The final exam

Testing:

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

Training:

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

Where did the M come from?

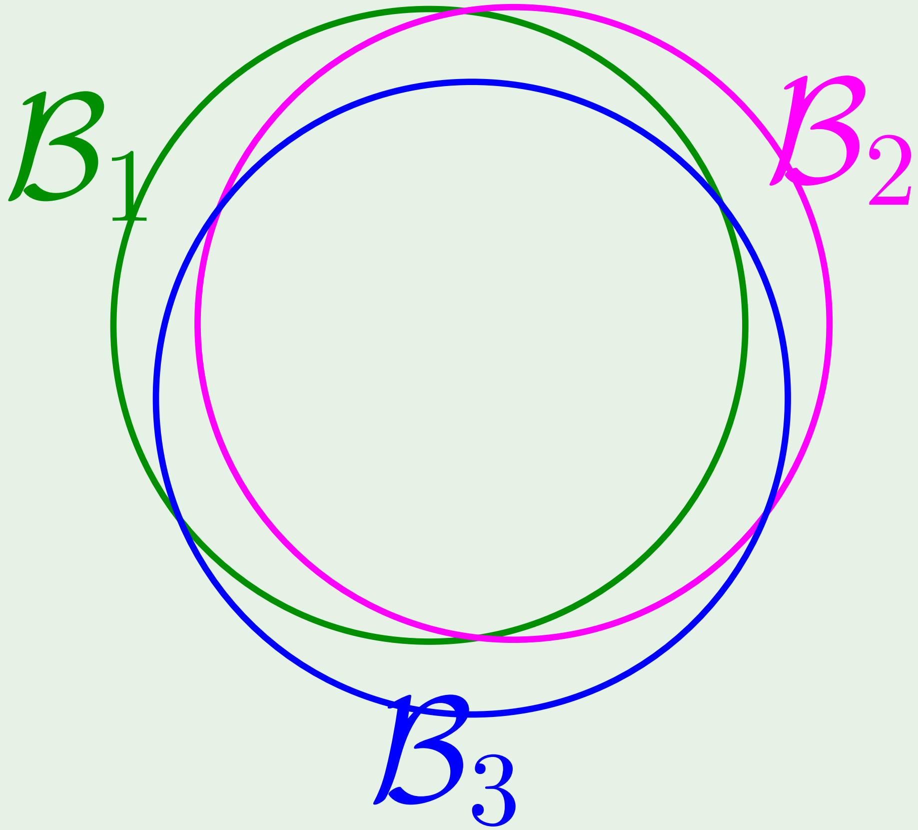
The $\mathcal{B}$ad events $\mathcal{B}_m$ are

$$\text{“}|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\text{”}$$

The union bound:

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

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



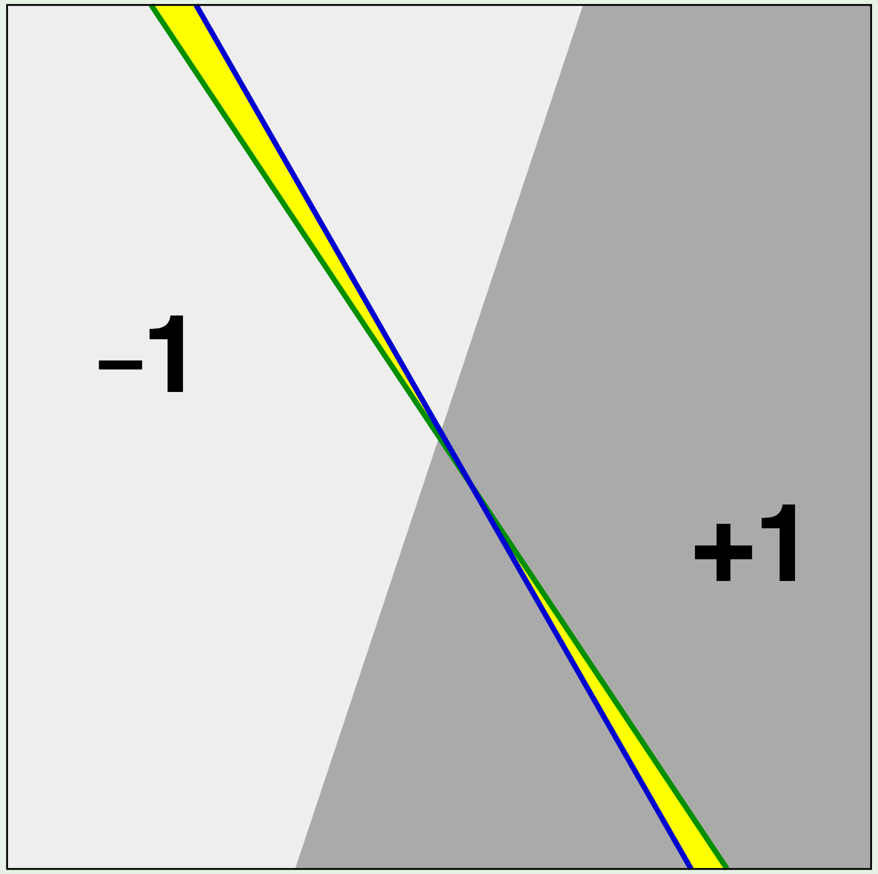
Can we improve on M?

