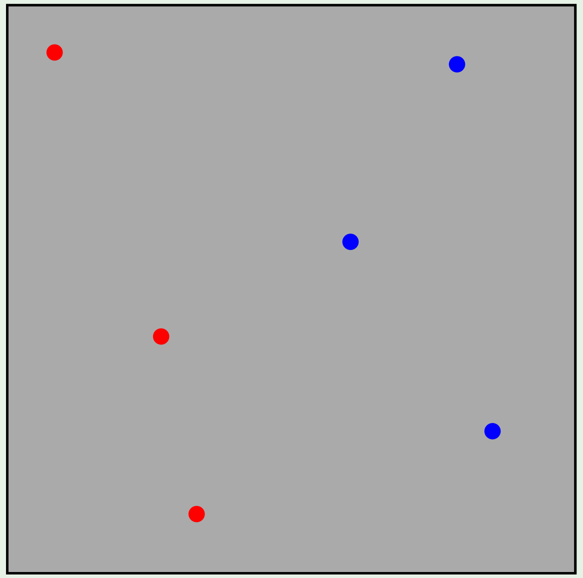
Review of Lecture 5

- Dichotomies

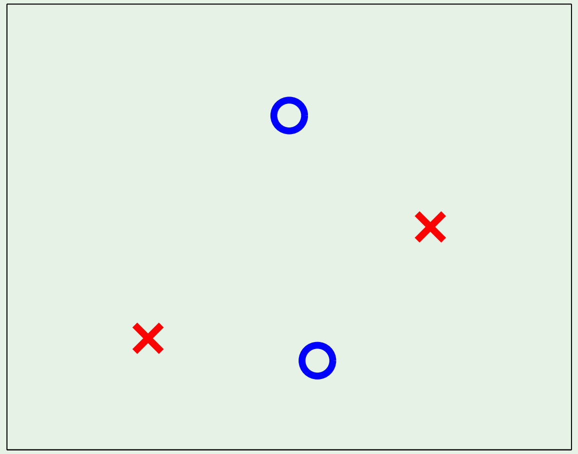
= hypotheses restricted to a finite set of points



- Growth function

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

- Break point



- Maximum # of dichotomies

resulting from the constraint of the break point (here k=2)

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

Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 6: **Theory of Generalization**

Sponsored by Caltech's Provost Office, E&AS Division, and IST • Thursday, April 19, 2012

Outline

- Proof that $m_{\mathcal{H}}(N)$ is polynomial (with a break point)

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

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

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

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

We want to bound m, and we do this with B(N, k) - the maximum number of dichotomies you can possibly have given there is a break point - this bound applies to any H. This is purely combinatorial, meaning we can avoid any consideration of input space or correlation between events etc.

Key quantity:

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

i.e max number of dichotomies on N points such that no subset of size k of the N points can be shattered.
The definition assumes a break point k, then tries to find the most dichotomies on N points without imposing any further restrictions. Since B(N, k) is the maximum, it will serve as an upper bound of mH(N).

Recursive bound on $B(N, k)$

Consider the following table:

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

B(N,k) is the maximum number of patterns we can get of
N points such that no k columns have all possible patterns (are shattered).

S1 contains rows which appear only once as far as x1 to xN-1 are concerned - the prefix (x1-xN-1) happens once and only has one extension (xN=+1 OR xN=-1)

S2 contains prefixes with both xN=+1 AND xN=-1 - we split each of these into subgroups S2+ and S2-

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

Estimating α and β

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

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

All rows (total alpha+beta) highlighted are different, (note S2+ and S2- have equal prefixes, so not different).

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

Also, on the original matrix we could not find all possible patterns on any k columns, so we also cannot on the highlighted matrix. If we could, then these k columns would feature all possible patterns in the original matrix, but we do not. So the smaller matrix has the same break point k.

