

Announcements



Assignments

- HW5
 - Due Mon, 10/26, 11:59 pm
- HW6
 - Out tomorrow
 - Due Mon, 11/2, 11:59 pm

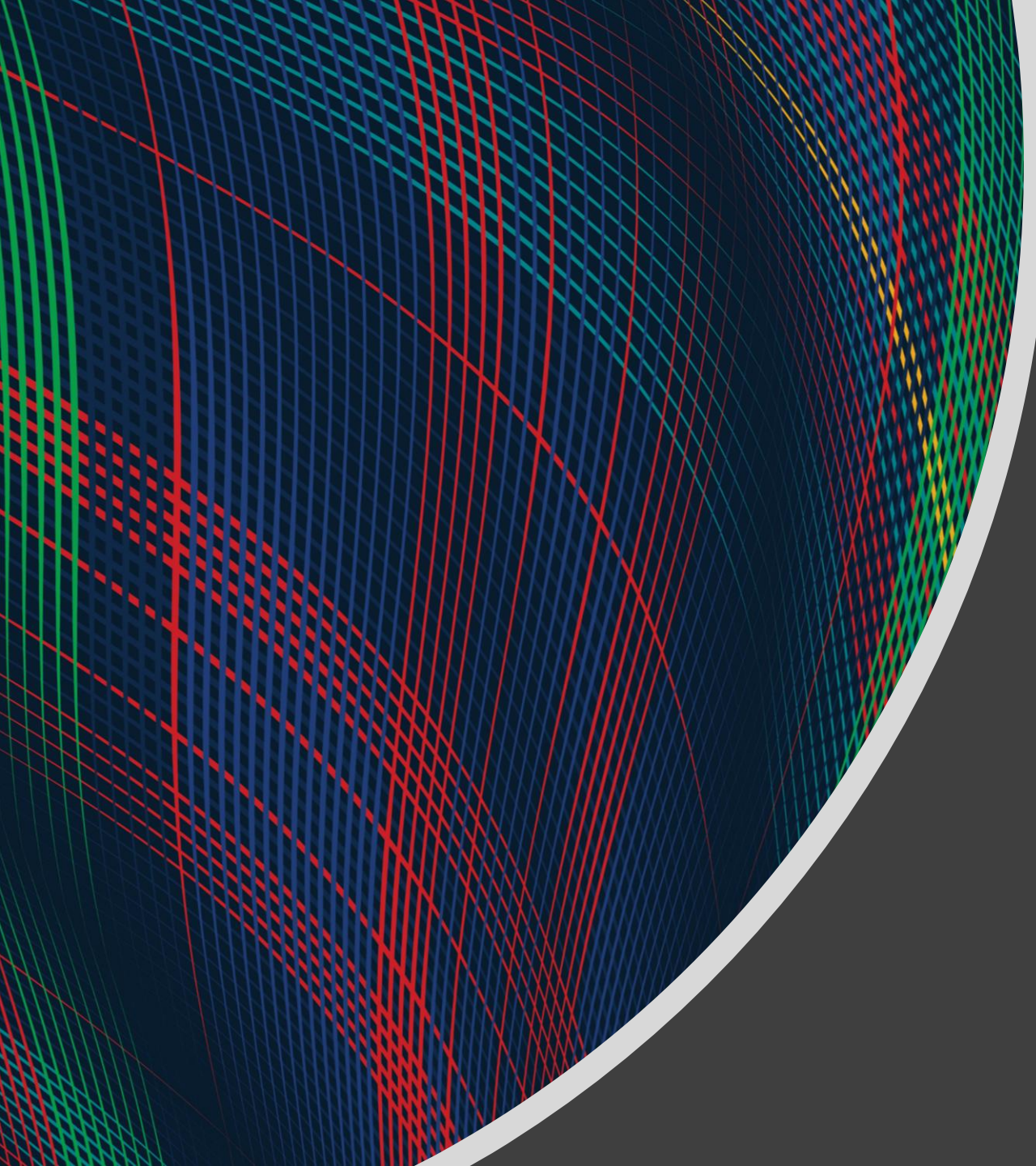
Midterm 2

- Mon, 11/9, during lecture

Videos

Fireside Chat about the CMU ML PhD Program

- Fri, 10/30, 8:00 pm
- See Piazza for details, including form to show interest

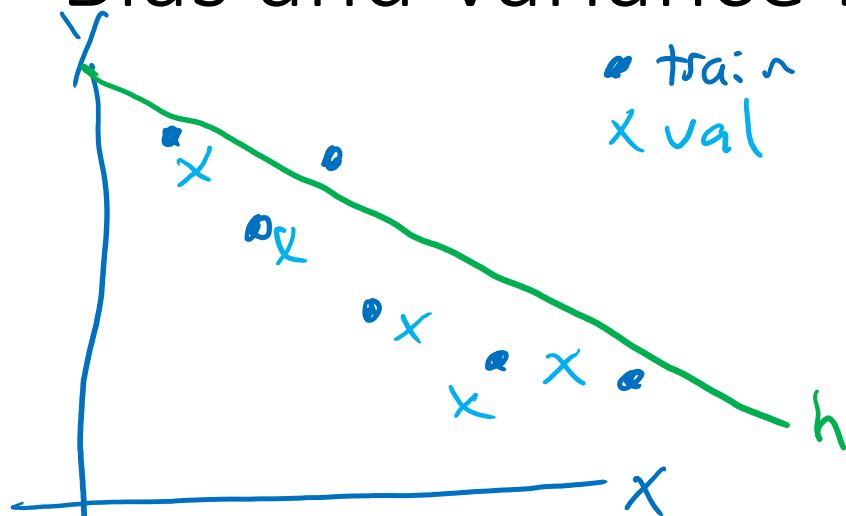


Introduction to Machine Learning

Learning Theory

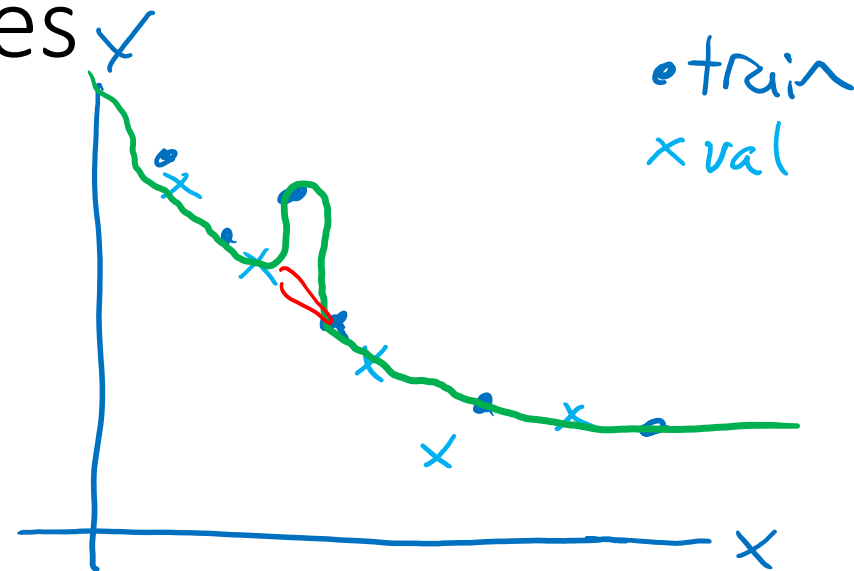
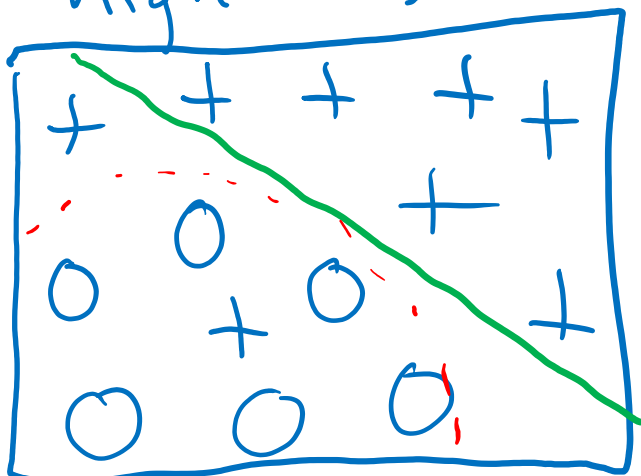
Instructor: Pat Virtue

