Our question: $\boxed{\text{Does } E_{in}(h) \text{ say anything about } E_{out}(h) \text{ ?}}$

Probability of a "bad" event (fixed h) (Hoeffding)

$$P\Big(|E_{in}(h) - E_{out}(h)| > \epsilon\Big) \le 2e^{-2\epsilon^2 N}$$

Probability of a "bad" event (g selected from a set of M hypothesis)

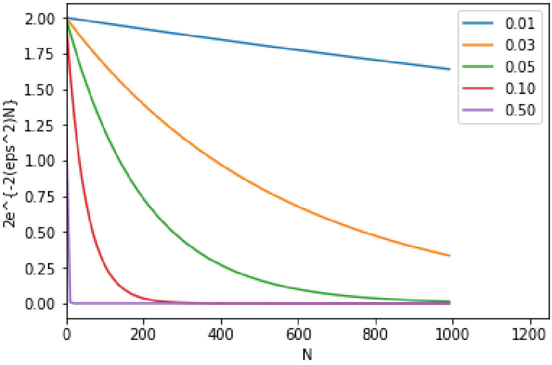
$$P\Big(\,|E_{in}(g) - E_{out}(g)| > \epsilon\,\Big) \le 2Me^{-2\epsilon^2 N}$$

Compare the experiment of **tossing one coin** N **times** with the experiment of **tossing** M **coins,** N **times each**. The chance of a coin resulting in N heads is much larger for the second case.

Recall: Bound variation in function of N

The smaller ϵ, the larger the number of samples needed to keep the probability of "bad" events small (Each color represents a different value of ϵ)

$$P\Big(|E_{in}(h) - E_{out}(h)| > \epsilon\Big) \leq 2e^{-2\epsilon^2 N}$$



$$P\Big(\big| E_{in}(g) - E_{out}(g) \big| > \epsilon \Big) \leq 2Me^{-2\epsilon^2 N}$$

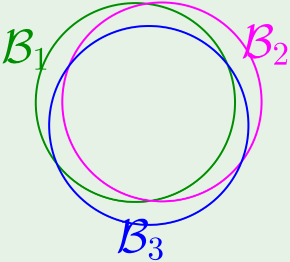
If M is infinite, the bound $2Me^{-2\epsilon^2 N}$ will be large (meaningless)

Can we replace M ?

Where did the M come from?

The $\mathcal{B}$ad events $\mathcal{B}_m$ are

$$\text{``}|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\text{''}$$



The union bound:

$\mathbb{P}[\mathcal{B}_1 \textbf{ or } \mathcal{B}_2 \textbf{ or } \cdots \textbf{ or } \mathcal{B}_M]$

$$\leq \underbrace{\mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + \cdots + \mathbb{P}[\mathcal{B}_M]}_{\text{no overlaps: } M \text{ terms}}$$

The choice of g from $\mathcal{H}$ is affected by D (training data)

Usually there are many similar hypothesis h_j that classify samples in D in the exact same way

If in such a group of hypothesis, there is one that corresponds to a "bad" event, would it not be reasonable to think that other similar hypothesis also correspond to "bad" event ?

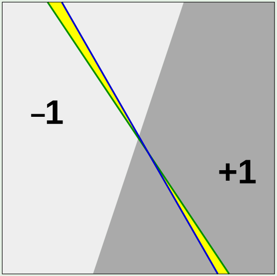
Can we improve on M?

Yes, bad events are *very* overlapping!

ΔE_{out}: change in $+1$ and -1 areas

ΔE_{in}: change in labels of data points

$$|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| \approx |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)|$$



To improve the bound, we will replace the *Union bound* with one that takes the overlap into consideration

For that, we will define a "number" that characterizes the complexity of $\mathcal{H}$

Important definitions:

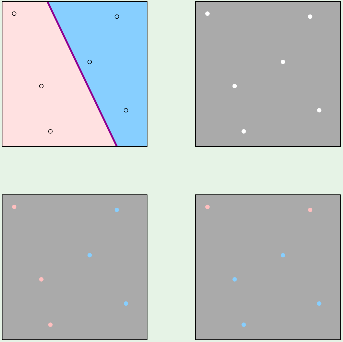
- dichotomy
- growth function
- break point (the "number")

What can we replace M with?

