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)

Slide credit: CMU MLD Matt Gormley

PAC Learning

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

true error train error

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

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:

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

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**.

Slide credit: CMU MLD Matt Gormley

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

Slide credit: CMU MLD Matt Gormley

SLT-style Corollaries

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

$|R(h) - \hat{R}(h)| \leq \varepsilon$

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…

Slide credit: CMU MLD Matt Gormley

Using a PAC bound

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

- Given ε and δ, yields sample complexity

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

- Given m and δ, yields error bound

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

Slide credit: CMU MLD Aarti Singh and Carlos Guestrin

Summary of PAC bounds for finite model classes

With probability ≥ 1-δ,

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

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

Haussler's bound

2) For all h $\in$ H

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

$$|R(h) - \hat{R}(h)|$$

Hoeffding's bound

Slide credit: CMU MLD Aarti Singh and Carlos Guestrin

PAC bound and Bias-Variance tradeoff

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

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

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

- Fixed |H|

Training size		
m small	small	large
m large	large	small

Slide credit: CMU MLD Aarti Singh and Carlos Guestrin

PAC bound and Bias-Variance tradeoff

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

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

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

- Fixed m

Model class	$\downarrow$	$\downarrow$
|H| large (complex)	small	large
|H| small (simple)	large	small

Slide credit: CMU MLD Aarti Singh and Carlos Guestrin

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

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

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

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

k = m	0	large (~ > ½)
k < m	>0	small (~ <½)

Slide credit: CMU MLD Aarti Singh and Carlos Guestrin

What about continuous hypothesis spaces?

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

- 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)!**

Slide credit: CMU MLD Aarti Singh and Carlos Guestrin

What about continuous hypothesis spaces?

|H| finite

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

|H| infinite

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

Instead of ln|H|

Slide credit: CMU MLD Aarti Singh and Carlos Guestrin

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

Slide credit: CMU MLD Matt Gormley

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…

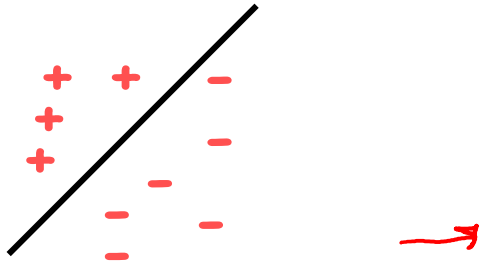
	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.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

Slide credit: CMU MLD Matt Gormley

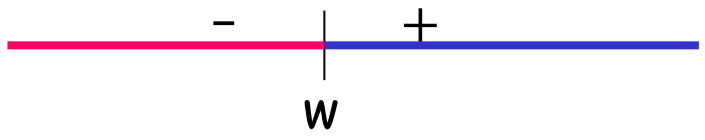
VC DIMENSION

What if H is infinite?

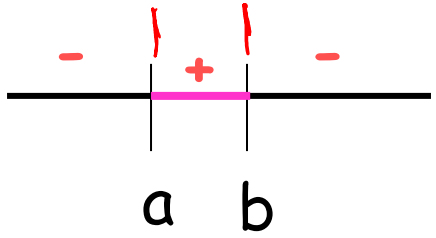
E.g., linear separators in R^d



E.g., thresholds on the real line



E.g., intervals on the real line



Slide credit: CMU MLD Nina Balcan

Shattering, VC-dimension

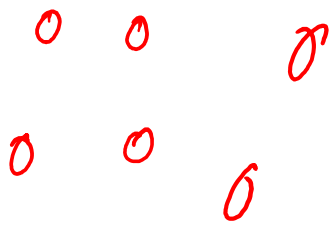
Definition:

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

hypotheses

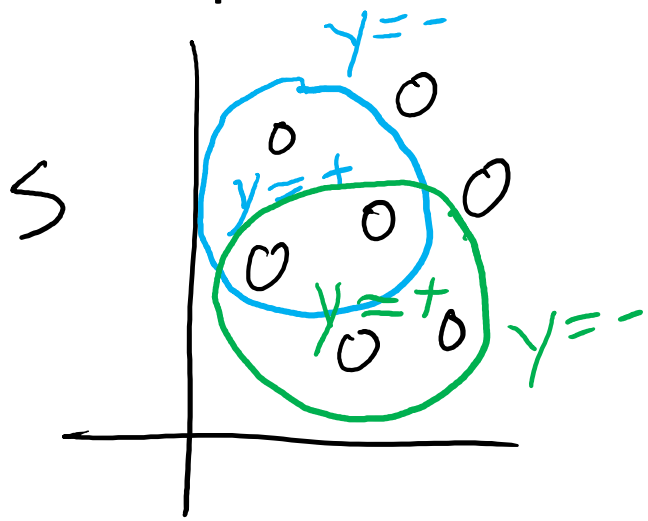
H shatters S if $|H[S]| = 2^{|S|}$.

A set of points S is shattered by H is 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.



Slide credit: CMU MLD Nina Balcan

Example: Shattering for Binary Classification



points $|S| = 7$

labellings of $S = \underline{\underline{2^{|S|}}}$

$\mathcal{H}$ = all circular decision boundaries

$\underline{\underline{H[S]}}$ = # splittings of S by $\mathcal{H}$

Slide credit: CMU MLD Matt Gormley

Piazza Poll 1

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

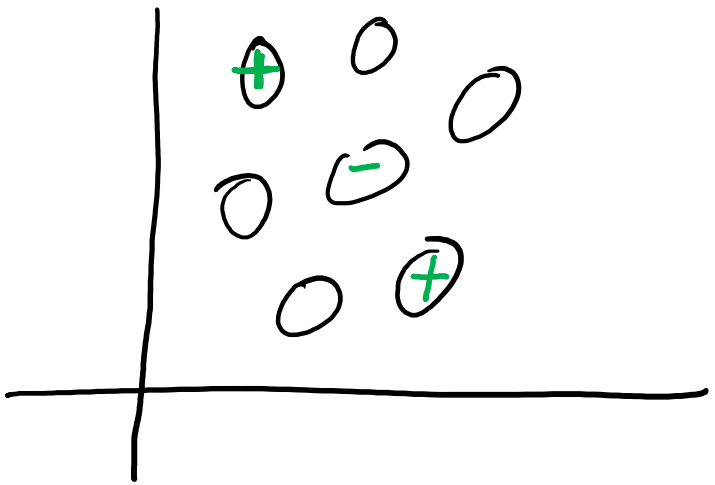
i.e. Does the number of splittings, $|\mathcal{H}[\mathcal{S}]|$, equal $2^{|\mathcal{S}|}$?

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

A. Yes

B. Calamity

C. No



$|H[S]| < 2^{|S|}$

Shattering, VC-dimension

Definition:

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

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

A set of points S is shattered by H is 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.

Slide credit: CMU MLD Nina Balcan

Shattering, VC-dimension

Definition:

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

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

A set of points S is shattered by H is 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 $\mathrm{VCdim(H)} = \infty$

Slide credit: CMU MLD Nina Balcan

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 $\mathrm{VCdim}(H) = \infty$

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 $\mathrm{VCdim}(H) \leq \log(|H|)$.

