Machine Learning

Lecture 7: Theory of Generalization (I)

王浩

信息科学与技术学院

Email: wanghao1@shanghaitech.edu.cn

本节内容

- Growth function
- Break point
- Shattering
- Bounding the growth function

Where we left...

Testing:

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

Training:

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

Infinite hypothesis set

Want:

- establish a finite quantity that replaces M

$$\mathbb{P}[\,|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2M_{\mathcal{H}}e^{-2\epsilon^2 N}$$

- justify the feasibility of learning for infinite M
- study $M_{\mathcal{H}}$ to understand its trade-off for "right" $\mathcal{H}$, just like M

How comes this M in finite case?

- Bad events $\mathcal{B}_i$: $|E_{in}(h_i) - E_{out}(h_i)| > \epsilon$
- to give freedom of choice: bound $\mathbb{P}[\mathcal{B}_1 \ or \ \mathcal{B}_2 \ or ... \mathcal{B}_M]$
- worst case: all $\mathcal{B}_i$ non-overlapping (union bound)

$$\mathbb{P}[\mathcal{B}_1 \ or \ \mathcal{B}_2 \ or ... \mathcal{B}_M] \leq \mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + ... + \mathbb{P}[\mathcal{B}_M]$$

Where did the M come from?

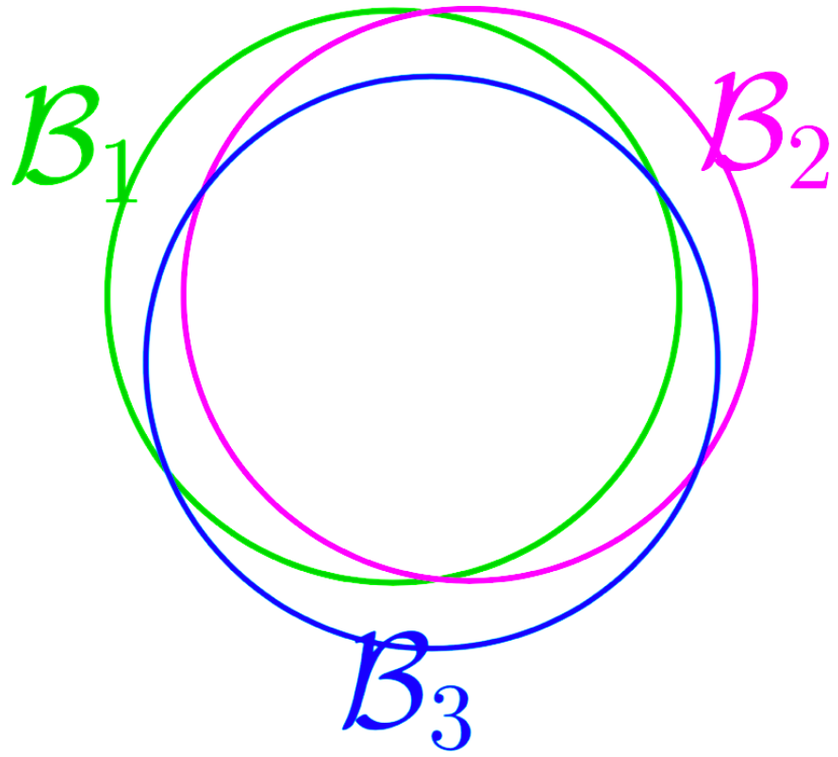
The $\mathcal{B}$ad events $\mathcal{B}_m$ are

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

The union bound:

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

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

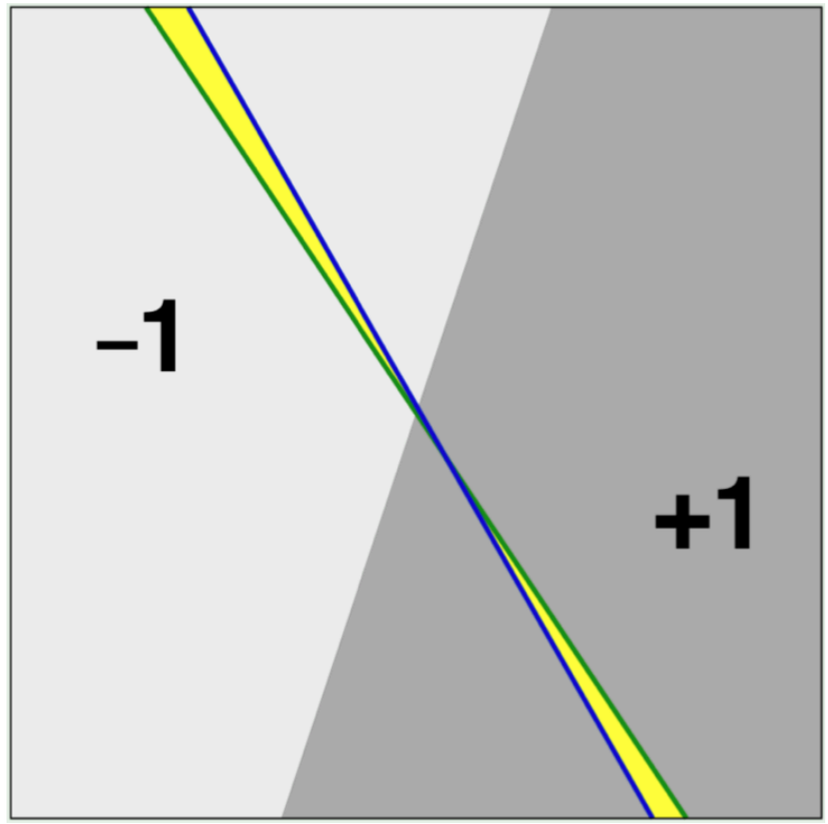


Can we improve on M?

Yes, bad events are very overlapping!

