

Lectures 5 and 6

CS436/536: Introduction to Machine Learning

Zhaohan Xi

Binghamton University

zxi1@binghamton.edu

Classifier Evaluation Metrics: Example

- Use the same confusion matrix, calculate the measure just introduced

Actual Class\Predicted class	cancer = yes	cancer = no	Total
cancer = yes	90	210	300
cancer = no	140	9560	9700
Total	230	9770	10000

- Sensitivity = $TP/P = 90/300 = 30\%$
- Specificity = $TN/N = 9560/9700 = 98.56\%$
- Accuracy = $(TP + TN)/All = (90+9560)/10000 = 96.50\%$
- Error rate = $(FP + FN)/All = (140 + 210)/10000 = 3.50\%$
- Precision = $TP/(TP + FP) = 90/(90 + 140) = 90/230 = 39.13\%$
- Recall = $TP/ (TP + FN) = 90/(90 + 210) = 90/300 = 30.00\%$
- F1 = $2 P \times R / (P + R) = 2 \times 39.13\% \times 30.00\% / (39.13\% + 30\%) = 33.96\%$

Hoeffding's Inequality for Learning (from finite \mathcal{H})

$$\mathbb{P}[|E_{in}(\mathbf{g}) - E_{out}(\mathbf{g})| > \epsilon] \leq 2|\mathcal{H}|e^{-2\epsilon^2 N}, \quad \text{for any } \epsilon > 0$$

Feasibility of Learning (with Finite Models)

- No Free Lunch: Cannot learn f *exactly* from \mathcal{D} over all \mathcal{X}
- But, Can learn f with high probability due to Hoeffding, if:
 - \mathcal{D} and the test data point are drawn i.i.d. from $P(\mathbf{x})$
 - \mathcal{H} is fixed and g is selected from \mathcal{H}

To achieve learning: i.e. select g from \mathcal{H} so that $E_{out}(g) \approx 0$, we must ensure:

(Step 1) $E_{out}(g) \approx E_{in}(g)$ -- Ensure $|\mathcal{H}|$ is small

Theorem. With probability at least $1 - \delta$, $E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \log \frac{2|\mathcal{H}|}{\delta}}$

(Step 2) $E_{in}(g) \approx 0$ -- Learning algorithm \mathcal{A}

To Infinity ...

How about infinite hypothesis set?

- e.g. perceptron model

To Infinity ...

Now on to Step 1: Ensure $E_{out}(g) \approx E_{in}(g)$

But Hoeffding's inequality only works for finite hypothesis set

Need to revisit model complexity

Hoeffding's Inequality Revisited

How did we arrive at Hoeffding's Inequality for Learning?

1. Consider the disjunction of “**bad**” events:

$$|E_{in}(h_1) - E_{out}(h_1)| > \epsilon$$

or $|E_{in}(h_2) - E_{out}(h_2)| > \epsilon$

...

or $|E_{in}(h_M) - E_{out}(h_M)| > \epsilon$

2. This must include the event $|E_{in}(g) - E_{out}(g)| > \epsilon$

where g is picked from \mathcal{H}

3. Derive an upper bound on the probability of this bad event

Using the Union Bound

$|\mathcal{H}|$ is overkill

- Want: Upper bound on \mathcal{B} ad events:

$$\mathcal{B}_g = \{|E_{out}(g) - E_{in}(g)| > \epsilon\}$$

$$\mathcal{B}_m = \{|E_{out}(h_m) - E_{in}(h_m)| > \epsilon\} \text{ for each } h_m \in \mathcal{H}$$

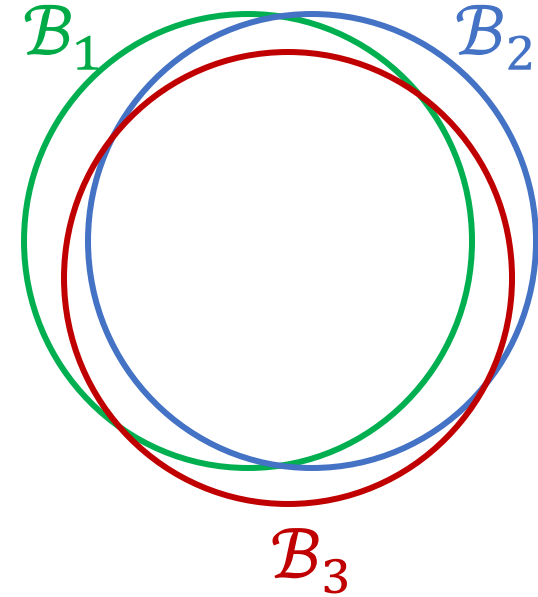
Since g is unknown, we use a worst-case union bound:

$$\mathbb{P}[\mathcal{B}_g] \leq \mathbb{P}[\text{any } \mathcal{B}_m] \leq \sum_{m=1}^{|\mathcal{H}|} \mathbb{P}[\mathcal{B}_m]$$

Each \mathcal{B}_m is an event: a set of outcomes to which a probability is assigned

These events may overlap: E.g. If h_1 is very similar to h_2

Then, the union bound is loose!



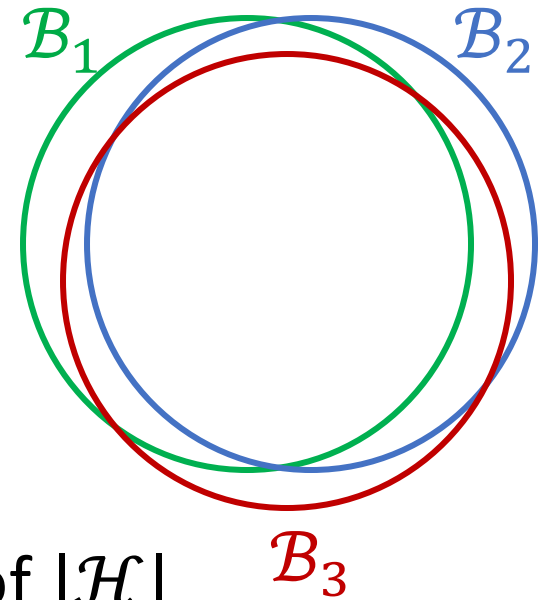
$|\mathcal{H}|$ fails to account for “similarity” of hypotheses

- If many $h_m \in \mathcal{H}$ are similar, then the corresponding events \mathcal{B}_m overlap

$\mathbb{P}[\mathcal{B}_g] \leq \mathbb{P}[\text{any } \mathcal{B}_m] \leq \sum_{m=1}^{|\mathcal{H}|} \mathbb{P}[\mathcal{B}_m]$ is an overestimate

Idea:

- If there are fewer than $|\mathcal{H}|$ “effective” hypotheses:
 - Replace $|\mathcal{H}|$ with a different measure of the “diversity” of $|\mathcal{H}|$
- to obtain a tighter upper bound on E_{out}