Bias and Variance Examples



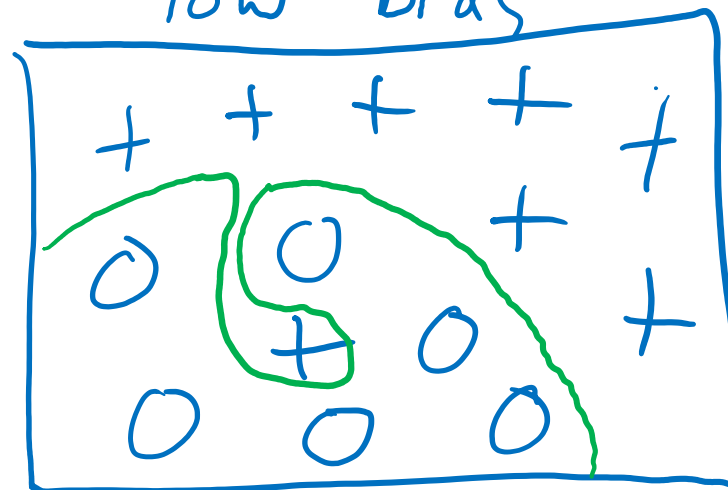
low variance

high bias



med variance

low bias



Previous Piazza Polls

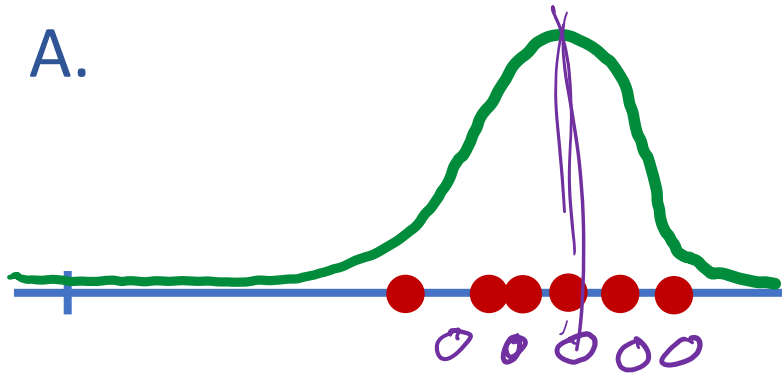
Poll 1: [SELECT TWO] Which have high variance? B D

Poll 2: [SELECT TWO] Which have high bias? C D

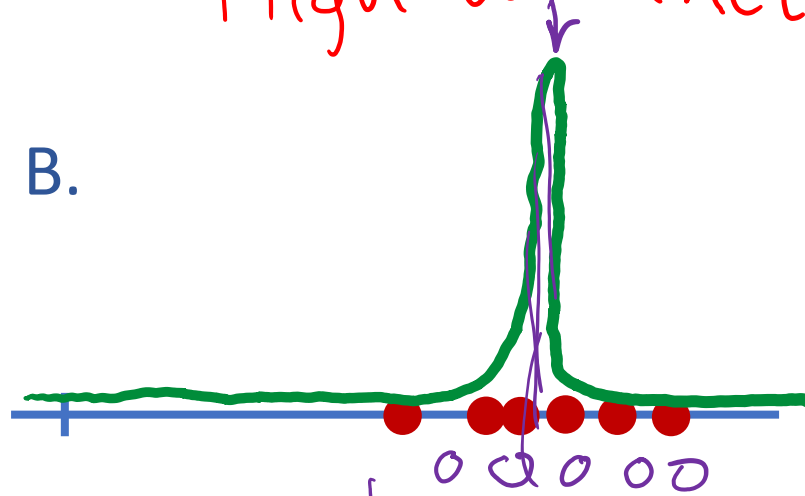
low Variance

High Variance

A.

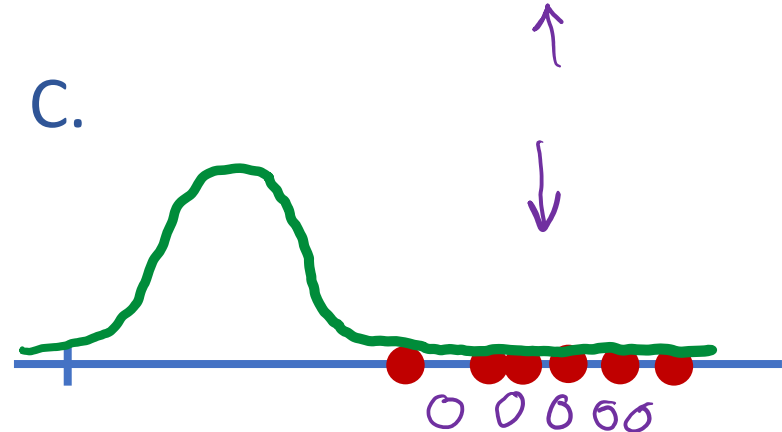


B.

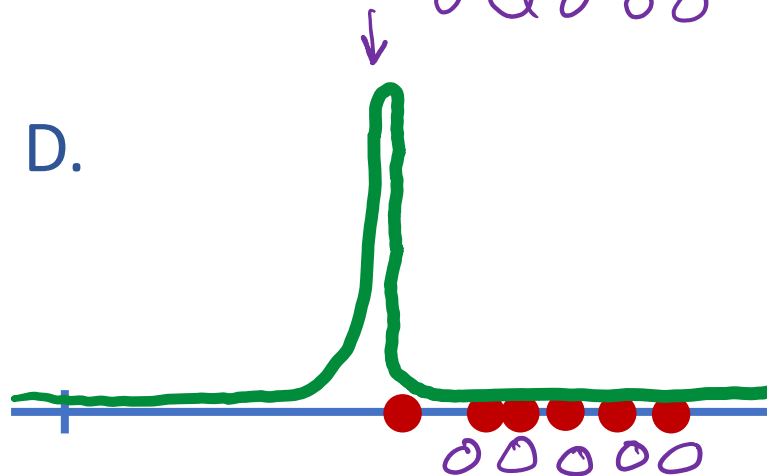


low bias

C.



D.