ΔE_{out} : change in $+1$ and -1 areas

ΔE_{out} : change in labels of data points



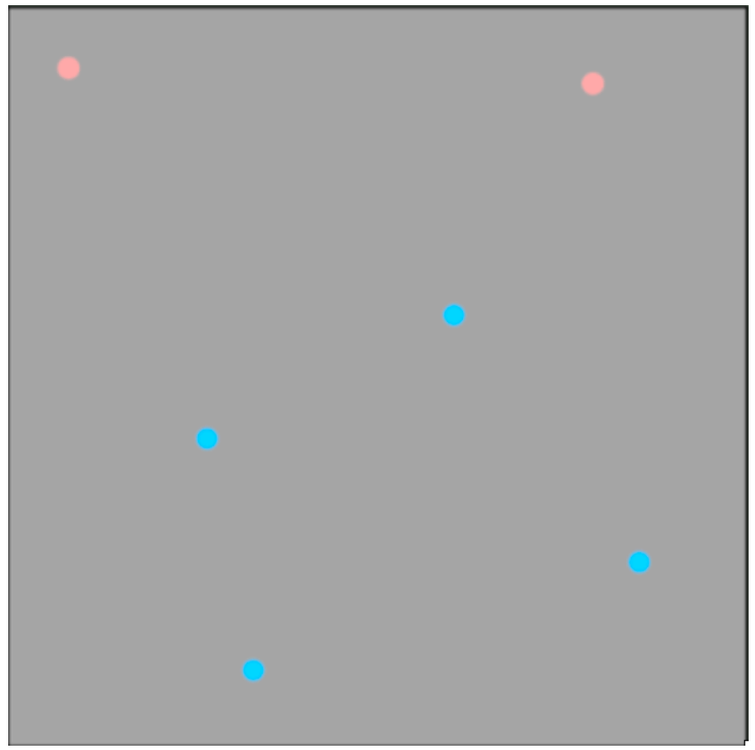
$$|E_{in}(h_1) - E_{out}(h_1)| \approx |E_{in}(h_2) - E_{out}(h_2)|$$

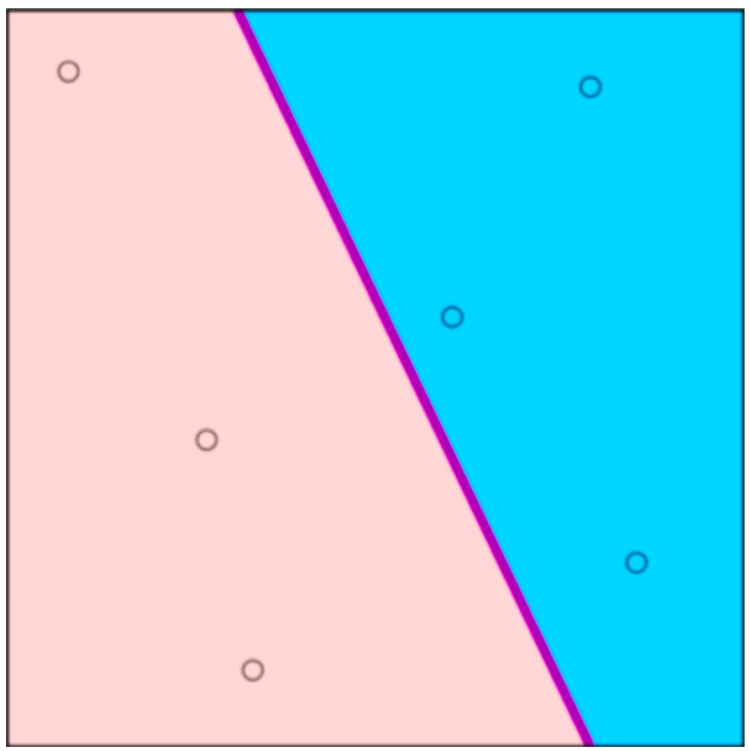
What can we replace M with?

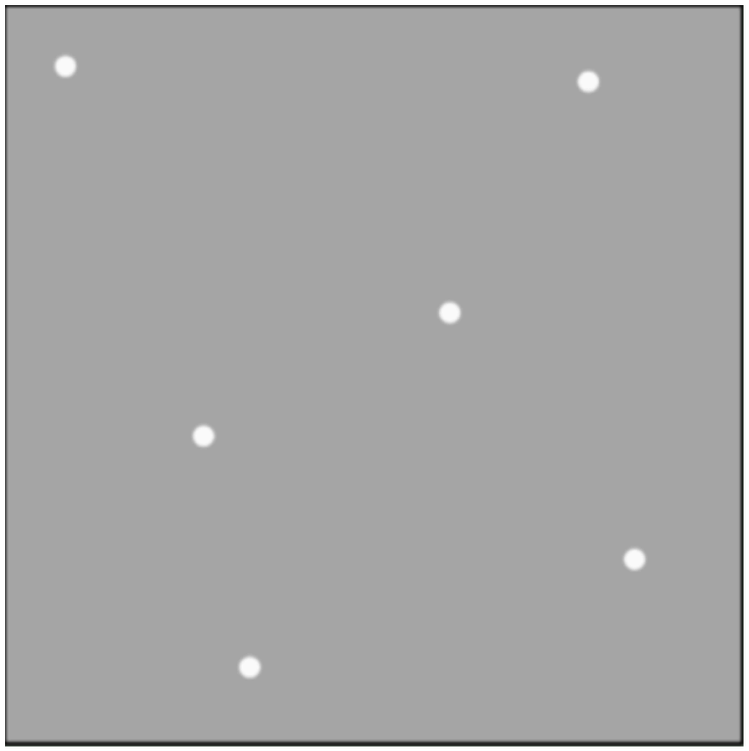
Instead of the whole input space,

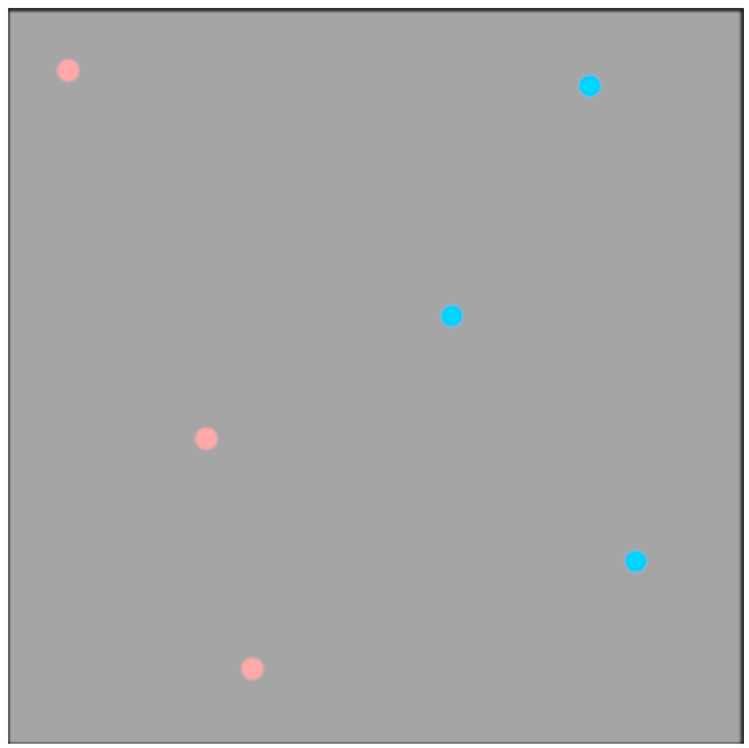
we consider a finite set of input points,

and count the number of dichotomies









Dichotomies: mini-hypotheses

- A hypothesis $h : \mathcal{X} \to \{-1, +1\}$
- A dichotomy $h : \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \to \{-1, +1\}^N$
- Number of hypotheses: $|\mathcal{H}|$ can be infinite
- Number of dichotomies $|\mathcal{H}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)|$ is at most 2^N
- Candidate for replacing M

