

# Announcements

## Assignments

- HW10 (programming + “written”)
  - Due Thu 4/30, 11:59 pm

## Final Exam

- Stay tuned to Piazza for details
- Date: Mon 5/11, 5:30 – 8:30 pm
- Format: “online assignment” in Gradescope
- Scope: Content before this week
- Practice exam: Out later this week
- Recitation this Friday: Review session

# Introduction to Machine Learning

## Learning Theory

Instructor: Pat Virtue

# Questions For Today

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)

# PAC Learning

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

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

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  $\epsilon$  and  $\delta$ , 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**.

# Sample Complexity Results

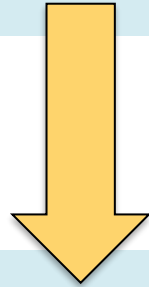
**Definition 0.1.** The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

**Four Cases we care about...**

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

# 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...*

# Using a PAC bound

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

- Given  $\epsilon$  and  $\delta$ , yields sample complexity

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

- Given  $m$  and  $\delta$ , yields error bound

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

# 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  $\geq 1 - \delta$

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

- 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  $\geq 1 - \delta$

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

- Fixed  $m$

Model class

$|H|$  large (complex)

$|H|$  small (simple)

small

large

large



small

# PAC bound for decision trees with $k$ leaves – Bias-Variance revisited

With prob  $\geq 1-\delta$   $\text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}$

With  $H_k \leq n^{k-1} 2^{2k-1}$ , we get

$$\text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{(k-1) \ln n + (2k-1) \ln 2 + \ln \frac{2}{\delta}}{2m}}$$

		
$k = m$	0	large ( $\sim > \frac{1}{2}$ )
$k < m$	$> 0$	small ( $\sim < \frac{1}{2}$ )

# What about continuous hypothesis spaces?

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

- Continuous model class (e.g. linear classifiers):
  - $|H| = \infty$
  - Infinite gap???
- **As with decision trees, complexity of model class only depends on maximum number of points that can be classified exactly (and not necessarily its size)!**

# What about continuous hypothesis spaces?

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

$$\text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + 8 \sqrt{\frac{VC(H) \left( \ln \frac{m}{VC(H)} + 1 \right) + \ln \frac{8}{\delta}}{2m}}$$

↓  
Instead of  $\ln |H|$

# 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...**

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

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

# **VC DIMENSION**

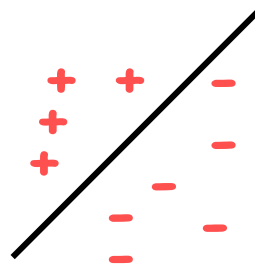




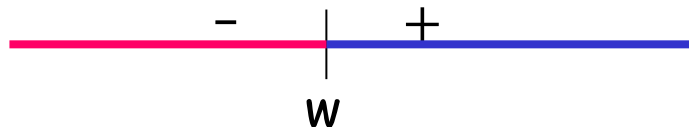
# What if $H$ is infinite?



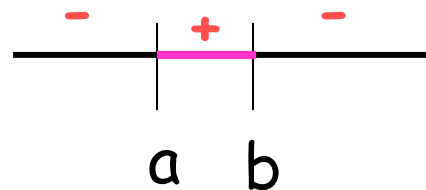
E.g., linear separators in  $\mathbb{R}^d$



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

# Piazza Poll 1

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  $\mathcal{S}$ ?

A.

B.

C.

# 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  $\mathcal{S}$



# 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.  $|\mathcal{S}| = d + 1$   $\mathcal{H}$  cannot 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.  $|\mathcal{S}| = d + 1$   $\mathcal{H}$  cannot 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  $\mathcal{S}$

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

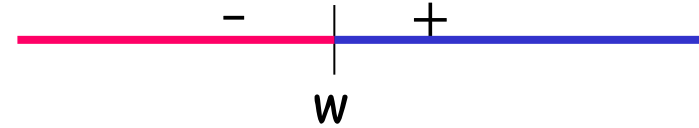
## Shattering

- 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

# Shattering, VC-dimension

If the VC-dimension is  $d$ , that means **there exists** a set of  $d$  points that can be shattered, but there is **no** set of  $d+1$  points that can be shattered.

