

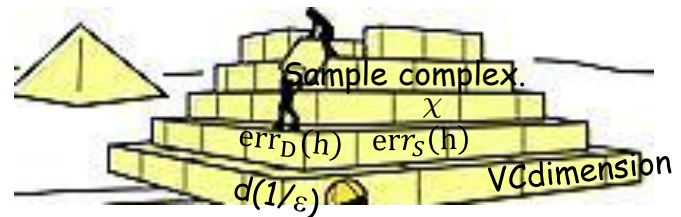
Machine Learning Theory II

Maria-Florina (Nina) Balcan

February 11th, 2015

Today's focus

1. SLT for infinite H
2. Model selection



Machine Learning Theory

1. Binary classification
2. Noise-free labeling

$$err_D(h) \leq err_S(h) + \varepsilon$$

?

Finite H

Infinite H

Realizable
 $c^* \in H$

Agnostic
 $c^* \notin H$

Realizable
 $c^* \in H$

Agnostic
 $c^* \notin H$

$$err_D(h) \leq err_S(h) + \sqrt{\frac{1}{2m} \ln \frac{2|H|}{\delta}}$$

$$err_D(h) \leq err_S(h) + \mathcal{O}\left(\sqrt{\frac{1}{m} \left(d \ln \frac{m}{d} + \ln \frac{1}{\delta}\right)}\right)$$

$$err_D(h) \leq \frac{1}{m} \left(\ln |H| + \ln \frac{1}{\delta} \right)$$

$$err_D(h) \leq \mathcal{O}\left(\frac{1}{m} \left(d \ln \frac{m}{d} + \ln \frac{1}{\delta}\right)\right)$$

Some concepts:

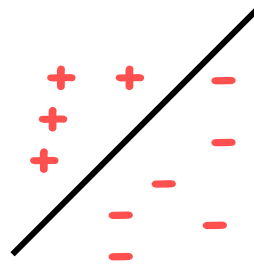
1. Consistent learner
2. Version space (VS)
3. ε -exhausted VS
4. PAC-learnable
5. Hoeffding inequality
6. Dichotomy
7. Shattering
8. Shattering coefficient (growth function)
9. VC-dimension
10. Sauer's lemma



What if H is infinite?

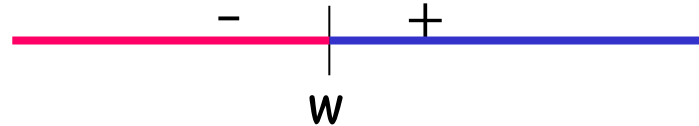


E.g., linear separators in \mathbb{R}^d



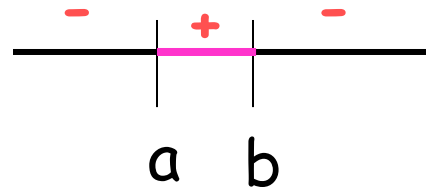
$$h(x) = \text{sign}(w^T x)$$

E.g., thresholds on the real line



$$h(x) = \text{sign}(x - w)$$

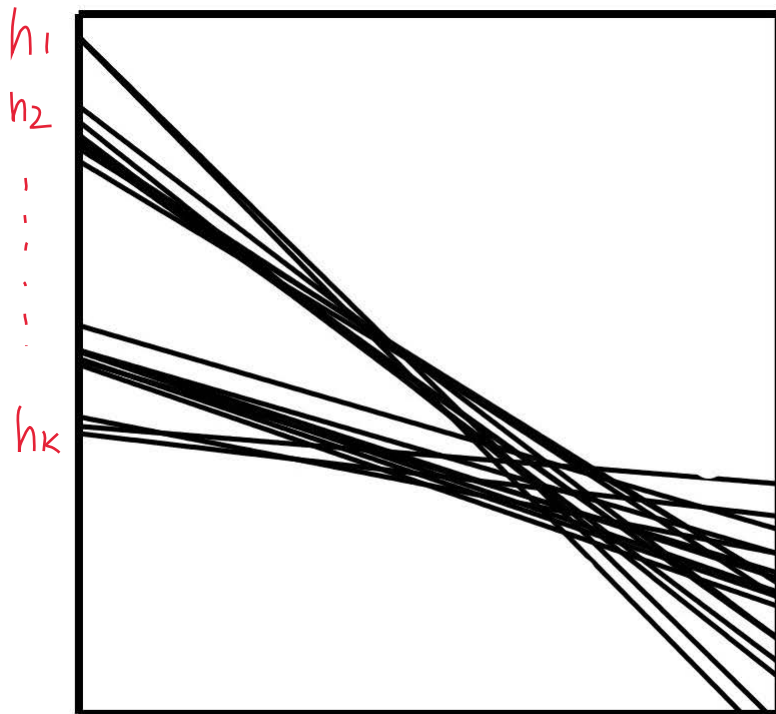
E.g., intervals on the real line



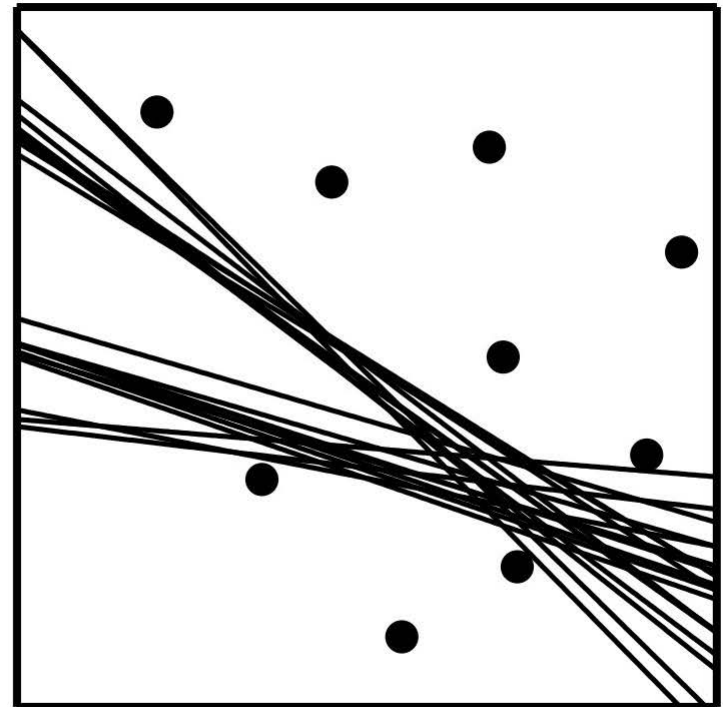
$$h(x) = \begin{cases} 1, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