The growth function of $\mathscr{H}$

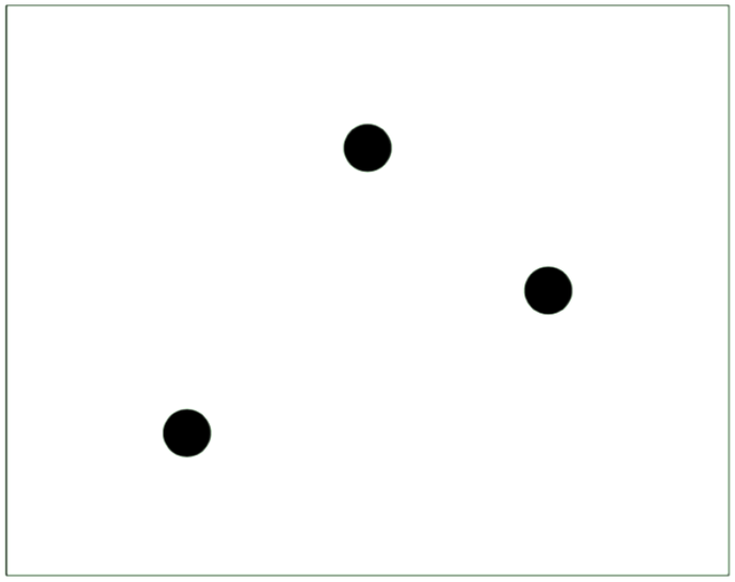
- The growth function counts the most dichotomies on any N points

$$m_{\mathscr{H}}(N) = \max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N \in \mathscr{X}} \left| \mathscr{H}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \right|$$

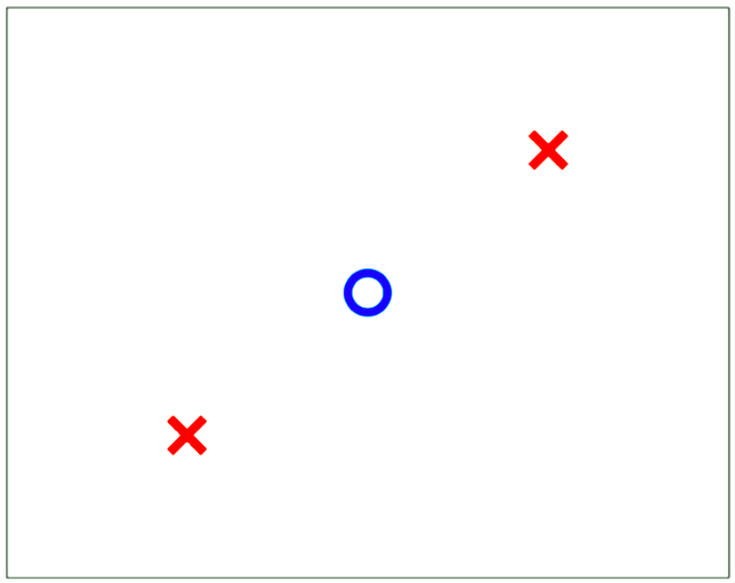
- The growth function satisfies

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

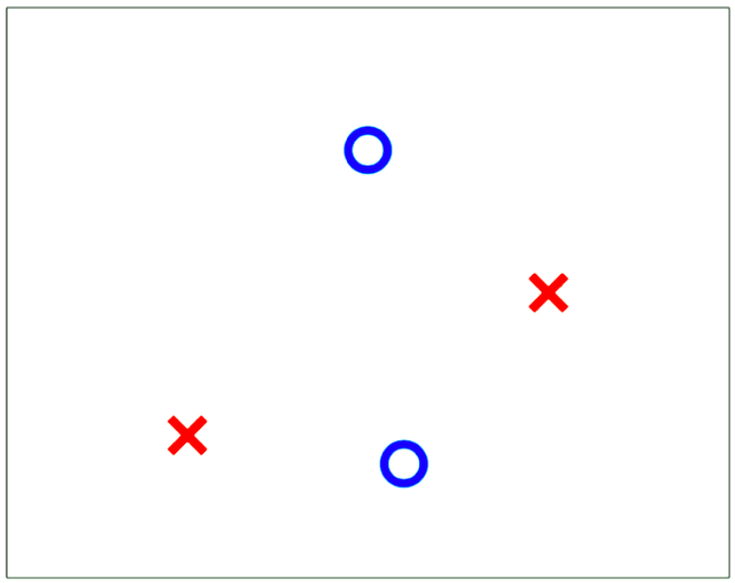
Applying $m_{\mathcal{H}}(N)$ definition — perceptrons



