Review of Lecture 6

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

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

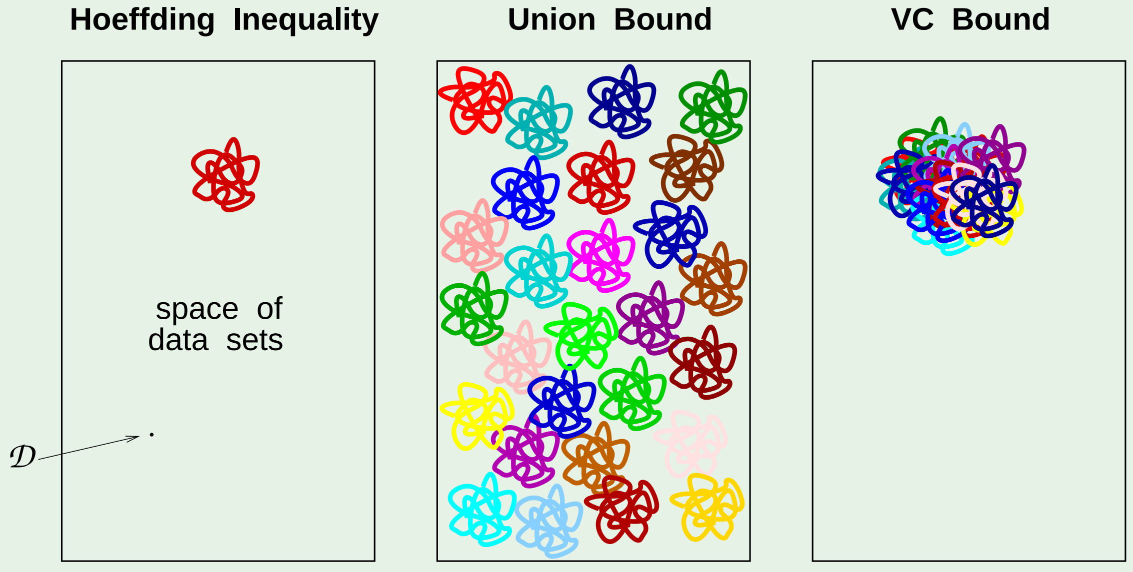
		k 1	2	3	4	5	6	..
	1	1	2	2	2	2	2	..
	2	1						
	3	1						
N	4	1						
	5	1						
	6	1						
	:	:						

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

Growth function characterizes the redundancy in the bad regions that we need to understand to be able to switch from Hoeffding to VC inequality.

- The VC Inequality

Can see the redundancy resulting from the fact that different hyp have, by and large, overlapping bad regions



(a) (b) (c)

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

$$\downarrow \quad \downarrow \quad \downarrow$$

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

VC inequality characterizes generalization of learning

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

i.e. the most points for which H is capable of generating all possible dichotomies. (H can fit these points in a binary function without compromise)

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

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

Note that if dVC=N, it does not say every arrangement of N points can be shattered - only one set of N points that can be shattered is sufficient for the above statement (this has always been the case in our analysis).

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

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

No smaller break point than k = dVC + 1 can exist, so H can shatter dVC points and also any subset of these points

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}:

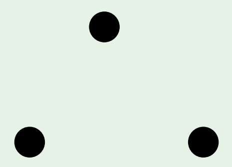
$$m_{\mathcal{H}}(N) \;\leq\; \underbrace{\sum_{i=0}^{d_{\mathrm{VC}}} \binom{N}{i}}_{\text{maximum power is } N^{d_{\mathrm{VC}}}}$$