Yes, bad events are *very* overlapping!

ΔE_{out}: change in $+1$ and -1 areas

ΔE_{in}: change in labels of data points

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

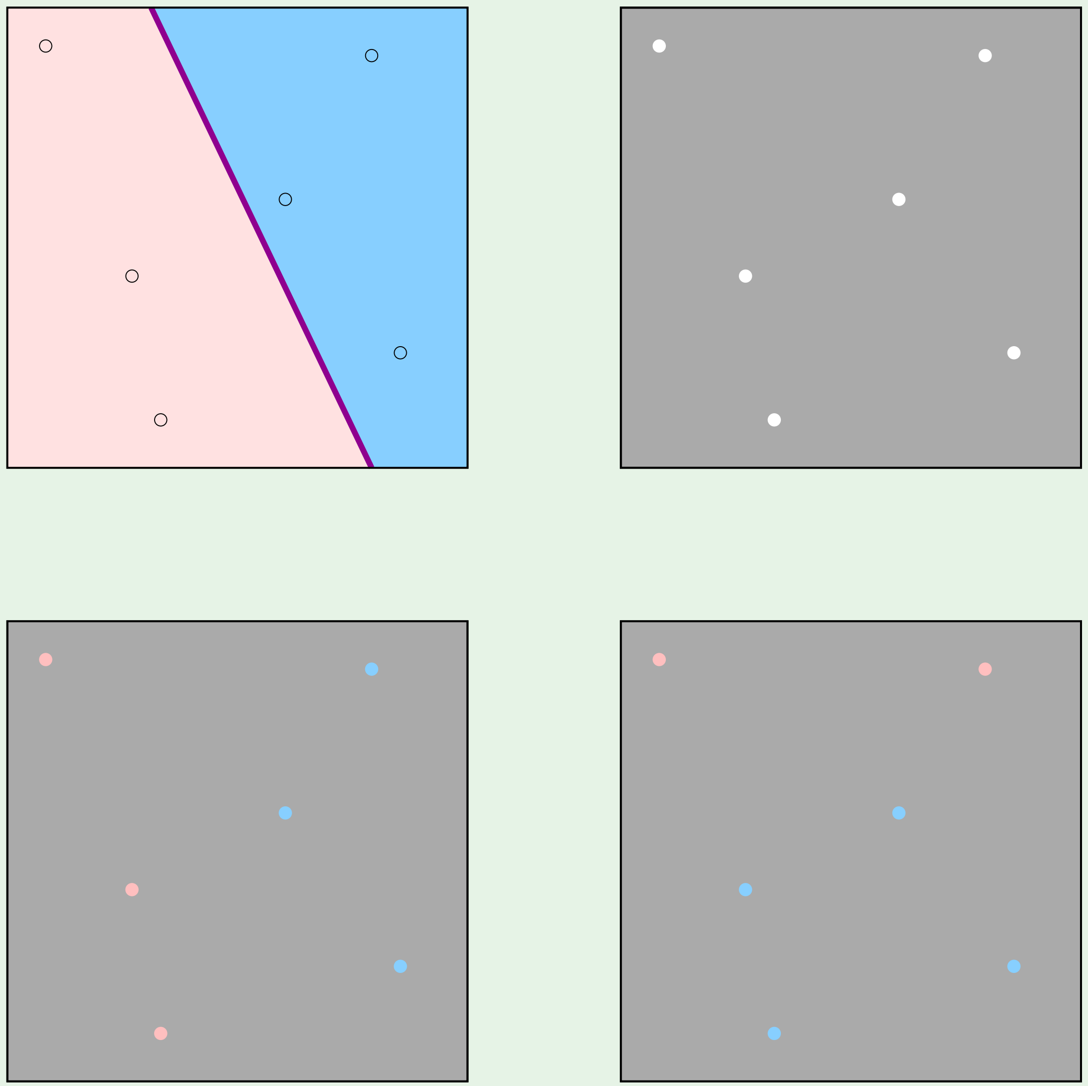


What can we replace M with?

Instead of the whole input space,

we consider a finite set of input points,