$N = 3$



$N = 3$



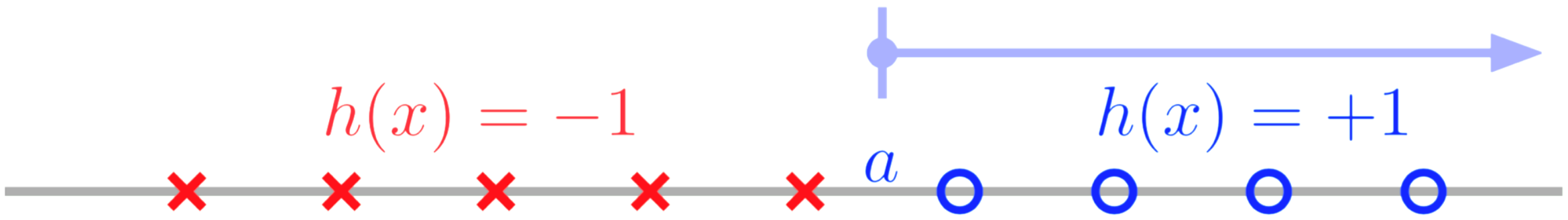
$N = 4$

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

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

Illustrative examples

Example 1: positive rays



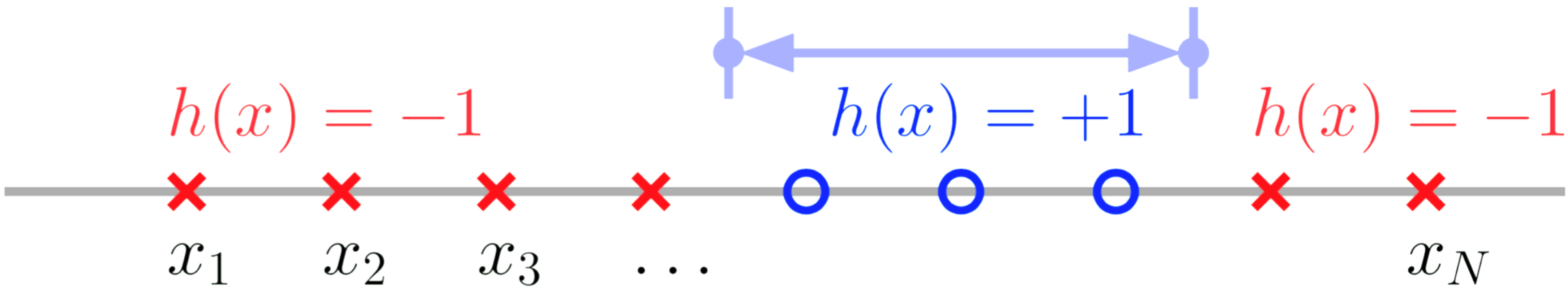
$\mathcal{H}$ is set of $h : \mathbb{R} \to \{-1, +1\}$

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

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

Illustrative examples

Example 2: positive intervals



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

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

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

Illustrative examples

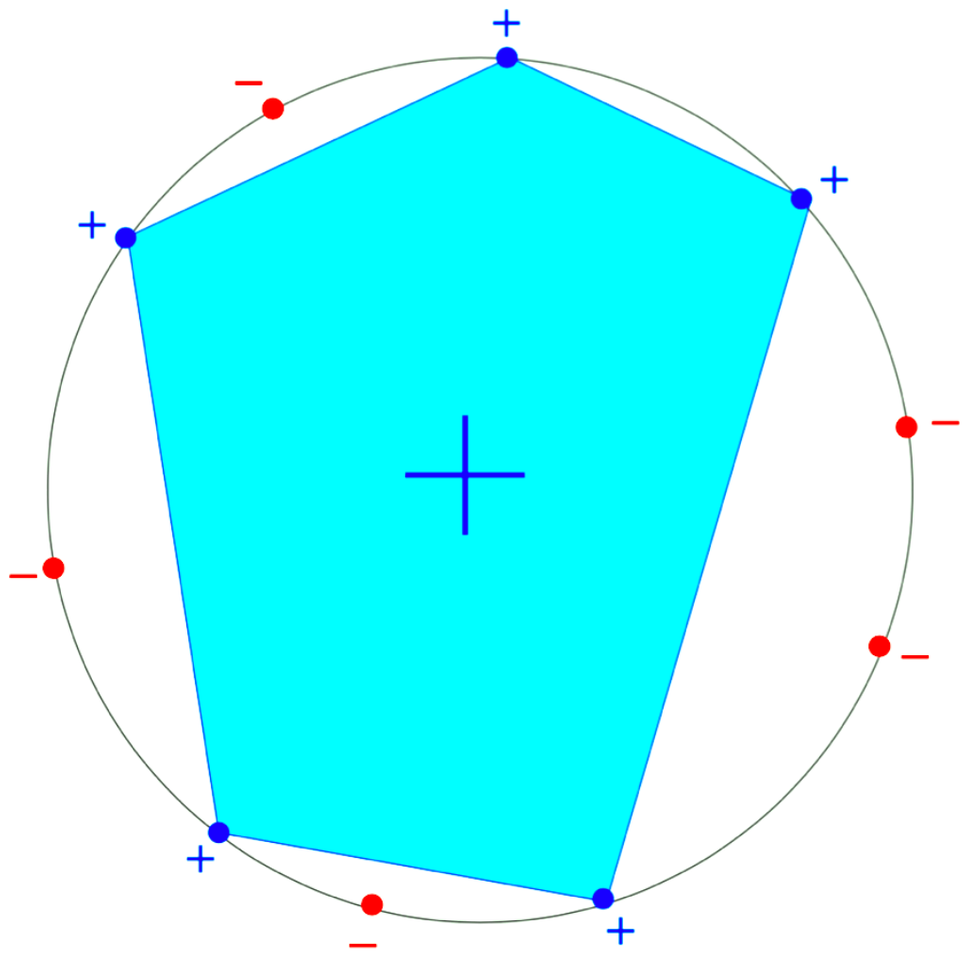
Example 3: convex sets

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

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

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

The N points are 'shattered' by convex sets



Different $\mathscr{H}$ has different growth functions...

The 3 growth functions

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

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

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

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

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

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

Back to the big picture...

