Learning Theory II

Aarti Singh and Eric Xing

Machine Learning 10-701/15-781
Nov 7, 2012

Slides courtesy: Carlos Guestrin

MACHINE LEARNING DEPARTMENT
Carnegie Mellon.
School of Computer Science

Summary of PAC bounds for finite hypothesis spaces

With probability $\geq 1-\delta$,

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

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

2) For all $h \in H$

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}$$

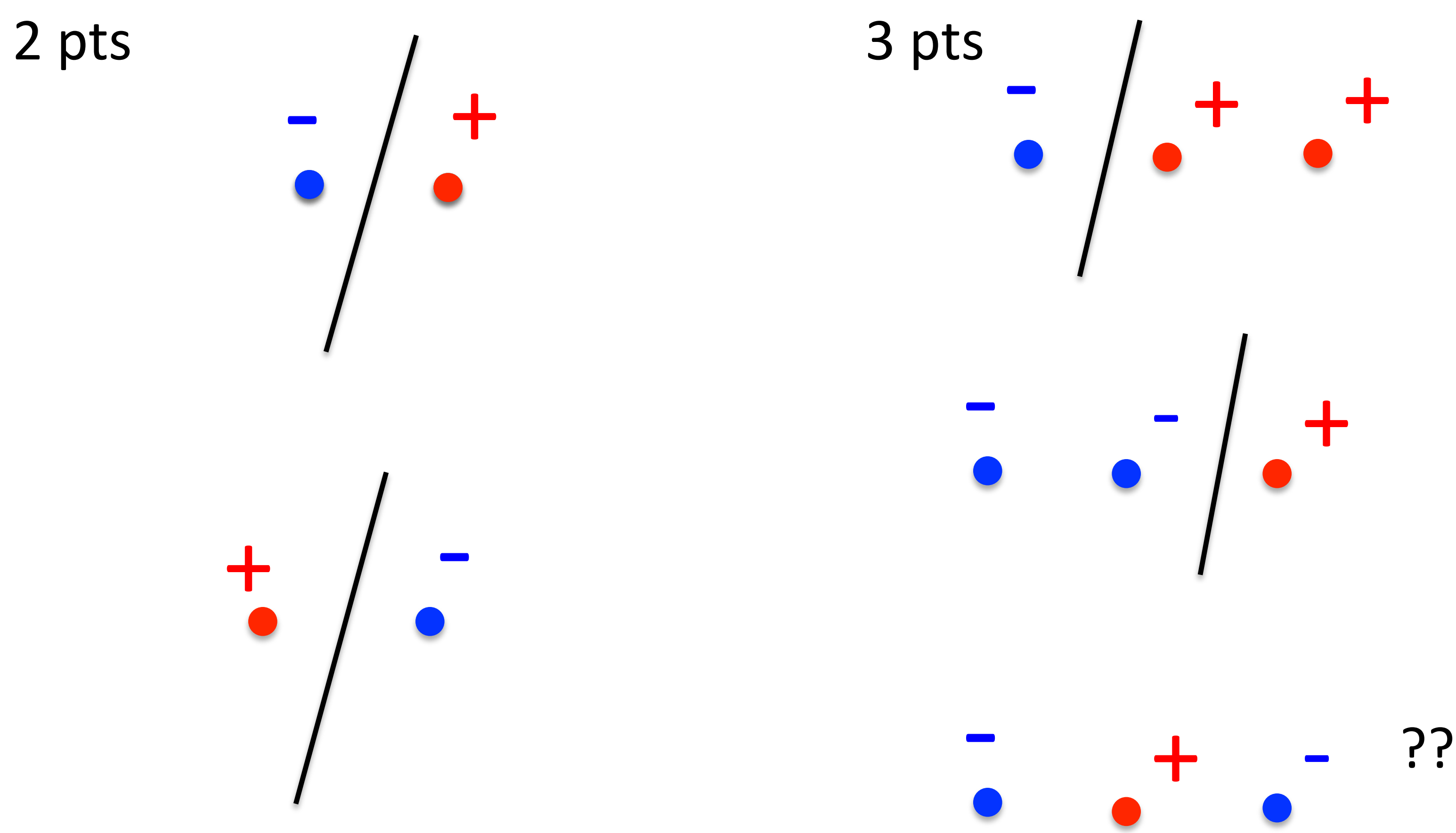
What about continuous hypothesis spaces?

