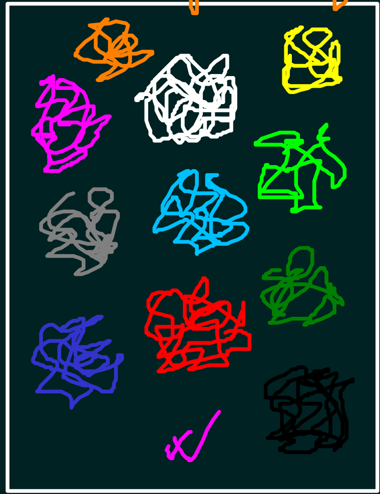
$$\Rightarrow m_{\mathcal{H}}(N) = B(N, k) \leq \left\{ \sum_{i=0}^{k-1} \binom{N}{i} \right\} \rightarrow O(N^{k-1})$$

polynomial

$\rightarrow$ Can $m_{\mathcal{H}}(N)$ be a good fn to replace M ??

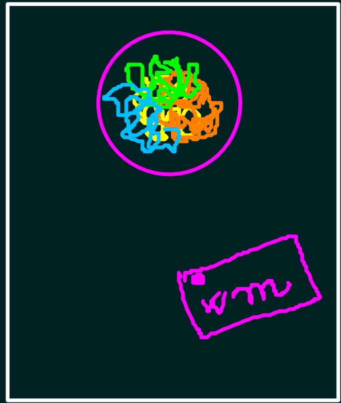
$\hookrightarrow$ Vapnik-Chervonenkis (VC) Bound $\quad \delta = 4\, m_{\mathcal{H}}(2N)\, e^{-\frac{1}{8}\epsilon^2 N}$

Hoeffding Inequality



Data points

$m_{\mathcal{H}}(N) \rightarrow$



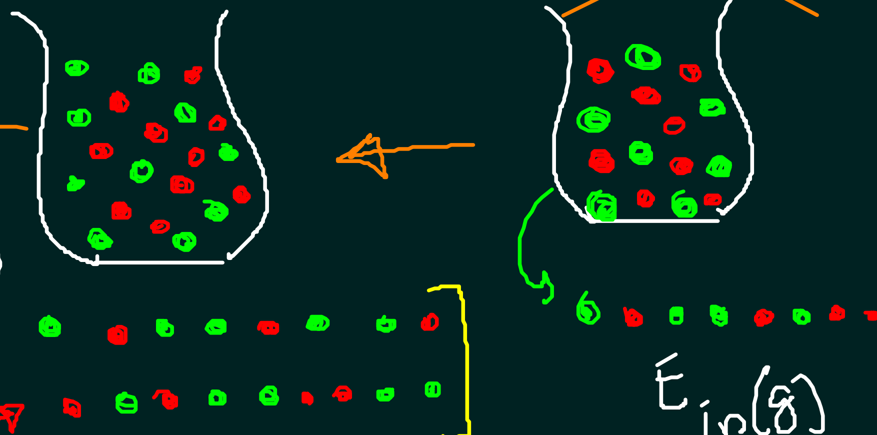
VC bound

$$\text{Prob}\left[\,|E_{out}(g) - E_{in}(g)| > \epsilon\,\right] \leq \;?$$

[N points you look through Holes] $\Leftarrow m_{\mathcal{H}}(N)$

$\boxed{E_{in}(g) \approx E'_{in}(g)}$ $\quad E_{in} \approx E_{out}$

~~$E_{out}(g)$~~



$E_{in}(g)$

$E'_{in}(g)$

$E_{in}(g)$

$$P\left[\,|E_{out}(g) - E_{in}(g)| > \epsilon\,\right] \leq 4 \cdot \boxed{m_{\mathcal{H}}(2N)} \cdot e^{-\frac{1}{8}\epsilon^2 N}$$

k

[VC-Inequality
VC-Bound]

$m_{\mathcal{H}}(N)$ poly

$2^k N^{k-1}$

Const

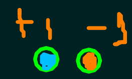
▶ VC-Dimension: $d_{vc}(\mathcal{H}) = k - 1$