Remember this inequality

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

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

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

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

Breaking point of $\mathscr{H}$

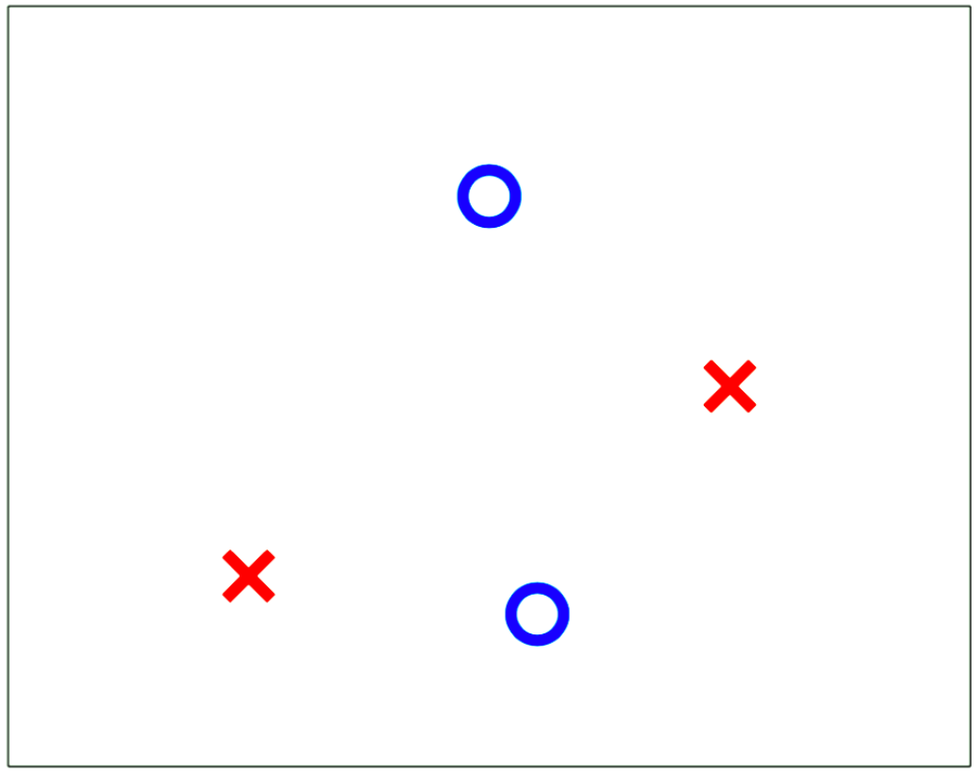
Definition:

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



Being able to achieve any labeling of a given set of points is also known as *shattering* the points

We say a classifier $h(\mathbf{x})$ can shatter points $\mathbf{x}^1, \ldots, \mathbf{x}^m$ if and only if: for all $y^1, \ldots, y^m$, $h(\mathbf{x})$ can achieve zero error on training data $(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^m, y^m)$

Breaking point of 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 $\implies$ $m_{\mathcal{H}}(N) = 2^N$

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

we mean polynomial "after" this break point, or, for sufficiently large N

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

- Proof that $m_{\mathcal{H}}(N)$ can replace M

Shattering and break point

If $\mathcal{H}$ is capable of generating all possible dichotomies on $\mathbf{x}_1, \cdots, \mathbf{x}_N$, then $\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \{-1, +1\}^N$ and we say that $\mathcal{H}$ can *shatter* $\mathbf{x}_1, \cdots, \mathbf{x}_N$. This signifies that $\mathcal{H}$ is as diverse as can be on this particular sample.

“打散”是对给定的特定数据集而言的。

Definition 2.3. *If no data set of size k can be shattered by $\mathcal{H}$, then k is said to be a break point for $\mathcal{H}$.*

”断点“是对任意”摆放“的k个数据点组成的数据集而言的。

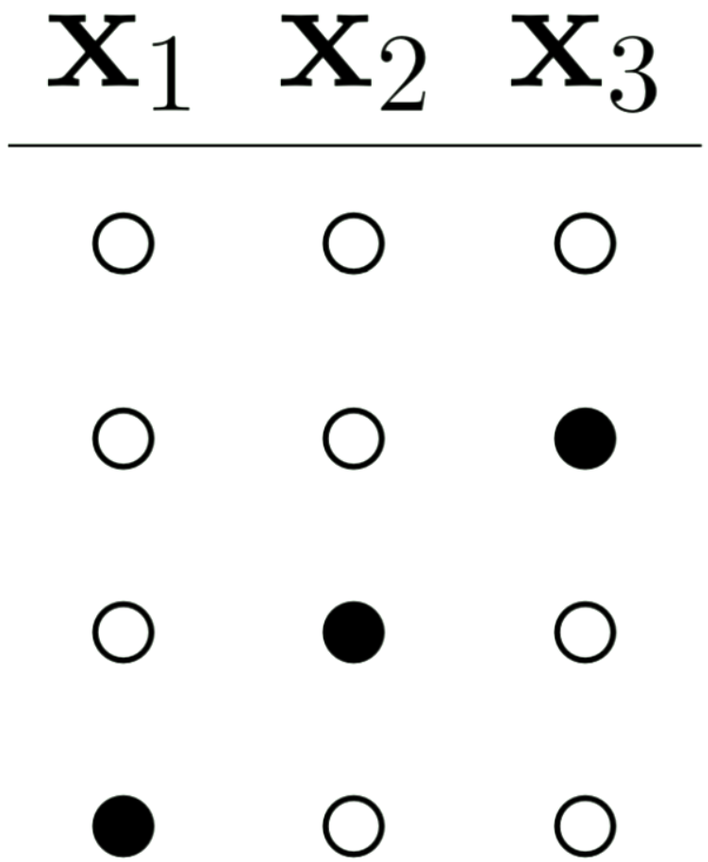
Definition 2.4. *$B(N, k)$ is the maximum number of dichotomies on N points such that no subset of size k of the N points can be shattered by these dichotomies.*

$B(N, k)$是为了考察在数据点多于k的时候，分类器还能产生多少”花样“而定义的数量。之所以要用到”maximum number”, 也是要考虑到不同类别的分类器虽然断点(最多)是k，但他们在N个数据点上的”花样“数目并不同。

Small example：

假设$k = 2$，现在给出$N = 3$,$k = 2$时$B(N, k)$的值

Claim $B(N, k)$ = 4 < 8..... How would you verify this?



Bounding $m_{\mathscr{H}}(N)$

To show: $m_{\mathcal{H}}(N)$ is polynomial

We show: $m_{\mathcal{H}}(N) \leq \cdots \leq \cdots \leq$ a polynomial

Key quantity: $B(N,k)$的定义是独立于$\mathscr{H}$的

$B(N,k)$: Maximum number of dichotomies on N points, with break point k

Recursive bound on $B(N,k)$

Consider the following table:

$B(N,k) = \alpha + 2\beta$

		# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
S_1		α	$+1$	$+1$	...	$+1$	$+1$
			-1	$+1$	...	$+1$	-1
			⋮	⋮	⋮	⋮	⋮
			$+1$	-1	...	-1	-1
			-1	$+1$	...	-1	$+1$
S_2	S_2^+	β	$+1$	-1	...	$+1$	$+1$
			-1	-1	...	$+1$	$+1$
			⋮	⋮	⋮	⋮	⋮
			$+1$	-1	...	$+1$	$+1$
			-1	-1	...	-1	$+1$
	S_2^-	β	$+1$	-1	...	$+1$	-1
			-1	-1	...	$+1$	-1
			⋮	⋮	⋮	⋮	⋮
			$+1$	-1	...	$+1$	-1
			-1	-1	...	-1	-1