and count the number of *dichotomies*



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

The growth function

The growth function counts the most dichotomies on any N points

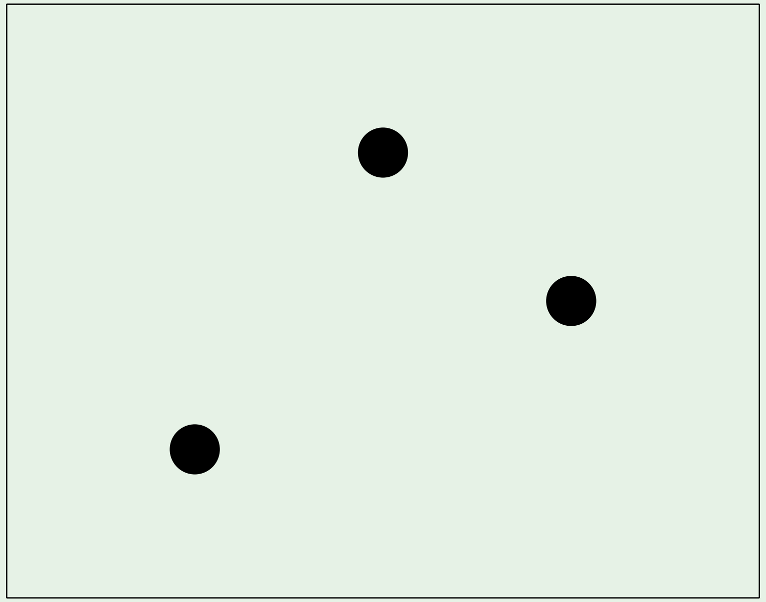
$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)|$$

The growth function satisfies:

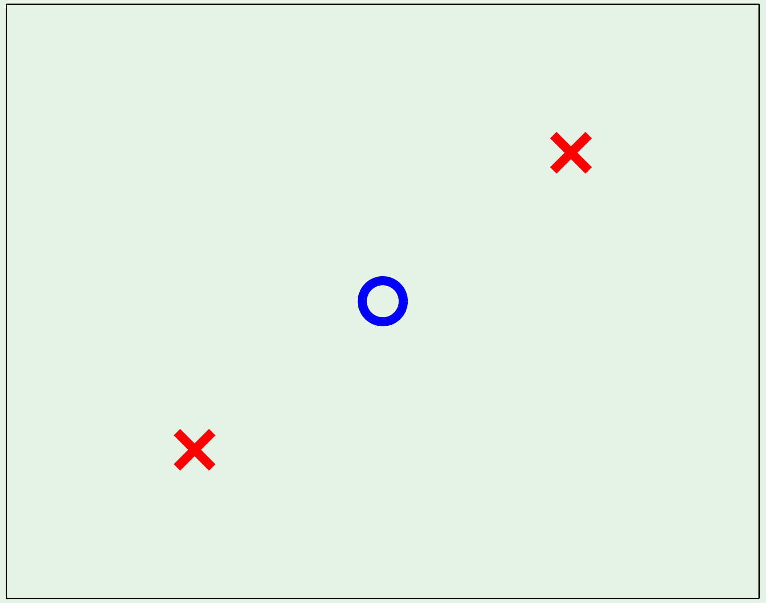
$$m_{\mathcal{H}}(N) \leq 2^N$$

Let's apply the definition.

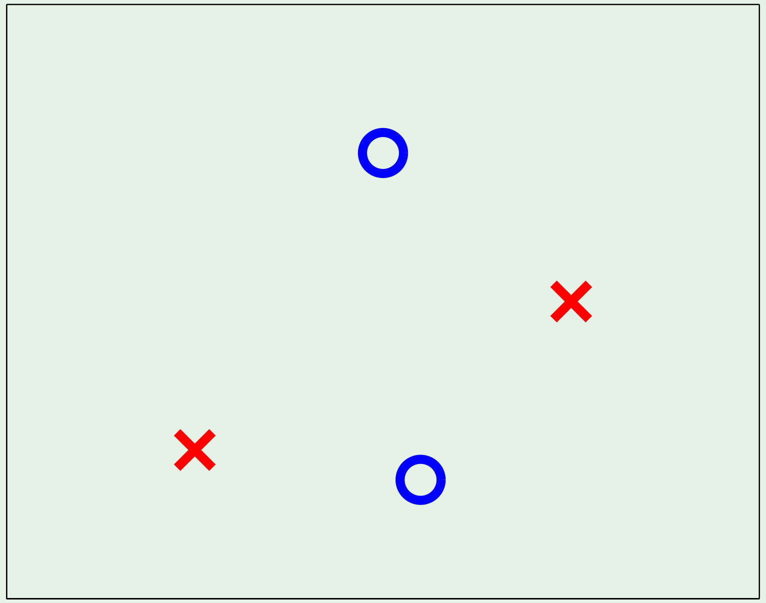
Applying $m_{\mathcal{H}}(N)$ definition - perceptrons



$N = 3$



$N = 3$



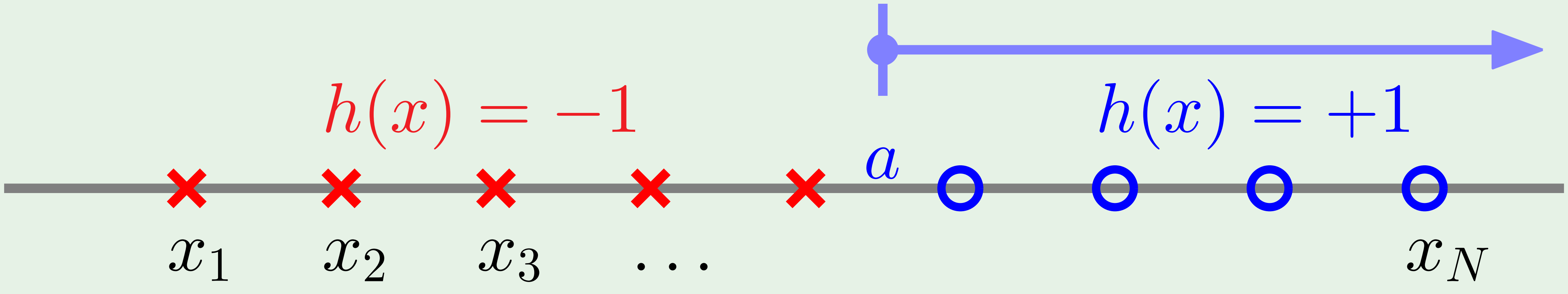
$N = 4$

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

Outline

- From training to testing
- Illustrative examples
- Key notion: break point
- Puzzle