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

- Continuous hypothesis space:
  - $|H| = \infty$
  - Infinite variance???

- **As with decision trees, complexity of hypothesis space only depends on maximum number of points that can be classified exactly (and not necessarily its size)!**

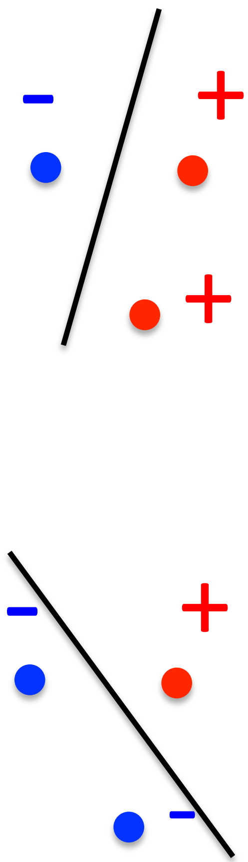
How many points can a linear boundary classify exactly? (1-D)



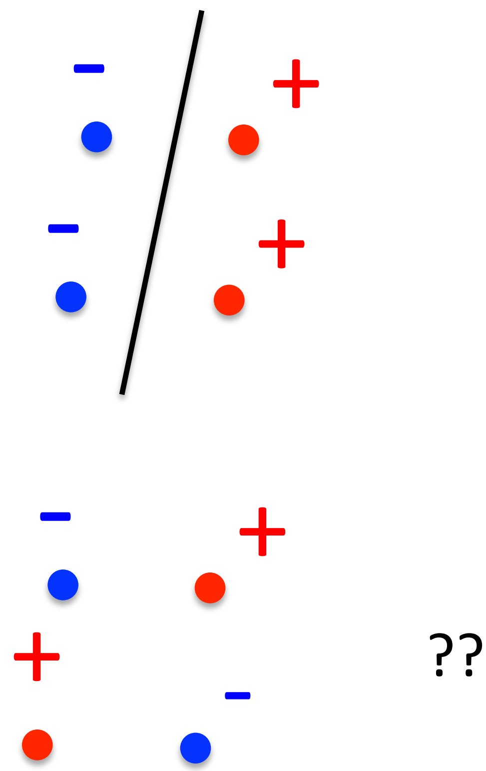
There exists placement s.t. all labelings can be classified

How many points can a linear boundary classify exactly? (2-D)

3 pts

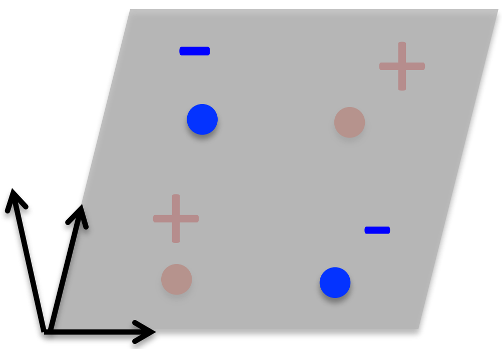


4 pts



There exists placement s.t. all labelings can be classified

How many points can a linear boundary classify exactly? (d-D)



d+1 pts

How many parameters in linear Classifier in d-Dimensions?

$$w_0 + \sum_{i=1}^{d} w_i x_i$$

d+1

There exists placement s.t. all labelings can be classified

PAC bound using VC dimension

- Number of training points that can be classified exactly is VC dimension!!!
  - Measures relevant size of hypothesis space, as with decision trees with k leaves

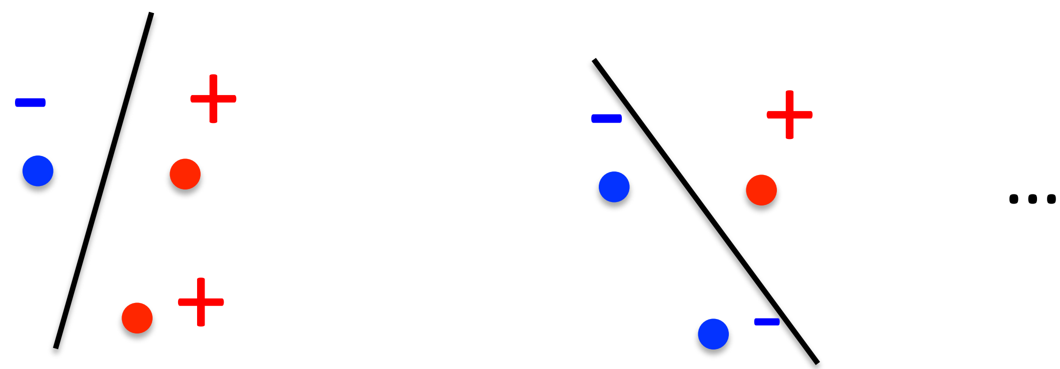
$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + 8\sqrt{\frac{VC(H)\left(\ln\frac{m}{VC(H)}+1\right)+\ln\frac{8}{\delta}}{2m}}$$