Estimating β by itself

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

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

No subset of size k -1 of the first N -1 points can
be shattered by the dichotomies in S2+. If there existed
such a subset, then taking the corresponding set of
dichotomies in S2- and adding the xN column to the data
points yields a subset of size k that is shattered, which we
know cannot exist in this table by definition of B(N, k).

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

Putting it together

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

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

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

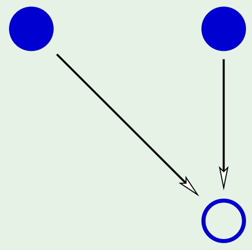
$$B(N,k) \le$$

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

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

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

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



		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	:		.		
	:	:	:	:			.	

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

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

Theorem:

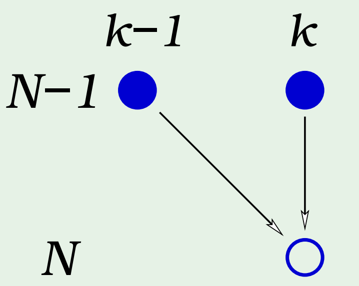
$$B(N,k) \le \sum_{i=0}^{k-1} \binom{N}{i}$$

1. **Boundary conditions:** easy

$$\begin{array}{cc|ccccccc}
 & & & & & k & & & \\
 & & 1 & 2 & 3 & 4 & 5 & 6 & .. \\
\hline
 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & .. \\
 & 2 & 1 & & & & & & \\
 & 3 & 1 & & & & & & \\
N & 4 & 1 & & & & & & \\
 & 5 & 1 & & & & & & \\
 & 6 & 1 & & & & & & \\
 & : & : & & & & & &
\end{array}$$

2. The induction step

$$
\begin{aligned}
\sum_{i=0}^{k-1}\binom{N}{i} &= \sum_{i=0}^{k-1}\binom{N-1}{i} + \sum_{i=0}^{k-2}\binom{N-1}{i} \ ? \\
&= 1+\sum_{i=1}^{k-1}\binom{N-1}{i} + \sum_{i=1}^{k-1}\binom{N-1}{i-1} \\
&= 1+\sum_{i=1}^{k-1}\left[\binom{N-1}{i} + \binom{N-1}{i-1}\right] \\
&= 1+\sum_{i=1}^{k-1}\binom{N}{i} = \sum_{i=0}^{k-1}\binom{N}{i} \ \checkmark
\end{aligned}
$$



It is polynomial!

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

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

The implication of this is that if H has a break point, we have waht we want to ensure good generalization: a polynomial bound on mH(N)

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 $k = 3$)

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

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

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

Outline

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

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

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:

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

Pictorial proof ☺

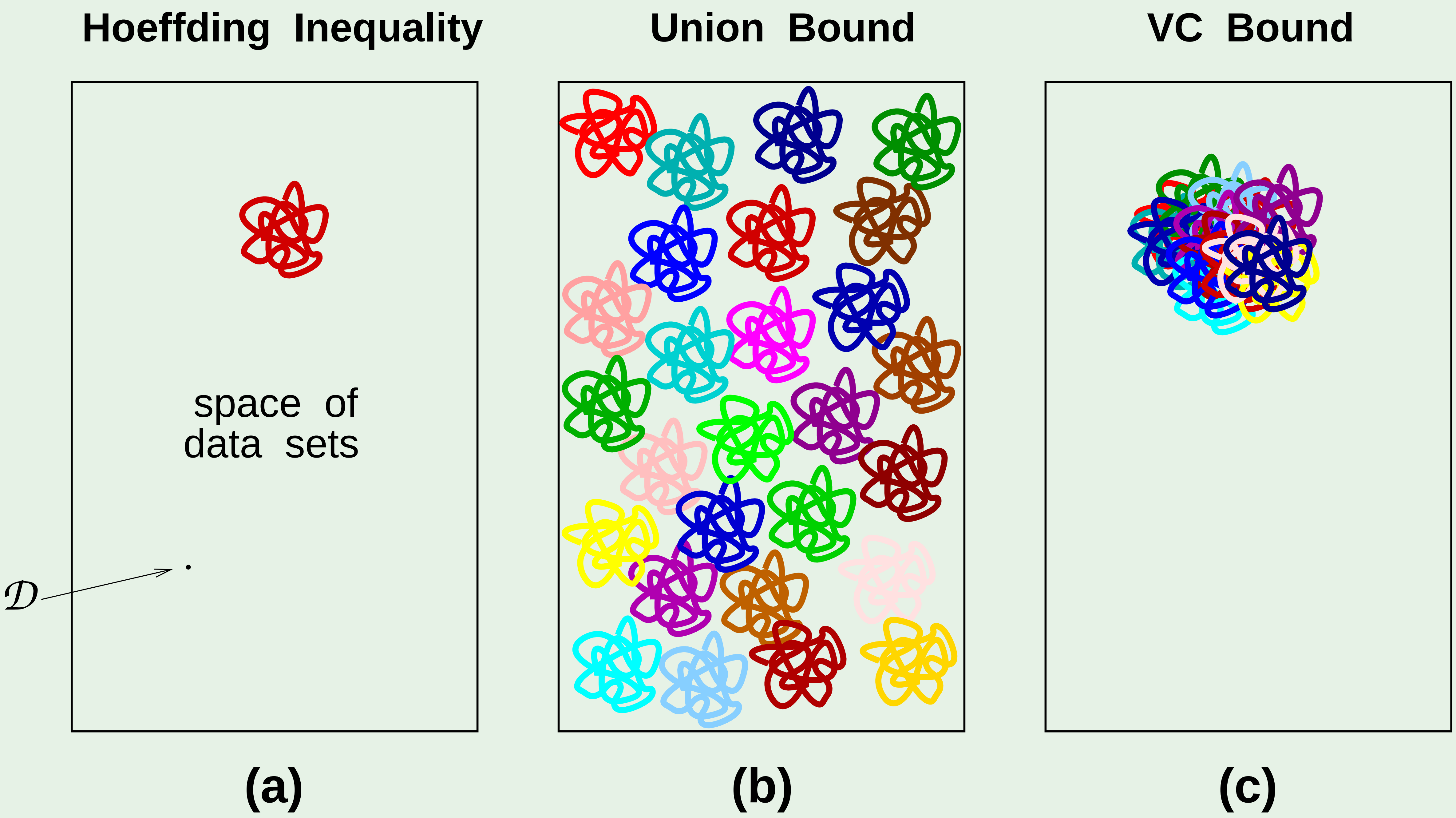
- How does $m_{\mathcal{H}}(N)$ relate to overlaps?

since M is created from the union bound which assumes disjoint hypotheses

- What to do about E_{out}?

