Review of Lecture 6

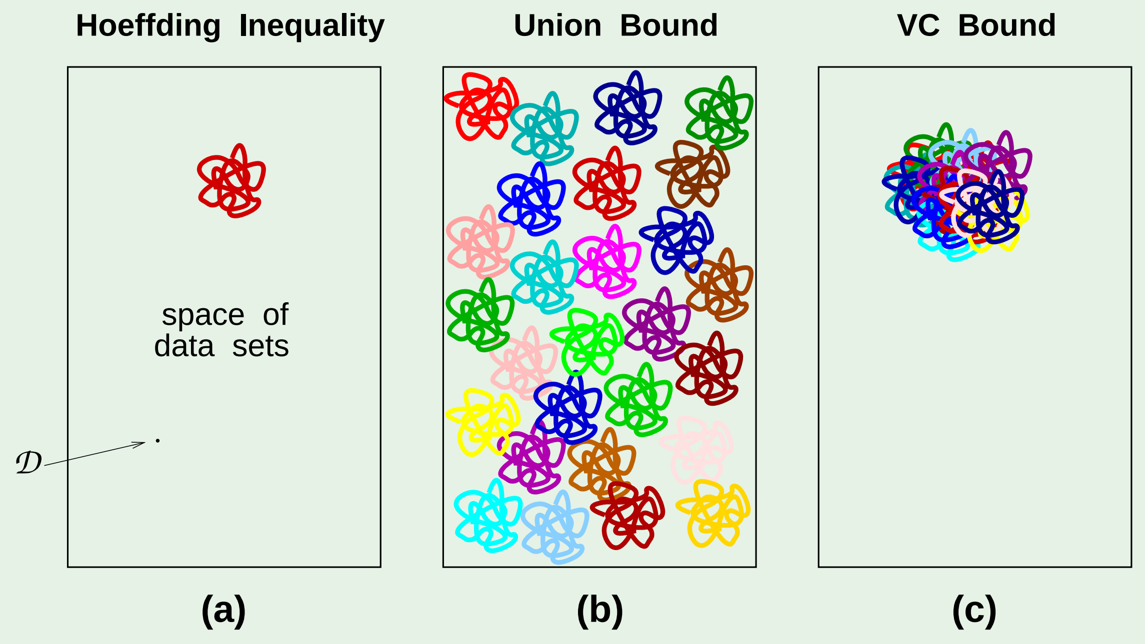
- $m_{\mathcal{H}}(N)$ is polynomial

if $\mathcal{H}$ has a break point k

		k						
		1	2	3	4	5	6	..
	1	1	2	2	2	2	2	..
	2	1						
	3	1						
N	4	1						
	5	1						
	6	1						
	:	:						

$$m_{\mathcal{H}}(N) \quad \leq \quad \underbrace{\sum_{i=0}^{k-1} \binom{N}{i}}_{\text{maximum power is } N^{k-1}}$$

- The VC Inequality



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

$$\downarrow \qquad \downarrow \qquad \downarrow$$

$$\downarrow \qquad \downarrow \qquad \downarrow$$

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

Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 7: The VC Dimension

Sponsored by Caltech's Provost Office, E&AS Division, and IST • Tuesday, April 24, 2012

Outline

- The definition
- VC dimension of perceptrons
- Interpreting the VC dimension
- Generalization bounds

Definition of VC dimension

The VC dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{\mathrm{VC}}(\mathcal{H})$, is

the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$

"the most points $\mathcal{H}$ can shatter"

$N \leq d_{\mathrm{VC}}(\mathcal{H}) \implies \mathcal{H}$ can shatter N points

$k > d_{\mathrm{VC}}(\mathcal{H}) \implies k$ is a break point for $\mathcal{H}$

The growth function

In terms of a break point k:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

In terms of the VC dimension d_{VC}:

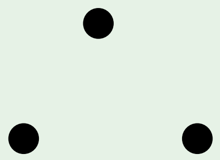
$$m_{\mathcal{H}}(N) \leq \underbrace{\sum_{i=0}^{d_{\mathrm{VC}}} \binom{N}{i}}_{\text{maximum power is } N^{d_{\mathrm{VC}}}}$$

Examples

- $\mathcal{H}$ is positive rays:

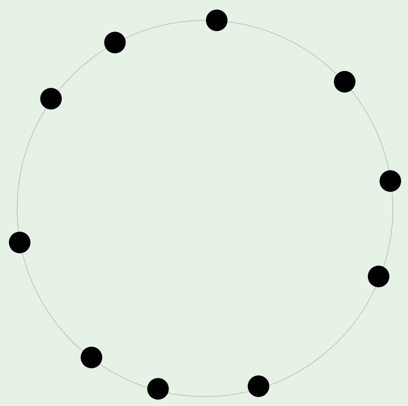
$$d_{\mathrm{VC}} = 1$$

- $\mathcal{H}$ is 2D perceptrons:



$$d_{\mathrm{VC}} = 3$$

- $\mathcal{H}$ is convex sets:

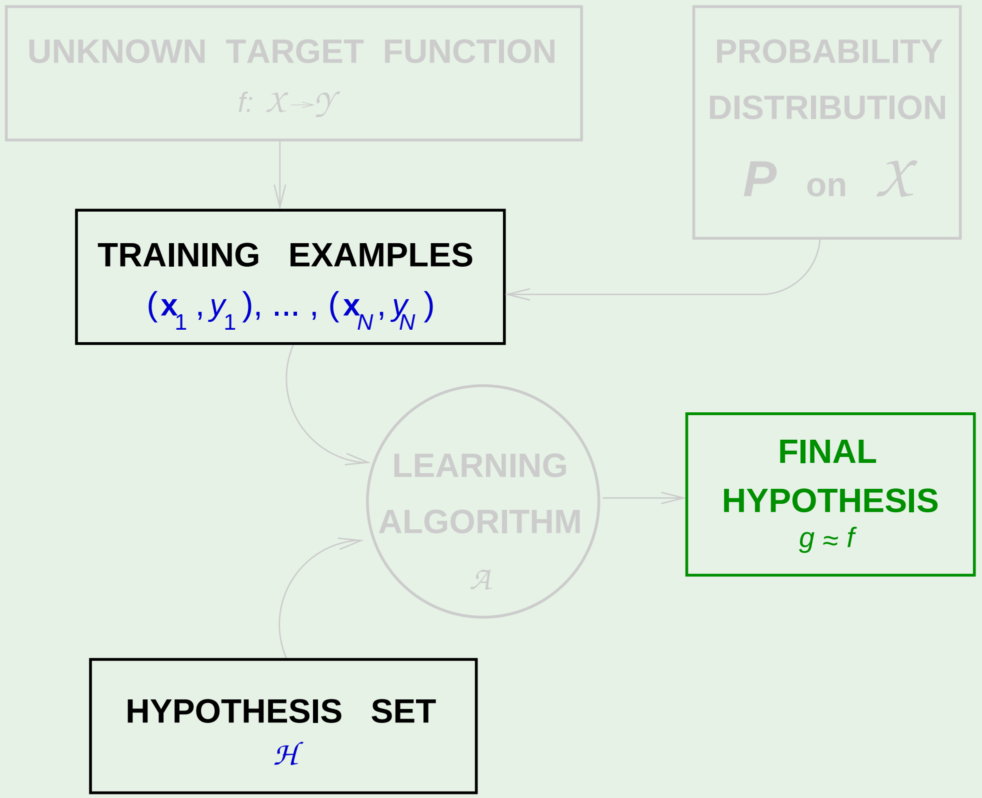


$$d_{\mathrm{VC}} = \infty$$

VC dimension and learning

$d_{\mathrm{VC}}(\mathcal{H})$ is finite $\Longrightarrow$ $g \in \mathcal{H}$ will generalize

- Independent of the learning algorithm
- Independent of the input distribution
- Independent of the target function