↳ largest N that can get SHATTERED by $\mathcal{H}$

→ $N \le d_{vc}(\mathcal{H}) \Rightarrow \mathcal{H}$ can SHATTER N points

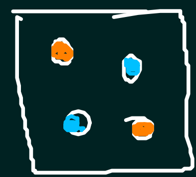
→ $k > d_{vc}(\mathcal{H}) \Rightarrow k$ is a break point for $\mathcal{H}$.

$$m_{\mathcal{H}}(N) \le \sum_{i=0}^{k-1} \binom{N}{i} = \sum_{i=0}^{d_{vc}} \binom{N}{i} \rightarrow O\left(N^{d_{vc}}\right)$$



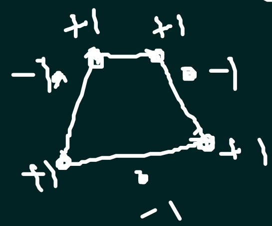
↳ Ex:

1. Positive Rays: $d_{vc} = 1$ $\quad k = 2$
2. Positive Interval: $d_{vc} = 2$ $\quad k = 3$ $\quad a \leftrightarrow b$
3. 2D-perceptrons: $\boxed{d_{vc} = 3}$ $\quad k = 4$



$m_{\mathcal{H}}(4) = 14$

4. Convex Set: $d_{vc} = \infty$ $\quad k = \infty$



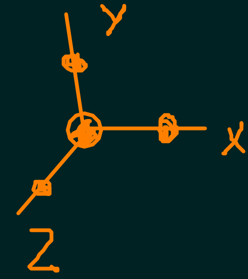
$g \approx f \Rightarrow g$ generalizes well

▷ A set of $N = d+1$ points in $\mathbb{R}^d$ shattered by Perceptron.

↳ $d_{vc}(\mathcal{H}) \geq d+1$ and $d_{vc}(\mathcal{H}) \leq d+1 \Rightarrow \boxed{d_{vc}(\mathcal{H}) = d+1}$

(1) ✓ (2) ✓

$$
\begin{array}{c} x_0 \\ \end{array}
\qquad
X = \begin{bmatrix} 1 & \text{———} \; x_1^T \; \text{———} \\ & \text{———} \; x_2^T \; \text{———} \\ & \vdots \\ & \text{———} \; x_{d+1}^T \; \text{———} \end{bmatrix}_{(d+1)\times(d+1)} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 0 & 1 \cdots & 0 \\ & & \vdots & & \\ 1 & 0 & 0 & 0 \cdots & 1 \end{bmatrix}
$$



(i) X is invertible ✓

X sq. ; $\det(x) =)$

any $y = \begin{bmatrix} y_1 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$ Can you find $\vec{w}$ satisfying $\text{sign}(Xw) = y$?

↳ We can shatter $d+1$ points

$\boxed{Xw = y} \Rightarrow w = X^{-1} y$

B.P = $d+2$??

$\boxed{d_{vc} \geq d+1}$

↳ d_{vc} is at least $d+1$

Let us take $\underline{d+2 \text{ points}}$, $\langle \overbrace{x_1}^{d+1}, \ldots, x_{d+2} \rangle$ $\quad [1 \cdot \ldots]$ $\uparrow$ x_0

$$x_j = \sum_{i \neq j} a_i x_i \leftarrow \text{Linear dependence} \rightarrow \text{(i)}$$

$[0 \cdot \ldots 0]$ $\quad$ not all a_i's are zero $\rightarrow$ (ii)

$\hookrightarrow$ x_i's with non-zero a_i's get $y_i = \text{sign}(a_i)$
and x_j gets $y_j = -1$ $\Big\}$ ONE Dich.

$$x_j = \sum_{i \neq j} a_i x_i \Rightarrow w^T x_j = \sum_{i \neq j} a_i w^T x_i > 0$$