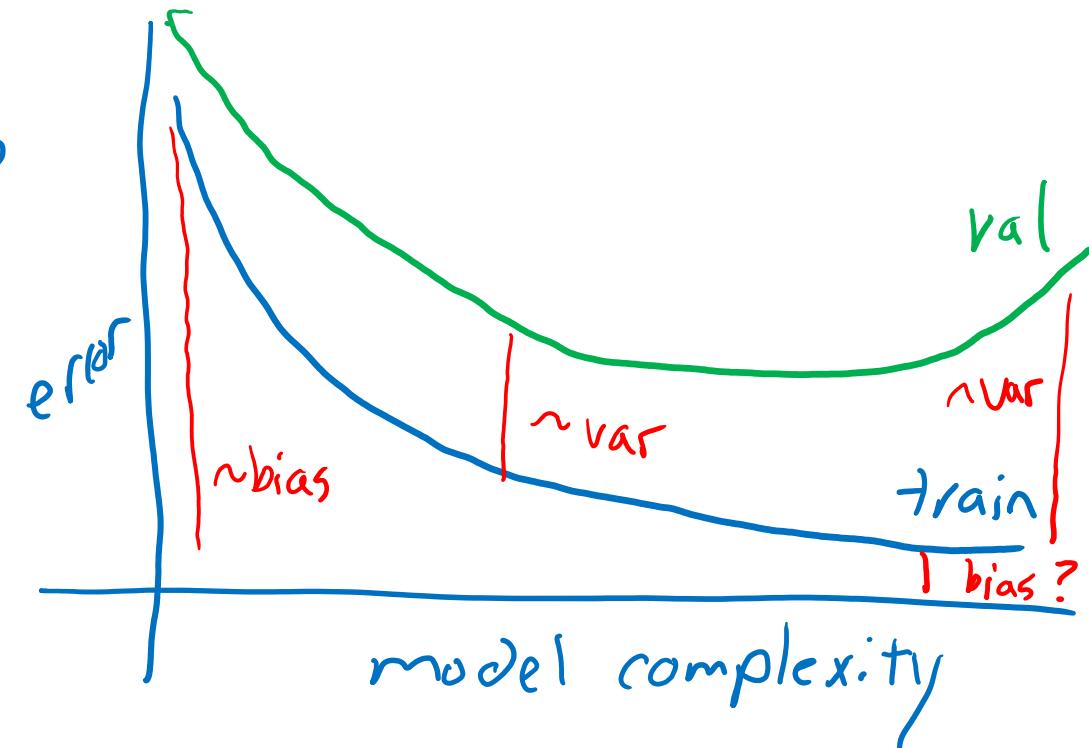


high bias

Bias and Variance Examples

Assume that $h^* \leq \underline{\underline{0.1}}$

	h_A	h_B
train err	0.25	0.1
val err	0.3	0.3
	high bias low var	low bias? high var



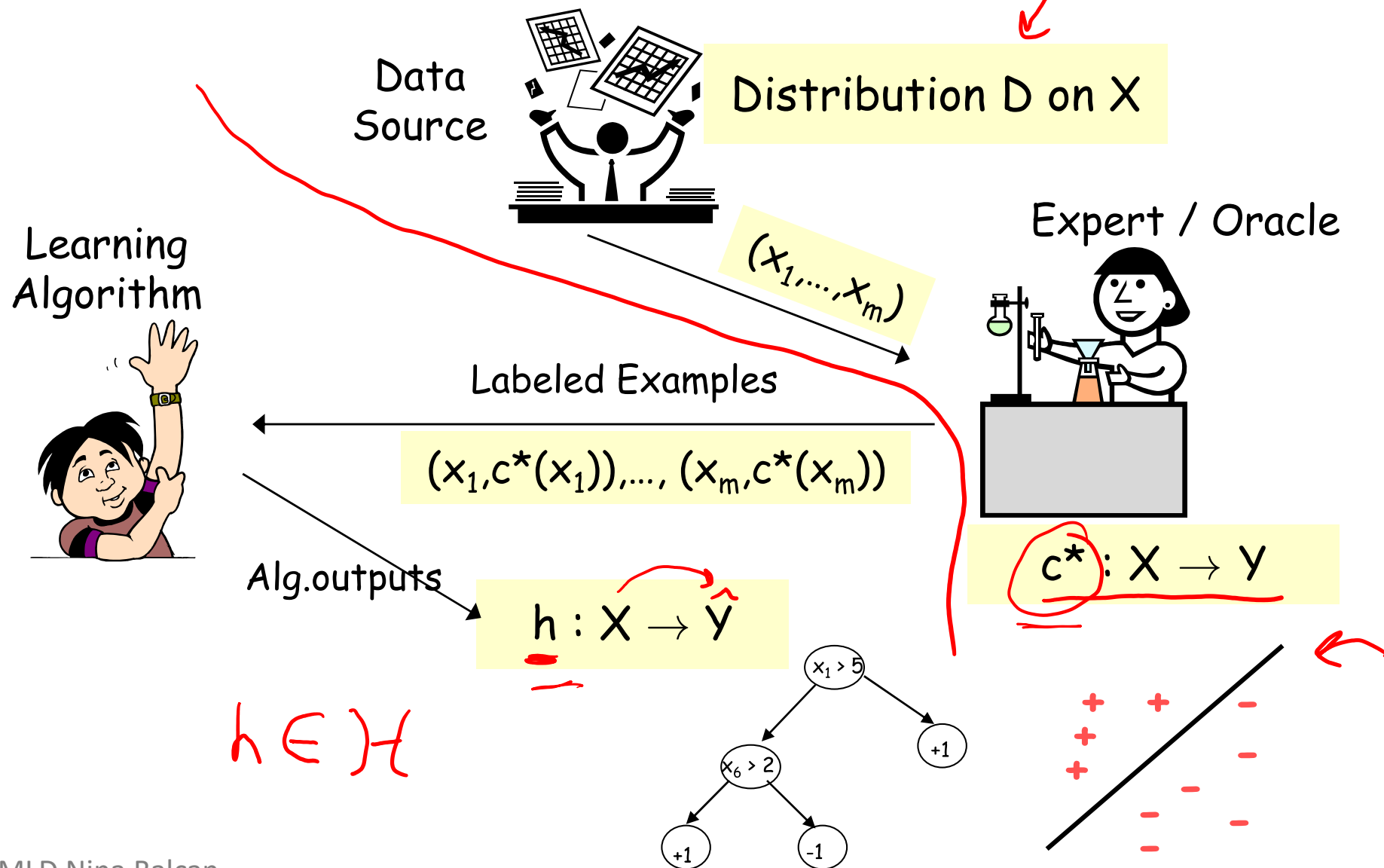
Questions

1. Given a classifier with zero training error, what can we say about **true error** (aka. generalization error)?
(Sample Complexity, Realizable Case)
2. Given a classifier with low training error, what can we say about **true error** (aka. generalization error)?
(Sample Complexity, Agnostic Case)
3. Is there a **theoretical justification for regularization** to avoid overfitting?
(Structural Risk Minimization)

$$\min J(\theta) + \lambda \|\theta\|$$

Model for Supervised Learning

$$x \sim p^*(x)$$



Risk, 0-1 Loss, and Error Rate

Risk is the expected loss over data points

0-1 loss means we have a cost of one when classify a point wrong

$$\ell(y, \hat{y}) = \begin{array}{cc} & \begin{array}{c} \hat{y}=0 \end{array} & \begin{array}{c} \hat{y}=1 \end{array} \\ \begin{array}{c} y=0 \end{array} & 0 & 1 \\ \begin{array}{c} y=1 \end{array} & 1 & 0 \end{array}$$

Risk for 0-1 loss is simply error rate

Two Types of Error

1. True Error (aka. **expected risk**)

$$\rightarrow R(h) = P_{\mathbf{x} \sim p^*(\mathbf{x})}(c^*(\mathbf{x}) \neq h(\mathbf{x}))$$

This quantity
is always
unknown

2. Train Error (aka. **empirical risk**)

$$\rightarrow \hat{R}(h) = P_{\mathbf{x} \sim \mathcal{S}}(c^*(\mathbf{x}) \neq h(\mathbf{x}))$$

$$= \frac{1}{N} \sum_{i=1}^N \mathbb{1}(c^*(\mathbf{x}^{(i)}) \neq h(\mathbf{x}^{(i)}))$$

$$= \frac{1}{N} \sum_{i=1}^N \mathbb{1}(y^{(i)} \neq h(\mathbf{x}^{(i)}))$$

We can
measure this
on the training
data

$\rightarrow$ where $\mathcal{S} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}_{i=1}^N$ is the training data set, and $\mathbf{x} \sim \mathcal{S}$ denotes that $\mathbf{x}$ is sampled from the empirical distribution.

PAC / SLT Model

1. Generate instances from *unknown* distribution p^*

$$\underline{\mathbf{x}^{(i)} \sim p^*(\mathbf{x}), \forall i} \quad (1)$$

2. Oracle labels each instance with *unknown* function c^*

$$y^{(i)} = c^*(\mathbf{x}^{(i)}), \forall i \quad (2)$$

3. Learning algorithm chooses hypothesis $h \in \mathcal{H}$ with low(est) training error, $\hat{R}(h)$

$$\hat{h} = \operatorname{argmin}_h \underline{\hat{R}(h)} \quad (3)$$

4. Goal: Choose an h with low generalization error $R(h)$

Three Hypotheses of Interest

The **true function** c^* is the one we are trying to learn and that labeled the training data:

$$y^{(i)} = c^*(\mathbf{x}^{(i)}), \forall i \quad (1)$$

The **expected risk minimizer** has lowest true error:

$$\rightarrow h^* = \operatorname{argmin}_{h \in \mathcal{H}} \underline{R(h)}$$

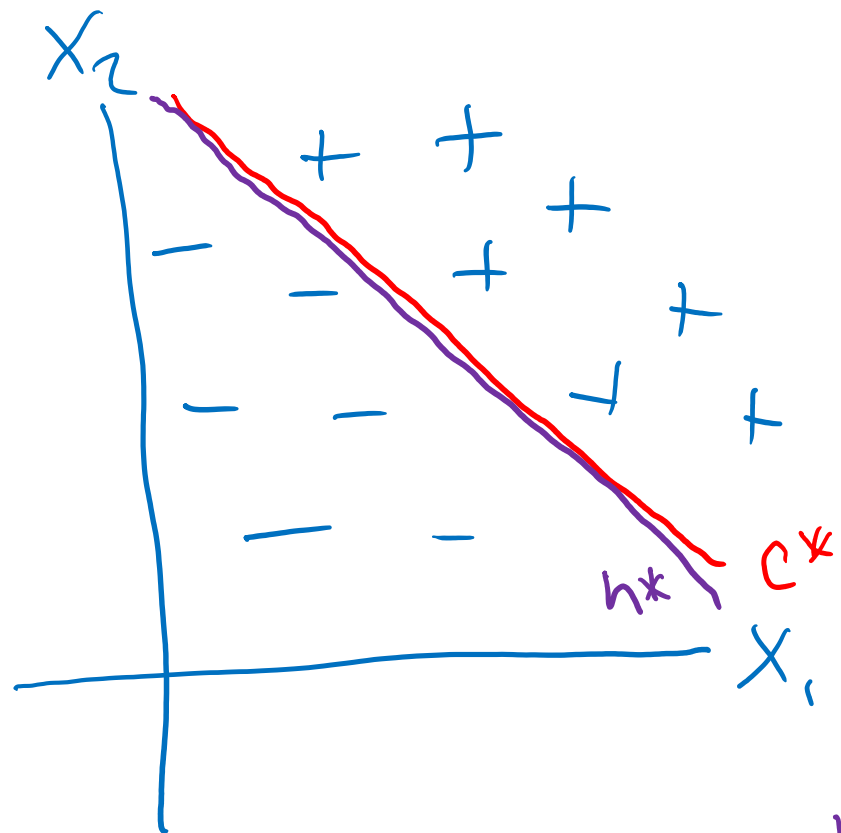
Question:
True or False:
 h^* and c^* are
always equal.

The **empirical risk minimizer** has lowest training error:

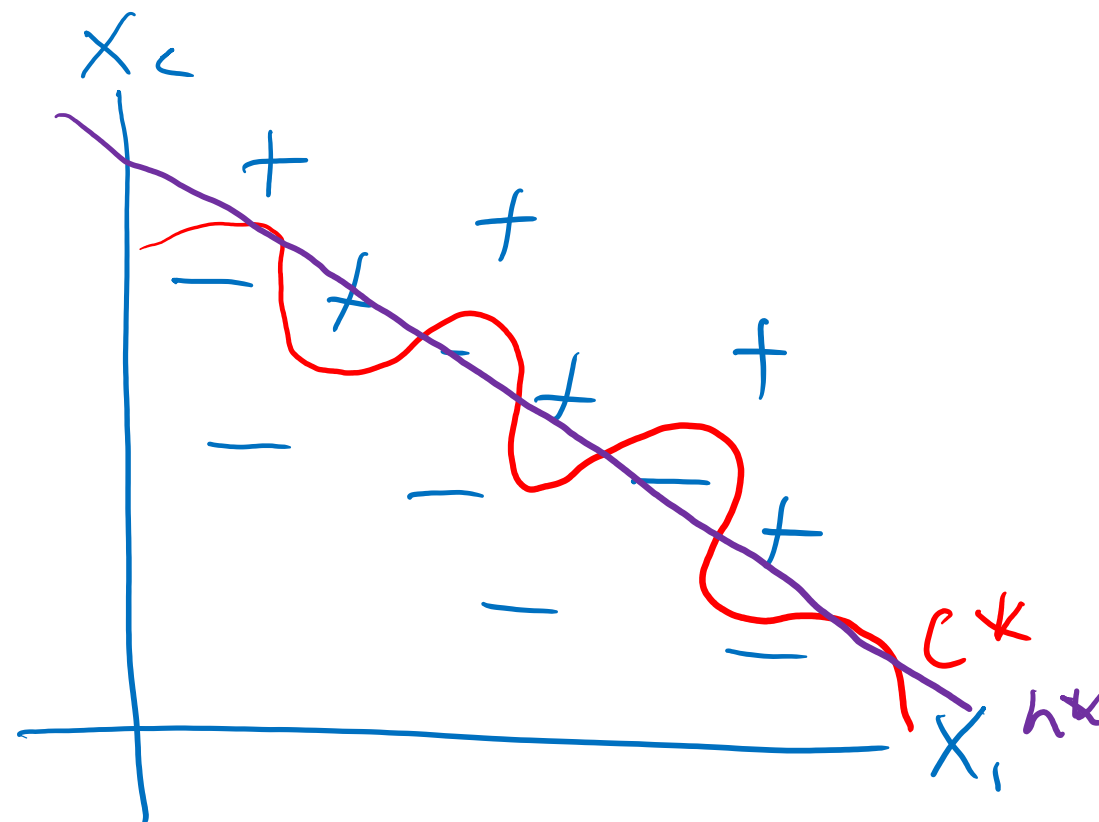
$$\rightarrow \hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}(h) \quad (3)$$