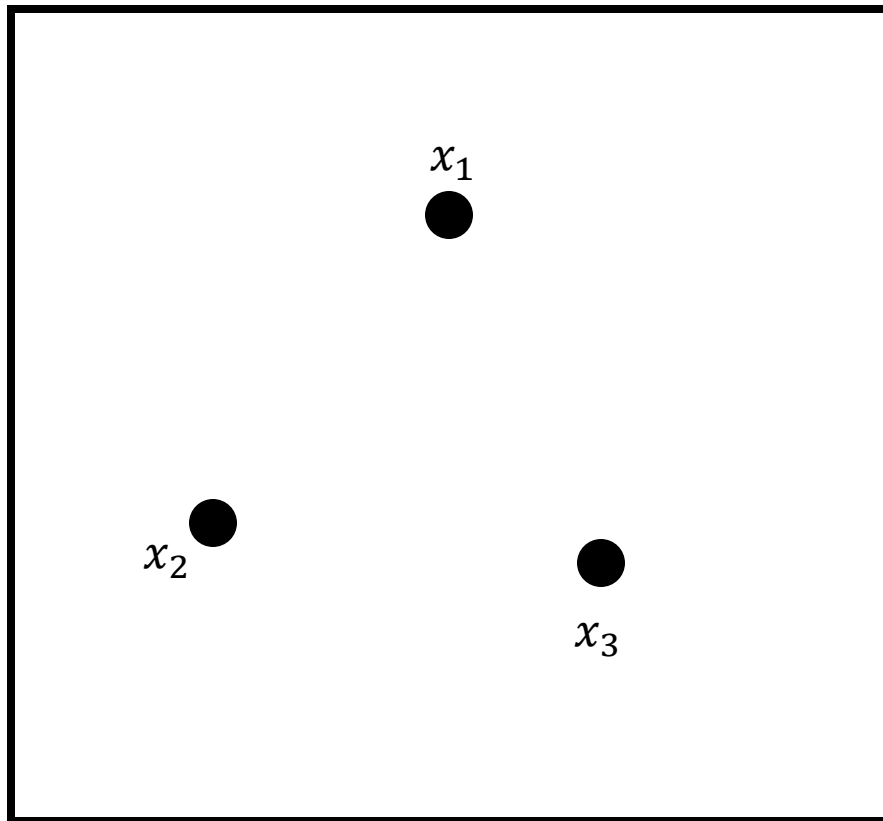


Thinking Aid: Dichotomies

Given a set of data point,

a dichotomy is any way to classify the data points into two classes

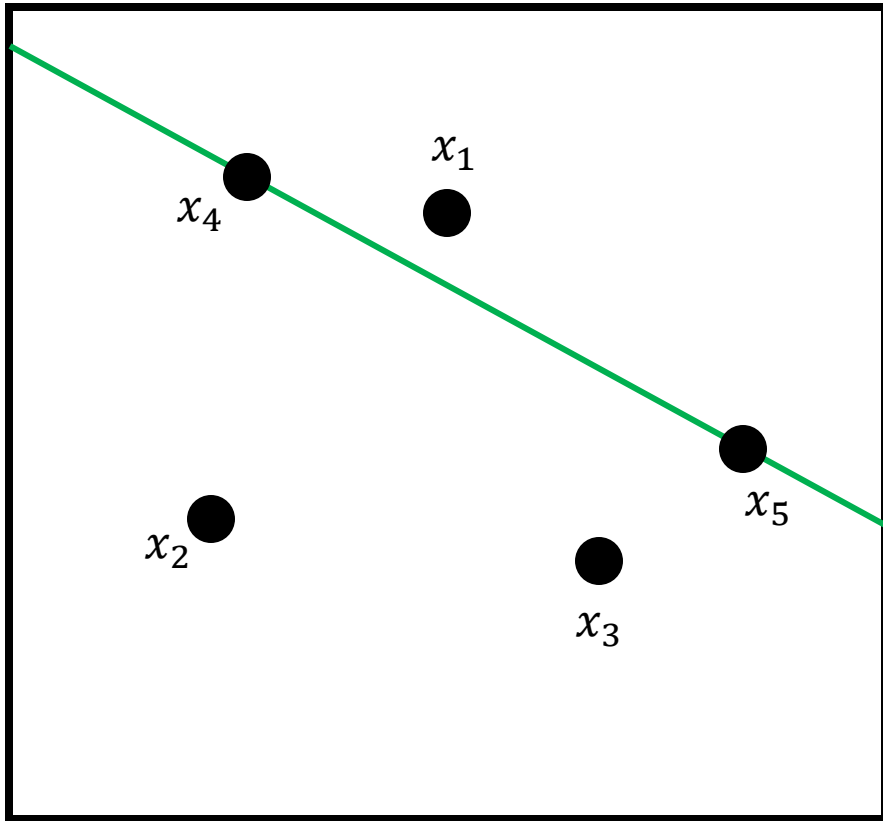
e.g. into two classes +1 and -1



How many possible dichotomies are there for this dataset?