An Effective Number of Hypotheses

$|H|$ only measures the maximum possible diversity of H



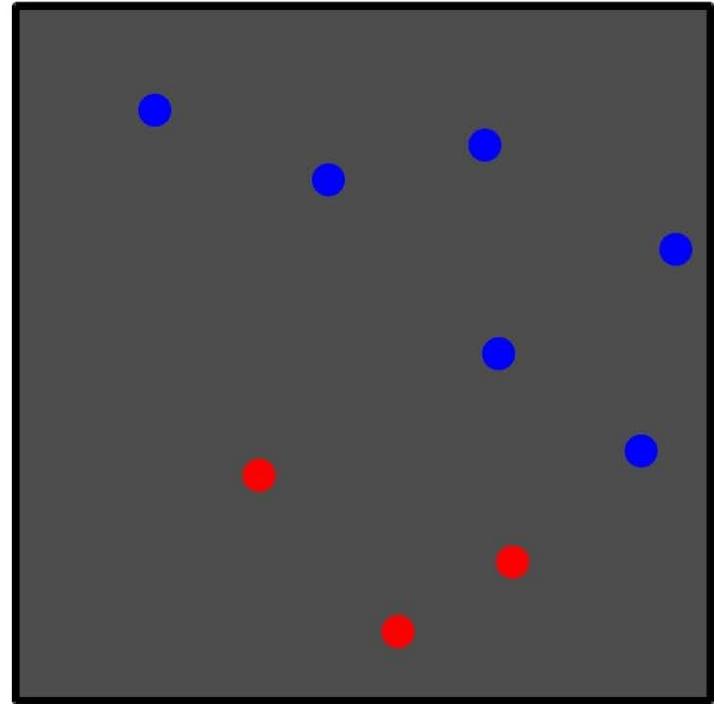
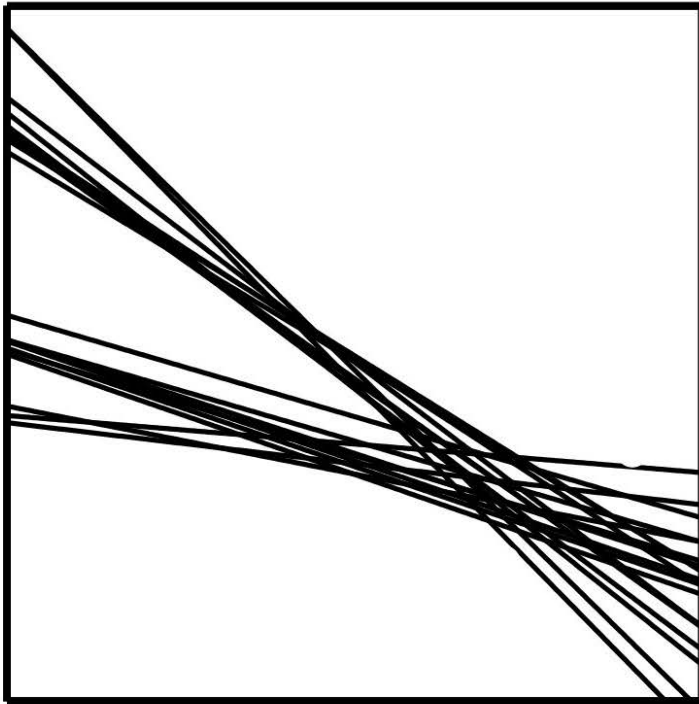
\mathcal{H}



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

An Effective Number of Hypotheses

$|H|$ only measures the maximum possible diversity of H



From the viewpoint of S , the entire H is just one **dichotomy**

An Effective Number of Hypotheses

$|H|$ only measures the maximum possible diversity of H

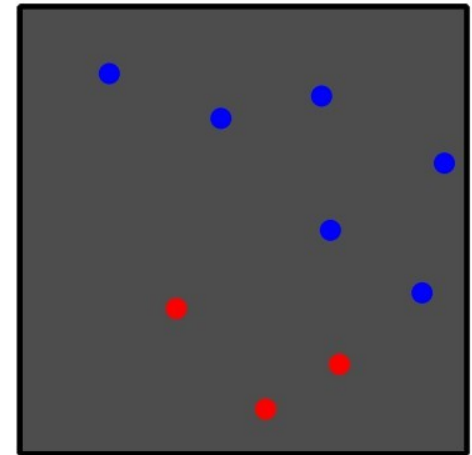
Given a dataset $S = \{x_1, \dots, x_m\}$,

$(h(x_1), \dots, h(x_m))$

A dichotomy of S

$h: X \rightarrow \{-1, +1\}$

1. If H is diverse, we get many different dichotomies.
2. If H contains many similar functions, we only get a few dichotomies.



dichotomy

The *shattering coefficient* quantifies this.

Sample Complexity: Infinite Hypothesis Spaces

- $H[m]$ - maximum number of ways to split m points using concepts in H ; i.e. $H[m] = \max_{|S|=m} |H[S]|$
 $c^* \in H$ $m \geq \frac{1}{\epsilon} (\ln |H| + \ln \frac{1}{\delta})$

Theorem For any class H , distrib. D , if the number of labeled examples seen m satisfies

$$m \geq \frac{2}{\epsilon} \left[\log_2(2H[2m]) + \log_2\left(\frac{1}{\delta}\right) \right]$$

then with probab. $1 - \delta$, all $h \in H$ with $err_D(h) \geq \epsilon$ have $err_S(h) > 0$.

Sauer's Lemma: $H[m] = O(m^{VCdim(H)})$

Theorem

$$m = O\left(\frac{1}{\epsilon} \left[\underline{VCdim(H)} \log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right) \right]\right)$$

labeled examples are sufficient so that with probab. $1 - \delta$, all $h \in H$ with $err_D(h) \geq \epsilon$ have $err_S(h) > 0$.