Slide credit: CMU MLD Nina Balcan

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

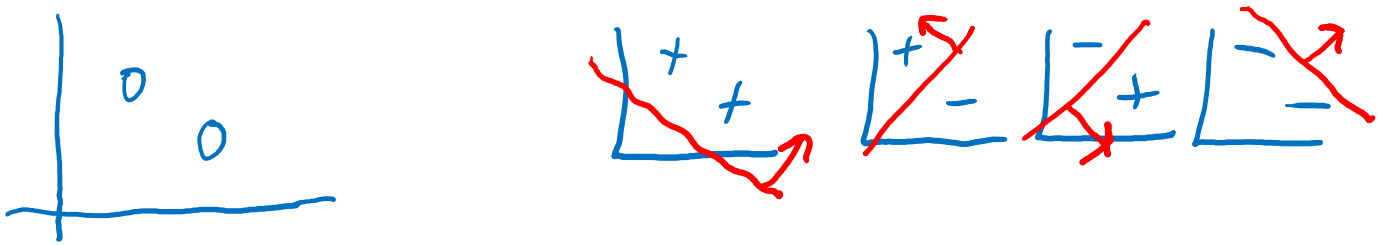
1

Pick one dataset (unlabelled) for d

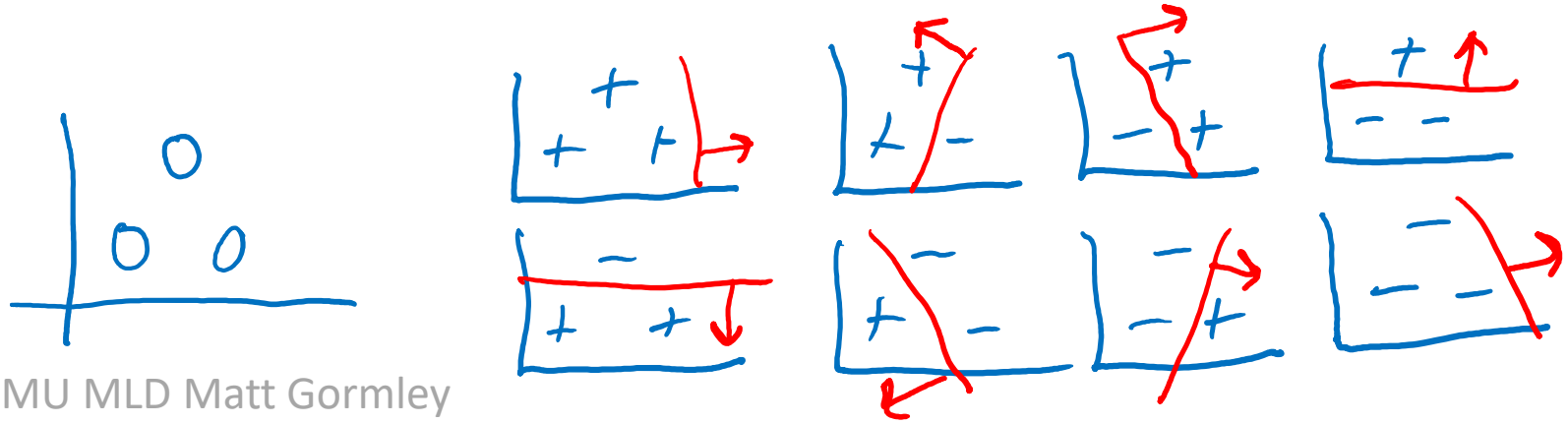
list all possible labellings of S

Show that we can shatter S

d=2



d=3



Slide credit: CMU MLD Matt Gormley

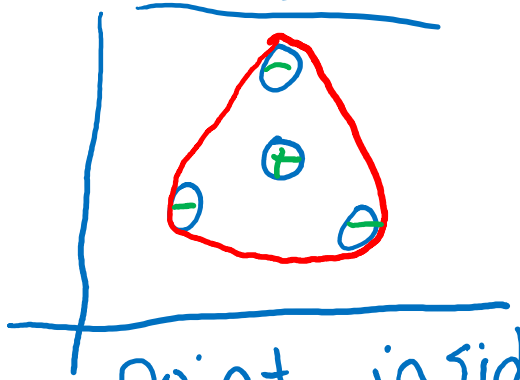
Example: VC Dimension for Linear Separators

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. $!\exists\, \mathcal{S} \in \mathcal{X}$ 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 $\mathcal{S}$