Instead of ln|H|

Shattering a set of points

Definition: a **dichotomy** of a set S is a partition of S into two disjoint subsets.

Definition: a set of instances S is **shattered** by hypothesis space H if and only if for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.

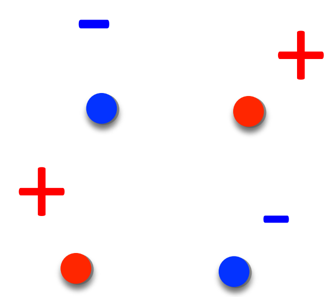


For all binary partitions of S into (S+,S-), there exists a classifier in H that classifies S+ as positive and S- as negative.

VC dimension

Definition: The **Vapnik-Chervonenkis dimension**, $VC(H)$, of hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H. If arbitrarily large finite sets of X can be shattered by H, then $VC(H) \equiv \infty$.

- You pick set of points
- Adversary assigns labels
- You find a hypothesis in H consistent with the labels



If VC(H) = k, then for all k+1 points, there exists a labeling that cannot be shattered (can’t find a hypothesis in H consistent with it)

PAC bound using VC dimension

- Number of training points that can be classified exactly is VC dimension!!!
  - Measures relevant size of hypothesis space, as with decision trees with k leaves
  - Bound for infinite dimension hypothesis spaces:

w.p. ≥ 1-δ

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

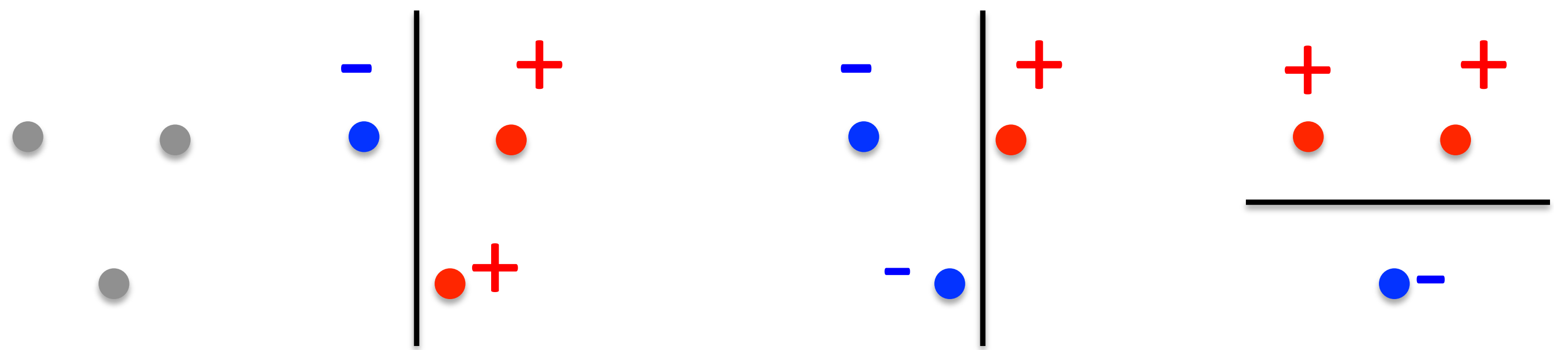
linear classifiers		
2D	large	small
10,000 D	small	large

Examples of VC dimension

- Linear classifiers:
  - VC(H) = d+1, for *d* features plus constant term

Another VC dim. example - What can we shatter?

- What's the VC dim. of decision stumps in 2d?



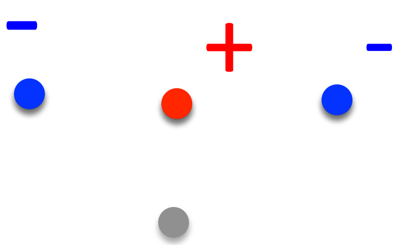
VC(H) ≥ 3

Another VC dim. example - What can't we shatter?