Effective number of hypotheses

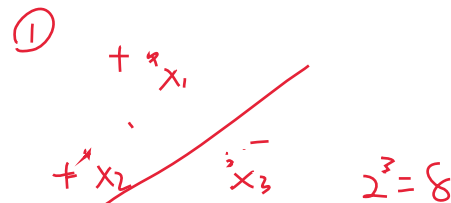
- $H[S]$ - the set of splittings of dataset S using concepts from H .
- $H[m]$ - max number of ways to split m points using concepts in H

$$H[m] = \max_{|S|=m} |H[S]|$$

$$H(S) = \{ (\underbrace{h(x_1), \dots, h(x_m)}_{\text{dichotomy}}) \mid h \in H \}, |H(S)| \leq 2^m$$

$$H[m] = \max_S \{ |H(S)| \mid |S|=m, \forall S \subseteq X \} \leq 2^m$$

H : linear separator



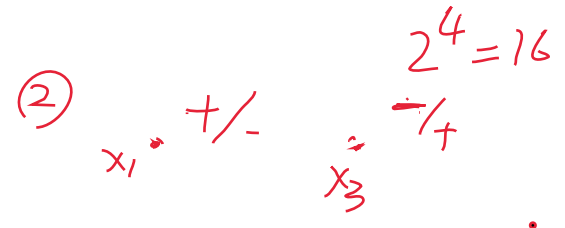
$$H(S) = \{+-+, ++-, \dots, ---\}$$

$$|H(S)| = 2^3 = 8$$

$$H[3] = 8$$



$$H[3] = 8$$



$$H(S) = \{++++, +++-, \dots, ----\}$$

$$16 - 2 = 14$$

$$H[4] = 14$$

Effective number of hypotheses

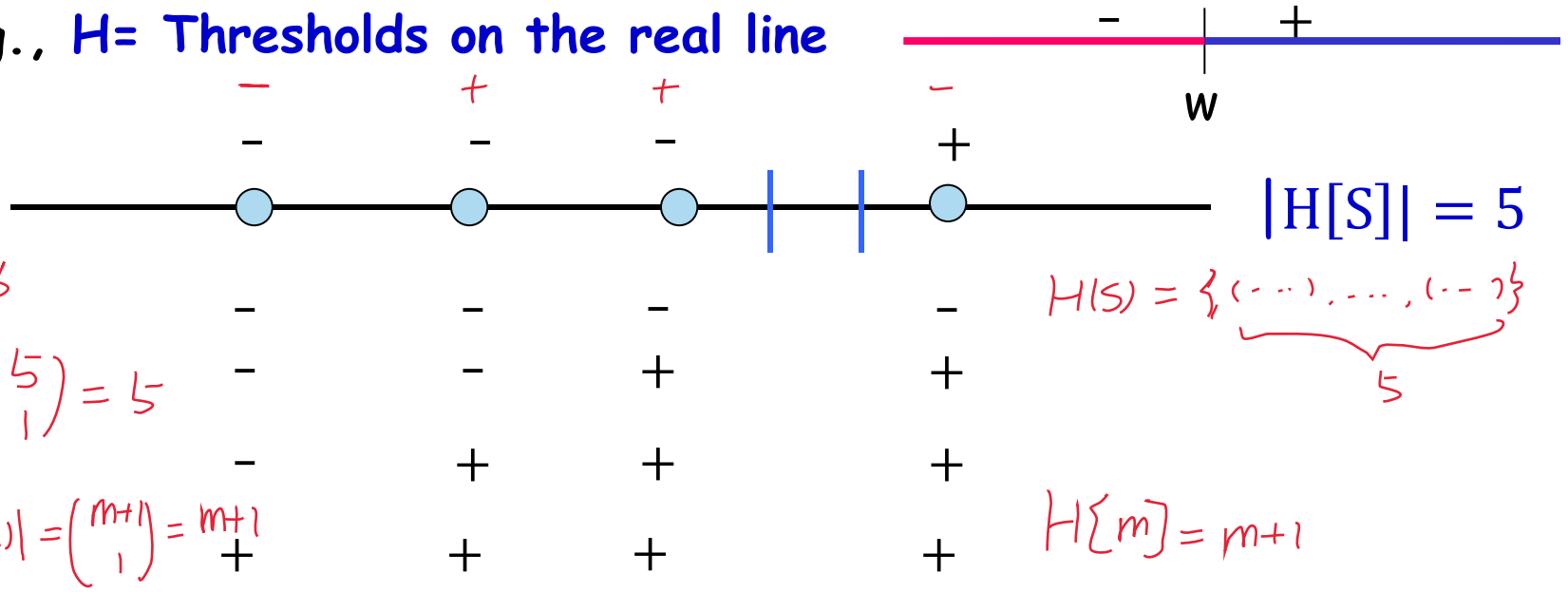
- $H[S]$ - the set of splittings of dataset S using concepts from H .
- $H[m]$ - max number of ways to split m points using concepts in H