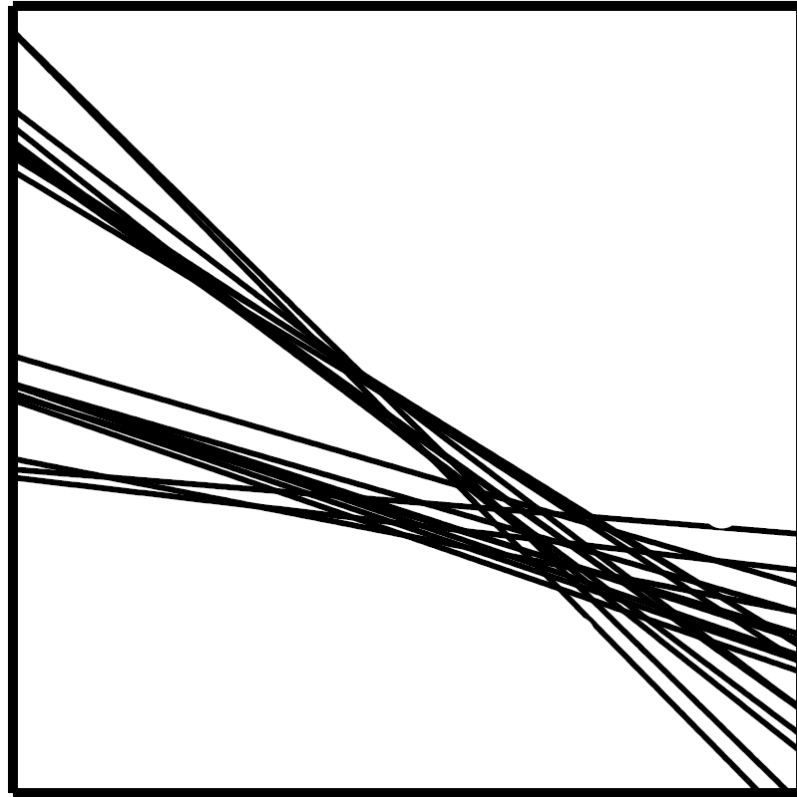
8

A Question About Lines

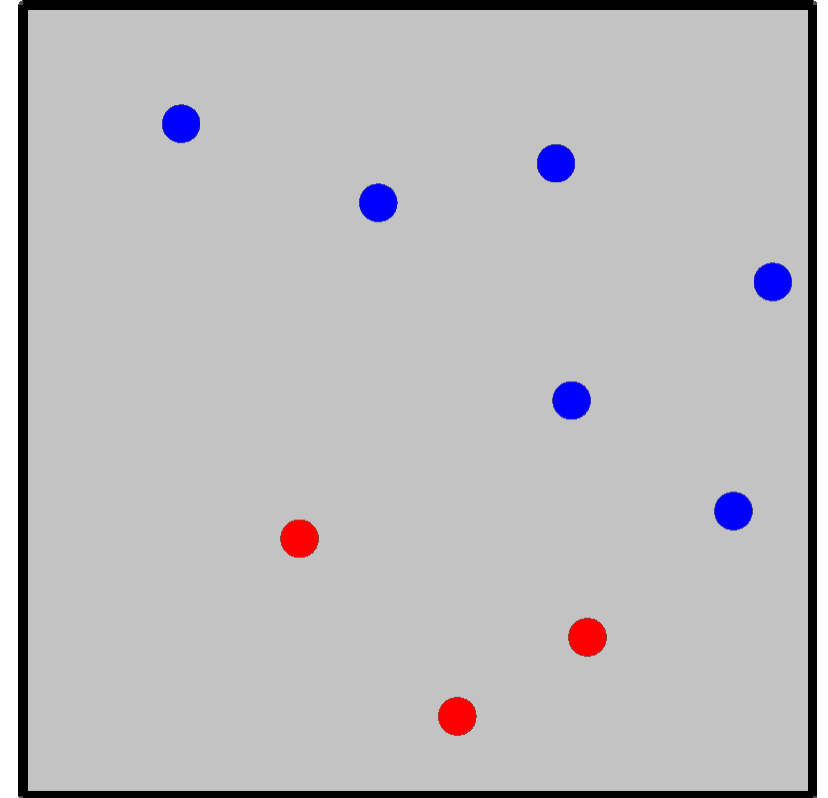


For each point, is $w^T x$
(w.r.t. the green line)
 $>$, $=$ or < 0 ?

Data Reveals the True Diversity of \mathcal{H}



\mathcal{H}



\mathcal{H} through the eyes of \mathcal{D}

For \mathcal{D} , \mathcal{H} is just one *dichotomy*

Effective Number of Hypotheses in \mathcal{H}

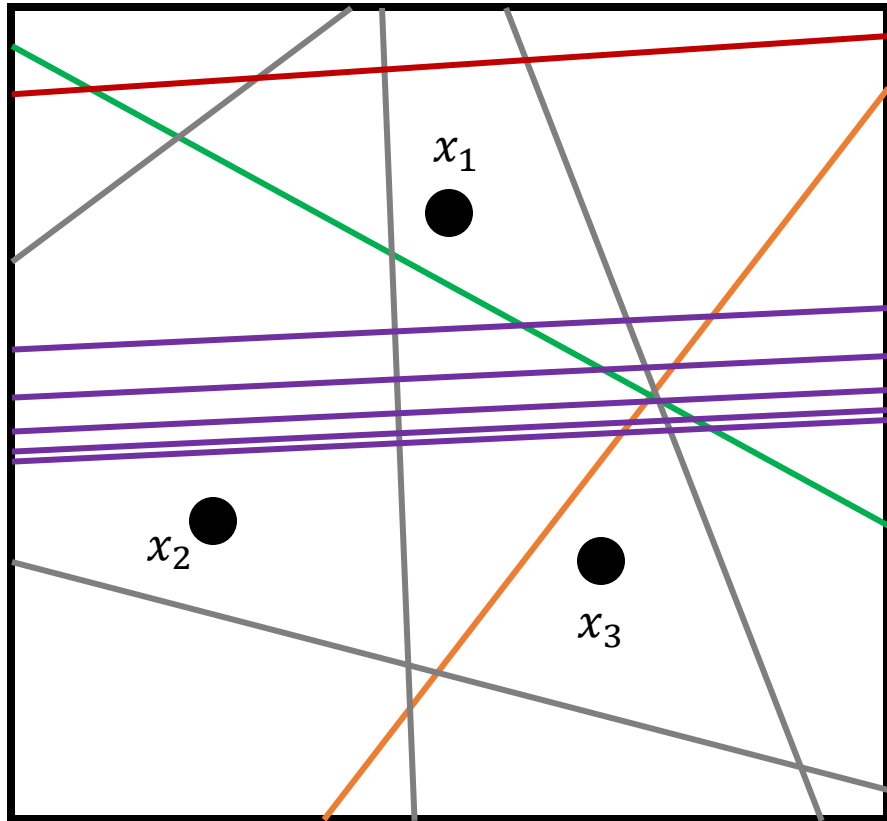
- Effective size of \mathcal{H} :

Diversity of \mathcal{H} : number of dichotomies it can implement

- Grows with N
- **growth function**: how does the effective size of \mathcal{H} grow with N ?
- $|\mathcal{H}|$ is the maximum possible diversity of \mathcal{H}

Bound the **growth** of the diversity of \mathcal{H} as a function of N ?

The Growth of the Effective Number of Hypotheses



8 dichotomies

	x_1	x_2	x_3
h_1	+1	-1	-1
h_2	-1	-1	+1
h_3			
h_4			
h_5			
h_6			
...			

The Growth Function $m_{\mathcal{H}}(N)$

- Effective number of hypotheses in \mathcal{H} through the eyes of \mathcal{D}

Definition: Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathcal{X}$.

The dichotomies generated by \mathcal{H} on these points are defined by

$$\mathcal{H}(\mathbf{x}_1, \dots, \mathbf{x}_N) = \left\{ \left(h(\mathbf{x}_1), \dots, h(\mathbf{x}_N) \right) \mid h \in \mathcal{H} \right\}$$

Definition: The Growth Function $m_{\mathcal{H}}(N)$

The largest set of dichotomies induced by \mathcal{H} :

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

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

Worst-case

Example Growth Functions $m_{\mathcal{H}}(N)$

	N					
	1	2	3	4	5	...
2D Perceptron	2	4	8	?	...	
	+-	++		
	-+	+-				
		-+				
		--				

- Ideally, want $m_{\mathcal{H}}(N)$ to grow slower than 2^N
- In general: A **break point** is any k for which $m_{\mathcal{H}}(k) < 2^k$

Example Growth Functions $m_{\mathcal{H}}(N)$

	N					
	1	2	3	4	5	...
2D Perceptron	2	4	8	14	...	
	+-	++		
	-+	+-				
		-+				
		--				