Piazza Poll 1

True or False, h^* and c^* are always equal.



$$h^* \in \mathcal{H}$$



PAC Learning

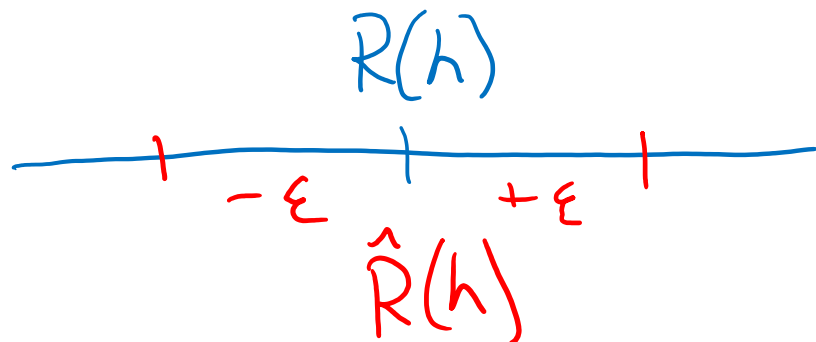
Can we bound $R(h)$ in terms of $\hat{R}(h)$?

Yes! PAC: Probably Approximately Correct

PAC learner yield $h \in \mathcal{H}$ that is approximately correct $R(h) \approx 0$
with high probability $p(R(h) \approx 0) \approx 1$

Definition: PAC Criterion:

$$P(\underbrace{|R(h) - \hat{R}(h)|}_{\text{small}} \leq \underbrace{\epsilon}_{\text{small}}) \geq 1 - \delta$$



Random? $\hat{R}(h)$

$$x_{\text{train}} \sim p^*(x)$$

PAC Learning



PAC Criterion

$$P(|R(h) - \hat{R}(h)| \leq \epsilon) \geq 1 - \delta$$

Definition: sample complexity is min. num. of training samples N s.t. the PAC criterion is satisfied for ϵ and δ

Definition: consistent hypothesis

a $h \in \mathcal{H}$ is consistent with training data
if $\hat{R}(h) = 0$

PAC Learning

The **PAC criterion** is that our learner produces a high accuracy learner with high probability:

$$\implies P(|R(h) - \hat{R}(h)| \leq \epsilon) \geq 1 - \delta \quad (1)$$

Suppose we have a learner that produces a hypothesis $h \in \mathcal{H}$ given a sample of N training examples. The algorithm is called **consistent** if for every ϵ and δ , there exists a positive number of training examples N such that for any distribution p^* , we have that:

$$P(|R(h) - \hat{R}(h)| > \epsilon) < \delta \quad (2)$$

The **sample complexity** is the minimum value of N for which this statement holds. If N is finite for some learning algorithm, then $\mathcal{H}$ is said to be **learnable**. If N is a polynomial function of $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ for some learning algorithm, then $\mathcal{H}$ is said to be **PAC learnable**.

PAC Learning

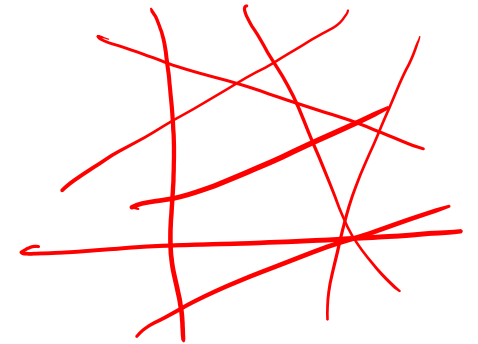
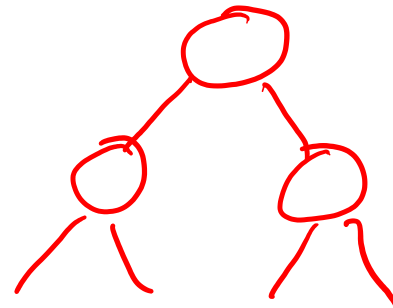
Four types of problems

Two cases for c^*

- Realizable: $c^* \in H$
- Agnostic: $c^* \in H$ or $c^* \notin H$

Two cases $|H|$

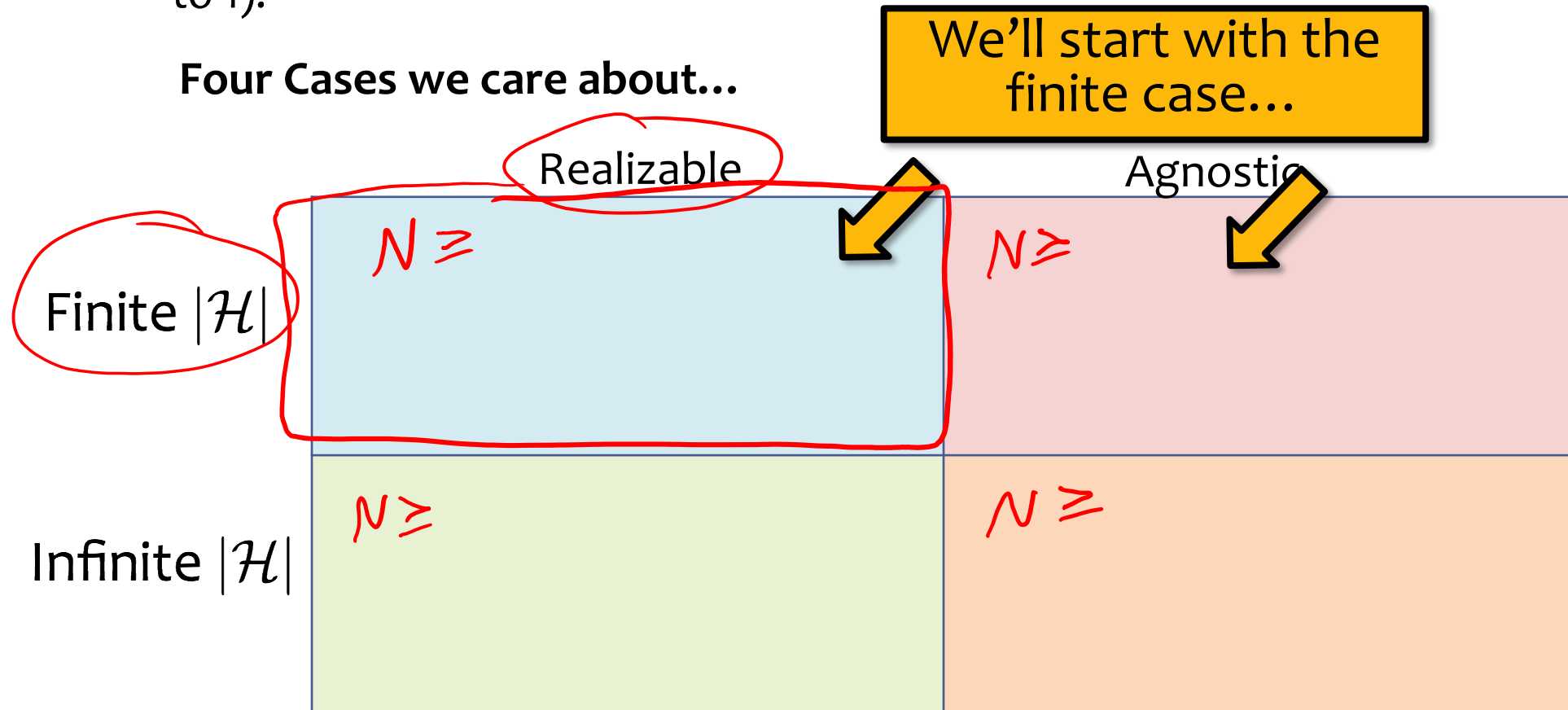
- $|H|$ is finite
- $|H|$ is infinite



Sample Complexity Results

Definition 0.1. The sample complexity of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...