VC dimension of perceptrons

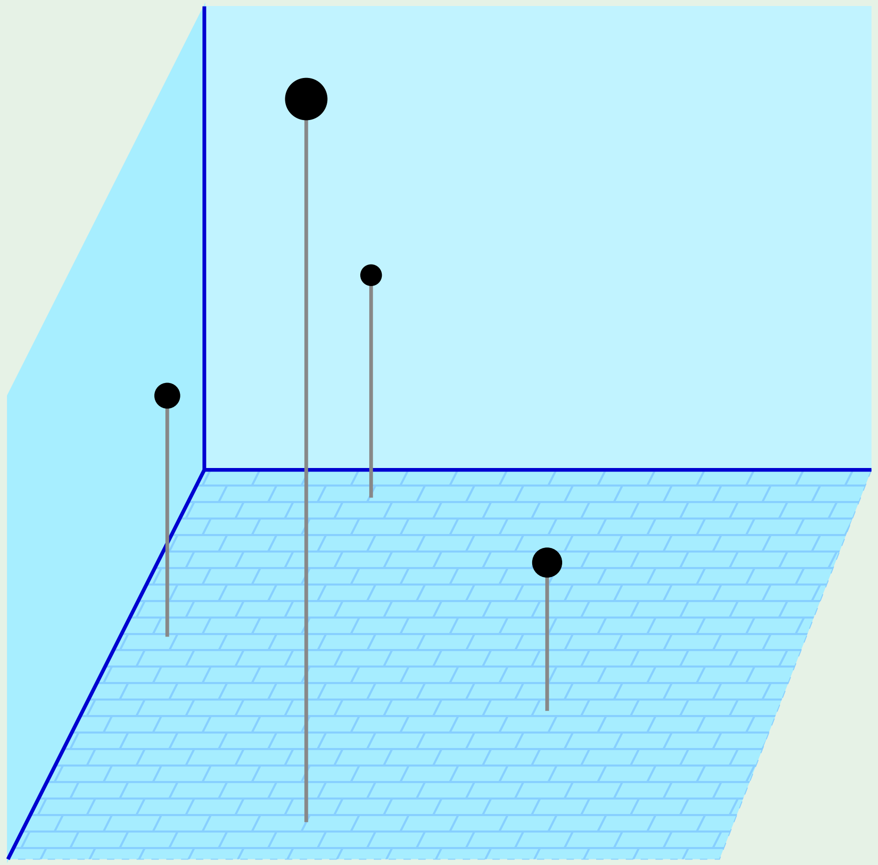
For $d = 2$, $d_{\text{VC}} = 3$

In general, $\boxed{d_{\text{VC}} = d + 1}$

We will prove two directions:

$$d_{\text{VC}} \leq d + 1$$

$$d_{\text{VC}} \geq d + 1$$



Here is one direction

A set of $N = d + 1$ points in $\mathbb{R}^d$ shattered by the perceptron:

$$\mathrm{X} = \begin{bmatrix} \text{—} \ \mathbf{x}_1^{\intercal} \ \text{—} \\ \text{—} \ \mathbf{x}_2^{\intercal} \ \text{—} \\ \text{—} \ \mathbf{x}_3^{\intercal} \ \text{—} \\ \vdots \\ \text{—} \mathbf{x}_{d+1}^{\intercal} \text{—} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & & 0 \\ & \vdots & & \ddots & 0 \\ 1 & 0 & \dots & 0 & 1 \end{bmatrix}$$

X is invertible

Can we shatter this data set?

For any $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$, can we find a vector $\mathbf{w}$ satisfying

$$\text{sign}(\mathrm{X}\mathbf{w}) = \mathbf{y}$$

Easy! Just make $\mathrm{X}\mathbf{w} = \mathbf{y}$

which means $\mathbf{w} = \mathrm{X}^{-1}\mathbf{y}$

We can shatter these $d+1$ points

This implies what?

[a] $d_{\text{VC}} = d+1$

[b] $d_{\text{VC}} \geq d+1$ ✓

[c] $d_{\text{VC}} \leq d+1$

[d] No conclusion

Now, to show that $d_{\text{vc}} \leq d + 1$

We need to show that:

[a] There are $d+1$ points we cannot shatter

[b] There are $d+2$ points we cannot shatter

[c] We cannot shatter *any* set of $d+1$ points

[d] We cannot shatter *any* set of $d+2$ points ✓

Take any $d+2$ points

For any $d+2$ points,

$$\mathbf{x}_1, \cdots, \mathbf{x}_{d+1}, \mathbf{x}_{d+2}$$

More points than dimensions $\implies$ we must have

$$\mathbf{x}_j = \sum_{i \neq j} a_i \ \mathbf{x}_i$$

where not all the a_i's are zeros

So?

$$\mathbf{x}_j = \sum_{i \neq j} a_i \ \mathbf{x}_i$$

Consider the following dichotomy:

$\mathbf{x}_i$'s with non-zero a_i get $\quad y_i = \text{sign}(a_i)$

and $\mathbf{x}_j$ gets $\quad y_j = -1$

No perceptron can implement such dichotomy!