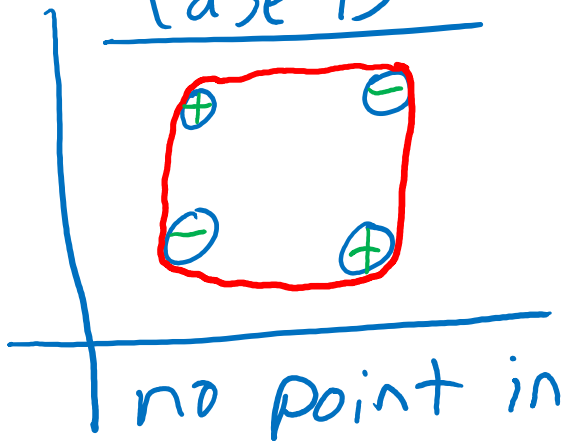
$d=4$

Case A



point inside convex hull

Case B



no point in convex hull

Slide credit: CMU MLD Matt Gormley

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. $!\exists\, \mathcal{S} \in \mathcal{X}$ 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 $\mathcal{S}$

Slide credit: CMU MLD Matt Gormley

Example: VC Dimension for Linear Separators

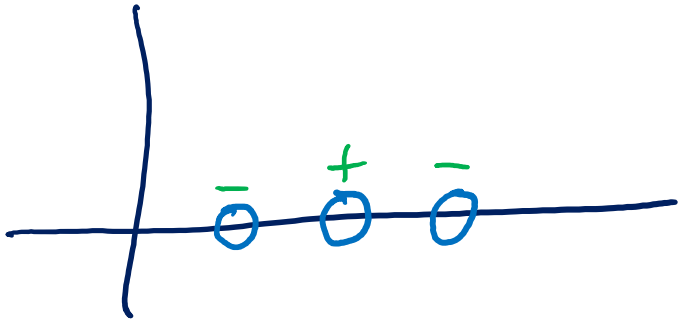
Consider $\mathcal{H}$ = linear separators in 2D. To prove $VC(\mathcal{H}) = d$:

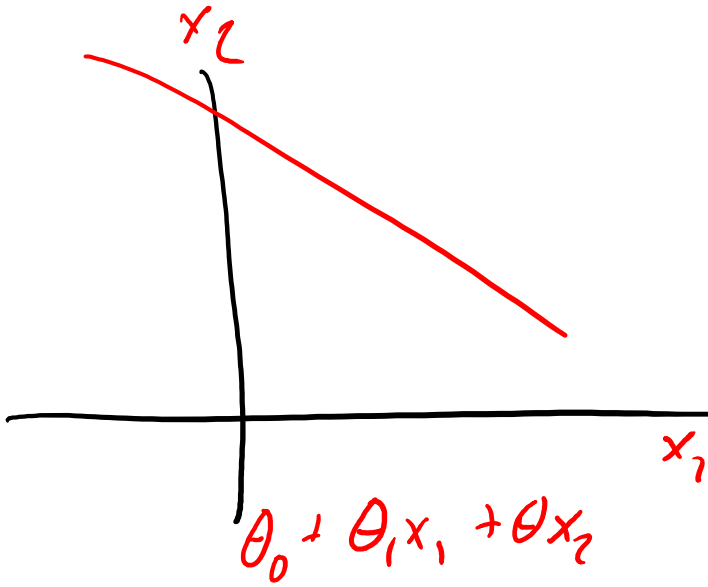
1. $\exists\, \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d$ and $\mathcal{H}$ shatters $\mathcal{S}$
2. $!\exists\, \mathcal{S} \in \mathcal{X}$ s.t. $|\mathcal{S}| = d + 1$ and $\mathcal{H}$ shatters $\mathcal{S}$