Example 1: positive rays

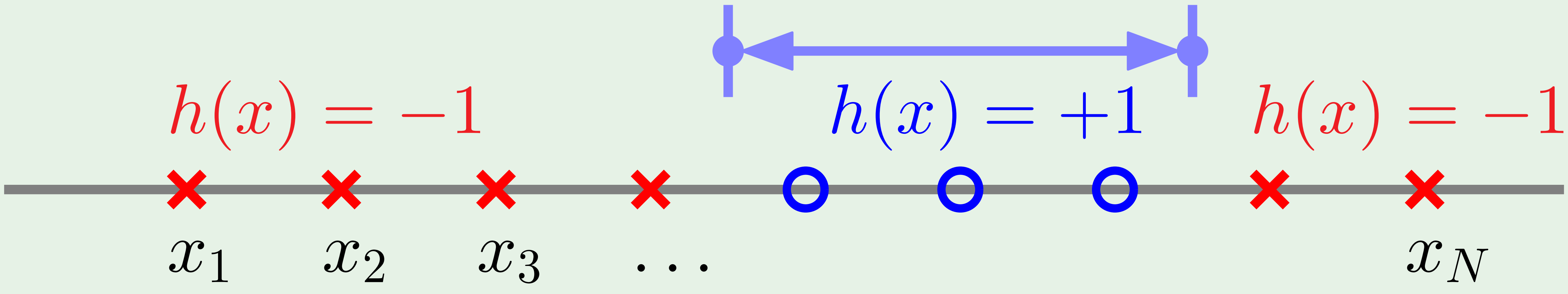


$\mathcal{H}$ is set of $h: \mathbb{R} \to \{-1, +1\}$

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

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

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

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

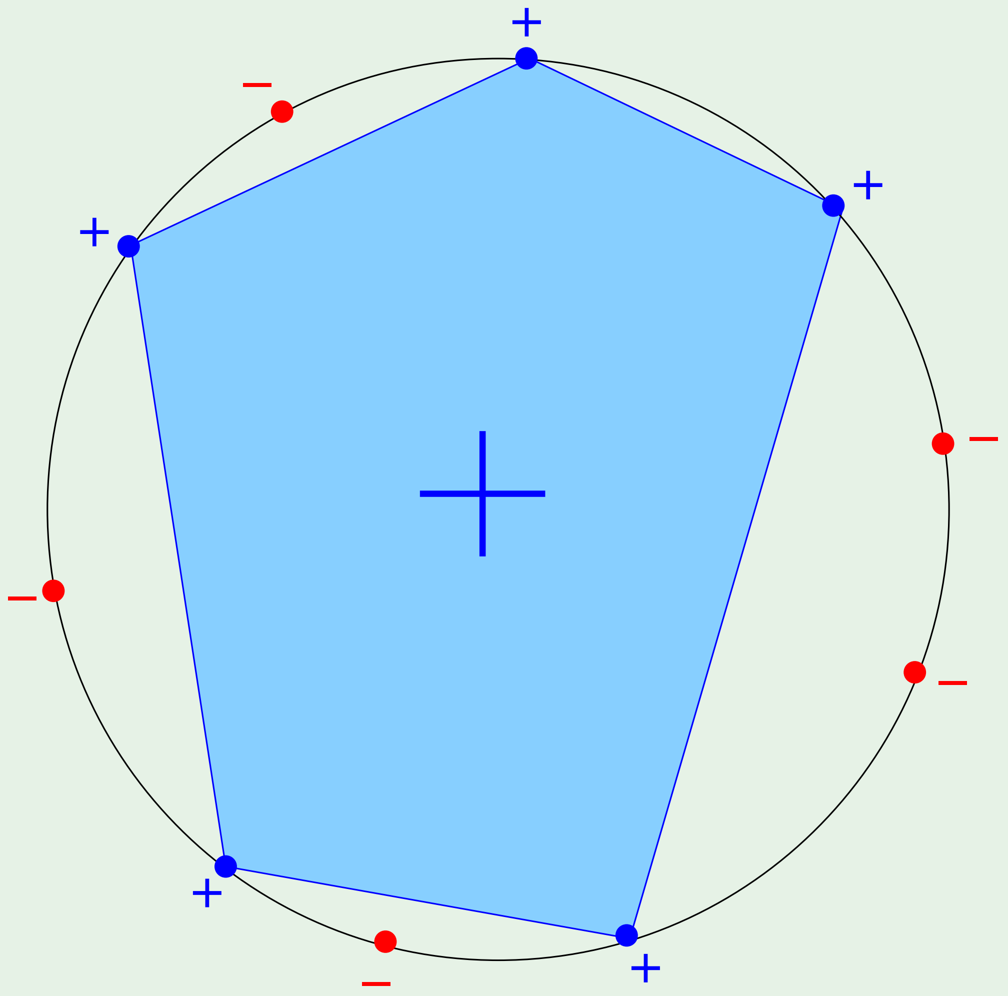
Example 3: convex sets

$\mathcal{H}$ 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:

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

Back to the big picture

Remember this inequality?

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

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?

Outline

- From training to testing
- Illustrative examples
- Key notion: **break point**
- Puzzle

Break point of $\mathcal{H}$