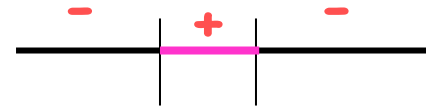
E.g.,  $H$  = Thresholds on the real line



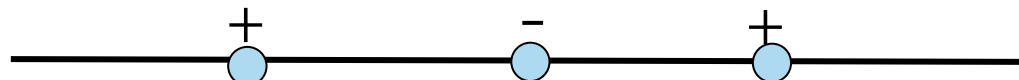
$$\text{VCdim}(H) = 1$$



E.g.,  $H$  = Intervals on the real line



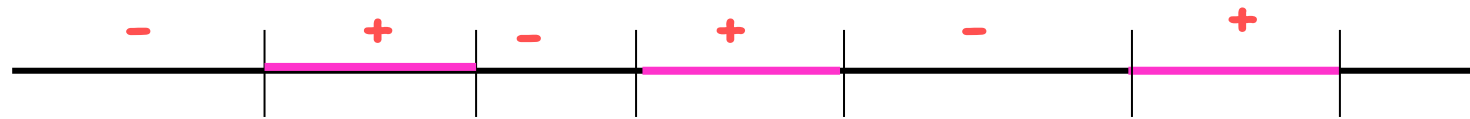
$$\text{VCdim}(H) = 2$$



# Shattering, VC-dimension

If the VC-dimension is  $d$ , that means **there exists** a set of  $d$  points that can be shattered, but there is **no** set of  $d+1$  points that can be shattered.

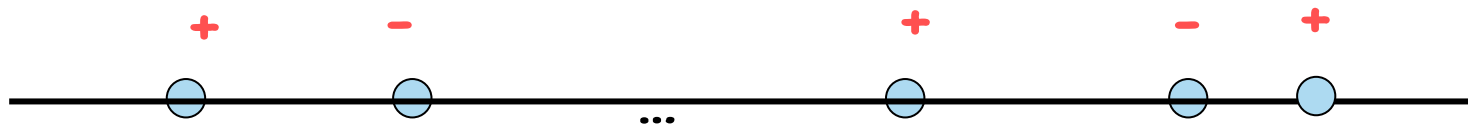
E.g.,  $H = \text{Union of } k \text{ intervals on the real line}$   $\text{VCdim}(H) = 2k$



$$\text{VCdim}(H) \geq 2k$$

A sample of size  $2k$  shatters  
(treat each pair of points as a  
separate case of intervals)

$$\text{VCdim}(H) < 2k + 1$$



# Sample Complexity Results

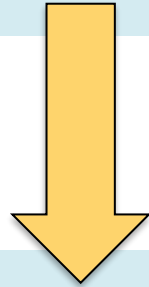
**Definition 0.1.** The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