$\forall$ labellings

But…

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





$VC(H) = 3$

$VC(H) = M + 1$

Slide credit: CMU MLD Matt Gormley

∃ vs. ∀

VCDim

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

Shattering

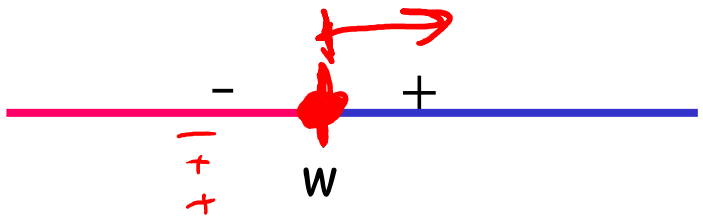
- Proving that a particular dataset can be **shattered** requires us to show that **for all (∀)** labelings of the dataset, our hypothesis class contains a hypothesis that can correctly classify it

Slide credit: CMU MLD Matt Gormley

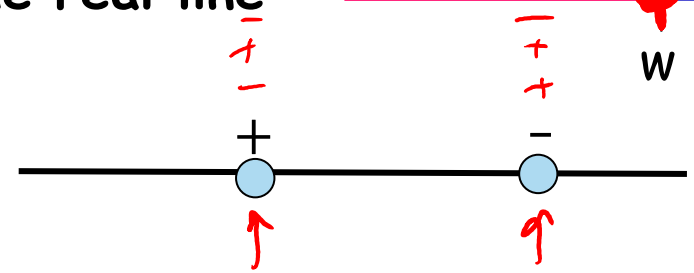
Shattering, VC-dimension

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.

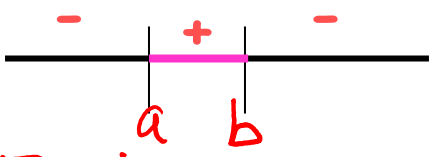
E.g., H= Thresholds on the real line



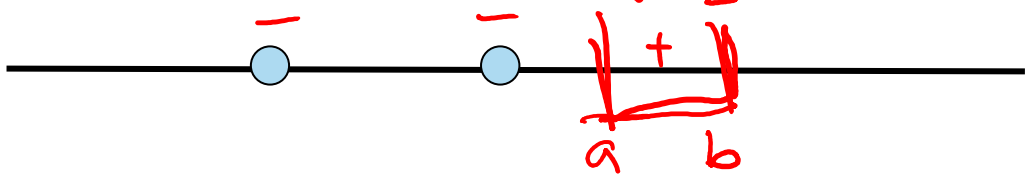
$\mathrm{VCdim(H)} = 1$

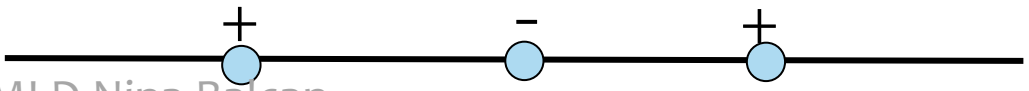


E.g., H= Intervals on the real line



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



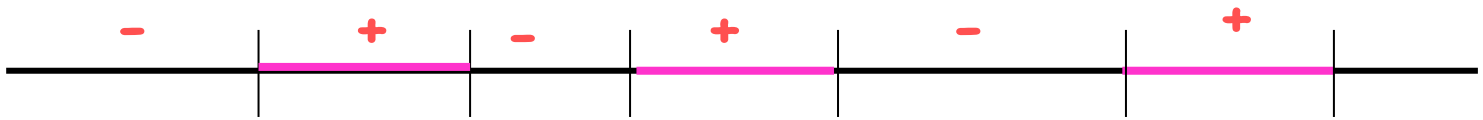


Slide credit: CMU MLD Nina Balcan

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.