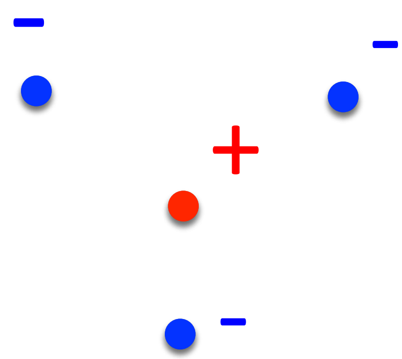
- What's the VC dim. of decision stumps in 2d?

  If VC(H) = 3, then for all placements of 4 pts, there exists a labeling that can't be shattered

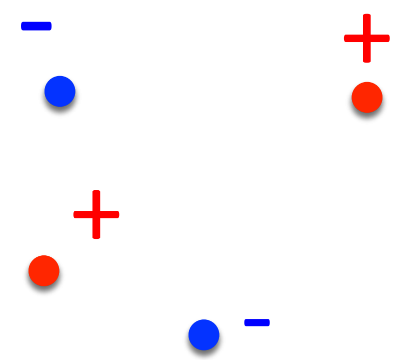
3 collinear



1 in convex hull of other 3



quadrilateral



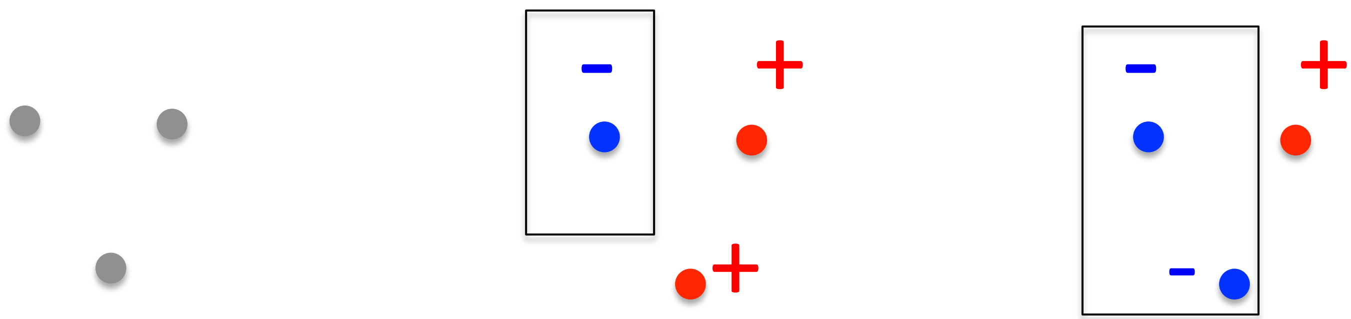
Examples of VC dimension

- Linear classifiers:
  - VC(H) = d+1, for *d* features plus constant term

- Decision stumps: VC(H) = d+1 (3 if d=2)

Another VC dim. example - What can we shatter?