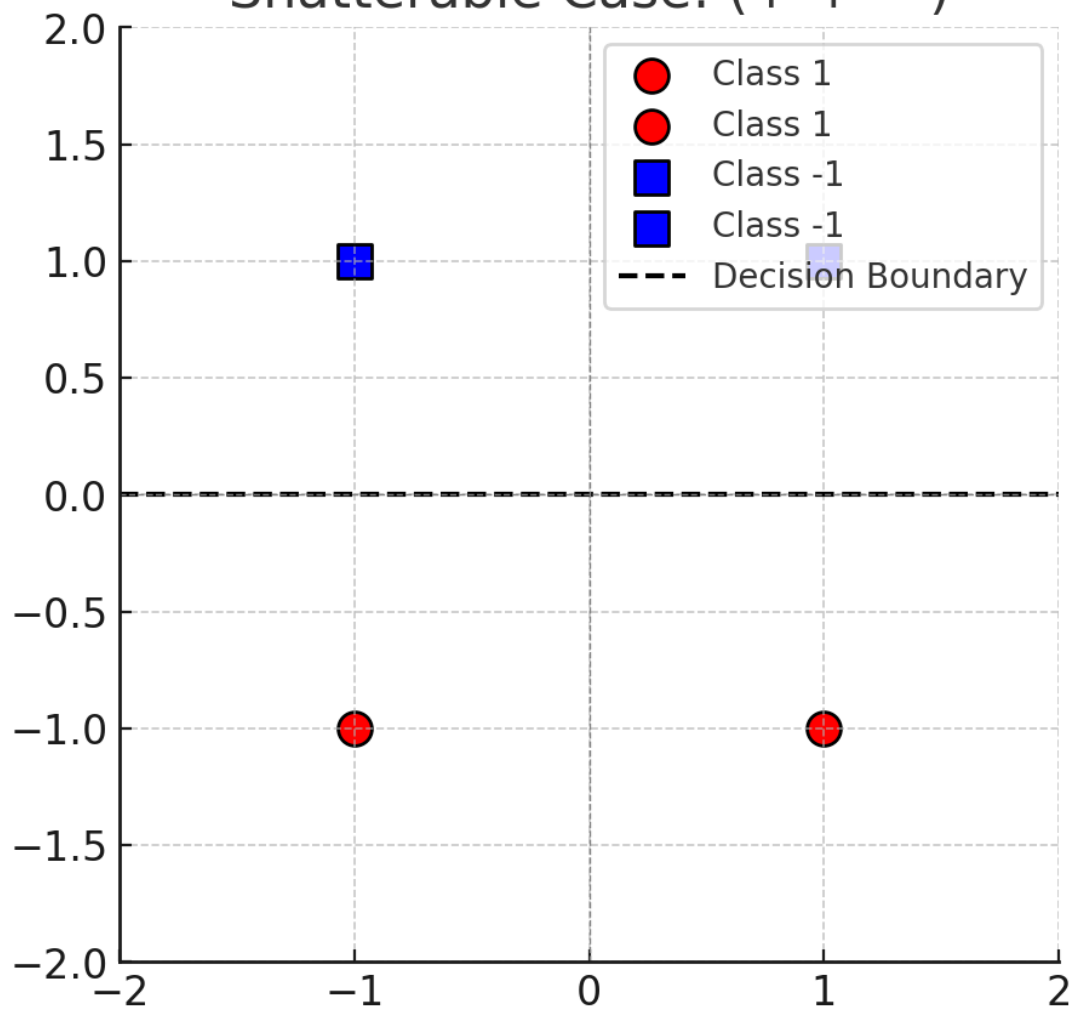
- Ideally, want $m_{\mathcal{H}}(N)$ to grow slower than 2^N
- In general: A **break point** is any k for which $m_{\mathcal{H}}(k) < 2^k$

Example Growth Functions $m_{\mathcal{H}}(N)$

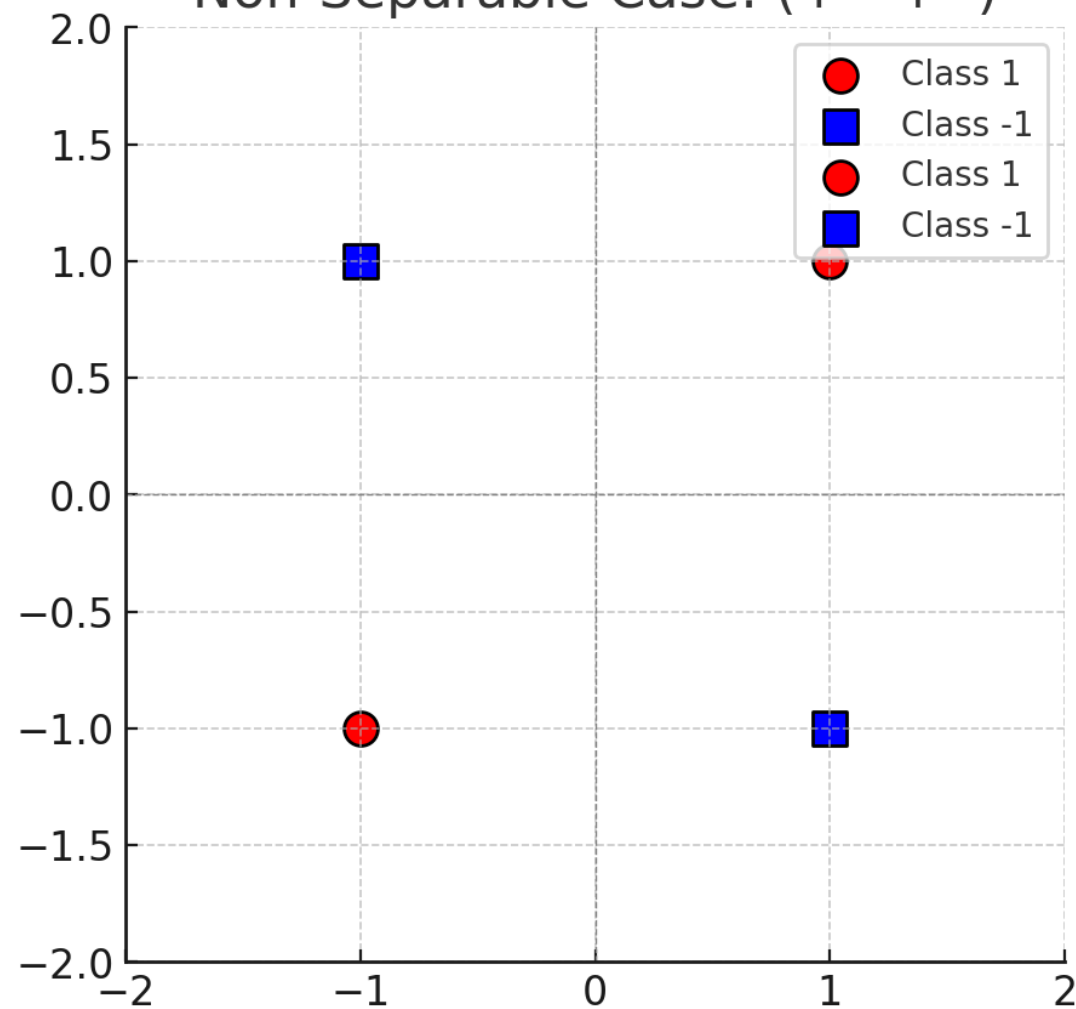
	N					
	1	2	3	4	5	...
2D Perceptron	2	4	8	14	...	

- **Why Can't a 2D Perceptron Shatter $N > 3$ Points?**
- A **2D perceptron** is a **linear classifier**, meaning it can only separate points using a **straight line**.

Shatterable Case: (+ + - -)



Non-Separable Case: (+ - + -)



On break points

For a given hypothesis set \mathcal{H} ,

To show that k is a break point, we must prove that:

- **For all sets of k points**, the hypothesis class **cannot** shatter all 2^k possible labeling.
- This means that there **exists at least one labeling** of k points that the hypothesis class **cannot realize**.

Example Growth Functions $m_{\mathcal{H}}(N)$

	N					
	1	2	3	4	5	...
1D Positive Ray	2	3	4	5		...