PAC Learning

PAC Criterion

$$P(|R(h) - \hat{R}(h)| \leq \epsilon) \geq 1 - \delta$$

Theorem 1: Sample Complexity (Realizable, Finite $|\mathcal{H}|$)

$$N \geq \frac{1}{\epsilon} \left[\log \underset{\uparrow}{|\mathcal{H}|} + \log \frac{1}{\delta} \right]$$

N examples is sufficient to ensure that
with Prob $1 - \delta$

all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $\underline{\underline{R(h) \leq \epsilon}}$
↑

PAC Learning

Proof of Theorem 1

See PAC Learning: Theorem 1 notes and video

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

	Realizable	Agnostic
Finite $ \mathcal{H} $	Thm. 1 $N \geq \frac{1}{\epsilon} [\log(\mathcal{H}) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	
Infinite $ \mathcal{H} $		

Piazza Poll 2

$$\frac{x_1}{(1-x_1)} \quad \text{---} \quad \text{---} \quad \text{---} \quad \frac{x_M}{(1-x_M)} \quad \leftarrow$$

Question:

Suppose H = class of conjunctions over $\mathbf{x}$ in $\{0,1\}^M$

Example hypotheses:

$$\rightarrow h(\mathbf{x}) = x_1 (1-x_3) x_5$$

$$\rightarrow h(\mathbf{x}) = x_1 (1-x_2) x_4 (1-x_5)$$

If $M = 10$, $\varepsilon = 0.1$, $\delta = 0.01$, how many examples suffice according to Theorem 1?

Answer:

A. $10^*(2*\ln(10)+\ln(100)) \approx 92$

B. $10^*(3*\ln(10)+\ln(100)) \approx 116$

C. $10^*(10*\ln(2)+\ln(100)) \approx 116$

D. $10^*(10*\ln(3)+\ln(100)) \approx 156$

E. $100^*(2*\ln(10)+\ln(10)) \approx 691$

F. $100^*(3*\ln(10)+\ln(10)) \approx 922$

G. $100^*(10*\ln(2)+\ln(10)) \approx 924$

H. $100^*(10*\ln(3)+\ln(10)) \approx 1329$

$$|H| = 3^{10}$$



Thm. 1 $N \geq \frac{1}{\epsilon} [\log(|\mathcal{H}|) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.

Piazza Poll 2

Question:

Suppose H = class of conjunctions over $\mathbf{x}$ in $\{0,1\}^M$

Example hypotheses:

$$h(\mathbf{x}) = x_1 (1-x_3) x_5$$

$$h(\mathbf{x}) = x_1 (1-x_2) x_4 (1-x_5)$$

If $M = 10$, $\epsilon = 0.1$, $\delta = 0.01$, how many examples suffice according to Theorem 1?

Answer:

- A. $10^*(2*\ln(10)+\ln(100)) \approx 92$
- B. $10^*(3*\ln(10)+\ln(100)) \approx 116$
- C. $10^*(10*\ln(2)+\ln(100)) \approx 116$
- D. $10^*(10*\ln(3)+\ln(100)) \approx 156$
- E. $100^*(2*\ln(10)+\ln(10)) \approx 691$
- F. $100^*(3*\ln(10)+\ln(10)) \approx 922$
- G. $100^*(10*\ln(2)+\ln(10)) \approx 924$
- H. $100^*(10*\ln(3)+\ln(10)) \approx 1329$

Thm. 1 $N \geq \frac{1}{\epsilon} [\log(|\mathcal{H}|) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

	Realizable	Agnostic
Finite $ \mathcal{H} $	Thm. 1 $N \geq \frac{1}{\epsilon} [\log(\mathcal{H}) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	Thm. 2 $N \geq \frac{1}{2\epsilon^2} [\log(\mathcal{H}) + \log(\frac{2}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $R(h) - \hat{R}(h) \leq \epsilon$.
Infinite $ \mathcal{H} $		

1. Bound is **inversely linear in epsilon** (e.g. halving the error requires double the examples)
2. Bound is **only logarithmic in $|\mathcal{H}|$** (e.g. quadrupling the hypothesis space only requires double the examples)

1. Bound is **inversely quadratic in epsilon** (e.g. halving the error requires 4x the examples)
2. Bound is **only logarithmic in $|\mathcal{H}|$** (i.e. same as Realizable case)

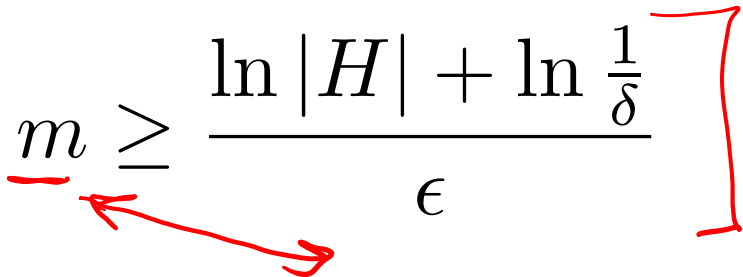
	Realizable	Agnostic
Finite $ \mathcal{H} $	Thm. 1 $N \geq \frac{1}{\epsilon} [\log(\mathcal{H}) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	Thm. 2 $N \geq \frac{1}{2\epsilon^2} [\log(\mathcal{H}) + \log(\frac{2}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $ R(h) - \hat{R}(h) \leq \epsilon$.
Infinite $ \mathcal{H} $		

Using a PAC bound

$$|H|e^{-m\epsilon} \leq \delta$$

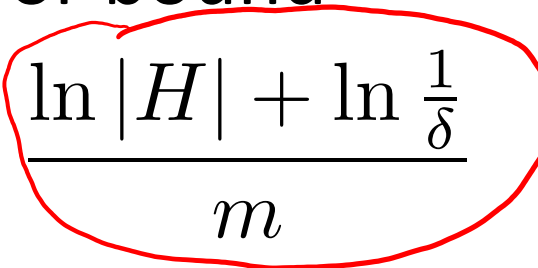
- Given ϵ and δ , yields sample complexity

#training data, $m \geq \frac{\ln |H| + \ln \frac{1}{\delta}}{\epsilon}$



- Given m and δ , yields error bound

error, $\epsilon \geq \frac{\ln |H| + \ln \frac{1}{\delta}}{m}$



Summary of PAC bounds for finite model classes

With probability $\geq 1-\delta$,

→ 1) For all $h \in H$ s.t. $\text{error}_{\text{train}}(h) = 0$,

$$\text{error}_{\text{true}}(h) \leq \varepsilon = \frac{\ln |H| + \ln \frac{1}{\delta}}{m}$$

Haussler's bound

→ 2) For all $h \in H$

$$|\text{error}_{\text{true}}(h) - \text{error}_{\text{train}}(h)| \leq \varepsilon = \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}$$

Hoeffding's bound

PAC bound and Bias-Variance tradeoff

$$P(|\text{error}_{\text{true}}(h) - \text{error}_{\text{train}}(h)| \geq \epsilon) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

- Equivalently, with probability $\geq 1 - \delta$

$$\text{error}_{\text{true}}(h) \leq \underbrace{\text{error}_{\text{train}}(h)}_{\text{bias}} + \underbrace{\sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}}_{\text{Variance}}$$

- Fixed $|H|$

Training size

m small

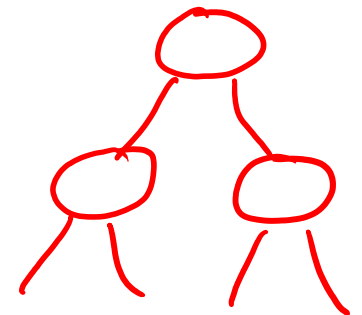
m large

small

large

large

small



PAC bound and Bias-Variance tradeoff

$$P(|\text{error}_{\text{true}}(h) - \text{error}_{\text{train}}(h)| \geq \epsilon) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

- Equivalently, with probability $\geq 1 - \delta$

$$\text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{\overbrace{\ln |H|}^{\text{bias}} + \overbrace{\ln \frac{2}{\delta}}^{\text{variance}}}{2m}}$$

- Fixed m

Model class		
H large (complex)	small	large
H small (simple)	large	small

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

	Realizable	Agnostic
Finite $ \mathcal{H} $	Thm. 1 $N \geq \frac{1}{\epsilon} [\log(\mathcal{H}) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	Thm. 2 $N \geq \frac{1}{2\epsilon^2} [\log(\mathcal{H}) + \log(\frac{2}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $ R(h) - \hat{R}(h) \leq \epsilon$.
<u>Infinite $\mathcal{H}$</u>	$VC(\mathcal{H})$	$VC(\mathcal{H})$

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

$c^* \in \mathcal{H}$

$c^* \in \mathcal{H}$ or $c^* \notin \mathcal{H}$

	Realizable	Agnostic
Finite $ \mathcal{H} $	Thm. 1 $N \geq \frac{1}{\epsilon} [\log(\mathcal{H}) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	Thm. 2 $N \geq \frac{1}{2\epsilon^2} [\log(\mathcal{H}) + \log(\frac{2}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $ R(h) - \hat{R}(h) \leq \epsilon$.
Infinite $ \mathcal{H} $	Thm. 3 $N = O(\frac{1}{\epsilon} [\text{VC}(\mathcal{H}) \log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})])$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	Thm. 4 $N = O(\frac{1}{\epsilon^2} [\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})])$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $ R(h) - \hat{R}(h) \leq \epsilon$.