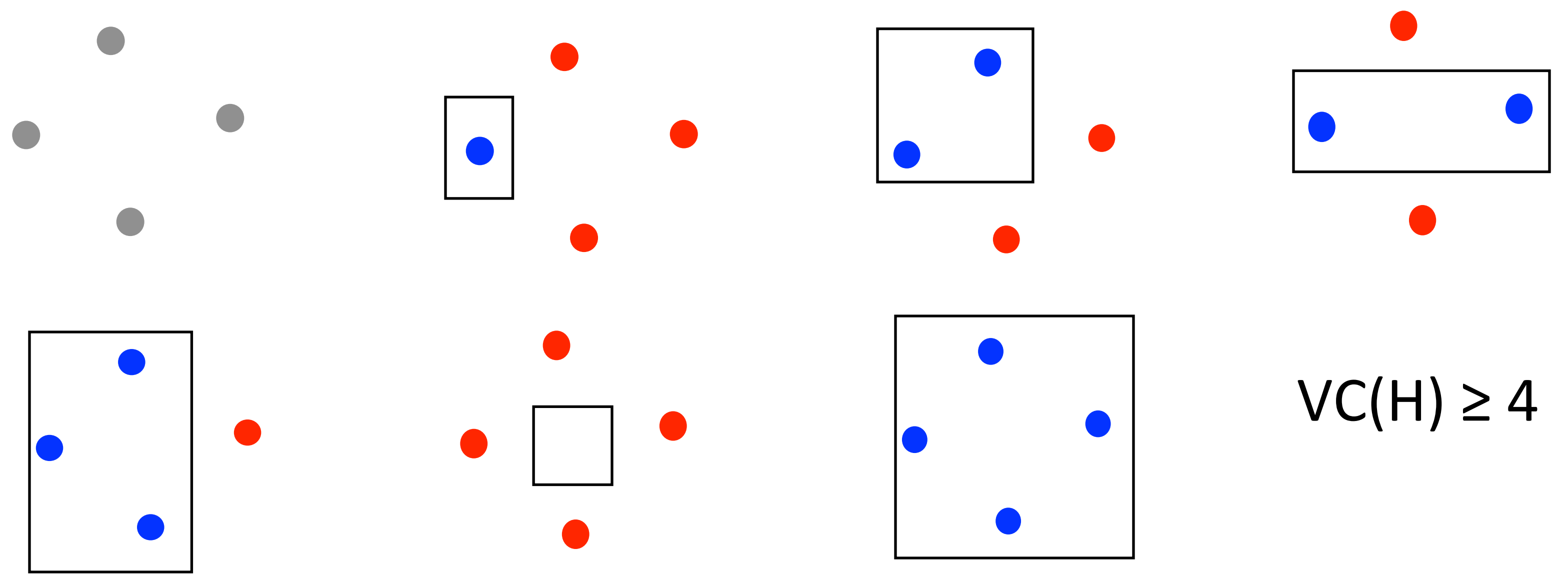
- What's the VC dim. of axis parallel rectangles in 2d?   sign(1- 2*$\mathbf{1}_{x \in \text{rectangle}}$)



VC(H) ≥ 3

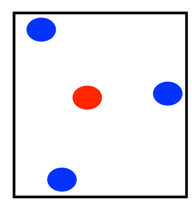
Another VC dim. example - What can't we shatter?

- What's the VC dim. of axis parallel rectangles in 2d? $\text{sign}(1 - 2 * \mathbf{1}_{x \in \text{rectangle}})$



$VC(H) \geq 4$

- Some placement of 4 pts can't be shattered

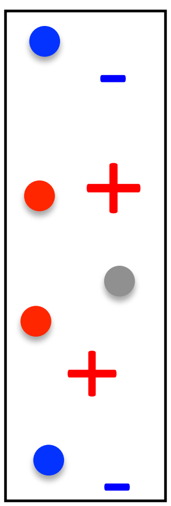


Another VC dim. example - What can't we shatter?

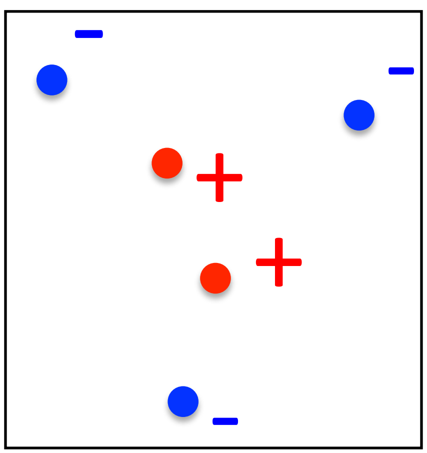
- What's the VC dim. of axis parallel rectangles in 2d? $\text{sign}(1- 2*\mathbf{1}_{x \in \text{rectangle}})$

If VC(H) = 4, then for all placements of 5 pts, there exists a labeling that can't be shattered

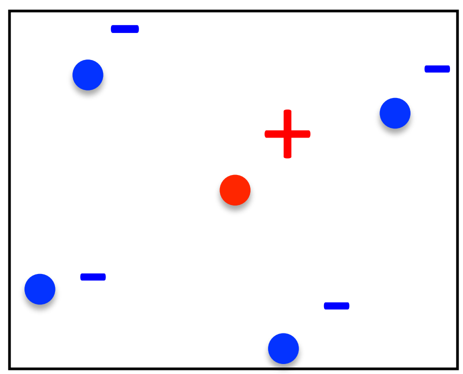
4 collinear



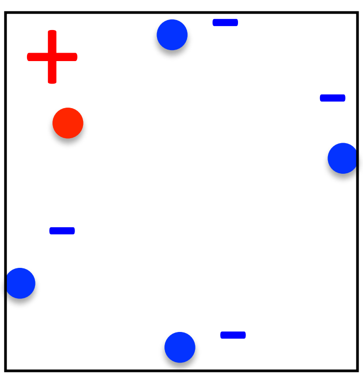
2 in convex hull of other 3



1 in convex hull of other 4



pentagon



Examples of VC dimension

- Linear classifiers:
  - VC(H) = d+1, for *d* features plus constant term

- Decision stumps: VC(H) = d+1

- Axis parallel rectangles: VC(H) = 2d (4 if d=2)