- The **1D Positive Ray** is a simple hypothesis class in **one-dimensional space**, where classification is determined based on a **threshold**. It is defined as:

$$h(x) = \begin{cases} -1, & x < threshold \\ +1, & x \geq threshold \end{cases}$$

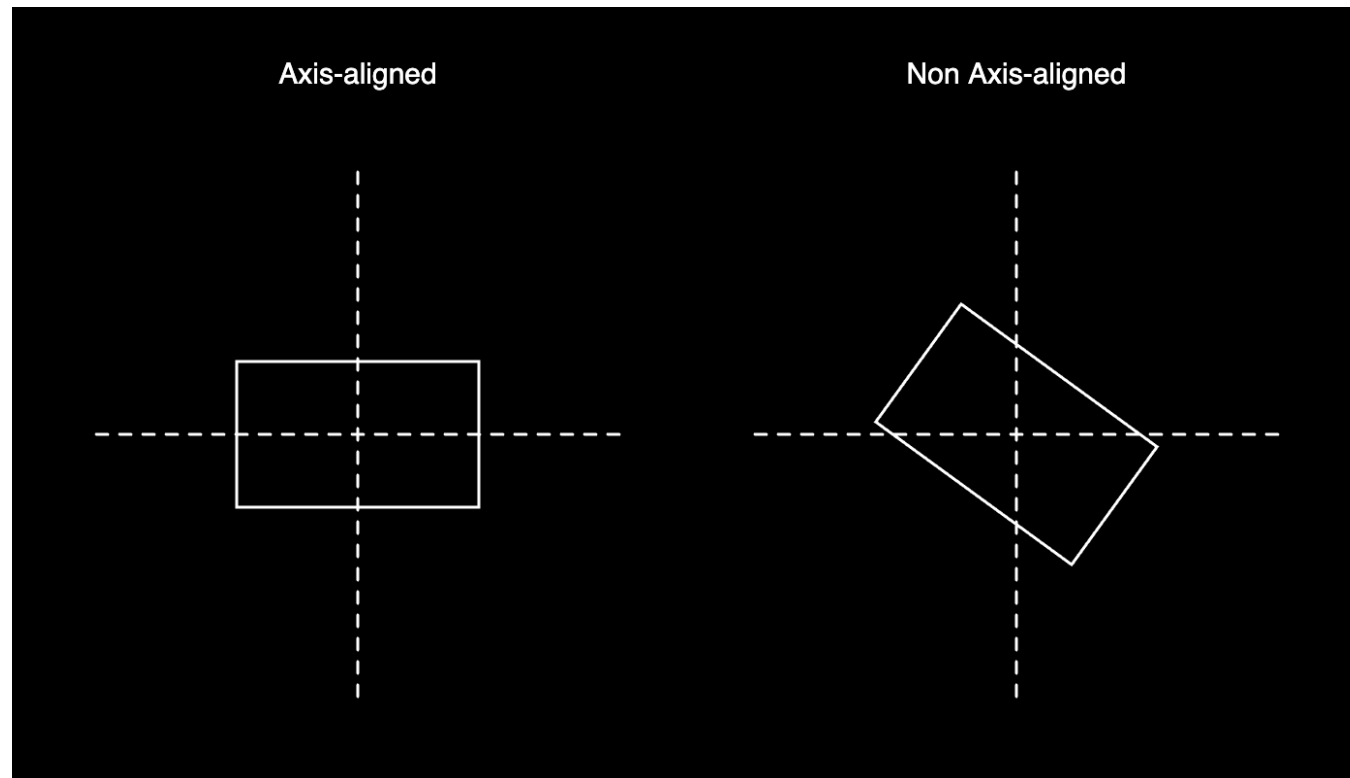
Example Growth Functions $m_{\mathcal{H}}(N)$

	N					
	1	2	3	4	5	...
1D Positive Ray	2	3	4	5		...

- $N=1$: (-) and (+)
- $N=2$: (--), (-+) and (++)
- $N=3$: (---), (--+), (-++), and (+++)
- **$m_{\mathcal{H}}(N) = N + 1$**

Example Growth Functions $m_{\mathcal{H}}(N)$


	N					
	1	2	3	4	5	...
2D (Axis-Aligned) Positive Rectangle	2	4	8	16	$< 2^5$...



Always Polynomial After Smallest Break Point

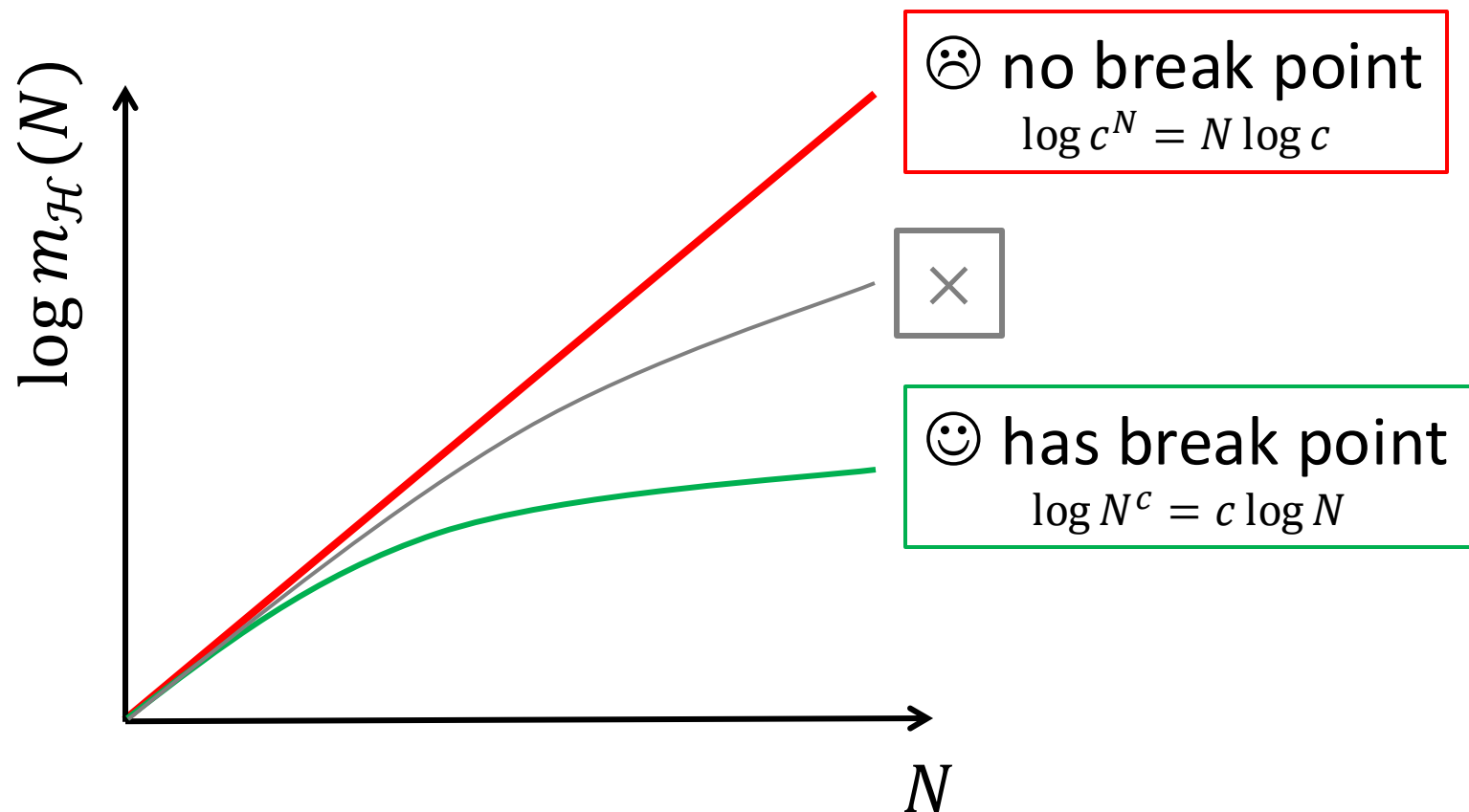
Theorem.