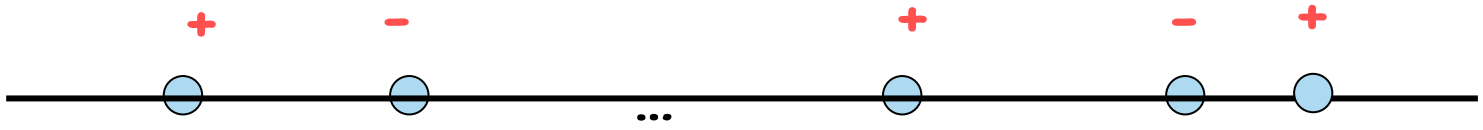
E.g., H= Union of k intervals on the real line $\mathrm{VCdim(H)} = 2\mathrm{k}$



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

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

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



Slide credit: CMU MLD Nina Balcan

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

Slide credit: CMU MLD Matt Gormley

SLT-style Corollaries

Thm. 1 $N \geq \frac{1}{\epsilon}\left[\log(|\mathcal{H}|) + \log(\frac{1}{\delta})\right]$ 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…

Slide credit: CMU MLD Matt Gormley

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 $\mathcal{H}$,

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

Slide credit: CMU MLD Matt Gormley

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 $\hat{R}(h) = 0$),

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

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

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

Slide credit: CMU MLD Matt Gormley

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 $\hat{R}(h) = 0$),

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

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

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

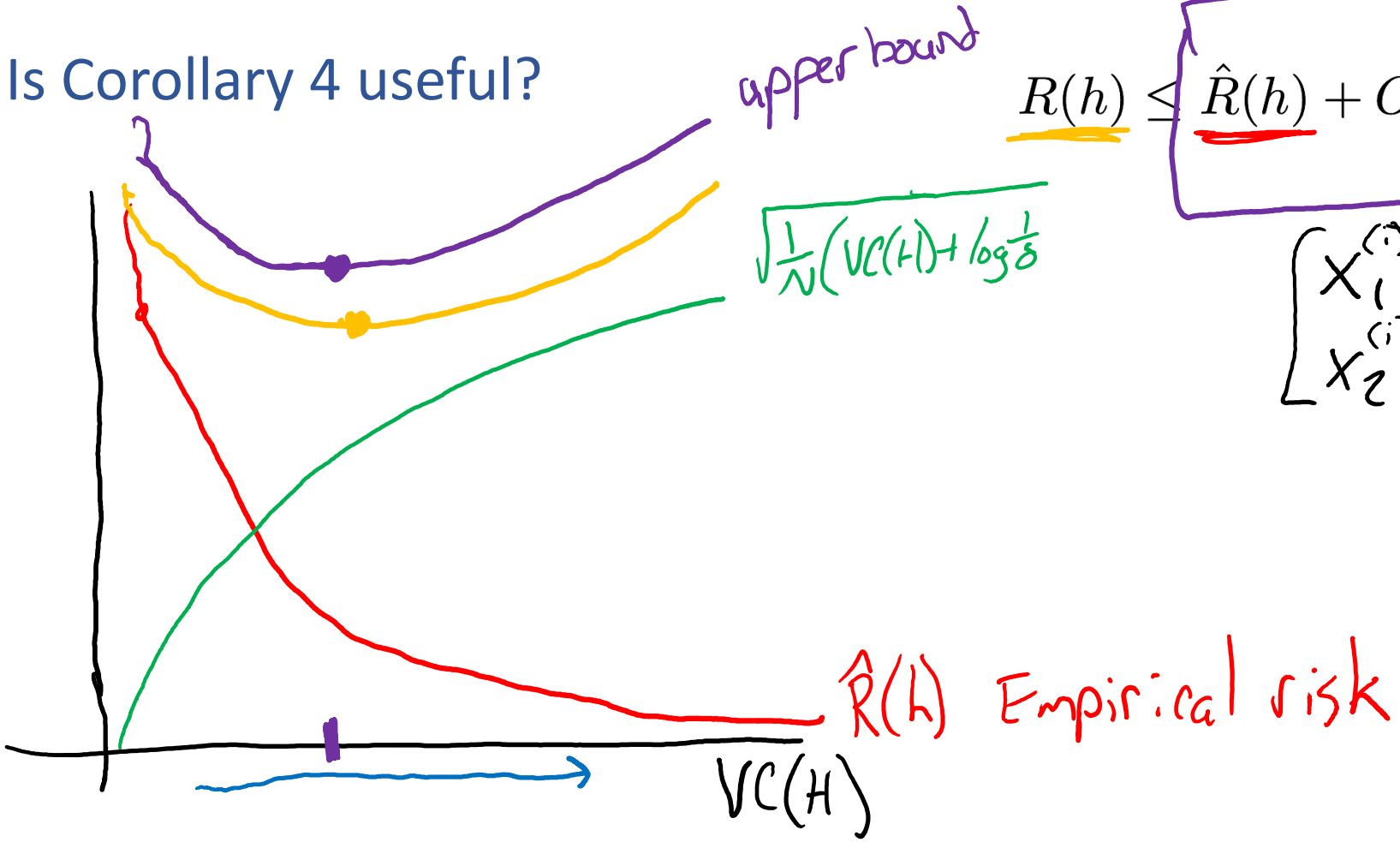
Should these corollaries inform how we do model selection?