Instead of the whole input space,

we consider a finite set of input points,

and count the number of *dichotomies*



Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ (N points)

Let $\mathcal{H}$ be a hypothesis space

Dichotomies generated by $\mathcal{H}$:
any bipartition of X as $X_{-1} \cup X_{+1}$ that agrees with a hypothesis $h \in \mathcal{H}$

$$\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \Big\{ \big(h(\mathbf{x}_1), h(\mathbf{x}_2), \ldots, h(\mathbf{x}_N) \big) \mid h \in \mathcal{H} \Big\}$$

Dichotomies: mini-hypotheses

A hypothesis $h : \mathcal{X} \to \{-1, +1\}$

A dichotomy $h : \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \to \{-1, +1\}$

Number of hypotheses $|\mathcal{H}|$ can be infinite

Number of dichotomies $|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)|$ is at most 2^N

Candidate for replacing M

Why the number of dichotomies $|\,\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)\,|$ is at most 2^N ?

If you consider another set of points, say, $X' = \{\mathbf{x'}_1, \mathbf{x'}_2, \ldots, \mathbf{x'}_N\}$,

- is $\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \mathcal{H}(\mathbf{x'}_1, \mathbf{x'}_2, \ldots, \mathbf{x'}_N)$?
- is $|\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)| = |\mathcal{H}(\mathbf{x'}_1, \mathbf{x'}_2, \ldots, \mathbf{x'}_N)|$?

The growth function

The growth function counts the <u>most</u> dichotomies on any N points

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)|$$

The growth function satisfies:

$$m_{\mathcal{H}}(N) \leq 2^N$$

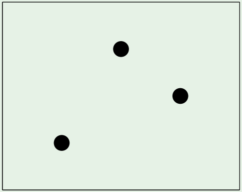
Let's apply the definition.

Growth function for the perceptron

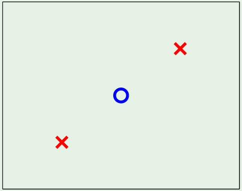
$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)|$$

$m_{\mathcal{H}}(3) = ?$

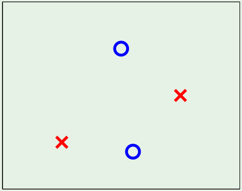
Applying $m_{\mathcal{H}}(N)$ definition - perceptrons



$N = 3$



$N = 3$



$N = 4$

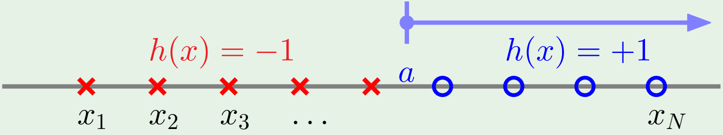
$m_{\mathcal{H}}(3) = 8$ $\qquad$ $m_{\mathcal{H}}(4) = 14$

It may not be easy to compute the growth function for an arbitrary hypothesis set.

Imagine doing that for perceptrons, for each value of N !!