$$H[m] = \max_{|S|=m} |H[S]|$$

$$H[m] \leq 2^m$$

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

$$h(x) = \text{sign}(x - w)$$



In general, if $|S|=m$ (all distinct), $|H[S]| = m + 1 \ll 2^m$

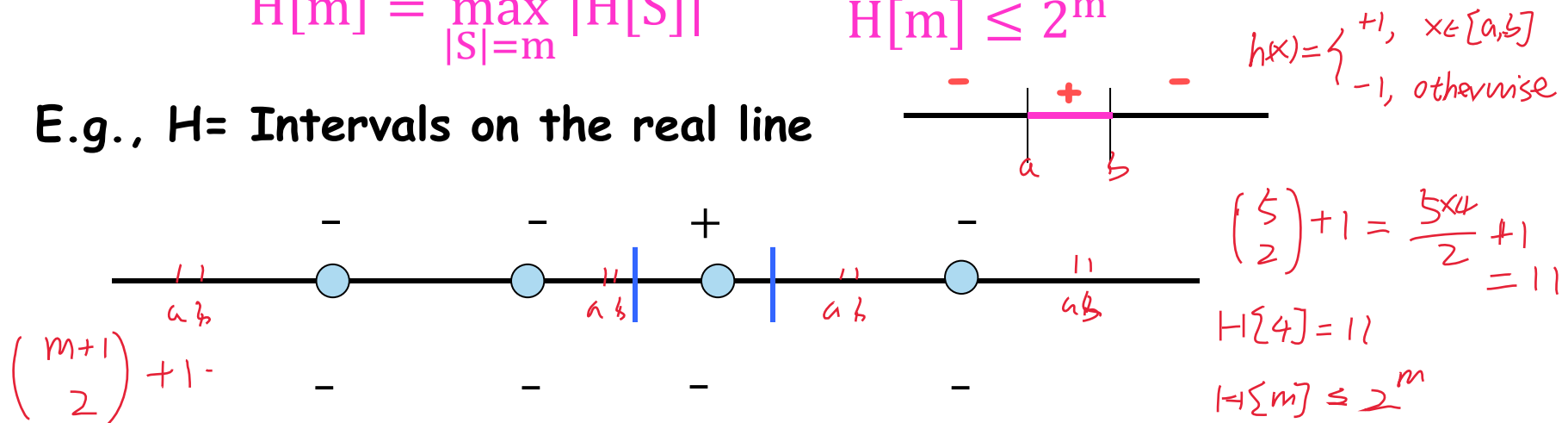
Effective number of hypotheses

- $H[S]$ - the set of splittings of dataset S using concepts from H .
- $H[m]$ - max number of ways to split m points using concepts in H

$$H[m] = \max_{|S|=m} |H[S]|$$

$$H[m] \leq 2^m$$

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



In general, $|S|=m$ (all distinct), $H[m] = \frac{m(m+1)}{2} + 1 = O(m^2) \ll 2^m$

There are $m+1$ possible options for the first part, m left for the second part, the order does not matter, so $\binom{m}{2} + 1$ (for empty interval).

Effective number of hypotheses

- $H[S]$ - the set of splittings of dataset S using concepts from H .
- $H[m]$ - max number of ways to split m points using concepts in H

$$H[m] = \max_{\substack{|S|=m \\ \forall S \subseteq X}} |H[S]| \quad H[m] \leq 2^m$$

Definition: H shatters S if $|H[S]| = 2^{|S|} = 2^m$
specific

Sample Complexity: Infinite Hypothesis Spaces

Realizable Case

$$c^* \in H, \quad m \geq \frac{1}{\epsilon} \left(|H| + \ln \frac{1}{\delta} \right)$$

probab. $1 - \delta$, $err_D(h) \leq \epsilon$ approximately PAC-learn

$H[m]$ - max number of ways to split m points using concepts in H

Theorem For any class H , distrib. D , if the number of labeled examples seen m satisfies

$$m \geq \frac{2}{\epsilon} \left[\log_2(2H[2m]) + \log_2\left(\frac{1}{\delta}\right) \right]$$

$$|H[m]| \leq 2^m$$

then with probab. $1 - \delta$, all $h \in H$ with $err_D(h) \geq \epsilon$ have $err_S(h) > 0$.
 $\underbrace{err_D(h) \geq \epsilon}_A \Rightarrow \underbrace{err_S(h) > 0}_B$
 $\neg B \Rightarrow \neg A$ $err_S(h) = 0 \Rightarrow err_D(h) < \epsilon$

- Not too easy to interpret sometimes hard to calculate exactly, but can get a good bound using "VC-dimension"

If $H[m] = 2^m$, then $m \geq \frac{m}{\epsilon} (\dots)$ ☹

- VC-dimension is roughly the point at which H stops looking like it contains all functions, so hope for solving for m .

Sample Complexity: Infinite Hypothesis Spaces

$H[m]$ - max number of ways to split m points using concepts in H

Theorem For any class H , distrib. D , if the number of labeled examples seen m satisfies

$$m \geq \frac{2}{\varepsilon} \left[\log_2(2H[2m]) + \log_2\left(\frac{1}{\delta}\right) \right]$$

then with probab. $1 - \delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Sauer's Lemma: $H[m] = O(m^{\overset{d}{VCdim(H)}})$ $\mapsto \{m\} = o(m^d)$

Theorem

$$m = O\left(\frac{1}{\varepsilon} \left[\overset{d}{VCdim(H)} \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right) \right]\right)$$

labeled examples are sufficient so that with probab. $1 - \delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Shattering, VC-dimension

Definition: H shatters S if $|H[S]| = 2^{|S|} = 2^m$

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, 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 . $|S| = m$

If arbitrarily large finite sets can be shattered by H , then VCdim(H) = ∞

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.

$$\text{VCdim}(H) \geq d$$

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

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

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

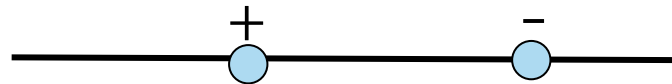
$$2^d \leq |H|$$

Shattering, VC-dimension

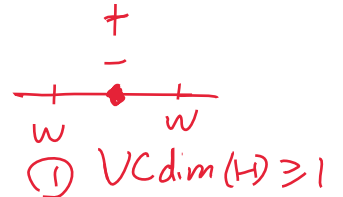
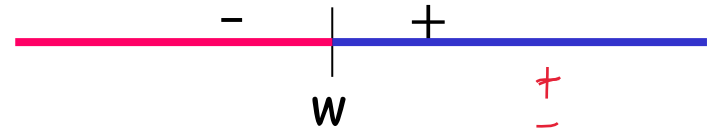
If the VC-dimension is d , that means **there exists** a set of d points that can be shattered, but there is **no** set of $d+1$ points that can be shattered.

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

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



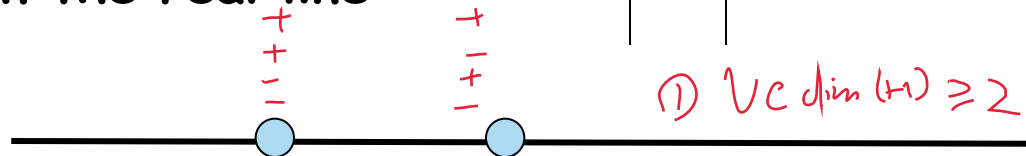
$$\textcircled{2} VCdim(H) < 2$$



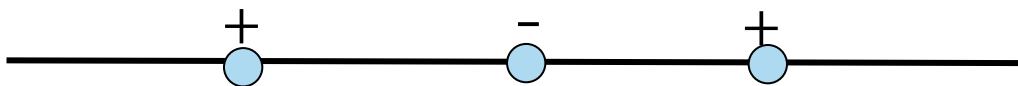
$$\textcircled{1} VCdim(H) \geq 1$$

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

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



$$\textcircled{1} VCdim(H) \geq 2$$



$$\textcircled{2} VCdim(H) < 3$$

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.