Slide credit: CMU MLD Matt Gormley

PAC Bounds and Model Selection

Is Corollary 4 useful?

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

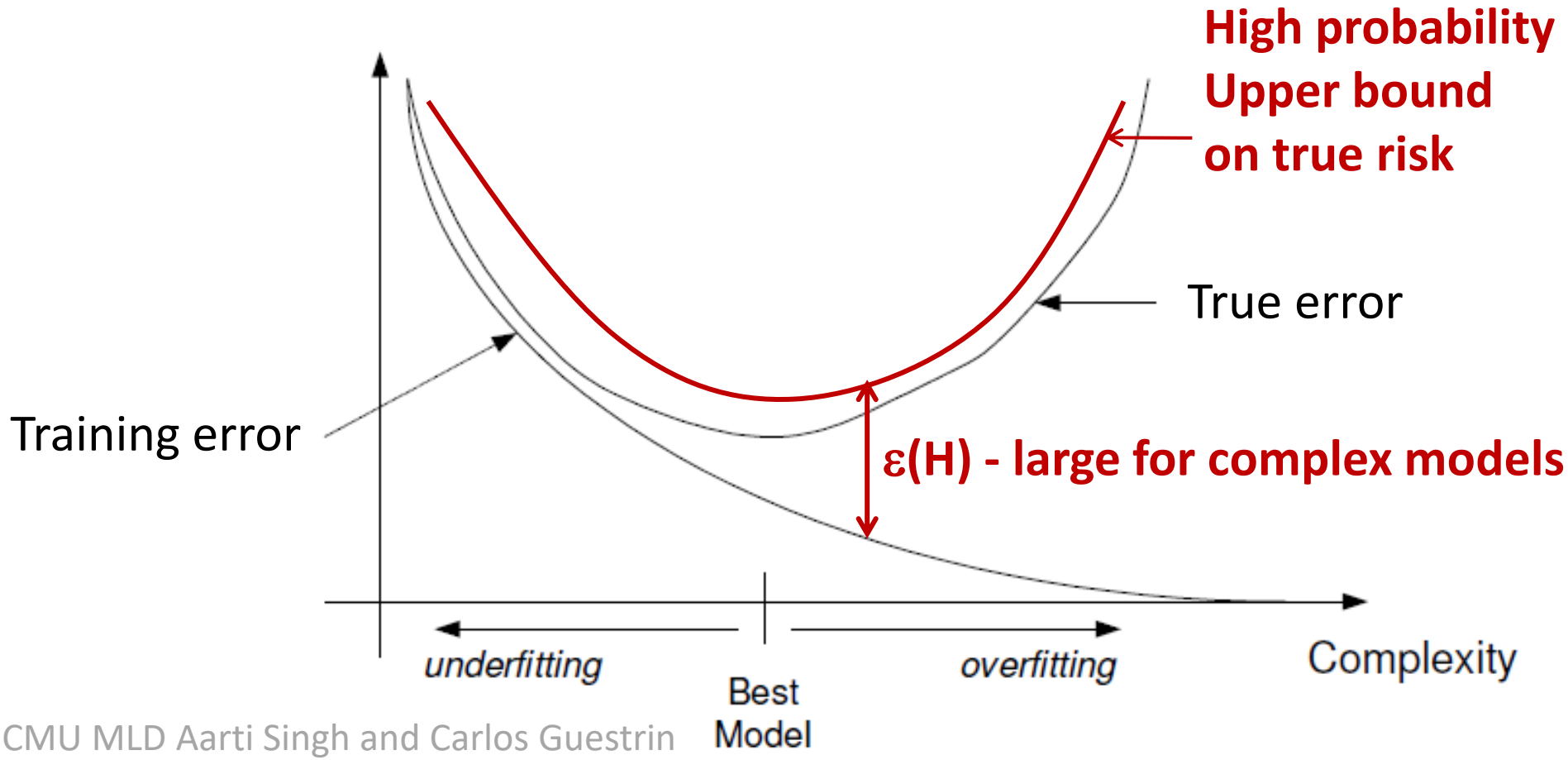


$$\begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \end{bmatrix} \xrightarrow{\phi(\vec{x})} \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}$$

PAC Bounds

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

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



Slide credit: CMU MLD Aarti Singh and Carlos Guestrin

Using PAC Bounds to pick a hypothesis and model selection

- Empirical Risk Minimization (ERM):

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

$\varepsilon_{D_{train}}$

- Structural Risk Minimization (SRM):

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