If $m_{\mathcal{H}}(k) < 2^k$ for some value of k , then

$$\begin{aligned} m_{\mathcal{H}}(N) &\leq \sum_{i=0}^{k-1} \binom{N}{i}, \quad \text{for all } N \\ &= O(N^{k-1}) \end{aligned}$$


Good news for $E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{\frac{1}{N} \log \frac{m_{\mathcal{H}}(N)}{\delta}}\right)$

Only Two Kinds of Hypothesis Sets: Good or Bad



- Upper bound on E_{out} does **not** approach E_{in} as N increases

- $E_{out} \approx E_{in}$ for sufficiently large N
“Learning is a process by which a system improves performance from experience”
- Herbert Simon

$$E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{\frac{1}{N} \log \frac{m_{\mathcal{H}}(N)}{\delta}}\right)$$

Feasibility of Learning (Revisited)

To achieve learning:

i.e. select g from \mathcal{H} so that $E_{out}(g) \approx 0$,

we must ensure:

(Step 1) $E_{out}(g) \approx E_{in}(g)$

[Generalization]

(Step 2) $E_{in}(g) \approx 0$

[Approximation]

The VC Dimension

Definition.

The *Vapnik-Chervonenkis dimension* of a hypothesis set \mathcal{H} , denoted by $d_{VC}(\mathcal{H})$ or simply d_{VC} , is the largest value of N , for which $m_{\mathcal{H}}(N) = 2^N$.

If $m_{\mathcal{H}}(N) = 2^N$ for all N , then $d_{VC}(\mathcal{H}) = \infty$.

Always Polynomial After Smallest Break Point

Theorem. (revisited) If $m_{\mathcal{H}}(k) < 2^k$ for some value of k , then

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}, \quad \text{for all } N$$

k is break point

$$k = d_{VC} + 1: \quad m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{VC}} \binom{N}{i}$$

And in fact:
$$m_{\mathcal{H}}(N) \leq N^{d_{VC}} + 1$$

Example Growth Functions $m_{\mathcal{H}}(N)$

	N					
	1	2	3	4	5	...
2D Perceptron	2	4	8	14		...
1D Positive Ray	2	3	4	5		...
2D (Axis-Aligned) Positive Rectangle	2	4	8	16	< 2 ⁵	...

The Vapnik-Chervonenkis Bound (VC Bound)

$$\mathbb{P}[|E_{in}(\mathbf{g}) - E_{out}(\mathbf{g})| > \epsilon] \leq 2|\mathcal{H}|e^{-2\epsilon^2 N}, \quad \text{for any } \epsilon > 0$$

$$\mathbb{P}[|E_{in}(\mathbf{g}) - E_{out}(\mathbf{g})| > \epsilon] \leq 4m_{\mathcal{H}}(2N)e^{-\frac{\epsilon^2 N}{8}}, \quad \text{for any } \epsilon > 0$$

$$\mathbb{P}[|E_{in}(\mathbf{g}) - E_{out}(\mathbf{g})| \leq \epsilon] \leq 1 - 4m_{\mathcal{H}}(2N)e^{-\frac{\epsilon^2 N}{8}}, \quad \text{for any } \epsilon > 0$$

$$E_{out}(\mathbf{g}) \leq E_{in} + \sqrt{\frac{8}{N} \log \frac{4m_{\mathcal{H}}(2N)}{\delta}}, \quad \text{with probability at least } 1 - \delta$$

The Vapnik-Chervonenkis Bound (VC Bound)

$$E_{out}(g) \leq E_{in} + \sqrt{\frac{8}{N} \log \frac{4m_{\mathcal{H}}(2N)}{\delta}}, \quad \text{with probability at least } 1 - \delta$$

where $m_{\mathcal{H}}(N) \sim N^k$ when \mathcal{H} is “good” i.e. has a break point

The tightest upper bound on $E_{out}(g)$ is obtained on the smallest break point k^*

By definition, $d_{VC} = k^* - 1$

The Vapnik-Chervonenkis Bound (VC Bound)

- $d_{VC}(\mathcal{H})$ is the *largest* value of N for which \mathcal{H} can implement all 2^N possible dichotomies on **some** data set of size N

i.e. where $m_{\mathcal{H}}(N) = 2^N$

- If $N > d_{VC}$, N is a break point and

$$m_{\mathcal{H}}(N) \leq N^{d_{VC}} + 1 \sim N^{d_{VC}}$$

$$E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{\frac{1}{N} \log \frac{m_{\mathcal{H}}(N)}{\delta}}\right) \Rightarrow E_{out}(g) \leq E_{in} + O\left(\sqrt{\frac{1}{N} \log \frac{N^{d_{VC}}}{\delta}}\right)$$

For fixed confidence level $1 - \delta$: $E_{out}(g) \leq E_{in} + O\left(\sqrt{\frac{d_{VC} \log N}{N}}\right)$