$\log |\mathcal{H}|$
 $\downarrow$
 $\text{VC}(\mathcal{H})$

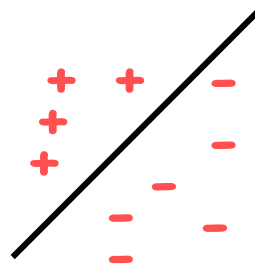
VC DIMENSION



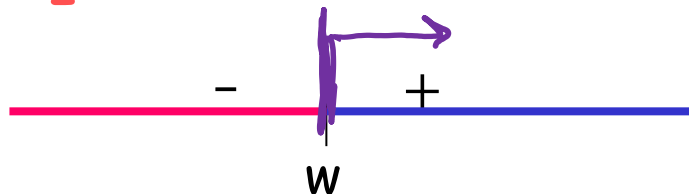
What if H is infinite?



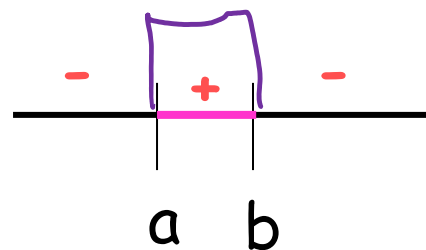
E.g., linear separators in $\mathbb{R}^d$



E.g., thresholds on the real line



E.g., intervals on the real line



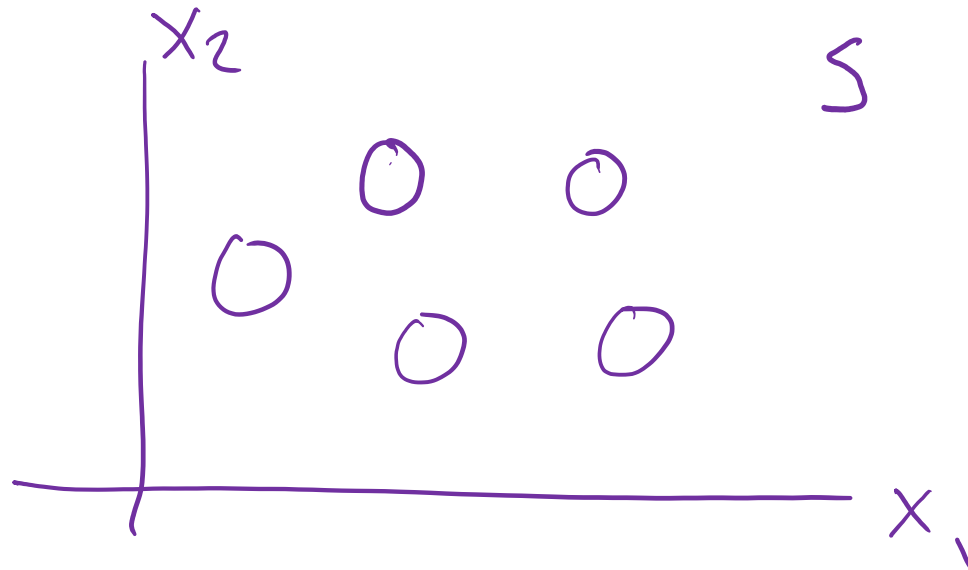
Shattering, VC-dimension

Definition:

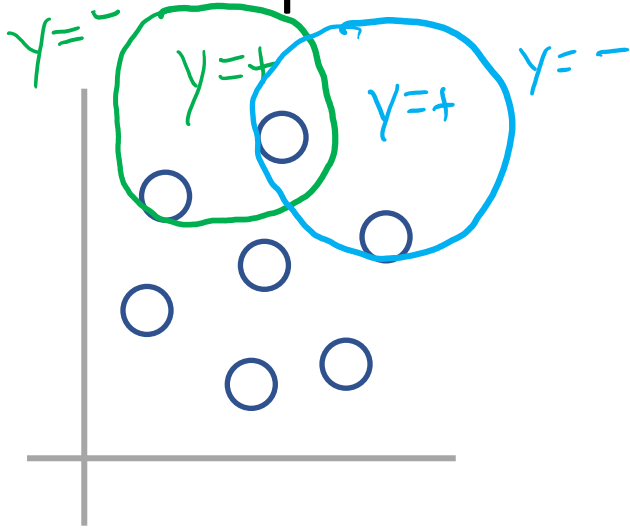
$H[S]$ - the set of splittings of dataset S using concepts from H .

H shatters S if $|H[S]| = 2^{|S|}$.

A set of points S is shattered by H if there are hypotheses in H that split S in all of the $2^{|S|}$ possible ways; i.e., all possible ways of classifying points in S are achievable using concepts in H .



Example: Shattering for Binary Classification



S

points $|S| = 7$

labelings $2^{|S|}$

H is all circular decision boundaries

$H[S]$ # splittings of S by H

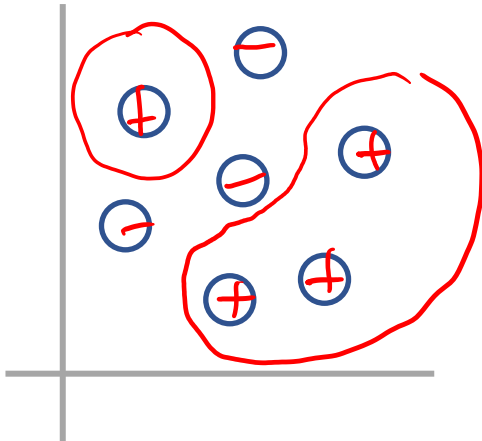
Piazza Poll 3

Does $\mathcal{H}$ shatter $\mathcal{S}$, where $\mathcal{H}$ = set of circular decision boundaries and $\mathcal{S}$ = set of 2D points?

i.e. Does the number of splittings, $|\mathcal{H}[\mathcal{S}]|$, equal $2^{|\mathcal{S}|}$?

i.e. Can a circular decision boundary perfectly separate any labelling of $\mathcal{S}$?

No



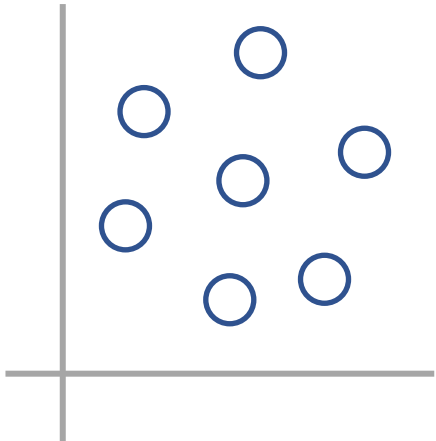
Piazza Poll 3

Does $\mathcal{H}$ shatter $\mathcal{S}$, where $\mathcal{H}$ = set of circular decision boundaries and $\mathcal{S}$ = set of 2D points?

i.e. Does the number of splittings, $|\mathcal{H}[\mathcal{S}]|$, equal $2^{|\mathcal{S}|}$?

i.e. Can a circular decision boundary perfectly separate any labelling of $\mathcal{S}$?

No.



Shattering, VC-dimension

Definition:

$H[S]$ - the set of splittings of dataset S using concepts from H .

H shatters S if $|H[S]| = 2^{|S|}$.

A set of points S is shattered by H if there are hypotheses in H that split S in all of the $2^{|S|}$ possible ways; i.e., all possible ways of classifying points in S are achievable using concepts in H .

Shattering, VC-dimension

Definition:

$H[S]$ - the set of splittings of dataset S using concepts from H .

H shatters S if $|H[S]| = 2^{|S|}$.

A set of points S is shattered by H if there are hypotheses in H that split S in all of the $2^{|S|}$ possible ways; i.e., all possible ways of classifying points in S are achievable using concepts in H .

Definition: VC-dimension (Vapnik-Chervonenkis dimension)

The **VC-dimension** of a hypothesis space H is the cardinality of the largest set S that can be shattered by H .

If arbitrarily large finite sets can be shattered by H , then $\text{VCdim}(H) = \infty$

Shattering, VC-dimension

Definition: VC-dimension (Vapnik-Chervonenkis dimension)

The **VC-dimension** of a hypothesis space H is the cardinality of the largest set S that can be shattered by H .

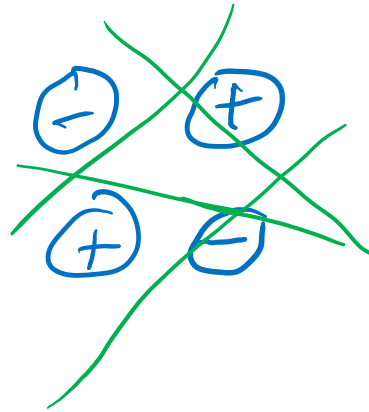
If arbitrarily large finite sets can be shattered by H , then $\text{VCdim}(H) = \infty$

To show that VC-dimension is d :

- - there exists a set of d points that can be shattered
- there is no set of $d+1$ points that can be shattered.

Fact: If H is finite, then $\text{VCdim}(H) \leq \log(|H|)$.