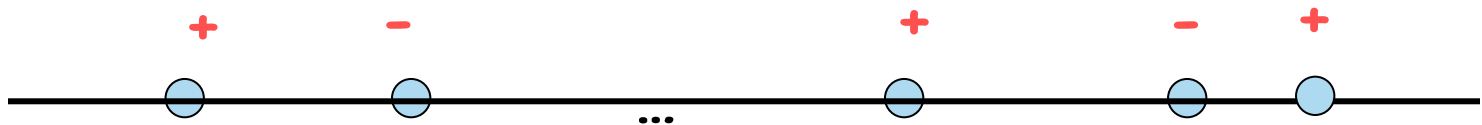
E.g., $H = \text{Union of } k \text{ intervals on the real line}$ $\text{VCdim}(H) = 2k$



① $\text{VCdim}(H) \geq 2k$

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

② $\text{VCdim}(H) < 2k + 1$



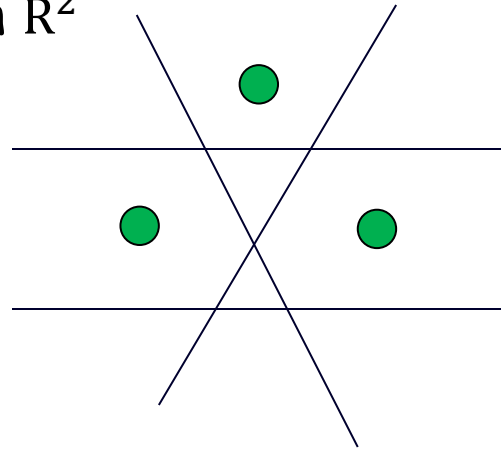
Shattering, VC-dimension

E.g., H = linear separators in \mathbb{R}^2

① $VCdim(H) \geq 3$

② $VCdim(H) < 4$

$\Rightarrow VCdim(H) = 3$



$2^3 = 8$

$\left. \begin{array}{ccc} + & + & + \\ + & + & - \\ & \vdots & \\ - & - & - \end{array} \right\} 8 \text{ dichotomy}$

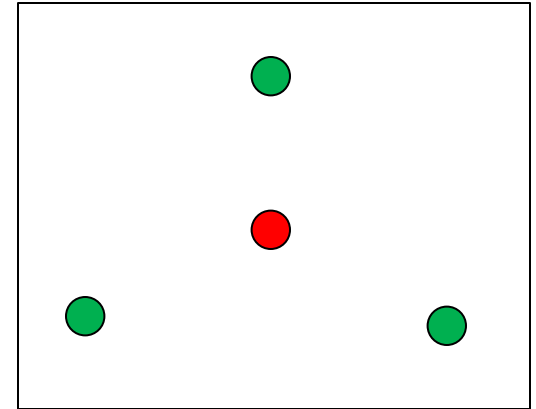
$\begin{array}{cc} + & - \\ \cdot & \cdot \\ - & + \\ \cdot & \cdot \end{array}$

Shattering, VC-dimension

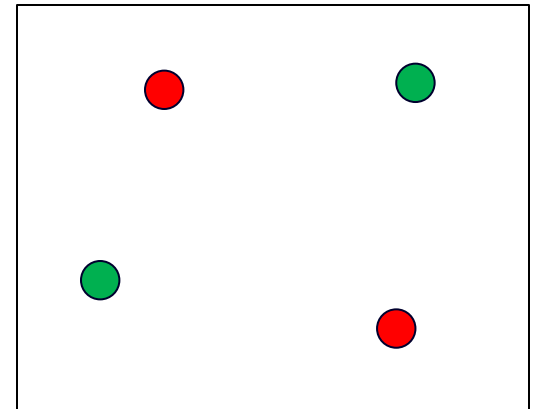
E.g., H = linear separators in \mathbb{R}^2

$$\text{VCdim}(H) < 4$$

Case 1: ~~one point inside the triangle~~ formed by the others. Cannot label inside point as positive and outside points as negative.

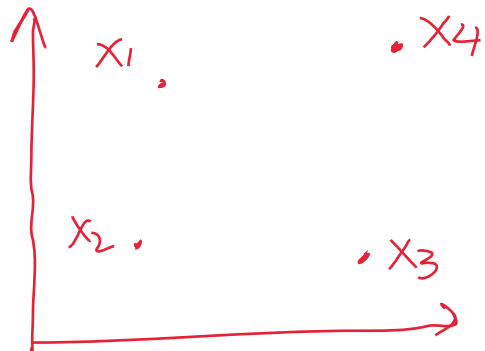


Case 2: all points on the boundary (convex hull). Cannot label two diagonally as positive and other two as negative.



Fact: VCdim of linear separators in \mathbb{R}^d is $d+1$

Today's Quiz



$$S = \{x_1, x_2, x_3, x_4\}$$

$$X^T = (x_1, x_2)$$

$$h(x) = \text{sign}(\beta^T \cdot X + \beta_0) \quad \# \text{params} = 3$$

$$x_1^2, x_2^2, x_1 x_2$$

H : quadratic separator

(#params = 6)

① $H(S) \stackrel{?}{=}$

② $H[m] \stackrel{?}{=}$

③ $\text{VC dim}(H) > 4$?

Sauer's Lemma

Sauer's Lemma:

Let $d = \text{VCdim}(H)$

- $m \leq d$, then $H[m] = 2^m$
- $m > d$, then $H[m] = O(m^d)$

Proof: induction on m and d . Cool combinatorial argument!

Hint: try proving it for intervals...

Sample Complexity: Infinite Hypothesis Spaces

Realizable Case

Theorem For any class H , distrib. D , if the number of labeled examples seen m satisfies

$$m \geq \frac{2}{\varepsilon} \left[\log_2(2H[2m]) + \log_2\left(\frac{1}{\delta}\right) \right]$$

then with probab. $1 - \delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Sauer's Lemma: $H[m] = O(m^{VCdim(H)})$

Theorem

$$m = O\left(\frac{1}{\varepsilon} \left[VCdim(H) \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right) \right]\right)$$

labeled examples are sufficient so that with probab. $1 - \delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Sample Complexity for Supervised Learning

Realizable Case

Consistent Learner

- Input: $S: (x_1, c^*(x_1)), \dots, (x_m, c^*(x_m))$
- Output: Find h in H consistent with S (if one exists).