Single Parameter to Characterize Complexity

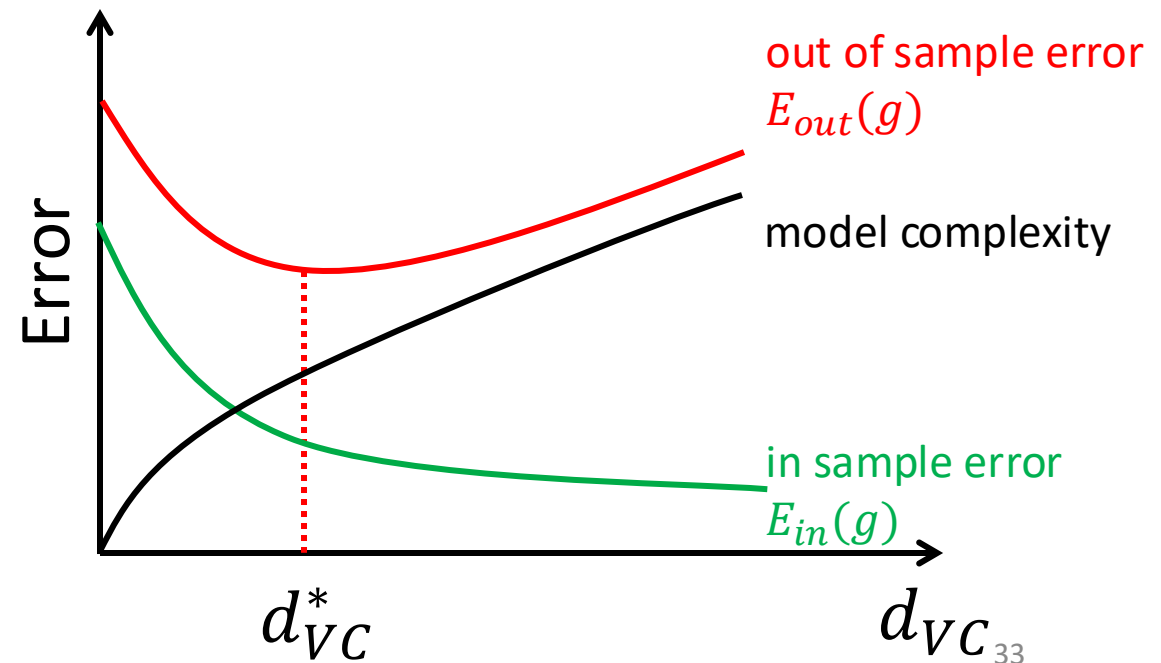
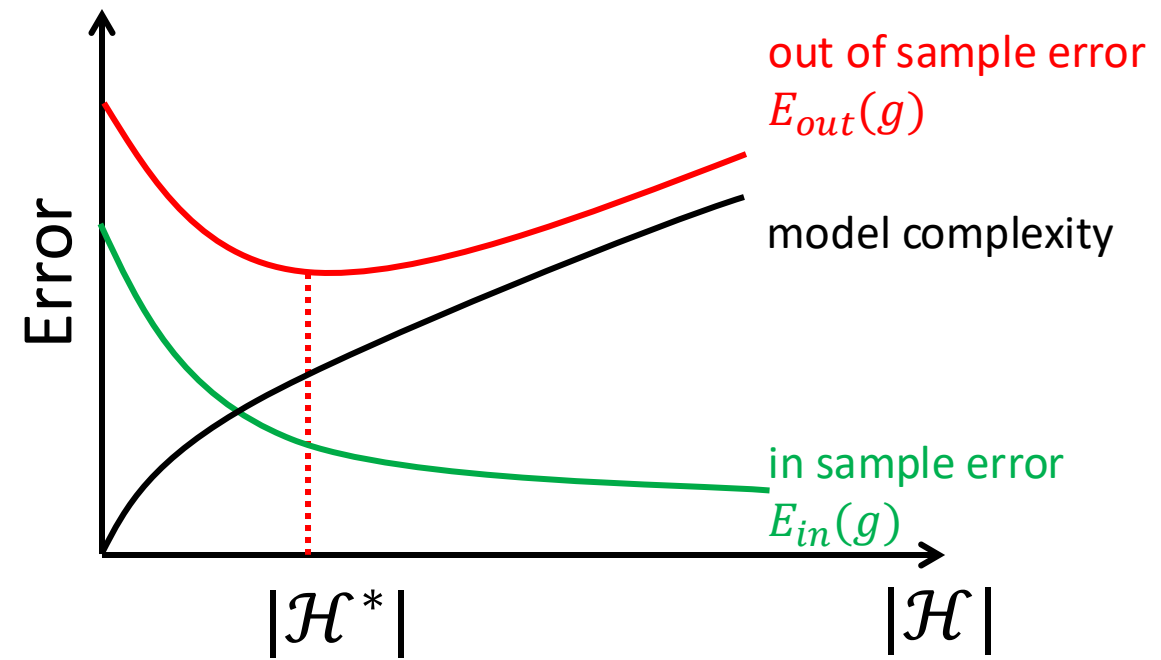
$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \log \frac{2|\mathcal{H}|}{\delta}}$$



$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \log \frac{4((2N)^{d_{VC}} + 1)}{\delta}}$$

Applicable to *infinite* \mathcal{H}

(with probability at least $1 - \delta$)



Sample Complexity

Recall VC Bound

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 4m_{\mathcal{H}}(2N)e^{-\frac{\epsilon^2 N}{8}}, \quad \text{for any } \epsilon > 0$$

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| \leq \epsilon] \leq 1 - 4m_{\mathcal{H}}(2N)e^{-\frac{\epsilon^2 N}{8}}, \quad \text{for any } \epsilon > 0$$

$$E_{out}(g) \leq E_{in} + \sqrt{\frac{8}{N} \log \frac{4m_{\mathcal{H}}(2N)}{\delta}}, \quad \text{with probability at least } 1 - \delta$$

Want: small ϵ , small δ

i.e. with high probability $(1 - \delta)$, want $|E_{in}(g) - E_{out}(g)|$ to be at most ϵ

How fast does N grow as we demand smaller and smaller ϵ , δ ?

Sample Complexity

- For fixed ϵ, δ ;

Want: $E_{out}(g) \leq E_{in} + \epsilon$ with probability at least $1 - \delta$

- $E_{out}(g) \leq E_{in} + \sqrt{\frac{8}{N} \log \frac{4m_{\mathcal{H}}(2N)}{\delta}}$

Want: $\sqrt{\frac{8}{N} \log \frac{4m_{\mathcal{H}}(2N)}{\delta}} \leq \epsilon$

- Need: $N \geq \frac{8}{\epsilon^2} \log \left(\frac{4m_{\mathcal{H}}(2N)}{\delta} \right)$

and using the upper bound on $m_{\mathcal{H}}(2N)$, Need: $N \geq \frac{8}{\epsilon^2} \log \left(\frac{4((2N)^{d_{VC}+1})}{\delta} \right)$

Sample Complexity: Solve for N

- Set error bar $\epsilon = \sqrt{\frac{8}{N} \log \frac{4((2N)^{d_{VC}} + 1)}{\delta}}$
- Solve for N using iterative methods

Sample Complexity: Solve for N

E.g. Suppose $d_{VC} = 3$ and

we want error bar of at most 0.1 with confidence 90%, i.e. $\epsilon = 0.1, \delta = 0.1$

$$\text{Need: } N \geq \frac{8}{0.1^2} \log \left(\frac{4(2N)^3 + 4}{0.1} \right)$$

Trying initial guess of $N = 1000$ in RHS, $N \geq \frac{8}{0.1^2} \log \left(\frac{4(2 \times 1000)^3 + 4}{0.1} \right) \approx 21,193$

Try new guess of $N = 21,193$ in RHS, and so on...