There are, however some simple hypothesis set for which we can write down the growth function in terms of N

Example 1: positive rays

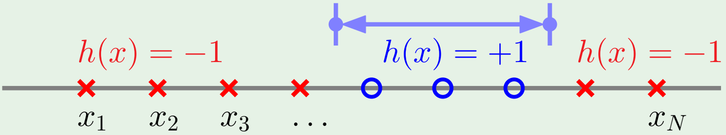


$\mathcal{H}$ is set of $h \colon \mathbb{R} \to \{-1, +1\}$

$h(x) = \text{sign}(x - a)$

$m_{\mathcal{H}}(N) = N + 1$

Example 2: positive intervals



$\mathcal{H}$ is set of $h \colon \mathbb{R} \to \{-1, +1\}$

Place interval ends in two of $N+1$ spots

$$m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

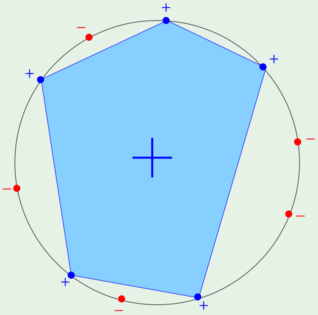
Example 3: convex sets

$\mathcal{H}$ is set of $h\colon \mathbb{R}^2 \to \{-1,+1\}$

$h(\mathbf{x}) = +1$ is convex

$m_{\mathcal{H}}(N) = 2^N$

The N points are 'shattered' by convex sets



The 3 growth functions

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

$$m_{\mathcal{H}}(N) = N + 1$$

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

$$m_{\mathcal{H}}(N) = \tfrac{1}{2}N^2 + \tfrac{1}{2}N + 1$$

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

$$m_{\mathcal{H}}(N) = 2^N$$

Why are we discussing growth functions ?

Back to the big picture

Remember this inequality?

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

What happens if $m_{\mathcal{H}}(N)$ replaces M?

$m_{\mathcal{H}}(N)$ polynomial $\quad\Longrightarrow\quad$ Good!

Just prove that $m_{\mathcal{H}}(N)$ is polynomial?

If the growth function is polynomial, the bound could be made arbitrarily small by choosing an adequate value of N.

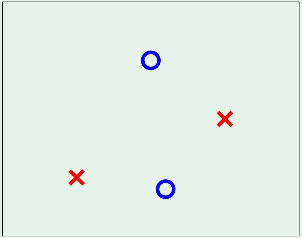
Do we need to compute the growth function value for each N ?

Break point of $\mathcal{H}$

Definition:

If no data set of size k can be shattered by $\mathcal{H}$, then k is a *break point* for $\mathcal{H}$

$$m_{\mathcal{H}}(k) \; < \; 2^k$$



For 2D perceptrons, $k = 4$

A bigger data set cannot be shattered either

Break point - the 3 examples

- Positive rays $m_{\mathcal{H}}(N) = N + 1$

break point $k = 2$

- Positive intervals $m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

break point $k = 3$

- Convex sets $m_{\mathcal{H}}(N) = 2^N$

break point $k =$'∞'

An exercise

Assume that for a certain hypothesis set $\mathcal{H}$ the break-point is 2

This means that $\mathcal{H}$ can not generate the $2^2 = 4$ possible dichotomies for any subset of two samples $\{x_1, x_2\}$.

Under such supposition, how many dichotomies are possible when we consider three samples $\{x_1, x_2, x_3\}$?

Puzzle

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
○	○	○
○	○	●
○	●	○
●	○	○

Main result

No break point $\implies$ $m_{\mathcal{H}}(N) = 2^N$

Any break point $\implies$ $m_{\mathcal{H}}(N)$ is **polynomial** in N

Summary

1. We started searching a replacement for M

$$P\left(\left|E_{in}(g) - E_{out}(g)\right| > \epsilon \right) \leq 2Me^{-2\epsilon^2 N}$$

2. **Dichotomies**: to deal with the issue of overlapping "bad" events.
   - The complexity of $\mathcal{H}$ is related to the number of dichotomies it can generate

3. **Growth function**: number of dichotomies for each N
   - Polynomial growth functions are good candidate for replacing M
   - Not always possible to write this function

4. **Break-point**: if it is finite, it means that the growth function is polynomial (to be demonstrated)

5. Next meeting

   - if there is a finite break-point, then the growth function is polynomial
   - it is valid to replace M with the growth function