锁定x_N，前面$N-1$个数据点上的”花样“必然有出现两次的（对应着x_N取$+1,-1$），有出现一次的

S_1, S_2^+, S_2^-都是此空间内的分类器

Estimating α and β

Focus on $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N-1}$ columns:

$$\alpha + \beta \leq B(N-1, k)$$

对应着这次特例，我们已经能在$N-1$个数据点上找到$\alpha+\beta$个"花样"，根据$B(N,k)$的定义，则肯定比这个数要大

	# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
S_1	α	$+1$	$+1$	...	$+1$	$+1$
		-1	$+1$	...	$+1$	-1
		⋮	⋮	⋮	⋮	⋮
		$+1$	-1	...	-1	-1
		-1	$+1$	...	-1	$+1$
S_2	β	$+1$	-1	...	$+1$	$+1$
S_2^+		-1	-1	...	$+1$	$+1$
		⋮	⋮	⋮	⋮	⋮
		$+1$	-1	...	$+1$	$+1$
		-1	-1	...	-1	$+1$
S_2	β	$+1$	-1	...	$+1$	-1
S_2^-		-1	-1	...	$+1$	-1
		⋮	⋮	⋮	⋮	⋮
		$+1$	-1	...	$+1$	-1
		-1	-1	...	-1	-1

Now, focus on the $S_2 = S_2^+ \cup S_2^-$ rows:

$$\beta \leq B(N-1, k-1)$$

对于$B(N-1,k-1)$我们确实不能直接计算，但是我们知道S_2^+中任意$k-1$个点不能被"打散"。因为一旦这被允许，也就是说明S_2^+中$k-1$个点能产生2^{k-1}个"花样"，那么把S_2^-拼上，再把x_N补回去，就能知道在此数据集上有k个数据点可以产生2^k个"花样"，和k为断点矛盾！！那么这个特例就证明了≤

这也是根据B的定义！！

		# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
S_1		α	$+1$	$+1$	...	$+1$	$+1$
			-1	$+1$	...	$+1$	-1
			⋮	⋮	⋮	⋮	⋮
			$+1$	-1	...	-1	-1
			-1	$+1$	...	-1	$+1$
S_2	S_2^+	β	$+1$	-1	...	$+1$	$+1$
			-1	-1	...	$+1$	$+1$
			⋮	⋮	⋮	⋮	⋮
			$+1$	-1	...	$+1$	$+1$
			-1	-1	...	-1	$+1$
	S_2^-	β	$+1$	-1	...	$+1$	-1
			-1	-1	...	$+1$	-1
			⋮	⋮	⋮	⋮	⋮
			$+1$	-1	...	$+1$	-1
			-1	-1	...	-1	-1

Putting it together

$$B(N,k) = \alpha + 2\beta$$

$$\alpha + \beta \leq B(N-1,k)$$

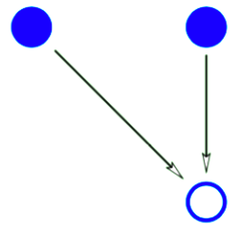
$$\beta \leq B(N-1,k-1)$$

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

		# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$
S_1		α	$+1$	$+1$	...	$+1$	$+1$
			-1	$+1$	...	$+1$	-1
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	...	-1	-1
			-1	$+1$	...	-1	$+1$
S_2	S_2^+	β	$+1$	-1	...	$+1$	$+1$
			-1	-1	...	$+1$	$+1$
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	...	$+1$	$+1$
			-1	-1	...	-1	$+1$
	S_2^-	β	$+1$	-1	...	$+1$	-1
			-1	-1	...	$+1$	-1
			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$+1$	-1	...	$+1$	-1
			-1	-1	...	-1	-1

Numerical computation of $B(N,k)$ bound

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



Baseline：

$B(N,1) = 1$，显然我们知道对于断点为1，多少个数据点都只能有1种"花样"，否则必然有某数据点上有2个以上的"花样"，矛盾！

$B(1,k) = 2$，断点很大，数据点很少。所以我们肯定最多能搞出2个"花样"。

		k						
		1	2	3	4	5	6	. .
	1	1	2	2	2	2	2	. .
	2	1	3	4	4	4	4	. .
	3	1	4	7	8	8	8	. .
N	4	1	5	11	. .	. .	. .	. .
	5	1	6	:	.			
	6	1	7	:		.		
	:	:	:	:			.	

虽然理论上这张表是$B(N,k)$的一个上界，但这个界还是比较准的

Analytic solution for $B(N,k)$ bound

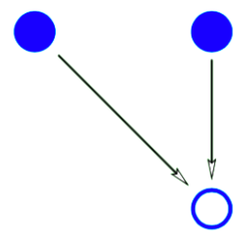
$$B(N,k) \leq B(N-1,k) + B(N-1,k-1)$$

Theorem:

$$B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

1. **Boundary conditions:** easy

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

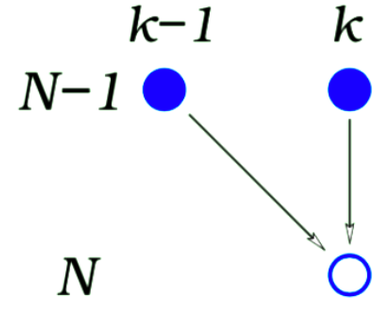


$$B(N,k) \le B(N-1,k) + B(N-1,k-1)$$

2. The induction step

根据归纳法假设前提，这两项的上界是下面这两项

$$\sum_{i=0}^{k-1} \binom{N-1}{i} + \sum_{i=0}^{k-2} \binom{N-1}{i} \; ?$$



$$= 1 + \sum_{i=1}^{k-1} \binom{N-1}{i} + \sum_{i=1}^{k-1} \binom{N-1}{i-1}$$

$$= 1 + \sum_{i=1}^{k-1} \left[\binom{N-1}{i} + \binom{N-1}{i-1} \right]$$

$$= 1 + \sum_{i=1}^{k-1} \binom{N}{i} = \sum_{i=0}^{k-1} \binom{N}{i} \; ✓$$

这两个上界之和小于（等于）这一项，那么$B(N,k)$肯定比它小了

We've derived a polynomial upper bound!!!