Example: VC Dimension for Linear Separators



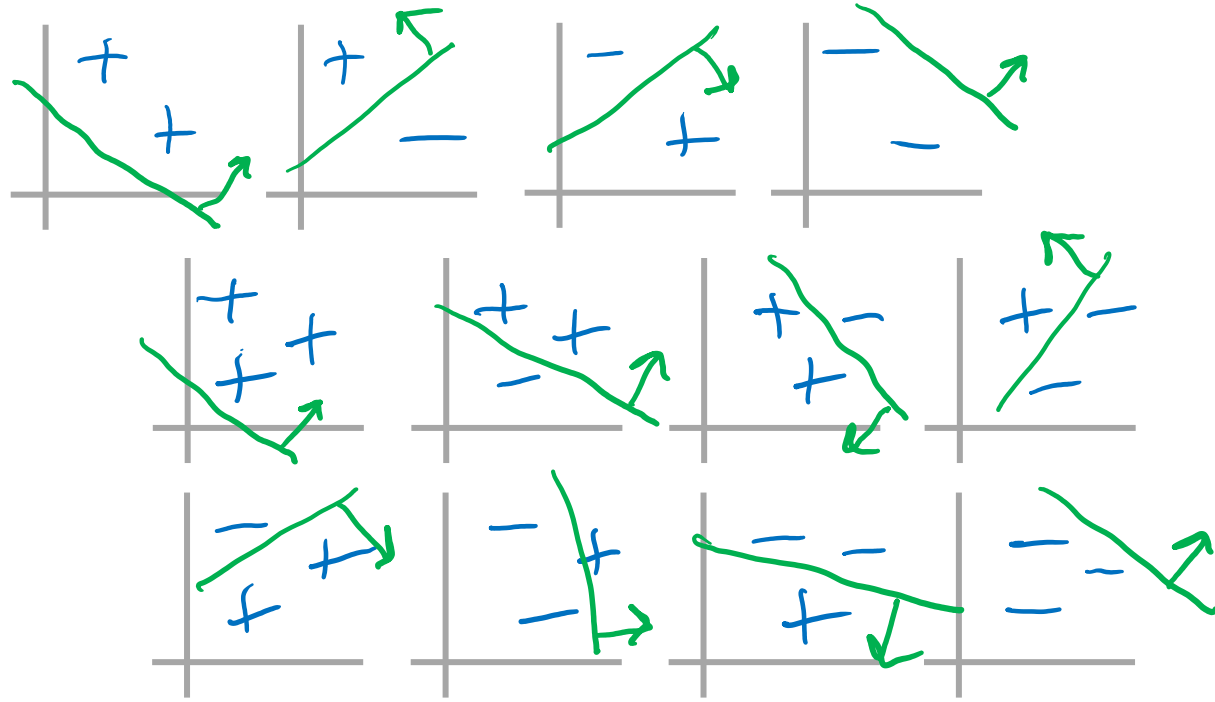
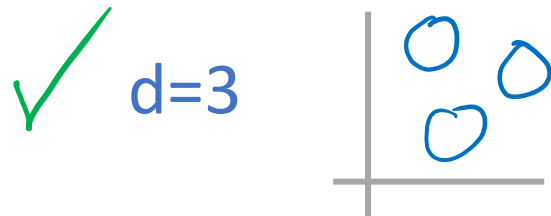
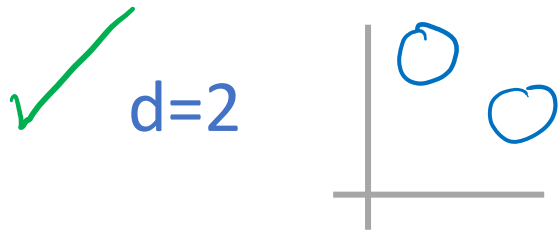
Consider $\mathcal{H}$ = linear separators in 2D. To prove $VC(\mathcal{H}) = d$:

1. $\exists \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d$ and $\mathcal{H}$ shatters $\mathcal{S}$
2. $\nexists \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d + 1$ and $\mathcal{H}$ shatters $\mathcal{S}$

1. Pick one (unlabeled) dataset for d

List all possible labelings of $\mathcal{S}$

Show that we can shatter $\mathcal{S}$

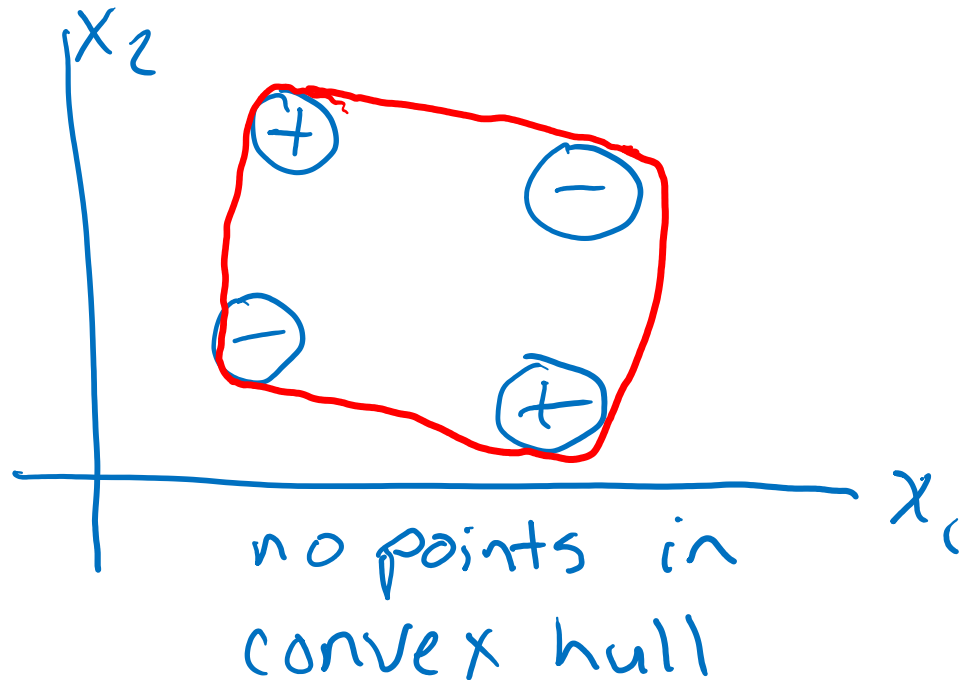
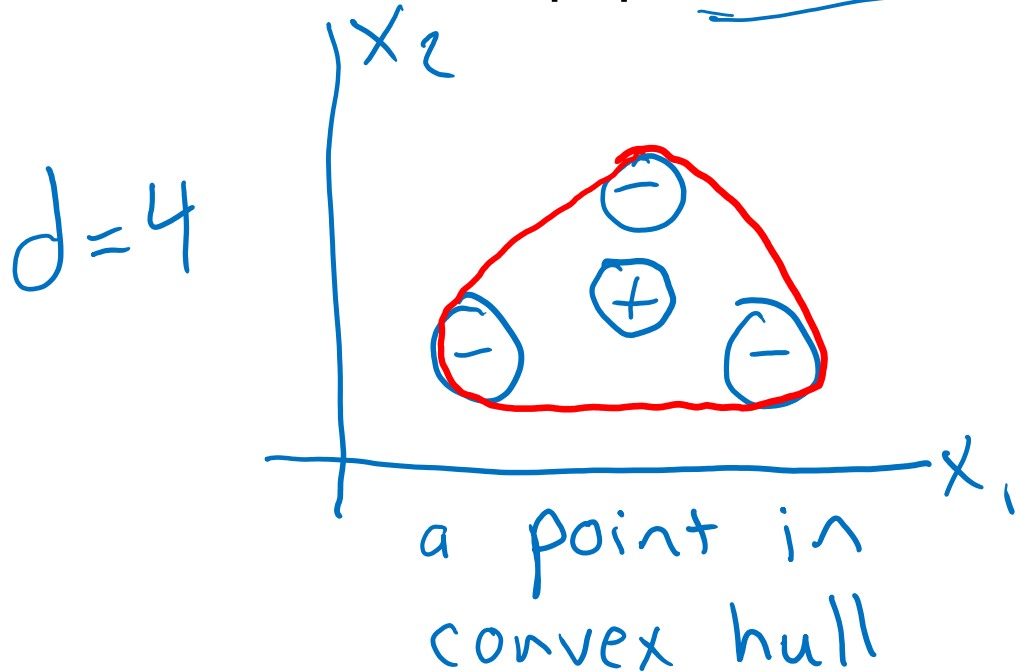


Example: VC Dimension for Linear Separators

Consider $\mathcal{H}$ = linear separators in 2D. To prove $VC(\mathcal{H}) = d$:

1. $\exists \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d$ and $\mathcal{H}$ shatters $\mathcal{S}$
2. $\nexists \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d + 1$ and $\mathcal{H}$ shatters $\mathcal{S}$
2. $\forall \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = \underline{d + 1}$ $\mathcal{H}$ cannot shatter $\mathcal{S}$

$$VC(\mathcal{H}: \text{lin sep } M=2) = 3$$



Example: VC Dimension for Linear Separators

Consider $\mathcal{H}$ = linear separators in 2D. To prove $VC(\mathcal{H}) = d$:

- 1. $\exists \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d$ and $\mathcal{H}$ shatters $\mathcal{S}$
- 2. $\nexists \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d + 1$ and $\mathcal{H}$ shatters $\mathcal{S}$

But...

Isn't there a dataset of size $d=3$ that can't be shattered?



$\exists$ vs. $\forall$

VCDim

- Proving **VC Dimension** requires us to show that **there exists ($\exists$)** a dataset of size d that can be shattered and that **there does not exist ($\nexists$)** a dataset of size $d+1$ that can be shattered

Shattering

- Proving that a particular dataset can be **shattered** requires us to show that **for all ($\forall$)** labelings of the dataset, our hypothesis class contains a hypothesis that can correctly classify it

Sample Complexity Results

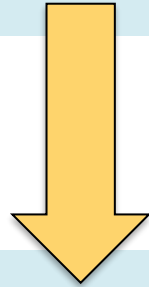
Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

	Realizable	Agnostic
Finite $ \mathcal{H} $	Thm. 1 $N \geq \frac{1}{\epsilon} [\log(\mathcal{H}) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	Thm. 2 $N \geq \frac{1}{2\epsilon^2} [\log(\mathcal{H}) + \log(\frac{2}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $ R(h) - \hat{R}(h) \leq \epsilon$.
Infinite $ \mathcal{H} $	Thm. 3 $N = O(\frac{1}{\epsilon} [\text{VC}(\mathcal{H}) \log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})])$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	Thm. 4 $N = O(\frac{1}{\epsilon^2} [\text{VC}(\mathcal{H}) \log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})])$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $ R(h) - \hat{R}(h) \leq \epsilon$.