Theorem

$$m \geq \frac{1}{\epsilon} \left[\ln(|H|) + \ln\left(\frac{1}{\delta}\right) \right]$$

Prob. over different
samples of m
training examples

labeled examples are sufficient so that with prob. $1 - \delta$, all $h \in H$ with $err_D(h) \geq \epsilon$ have $err_S(h) > 0$.

Linear in $1/\epsilon$

Theorem

$$m = O\left(\frac{1}{\epsilon} \left[VCdim(H) \log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right) \right]\right)$$

labeled examples are sufficient so that with probab. $1 - \delta$, all $h \in H$ with $err_D(h) \geq \epsilon$ have $err_S(h) > 0$.

Sample Complexity: Infinite Hypothesis Spaces

Realizable Case

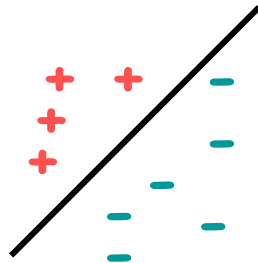
Theorem

$$m = O\left(\frac{1}{\varepsilon} \left[VCdim(H) \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right) \right]\right)$$

labeled examples are sufficient so that with probab. $1 - \delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

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

$$VCdim(H) = d+1$$



$$m = O\left(\frac{1}{\varepsilon} \left[d \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right) \right]\right)$$

Sample complexity linear in d

So, if double the number of features, then I only need roughly twice the number of samples to do well.

Practical rule of thumb: $VCdim(H) \sim \# \text{free parameters of } H$

What if $c^* \notin H$?



Sample Complexity: Uniform Convergence

Agnostic Case

Empirical Risk Minimization (ERM)

- Input: $S: (x_1, c^*(x_1)), \dots, (x_m, c^*(x_m))$
- Output: Find h in H with smallest $err_S(h)$

Theorem

$$m \geq \frac{1}{2\epsilon^2} \left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right) \right]$$

labeled examples are sufficient s.t. with probab. $\geq 1 - \delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \epsilon$.

$1/\epsilon^2$ dependence [as opposed to $1/\epsilon$ for realizable]

Theorem

$$m = O\left(\frac{1}{\epsilon^2} \left[VCdim(H) + \log\left(\frac{1}{\delta}\right) \right]\right)$$

labeled examples are sufficient so that with probab. $1 - \delta$, all $h \in H$ with $|err_D(h) - err_S(h)| \leq \epsilon$.

Sample Complexity: Finite Hypothesis Spaces

Agnostic Case

1) How many examples suffice to get UC whp (so success for ERM).

Theorem

$$m \geq \frac{1}{2\varepsilon^2} \left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right) \right]$$

$1/\varepsilon^2$ dependence [as opposed to $1/\varepsilon$ for realizable], but get for something stronger.

labeled examples are sufficient s.t. with probab. $\geq 1 - \delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$.

2) Statistical Learning Theory style:

With prob. at least $1 - \delta$, for all $h \in H$:

$\sqrt{\frac{1}{m}}$ as opposed to $\frac{1}{m}$ for realizable

$$err_D(h) \leq err_S(h) + \sqrt{\frac{1}{2m} \left(\ln(2|H|) + \ln\left(\frac{1}{\delta}\right) \right)}.$$

Sample Complexity: Infinite Hypothesis Spaces

Agnostic Case

1) How many examples suffice to get UC whp (so success for ERM).

Theorem

$$m = O\left(\frac{1}{\epsilon^2} \left[VCdim(H) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

labeled examples are sufficient so that with probab. $1 - \delta$, all $h \in H$ with $|err_D(h) - err_S(h)| \leq \epsilon$.

2) Statistical Learning Theory style:

With prob. at least $1 - \delta$, for all $h \in H$:

$$err_D(h) \leq err_S(h) + O\left(\sqrt{\frac{1}{2m} \left(VCdim(H) \ln\left(\frac{em}{VCdim(H)}\right) + \ln\left(\frac{1}{\delta}\right)\right)}\right).$$

VCdimension Generalization Bounds

E.g.,
$$\text{err}_D(h) \leq \text{err}_S(h) + O\left(\sqrt{\frac{1}{2m} \left(\text{VCdim}(H) \ln\left(\frac{em}{\text{VCdim}(H)}\right) + \ln\left(\frac{1}{\delta}\right) \right)}\right).$$

VC bounds: distribution independent bounds



- *Generic*: hold for **any concept class** and **any distribution**.

[nearly tight in the WC over choice of D]



- Might be very loose specific distr. that are more benign than the worst case....
- Hold only for binary classification; we want bounds for fns approximation in general (e.g., multiclass classification and regression).