Why?

$$\mathbf{x}_j \;=\; \sum_{i \neq j} a_i\, \mathbf{x}_i \quad \Longrightarrow \quad \mathbf{w}^\intercal \mathbf{x}_j \;=\; \sum_{i \neq j} a_i\, \mathbf{w}^\intercal \mathbf{x}_i$$

If $\;y_i = \text{sign}(\mathbf{w}^\intercal \mathbf{x}_i) = \text{sign}(a_i)$, then $\;a_i\, \mathbf{w}^\intercal \mathbf{x}_i \;>\; 0$

This forces

$$\mathbf{w}^\intercal \mathbf{x}_j = \sum_{i \neq j} a_i\, \mathbf{w}^\intercal \mathbf{x}_i \;>\; 0$$

Therefore, $\quad y_j = \;\text{sign}(\mathbf{w}^\intercal \mathbf{x}_j) = +1$

Putting it together

We proved $d_{\mathrm{VC}} \leq d+1$ and $d_{\mathrm{VC}} \geq d+1$

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

What is $d+1$ in the perceptron?

It is the number of parameters $w_0, w_1, \cdots, w_d$

Outline

- The definition
- VC dimension of perceptrons
- Interpreting the VC dimension
- Generalization bounds

1. Degrees of freedom

Parameters create degrees of freedom

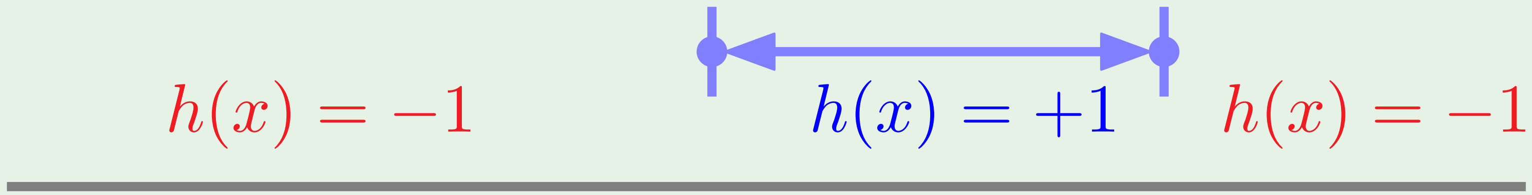
# of parameters: **analog** degrees of freedom

d_{VC}: equivalent **'binary'** degrees of freedom

The usual suspects

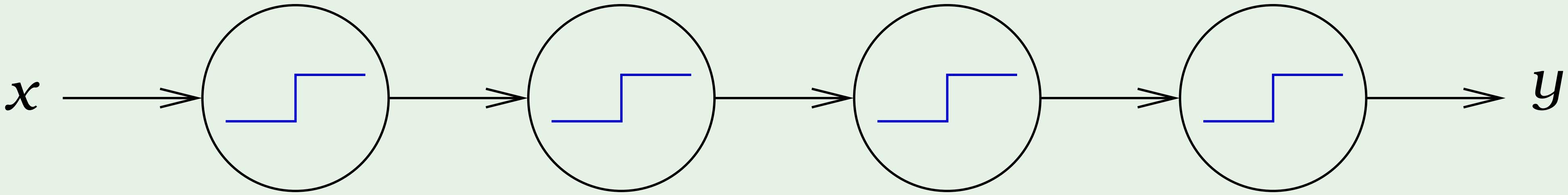
Positive rays ($d_{\text{VC}} = 1$):

Positive intervals ($d_{\text{VC}} = 2$):



Not just parameters

Parameters may not contribute degrees of freedom:



d_{VC} measures the **effective** number of parameters

2. Number of data points needed

Two small quantities in the VC inequality:

$$\mathbb{P}\left[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\right] \le \underbrace{4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}}_{\delta}$$

If we want certain ϵ and δ, how does N depend on d_{VC}?