SLT-style Corollaries

Thm. 1 $N \geq \frac{1}{\epsilon} \left[\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right) \right]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.



Solve the inequality in Thm.1 for epsilon to obtain Corollary 1

Corollary 1 (Realizable, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1 - \delta)$, for any h in $\mathcal{H}$ consistent with the training data (i.e. $\hat{R}(h) = 0$),

$$R(h) \leq \frac{1}{N} \left[\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right) \right]$$

We can obtain similar corollaries for each of the theorems...

SLT-style Corollaries

Corollary 1 (Realizable, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1 - \delta)$, for any h in $\mathcal{H}$ consistent with the training data (i.e. $\hat{R}(h) = 0$),

$$R(h) \leq \frac{1}{N} \left[\ln(|\mathcal{H}|) + \ln \left(\frac{1}{\delta} \right) \right]$$

Corollary 2 (Agnostic, Finite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1 - \delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2N} \left[\ln(|\mathcal{H}|) + \ln \left(\frac{2}{\delta} \right) \right]}$$

SLT-style Corollaries

Corollary 3 (Realizable, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1 - \delta)$, for any hypothesis h in $\mathcal{H}$ consistent with the data (i.e. with $\hat{R}(h) = 0$),

$$R(h) \leq O \left(\frac{1}{N} \left[\text{VC}(\mathcal{H}) \ln \left(\frac{N}{\text{VC}(\mathcal{H})} \right) + \ln \left(\frac{1}{\delta} \right) \right] \right) \quad (1)$$

Corollary 4 (Agnostic, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1 - \delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + O \left(\sqrt{\frac{1}{N} \left[\text{VC}(\mathcal{H}) + \ln \left(\frac{1}{\delta} \right) \right]} \right) \quad (2)$$

SLT-style Corollaries

Corollary 3 (Realizable, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1 - \delta)$, for any hypothesis h in $\mathcal{H}$ consistent with the data (i.e. with $\hat{R}(h) = 0$),

$$R(h) \leq O \left(\frac{1}{N} \left[\text{VC}(\mathcal{H}) \ln \left(\frac{N}{\text{VC}(\mathcal{H})} \right) + \ln \left(\frac{1}{\delta} \right) \right] \right) \quad (1)$$

Corollary 4 (Agnostic, Infinite $|\mathcal{H}|$). For some $\delta > 0$, with probability at least $(1 - \delta)$, for all hypotheses h in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + O \left(\sqrt{\frac{1}{N} \left[\text{VC}(\mathcal{H}) + \ln \left(\frac{1}{\delta} \right) \right]} \right) \quad (2)$$



Should these corollaries inform
how we do model selection?

PAC Bounds and Model Selection

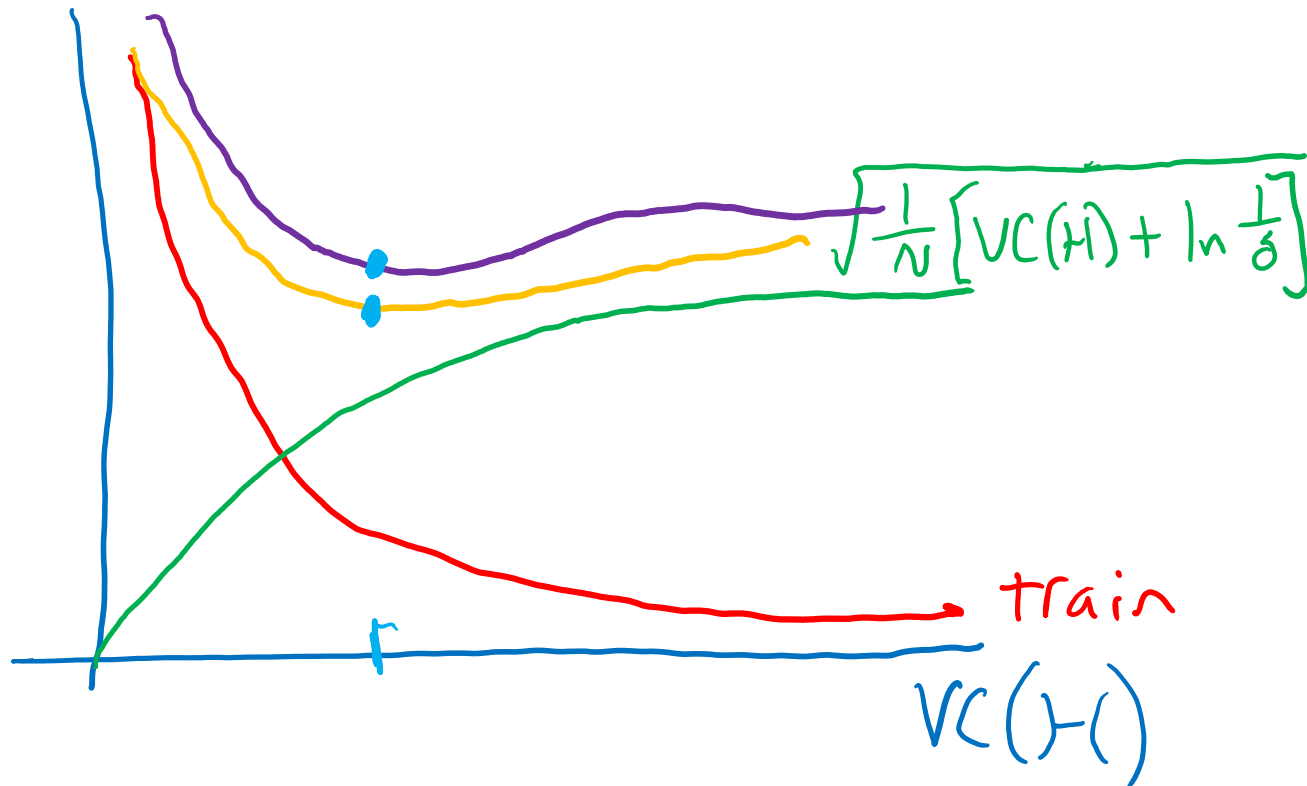
Is Corollary 4 useful?

$$VC(\mathcal{H}: \text{linear sep}) = M + 1$$

Variance

$$\underline{R(h)} \leq \underbrace{\hat{R}(h)}_{\text{Bias}} + O\left(\underbrace{\sqrt{\frac{1}{N} \left[VC(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right) \right]}}_{M=5}\right)$$

$M=2$



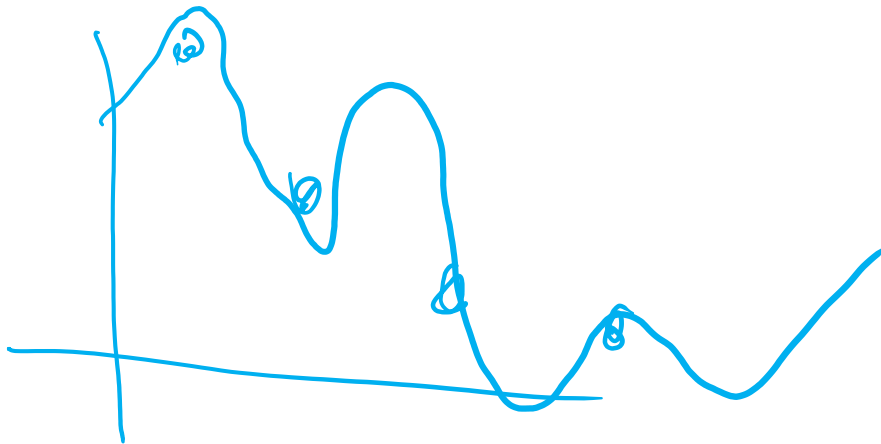
$$\begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \xrightarrow{\phi} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix} \quad \begin{matrix} \theta_0 \\ \theta_1 \quad X \\ \theta_2 \\ \theta_3 \\ \theta_4 \quad X \\ \theta_5 \end{matrix}$$

PAC Bounds and Regularization

Example: Linear separator in $\mathbb{R}^M$

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N} \left[\text{VC}(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right) \right]}\right)$$

$$\operatorname{argmin} J(\theta) + \lambda \|\theta\|,$$



Questions For Today

1. Given a classifier with zero training error, what can we say about generalization error?
(Sample Complexity, Realizable Case)
2. Given a classifier with low training error, what can we say about generalization error?
(Sample Complexity, Agnostic Case)
3. Is there a theoretical justification for regularization to avoid overfitting?
(Structural Risk Minimization)

PAC Learning Objectives

You should be able to...

- Identify the properties of a learning setting and assumptions required to ensure low generalization error
- Distinguish true error, train error, test error
- Define PAC and explain what it means to be approximately correct and what occurs with high probability
- Apply sample complexity bounds to real-world learning examples
- Distinguish between a large sample and a finite sample analysis
- Theoretically motivate regularization