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

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

$\|\theta\|$

where λ is chosen by cross-validation

Slide credit: CMU MLD Aarti Singh and Carlos Guestrin

PAC Bounds and Regularization

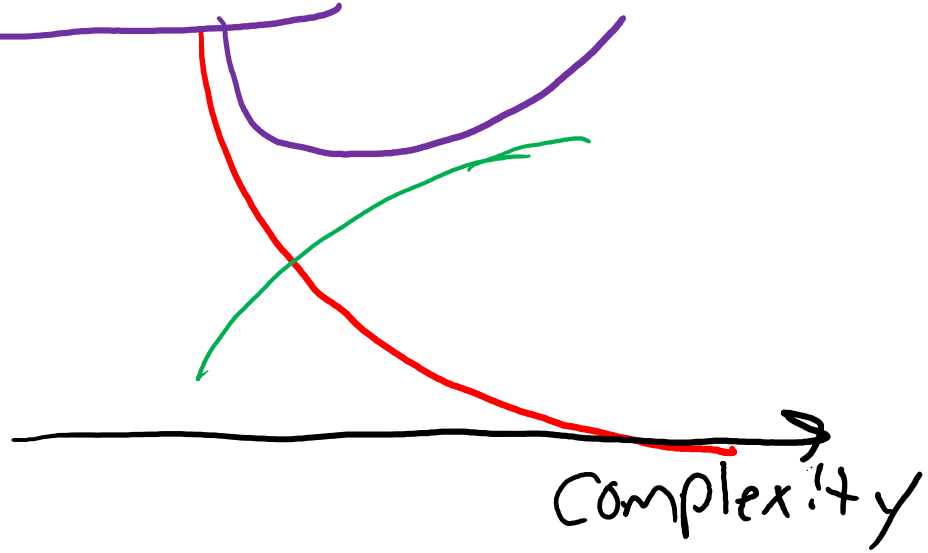
Example: Linear separator in $\mathbb{R}^M$

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

$$VC(H) = M+1$$

$$r(\vec{\theta}) = \|\vec{\theta}\|_1 = \sum_{j=1}^{M} |\theta_j|$$

$$\hat{\theta} = \underset{\vec{\theta}}{\operatorname{argmin}} \; J(\vec{\theta}) + r(\vec{\theta})$$



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)

$$R(h) \leq \hat{R}(h) + r(\theta)$$

Slide credit: CMU MLD Matt Gormley

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

Slide credit: CMU MLD Matt Gormley