- 1 Nearest Neighbor: VC(H) = ∞

VC dimension and size of hypothesis space

- To be able to shatter m points, how many hypothesis do we need?

2^m labelings ⇨ $|H| \geq 2^m$

Given |H| hypothesis can hope to shatter max $m = \log_2 |H|$ points

$$VC(H) \leq \log_2 |H|$$

So VC bound is tighter.

Summary of PAC bounds

With probability $\geq 1-\delta$,

1) for all $h \in H$ s.t. $error_{train}(h) = 0$,

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

2) for all $h \in H$,

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}$$

Finite hypothesis space

3) for all $h \in H$,

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = 8\sqrt{\frac{VC(H)\left(\ln\frac{m}{VC(H)} + 1\right) + \ln\frac{8}{\delta}}{2m}}$$

Infinite hypothesis space

Using PAC bound to pick a hypothesis

- Empirical Risk Minimization (ERM)

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

$$\mathrm{error}_{\mathrm{true}}(\widehat{h}) \leq \mathrm{error}_{\mathrm{train}}(\widehat{h}) + \epsilon \qquad\qquad w.p. \geq 1 - \delta$$

$$= \min_{h \in H} \ \mathrm{error}_{\mathrm{train}}(h) + \epsilon$$

$$\leq \min_{h \in H} \ \mathrm{error}_{\mathrm{true}}(h) + 2\epsilon$$

- If training error is best possible in H, then true error is also close to best possible in H (with high probability)

Using PAC bound for model selection

- Structural Risk Minimization (SRM)

model spaces $H_1, H_2, \ldots, H_k, \ldots$ of increasing complexity

$|H_1| \leq |H_2| \leq \ldots \leq |H_k| \leq \ldots$ OR

$VC(H_1) \leq VC(H_2) \leq \ldots \leq VC(H_k) \leq \ldots$

For each hypothesis space H_k, we know with probability $\geq 1-\delta_k$, for all $h \in H_k$

$error_{true}(h) \leq error_{train}(h) + \varepsilon(H_k)$ depends on $|H_k|$ or $VC(H_k)$

As complexity k increases, $error_{train}$ goes down but $\varepsilon(H_k)$ goes up – *Bias variance tradeoff*

Using PAC bound for model selection

- Structural Risk Minimization (SRM)

ERM within each model space

$$\widehat{h}_{\mathsf{k}} = \arg\min_{h \in H_{\mathsf{k}}} \ \text{error}_{\text{train}}(h)$$

Choose model space (minimize upper bound on true error)