For a given $\mathcal{H}$, the break point k is fixed

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

我们根本不需要知道$\mathscr{H}$具体是什么，只需要知道他有"断点"即可！！

很多常见的分类器是有断点的。这就和VC维度有关了。

Three examples

$$\sum_{i=0}^{k-1} \binom{N}{i}$$

- $\mathcal{H}$ is **positive rays**: (break point $k = 2$)

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

- $\mathcal{H}$ is **positive intervals**: (break point $k = 3$)

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

- $\mathcal{H}$ is **2D perceptrons**: (break point $k = 4$)

$$m_{\mathcal{H}}(N) = \ ? \quad \leq \quad \frac{1}{6}N^3 + \frac{5}{6}N + 1$$

What we want

Instead of:

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

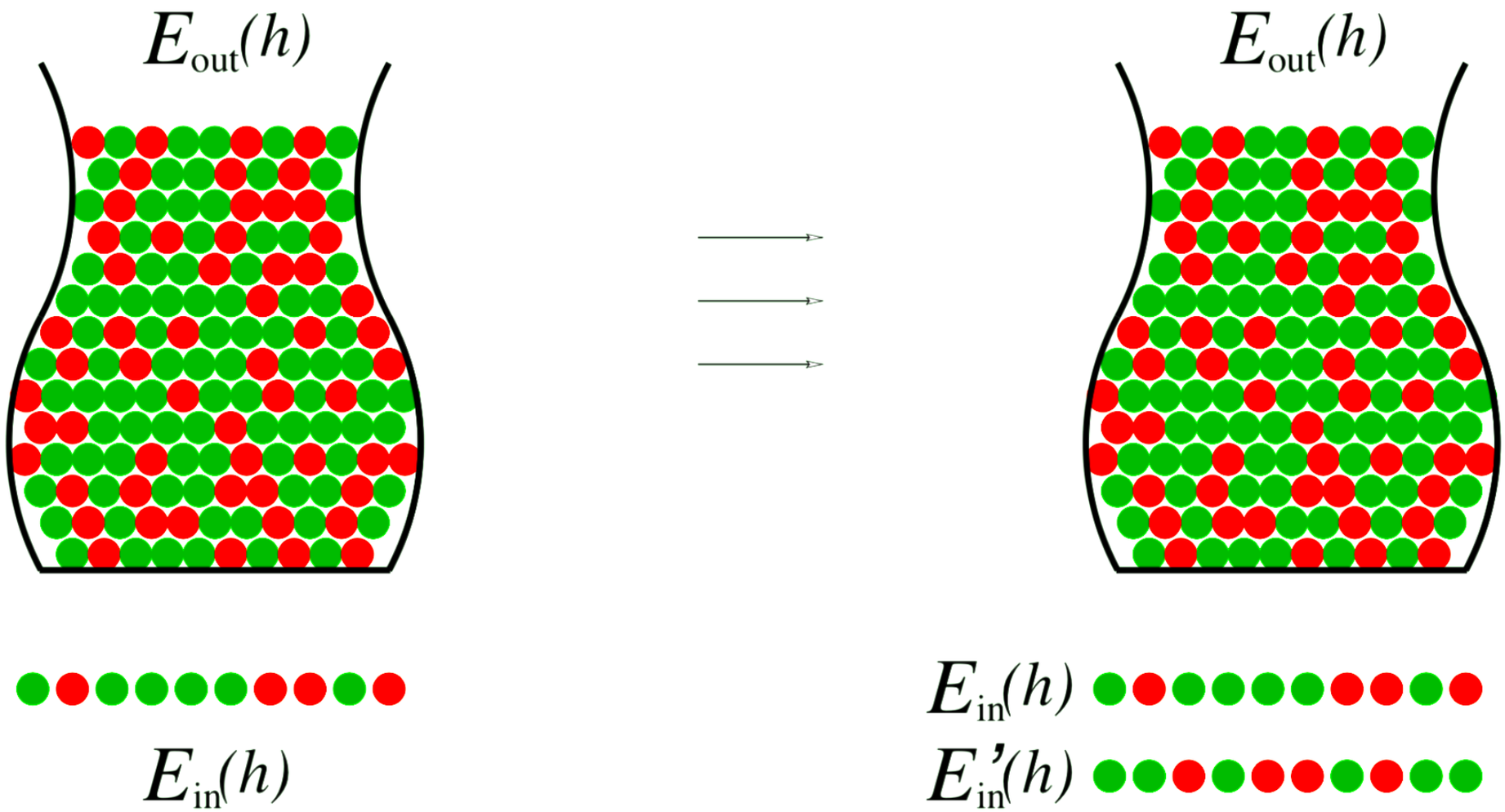
We want:

$$\mathbb{P}[\ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\ ] \ \leq \ 2 \ m_{\mathcal{H}}(N) \ e^{-2\epsilon^2 N}$$

完美达成目标！！！但还缺最后一步！！！

这个位置真的放的是$m_{\mathcal{H}}(N)$吗？虽然有限元素时没毛病

What to do about E_{out}



Putting it together

Not quite:

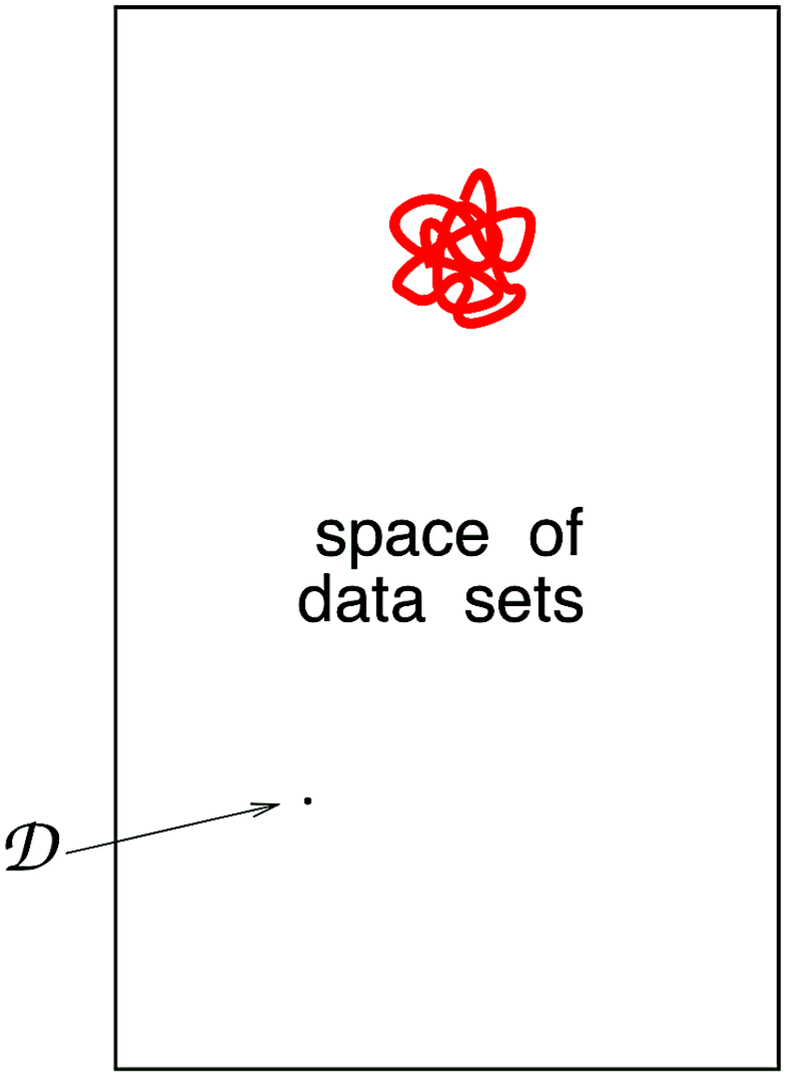
$$\mathbb{P}[\ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\ ] \ \leq\ 2\ \ m_{\mathcal{H}}(\ N\ )\ \ e^{-\,2\,\epsilon^2 N}$$