Examples

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

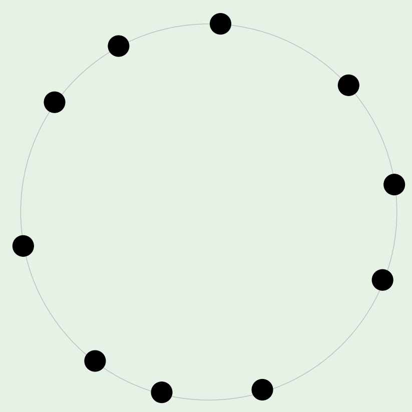
$$d_{\mathrm{VC}} = 1$$

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



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

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

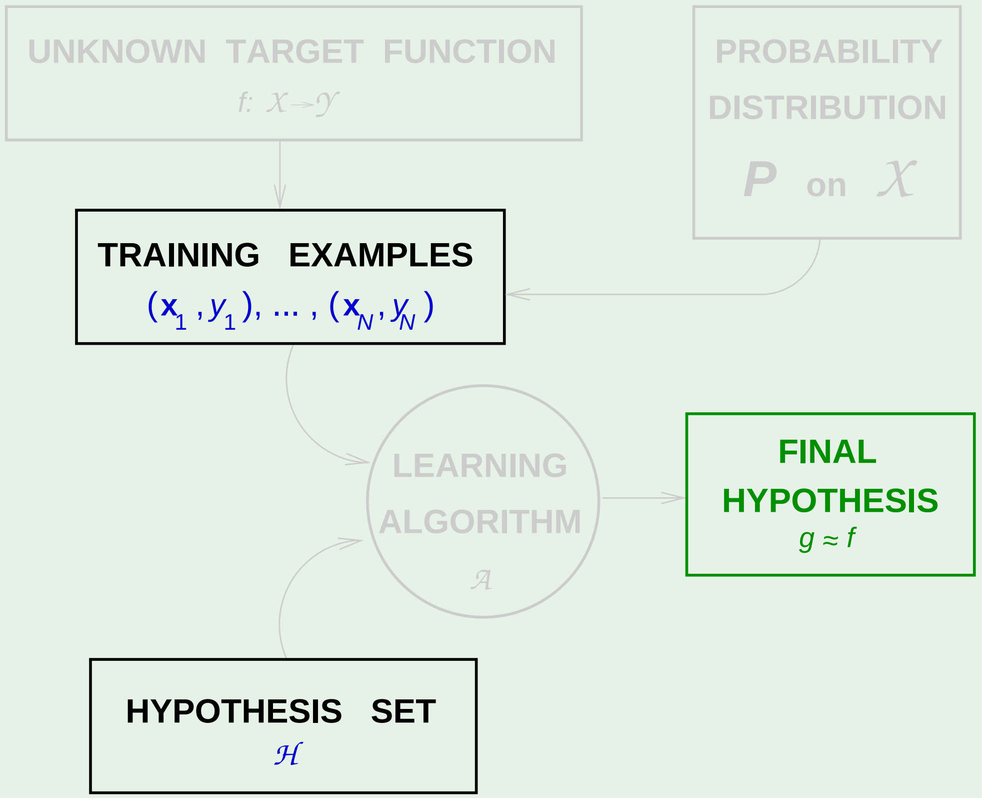


$$d_{\mathrm{VC}} = \infty$$

VC dimension and learning

$d_{\rm VC}(\mathcal{H})$ 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

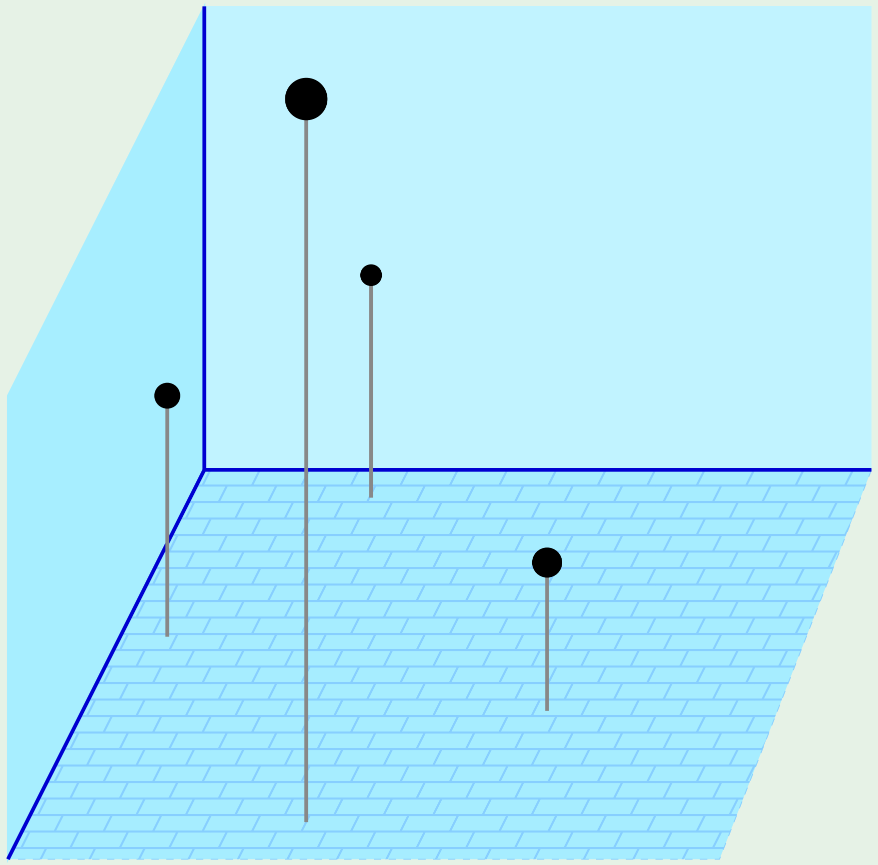
For $d = 2$, $d_{\mathrm{VC}} = 3$

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

We will prove two directions:

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

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



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 dichotomy that is picked from y, we want to show that we can find a perceptron that can realize this, and therefore we show that we can shatter the 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] $d_{\mathrm{VC}} = d+1$

[b] $d_{\mathrm{VC}} \geq d+1$ ✓

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

[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 $\Longrightarrow$ we must have

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

if there are more vectors than dimensions, they must be linearly dependant

where not all the a_i's are zeros

since x0 = 1 for each point due to the bias/threshold term, so not all ai's can be zero

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 $\quad y_i = \text{sign}(a_i)$

xi's with zero ai get +/- 1, we will not consider them

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

No perceptron can implement such dichotomy!

Why?

$$\mathbf{x}_j = \sum_{i \neq j} a_i \, \mathbf{x}_i \quad \Longrightarrow \quad \mathbf{w}^\mathsf{T}\mathbf{x}_j = \sum_{i \neq j} a_i \, \mathbf{w}^\mathsf{T}\mathbf{x}_i$$ (1)

we asserted this dichotomy in the last slide

If $y_i = \text{sign}(\mathbf{w}^\mathsf{T}\mathbf{x}_i) = \text{sign}(a_i)$, then $a_i \, \mathbf{w}^\mathsf{T}\mathbf{x}_i > 0$ (2) since ai and w*xi must have the same sign

This forces $$\mathbf{w}^\mathsf{T}\mathbf{x}_j = \sum_{i \neq j} a_i \, \mathbf{w}^\mathsf{T}\mathbf{x}_i > 0$$ combining (1) and (2)

Therefore, $y_j = \text{sign}(\mathbf{w}^\mathsf{T}\mathbf{x}_j) = +1$

which contradicts the dichotomy - therefore we cannot shatter (for any set we choose) a set of d+2 points with a d-dimensional perceptron

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

One can view the VC dimension as measuring the 'effective' number of parameters. The more parameters a model has, the more diverse its hypothesis set is, which is reflected in a larger value of the growth function mH(N) . In the case of perceptrons, the effective parameters correspond to explicit parameters in the model, namely w0, w1, ..., wd. In other models, the effective parameters may be less obvious or implicit. The VC dimension measures these effective parameters or 'degrees of freedom' that enable the model to express a diverse set of hypotheses.

1. Degrees of freedom

Parameters create degrees of freedom

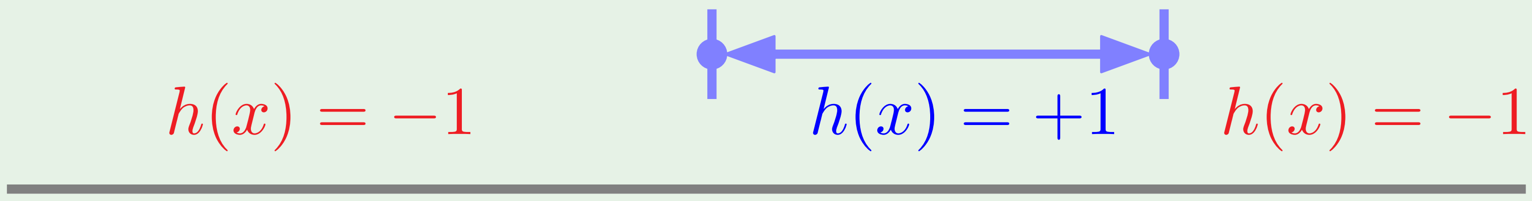
# of parameters: **analog** degrees of freedom

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

The usual suspects

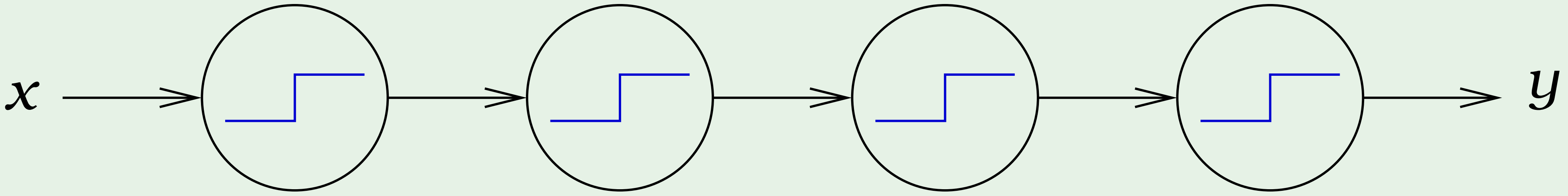
Positive rays ($d_{\text{VC}} = 1$):

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



Not just parameters

Parameters may not contribute degrees of freedom:



perceptrons 2,3,4 are redundant, so only 2 effective parameters rather than 8

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}?