Let us look at $$N^{d}e^{-N}$$

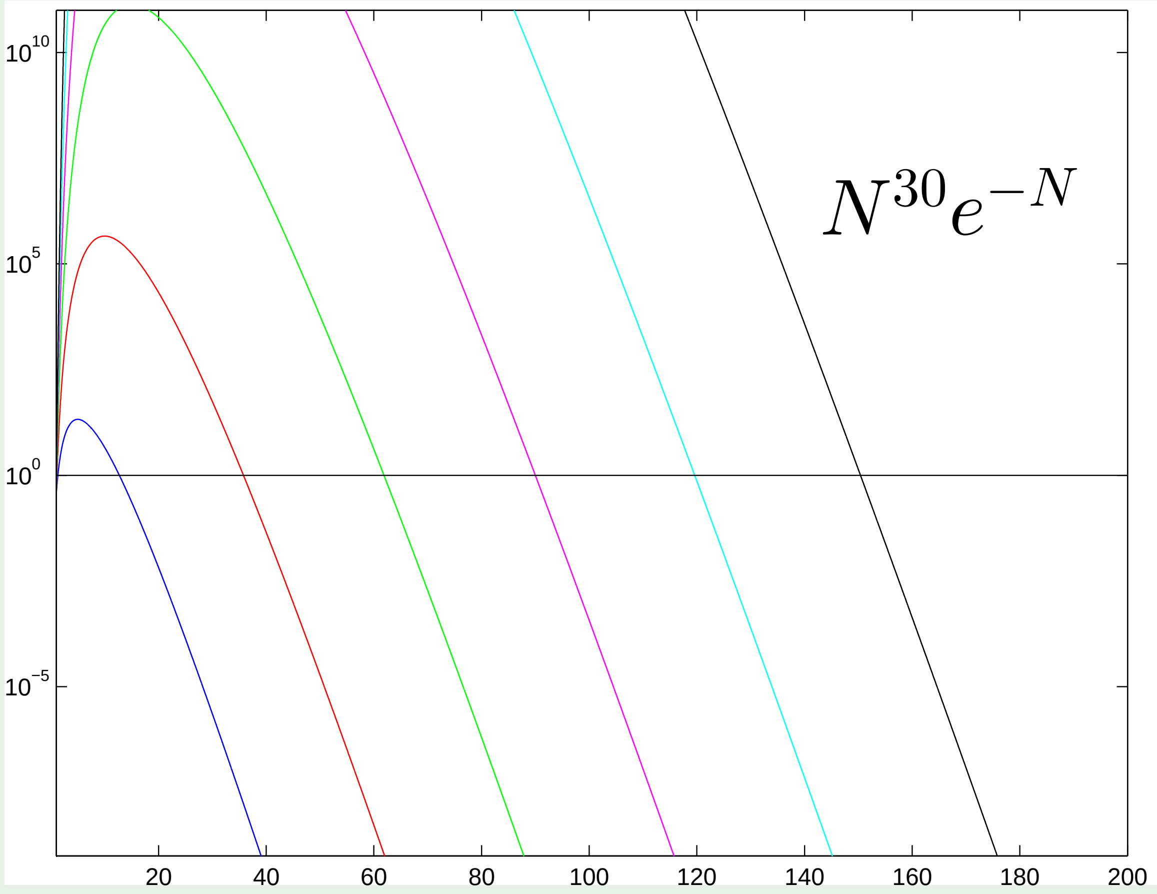
$N^{d}e^{-N}$

Fix $N^{d}e^{-N}$ = small value

How does N change with d?

Rule of thumb:

$$N \;\geq\; 10\, d_{\mathrm{VC}}$$



Outline

- The definition
- VC dimension of perceptrons
- Interpreting the VC dimension
- Generalization bounds

Rearranging things

Start from the VC inequality:

$$\mathbb{P}\left[|E_{\text{out}} - E_{\text{in}})| > \epsilon\right] \;\le\; \underbrace{4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}}_{\delta}$$

Get ϵ in terms of δ:

$$\delta = 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N} \;\Longrightarrow\; \epsilon = \underbrace{\sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}}_{\Omega}$$

With probability $\ge 1-\delta$, $\quad |E_{\text{out}} - E_{\text{in}}| \le \Omega(N, \mathcal{H}, \delta)$

Generalization bound

With probability $\geq 1 - \delta$, $\quad E_{\text{out}} - E_{\text{in}} \leq \Omega$

$$\Longrightarrow$$

With probability $\geq 1 - \delta$,

$$E_{\text{out}} \leq E_{\text{in}} + \Omega$$