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

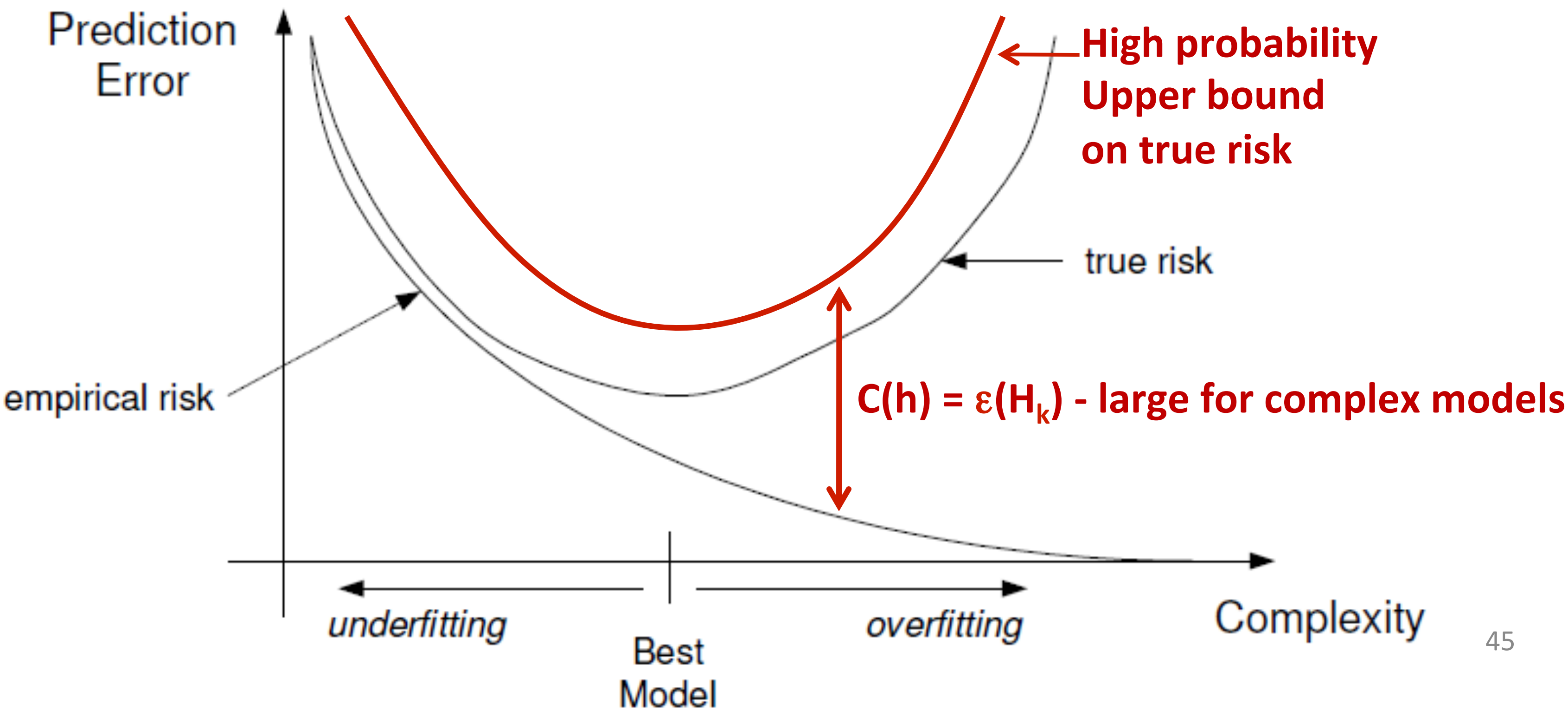
Final hypothesis

$$\widehat{h} = \widehat{h}_{\widehat{k}}$$

Using PAC bound for model selection

- Structural Risk Minimization (SRM)

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



Using PAC bound for model selection

- How good is the final hypothesis picked by SRM relative to best hypothesis in the best class k*?

$$
\begin{aligned}
\text{error}_{\text{true}}(\widehat{h}) &= \text{error}_{\text{true}}(\widehat{h}_{\widehat{k}}) \\
&\leq \text{error}_{\text{train}}(\widehat{h}_{\widehat{k}}) + \epsilon(H_{\widehat{k}}) \\
&= \min_k \left\{ \text{error}_{\text{train}}(\widehat{h}_k) + \epsilon(H_k) \right\} \\
&= \min_k \left\{ \min_{h \in H_k} \text{error}_{\text{train}}(h) + \epsilon(H_k) \right\} \\
&\leq \min_k \left\{ \underbrace{\min_{h \in H_k} \text{error}_{\text{true}}(h)}_{\text{Bias}} + \underbrace{2\epsilon(H_k)}_{\text{Variance}} \right\} \\
&= \min_{h \in H_{k^*}} \text{error}_{\text{true}}(h) + 2\epsilon(H_{k^*})
\end{aligned}
$$

$w.p. \geq 1 - \delta$

$\delta = \sum_k \delta_k$

Using PAC bound for model selection

- What if we picked the hypothesis using ERM over the union of all spaces $U_k H_k$?

$$\widehat{h} = \arg\min_{h \in H_{1,..,k,...}} \text{error}_{\text{train}}(h)$$

What you need to know

- PAC bounds on true error in terms of empirical/training error and complexity of hypothesis space
- Complexity of the classifier depends on number of points that can be classified exactly
  - Finite case – Number of hypothesis
  - Infinite case – VC dimension
- Bias-Variance tradeoff in learning theory
- Empirical and Structural Risk Minimization
- Other bounds – Margin based, Mistake bounds, …
- But often bounds too loose in practice