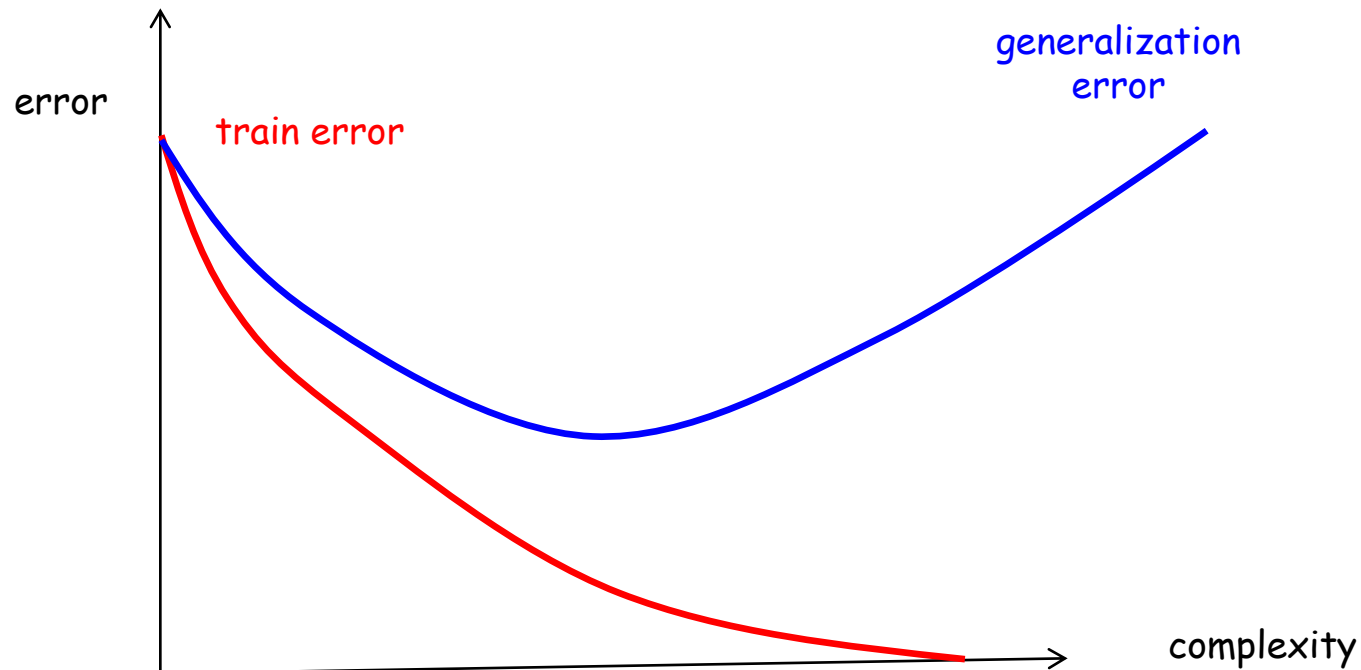
Can we use our bounds for
model selection?



True Error, Training Error, Overfitting

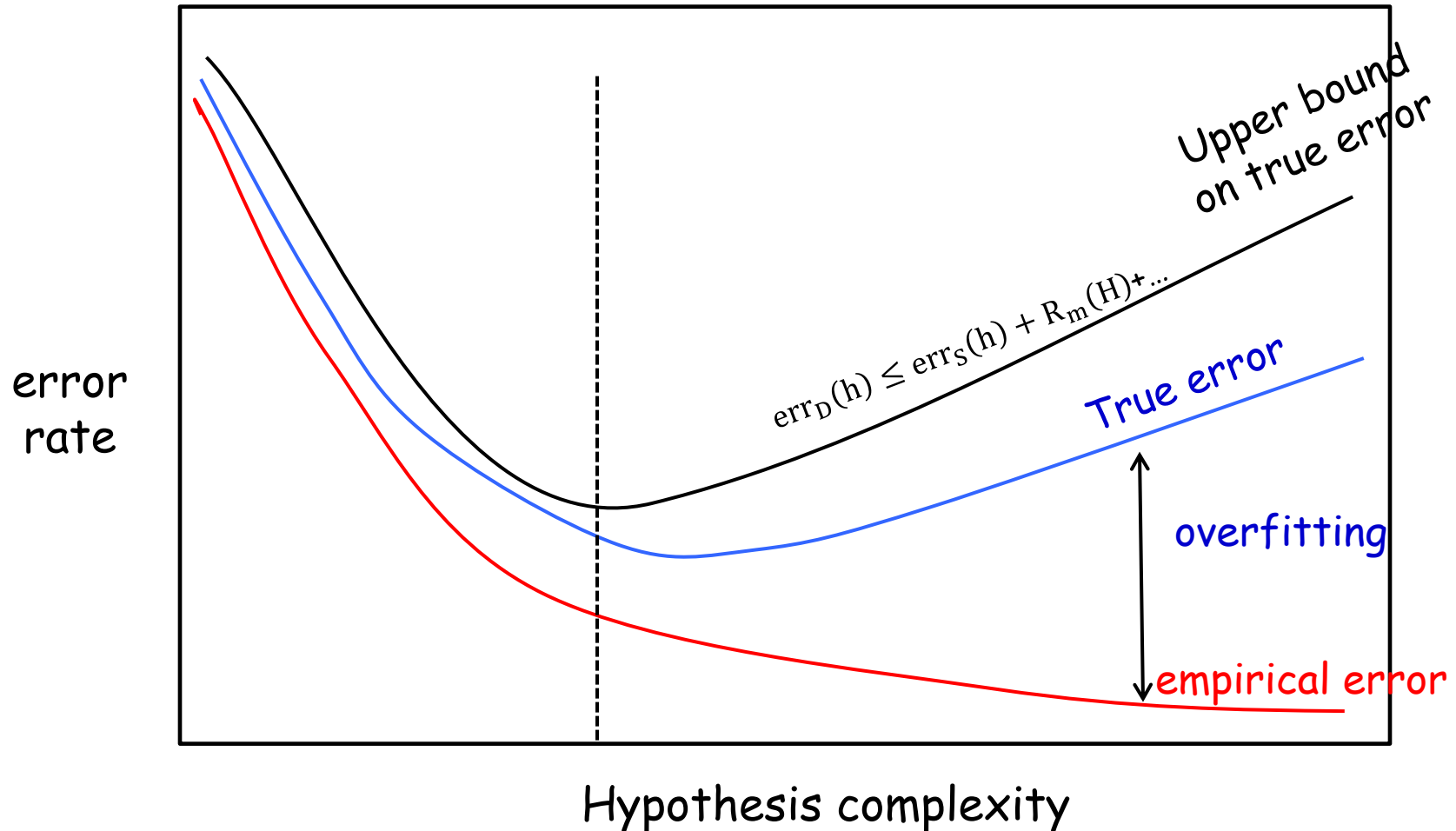
Model selection: trade-off between decreasing training error and keeping H simple.

$$\text{err}_D(h) \leq \text{err}_S(h) + R_m(H) + \dots$$



Structural Risk Minimization (SRM)

$$H_1 \subseteq H_2 \subseteq H_3 \subseteq \dots \subseteq H_i \subseteq \dots$$