**Four Cases we care about...**

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

# 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

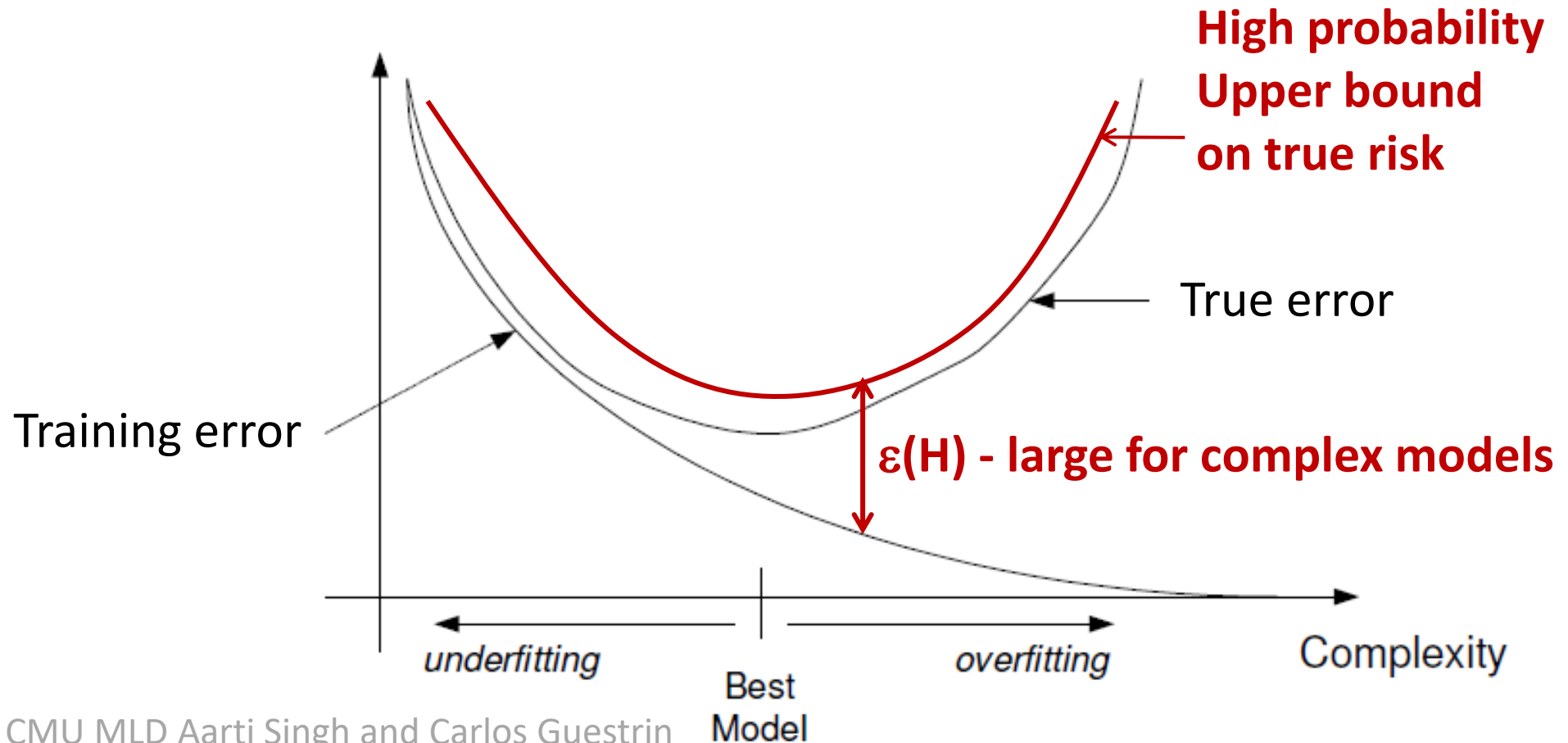
Is Corollary 4 useful?

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

# PAC Bounds

With probability  $\geq 1-\delta$ , for all  $h \in H$ ,

$$|\text{error}_{\text{true}}(h) - \text{error}_{\text{train}}(h)| \leq \varepsilon(H)$$



# Using PAC Bounds to pick a hypothesis and model selection

- Empirical Risk Minimization (ERM):

$$\hat{h} = \arg \min_{h \in H} \text{error}_{\text{train}}(h)$$

- Structural Risk Minimization (SRM):

$$\hat{k} = \arg \min_{k \geq 1} \{ \text{error}_{\text{train}}(\hat{h}_k) + \epsilon(H_k) \}$$

- Provide insights, but often too loose in practice:  
optimize

$$\hat{k} = \arg \min_{k \geq 1} \{ \text{error}_{\text{train}}(\hat{h}_k) + \lambda \epsilon(H_k) \}$$

where  $\lambda$  is chosen by cross-validation

# PAC Bounds and Regularization

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

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

# 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