but rather:

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

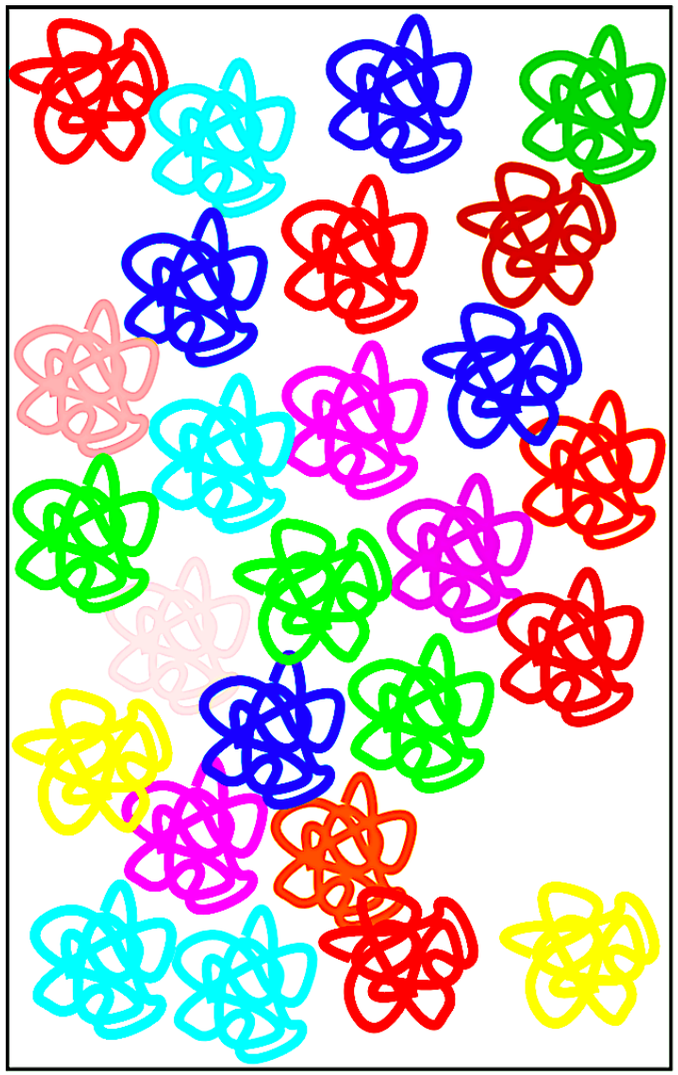
The Vapnik-Chervonenkis Inequality

Hoeffding Inequality



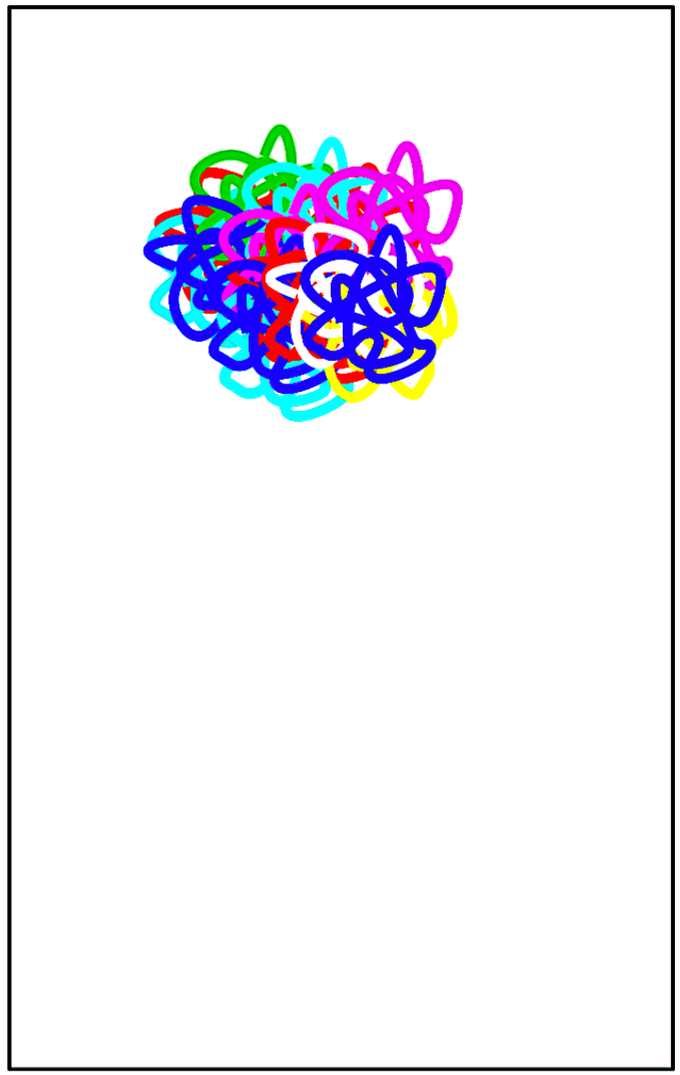
(a)

Union Bound



(b)

VC Bound



(c)