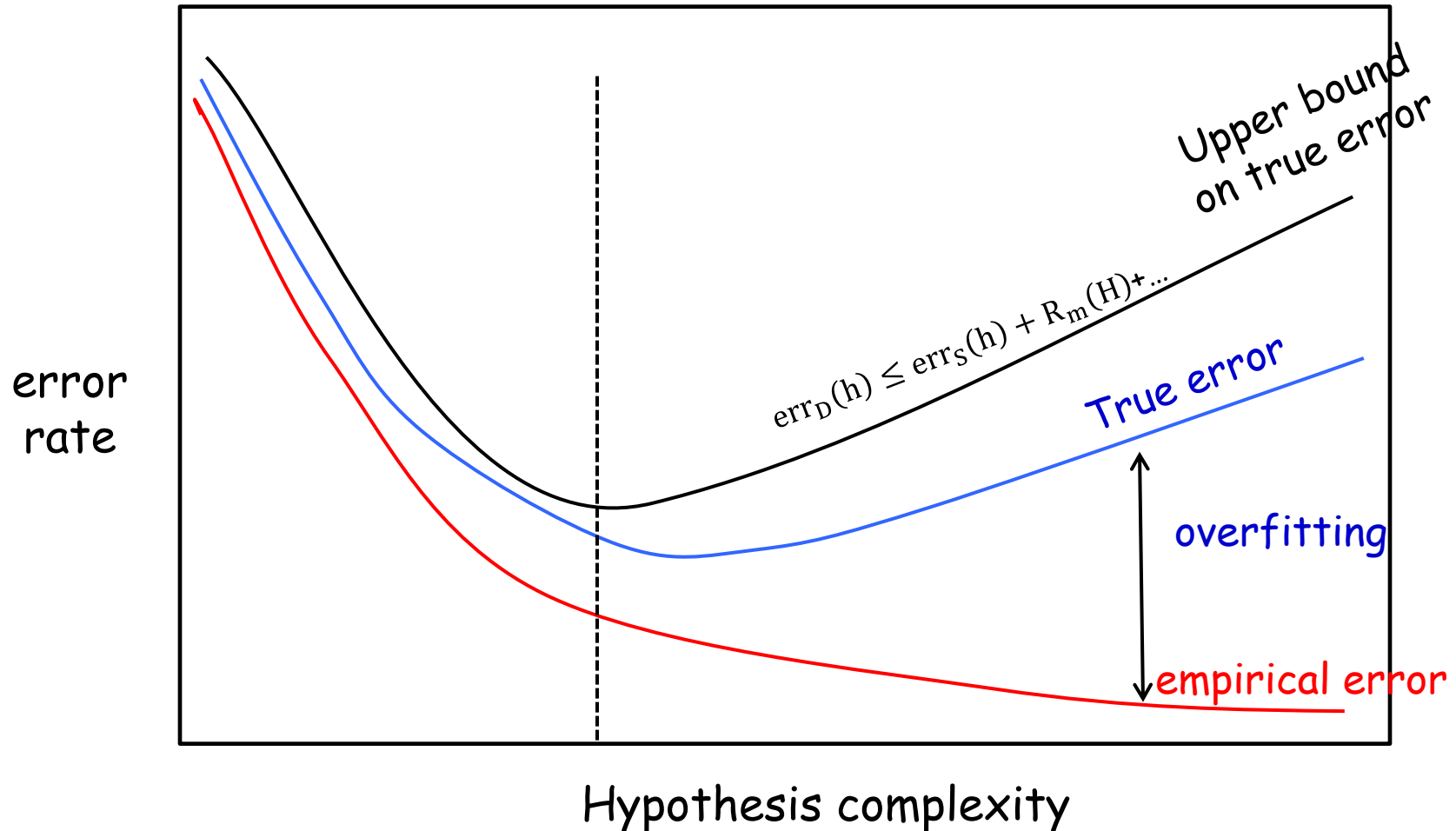


What happens if we increase m ?

Black curve will stay close to the red curve for longer, everything shifts to the right...

Structural Risk Minimization (SRM)

$$H_1 \subseteq H_2 \subseteq H_3 \subseteq \dots \subseteq H_i \subseteq \dots$$



Structural Risk Minimization (SRM)

- $H_1 \subseteq H_2 \subseteq H_3 \subseteq \dots \subseteq H_i \subseteq \dots$
- $\hat{h}_k = \operatorname{argmin}_{h \in H_k} \{\operatorname{err}_S(h)\}$
As k increases, $\operatorname{err}_S(\hat{h}_k)$ goes down but complex. term goes up.
- $\hat{k} = \operatorname{argmin}_{k \geq 1} \{\operatorname{err}_S(\hat{h}_k) + \operatorname{complexity}(H_k)\}$
Output $\hat{h} = \hat{h}_{\hat{k}}$

Claim: W.h.p., $\operatorname{err}_D(\hat{h}) \leq \min_{k^*} \min_{h^* \in H_{k^*}} [\operatorname{err}_D(h^*) + 2\operatorname{complexity}(H_{k^*})]$

Proof:

- We chose \hat{h} s.t. $\operatorname{err}_S(\hat{h}) + \operatorname{complexity}(H_{\hat{k}}) \leq \operatorname{err}_S(h^*) + \operatorname{complexity}(H_{k^*})$.
- Whp, $\operatorname{err}_D(\hat{h}) \leq \operatorname{err}_S(\hat{h}) + \operatorname{complexity}(H_{\hat{k}})$.
- Whp, $\operatorname{err}_S(h^*) \leq \operatorname{err}_D(h^*) + \operatorname{complexity}(H_{k^*})$.

Techniques to Handle Overfitting

- **Structural Risk Minimization (SRM).** $H_1 \subseteq H_2 \subseteq \dots \subseteq H_i \subseteq \dots$
Minimize gener. bound: $\hat{h} = \operatorname{argmin}_{k \geq 1} \{ \operatorname{err}_S(\hat{h}_k) + \operatorname{complexity}(H_k) \}$
 - Often computationally hard....
 - Nice case where it is possible: M. Kearns, Y. Mansour, ICML'98, "A Fast, Bottom-Up Decision Tree Pruning Algorithm with Near-Optimal Generalization"
- **Regularization:** general family closely related to SRM
 - E.g., SVM, regularized logistic regression, etc.
 - minimizes expressions of the form: $\operatorname{err}_S(h) + \lambda ||h||^2$
 - Some norm when H is a vector space; e.g., L_2 norm
- **Cross Validation:** Picked through cross validation
 - Hold out part of the training data and use it as a proxy for the generalization error

What you should know

- Notion of sample complexity.
- Understand reasoning behind the simple sample complexity bound for finite H .
- Shattering, VC dimension as measure of complexity, Sauer's lemma, form of the VC bounds.
- Model Selection, Structural Risk Minimization.