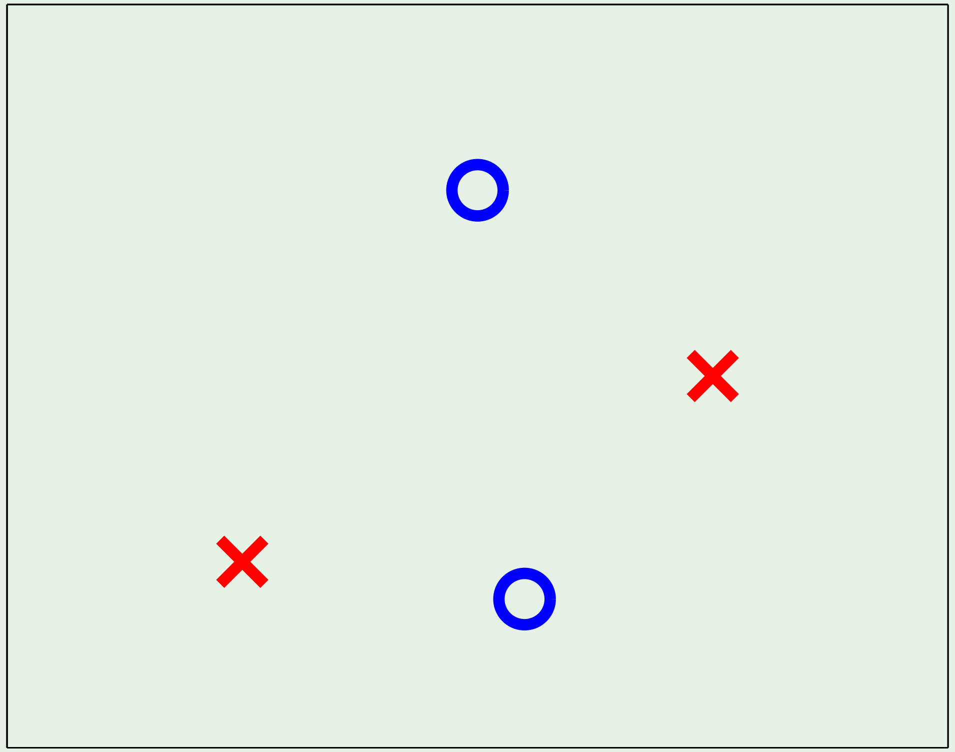
Definition:

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

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

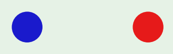
For 2D perceptrons, $k = 4$

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 $k =$'∞'

Main result

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

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

Rearranging things

Start from the VC inequality:

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

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 $\geq 1-\delta$, $\qquad |E_{\text{out}} - E_{\text{in}}| \leq \Omega(N, \mathcal{H}, \delta)$

Generalization bound

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

$$\Longrightarrow$$

With probability $\geq 1 - \delta$,

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