converges to $N \approx 30,000$

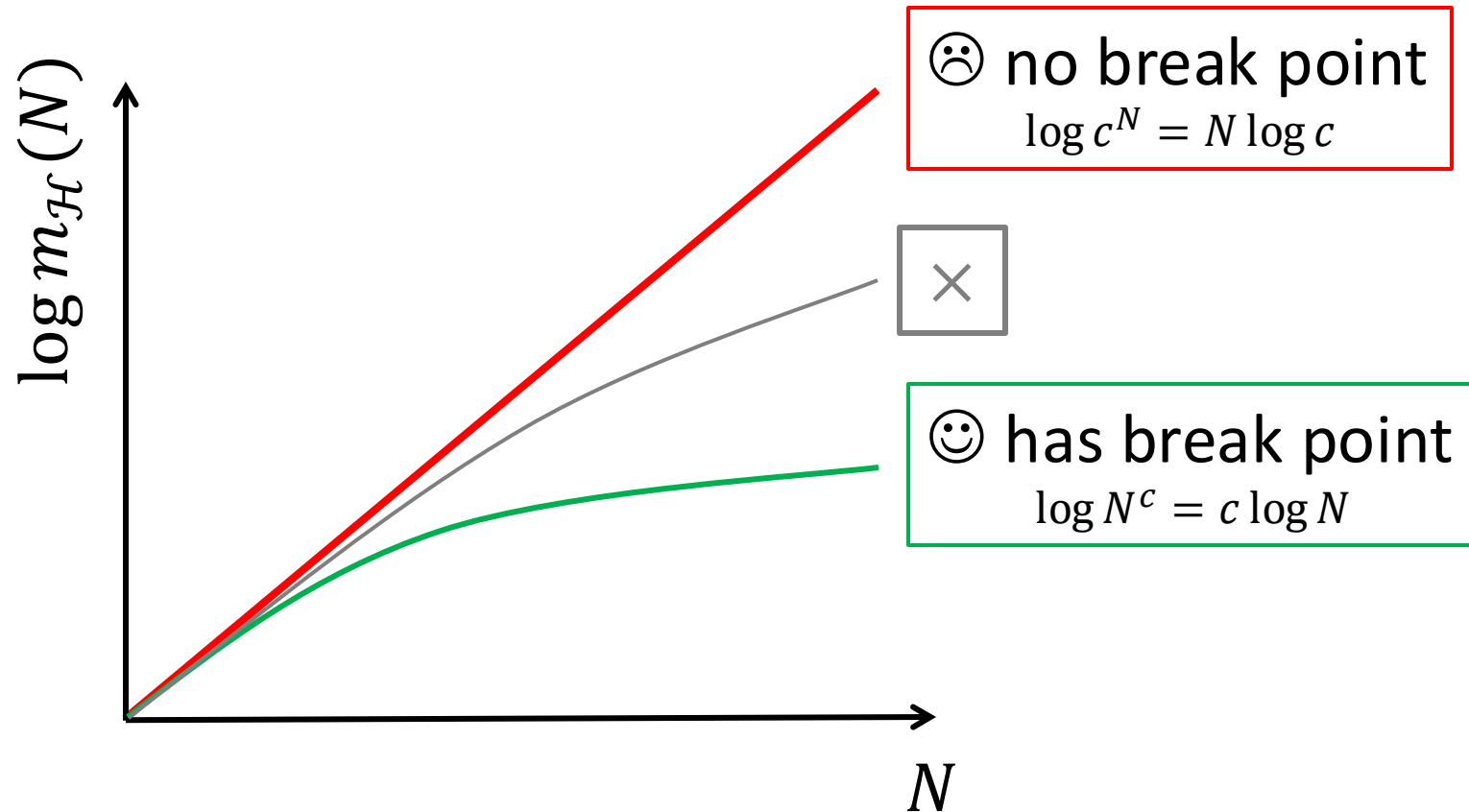
For $d_{VC} = 4$, $N \approx 40,000$

$N \propto d_{VC}$, but gross overestimate. Rule of Thumb: $N \approx 10 \times d_{VC}$

In Practice

- VC Bound is still a loose upper bound
 - $m_{\mathcal{H}}(N)$ is the worst-case number of dichotomies \mathcal{H} can implement
 - The polynomial bound on $m_{\mathcal{H}}(N)$ is also loose
- But it is a useful guide in *model selection*
 - models with small d_{VC} are good
- Rule of Thumb: Need $10 \times d_{VC}$ samples generalize well

Only Two Kinds of Hypothesis Sets: Good or Bad



- Upper bound on E_{out} does **not** approach E_{in} as N increases

- $E_{out} \approx E_{in}$ for sufficiently large N
“Learning is a process by which a system improves performance from experience”
- Herbert Simon

$$E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{\frac{1}{N} \log \frac{m_{\mathcal{H}}(N)}{\delta}}\right)$$

VC Analysis

By our choice of \mathcal{H} ,

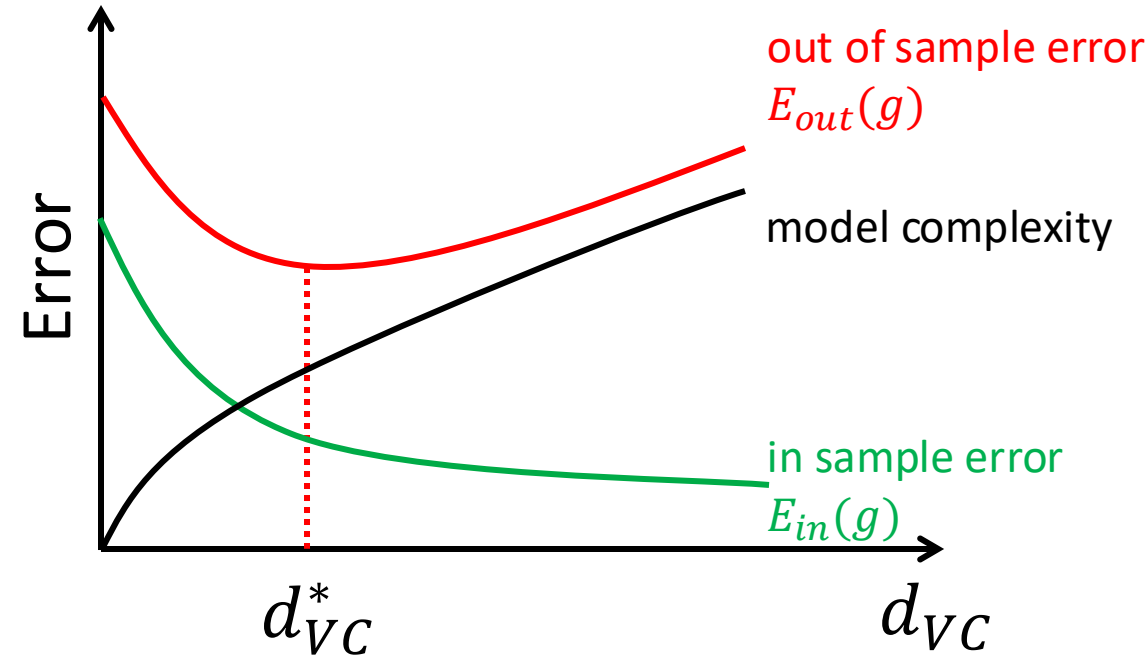
1. How well can we approximate the target function in-sample (E_{in})?

$d_{VC} \uparrow \Rightarrow E_{in} \approx 0$: Higher chance of approximating target function on data

2. How well does it generalize? (How close is E_{in} to E_{out} ?)

$d_{VC} \downarrow \Rightarrow E_{in} \approx E_{out}$: Higher chance of generalizing to out of data

VC Analysis: Approximation vs. Generalization



$d_{VC} \uparrow \Rightarrow E_{in} \approx 0$: Higher chance of approximating target function on data

$d_{VC} \downarrow \Rightarrow E_{in} \approx E_{out}$: Higher chance of generalizing to out of data

Depends only on \mathcal{H} ; Independent of $f, P(\mathbf{x}), \mathcal{A}$ (the learning algorithm) 41

Takeaways

d_{VC} gives us a single parameter to characterize the “complexity” of a model

This determines:

- the “sample complexity” for learning (see above)

We must use a combination of

- theory (e.g. VC analysis) and

- empirical techniques (more soon, e.g. regularization)

- to determine the sweet spot of the approximation – generalization tradeoff