$\rightarrow 0$

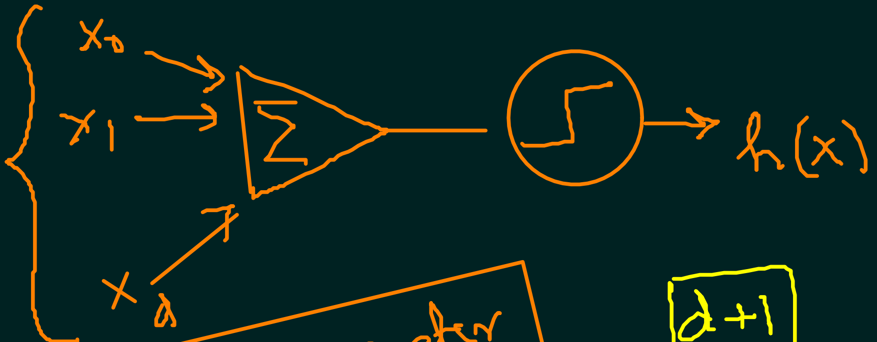
$$y_i = \text{sign}(a_i) \checkmark \rightarrow$$
$$= \text{sign}(w^T x_i) \checkmark$$

$$\boxed{d_{VC} = d+1}$$

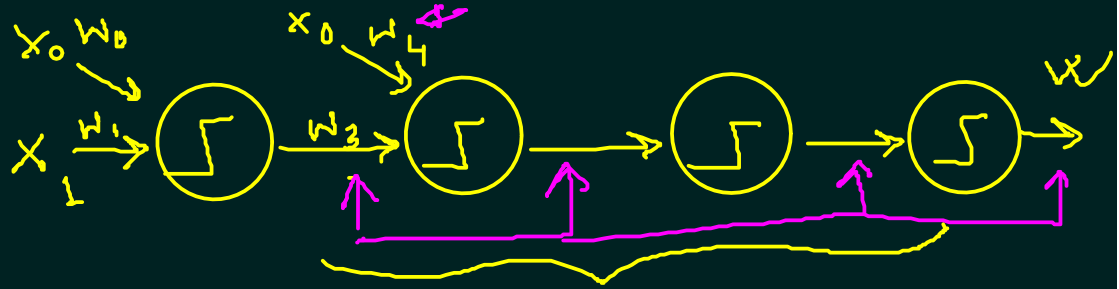
$\underline{O(N^d e^{-N})}$

$$y_j = \text{sign}(w^T x_j) = +1$$

$\underline{\text{Contradiction}}$



$\boxed{d+1 \text{ parameter}}$ $\boxed{d+1}$



\# effective parameters $\propto d_{vc}$

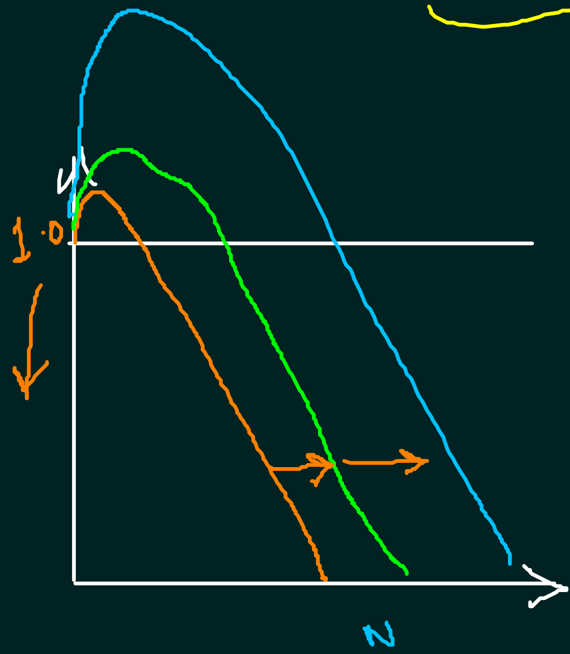
$$① \; P[\underbrace{|E_{out}(g) - E_{in}(g)| > \epsilon}^{\text{Bad}}] \leq \underbrace{4\, m_{\mathcal{H}}(2N)\, e^{-\frac{1}{8}\epsilon^2 N}}_{\delta \approx O(N^d e^{-N})}$$