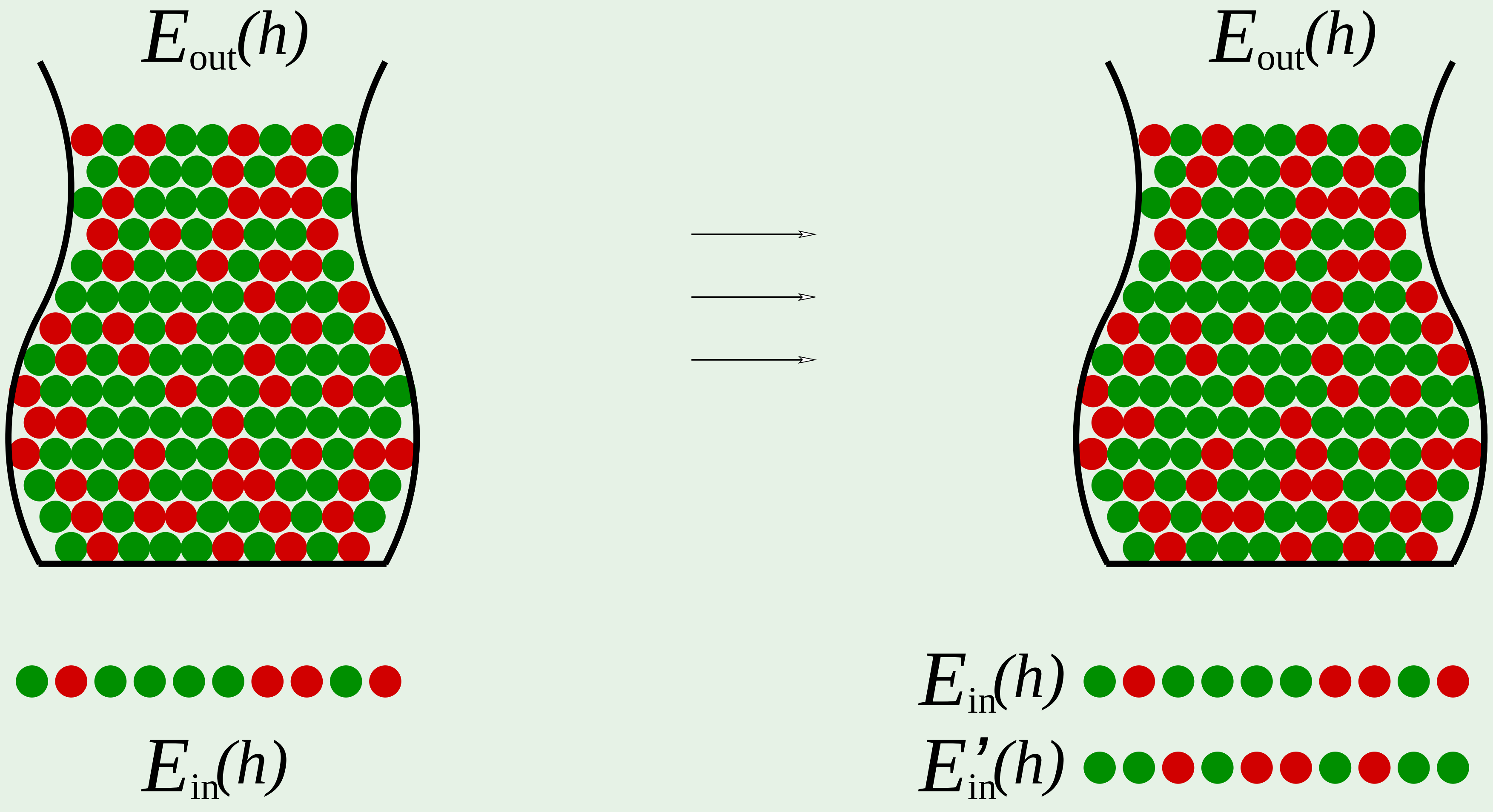
since the growth function relies on a finite sample (and the subsequent dichotomies), so it will handle the Ein aspect of Hoeffding. However Eout relates to the performance over the entire input space X and so we are dealing with full hypotheses, not dichotomies, so we lose the benefit of the growth function m.

- Putting it together



(a) For a given hypothesis, the colored points correspond to data sets where Ein does not generalize well to Eout · The Hoeffding Inequality guar antees a small colored area. (b) For several hypotheses, the union bound assumes no overlaps, so the total colored area is large. (c) The VC bound keeps track of overlaps, so it estimates the total area of bad generalization to be relatively small.

What to do about E_{out}



Ein and Ein' track eachother since they both track Eout (even if their tracking is looser) - e.g. you expect two polls of equal N to have close results to eachother. Like how the tracking of Eout and Ein become looser as the number of hypotheses increased (from M in Hoeffding), it also happens with Ein and Ein'. If we characterize this using the two samples only, no longer appealing to Eout, we are completely in the realm of dichotomies (instead of hypotheses on X) and, although the sample is bigger (2N), we can define a growth function on them - see next slide.

Putting it together

Not quite:

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

but rather:

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

The Vapnik-Chervonenkis Inequality