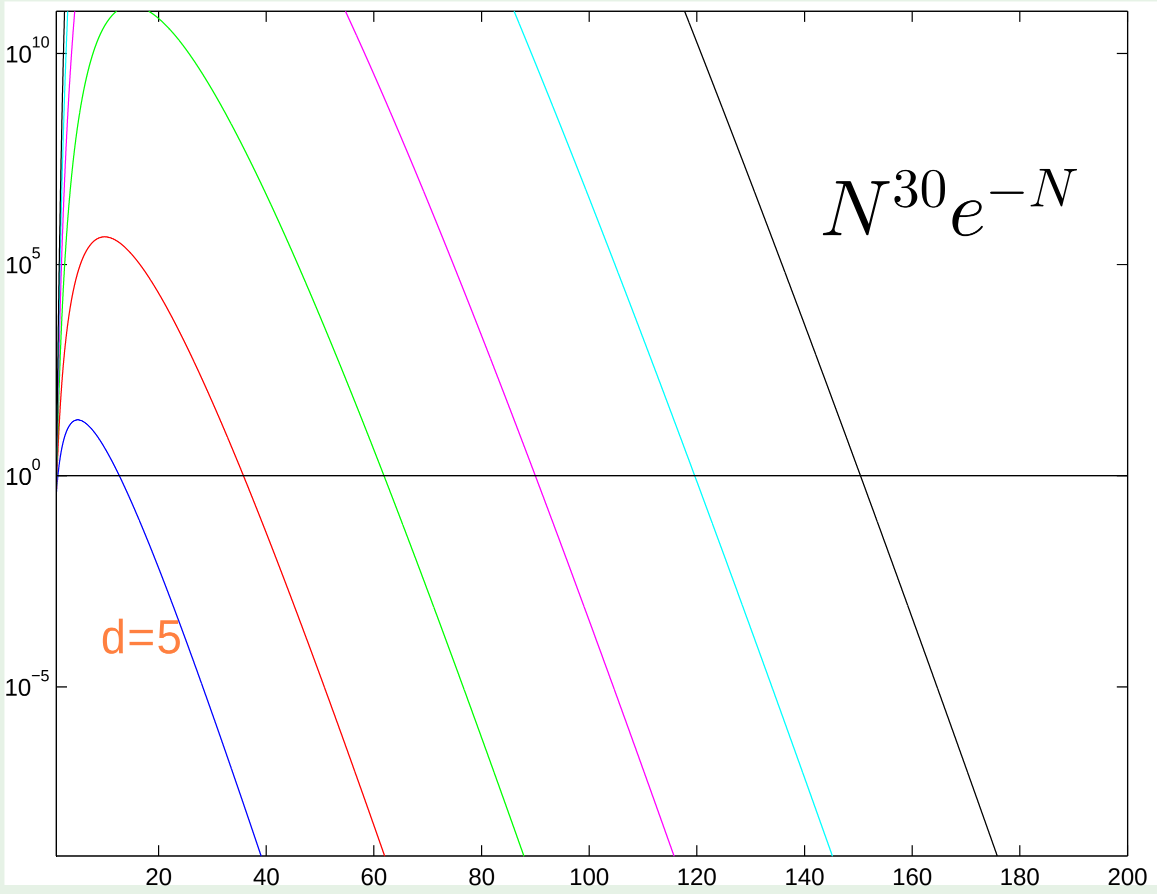
Let us look at $$N^{d}e^{-N}$$

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

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

How does N change with d?

P



N

for a large range of delta and epsilon, and large range of practical applications

Rule of thumb:

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

Practical observation: the actual quantity we are trying to bound follows the same monotonicity as the bound.- so in using bigger dVC, the quantities you get are bigger to achieve a certain level of performance and actually close to proportional. In spite of the fact we cannot get an exact value due to the bound, the relative aspect of the VC dimension holds - so a larger VC dimension requires more examples.

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:

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

so bigger dVC, larger omega and therefore larger epsilon, so worse generalization - conversely more data we have a smaller omega

good event

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

Generalization bound

generalization error

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

Ein is generally smaller than Eout as that is the quantity we are deliberately minimizing

$$\Longrightarrow$$

With probability $\geq 1 - \delta$,

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