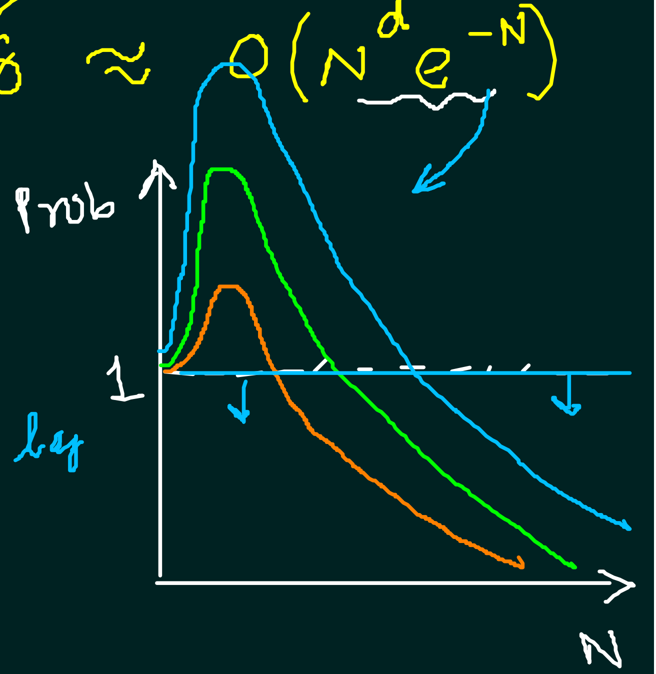
Rule: $N \geq 10\, d_{vc}$

$$P[|E_{out} - E_{in}| < \epsilon]$$

$$= 1 - P[|E_{out} - E_{in}| > \epsilon]$$

$$\Rightarrow P[[f(x) \neq h(x)] < \epsilon] \geq 1 - \delta \longrightarrow \text{Probably Approximately Correct learning}$$





$\leftarrow$ log

Leslie Valiant $\leftarrow$ $\boxed{\text{PAC}}$

$$4\, m_{\mathcal{H}}(2N)\, e^{-\frac{1}{8}\epsilon^2 N} = \delta$$

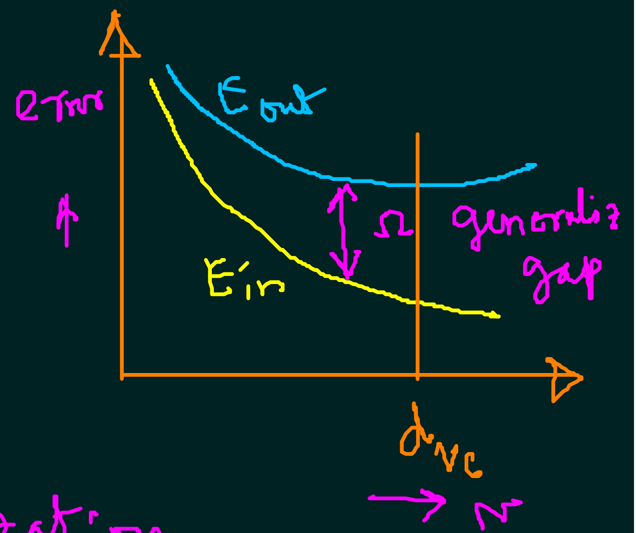
$$\Rightarrow \epsilon = \underbrace{\sqrt{\frac{8}{N} \ln \frac{4\, m_{\mathcal{H}}(2N)}{\delta}}}_{\Omega(N,\, \mathcal{H},\, \delta)}$$

error that generalization will make given δ, N,

$$|E_{out} - E_{in}| \leq \Omega \quad \text{(gen. bound)}$$

$$\Rightarrow E_{out} - E_{in} \leq \Omega$$

Generalization Bound



$$E_{in} \approx E_{out} \Rightarrow \boxed{E_{in} + \Omega \approx E_{out}}$$

$\rightarrow N \rightarrow$ reasonably low $\Rightarrow$ [poly time Train]

$\boxed{\text{PAC}}$