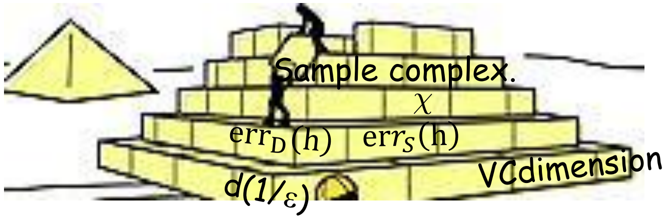
Machine Learning Theory II

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

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

?

Finite H

Realizable $c^* \in H$

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

Agnostic $c^* \notin H$

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

overestimated.

Infinite H

Realizable $c^* \in H$

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

Agnostic $c^* \notin H$

$$O\left(\sqrt{\frac{d}{m}\left(\ln\frac{m}{d}+\ln\frac{1}{\delta}\right)}\right)$$

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

$d \simeq$ #parameters

VC-dim.

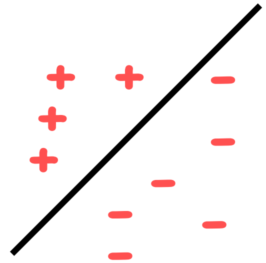
Some concepts:

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

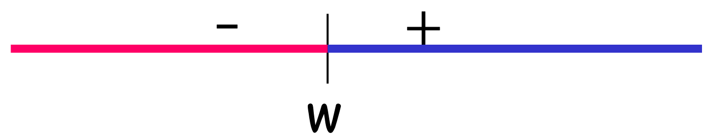
What if H is infinite?

E.g., linear separators in R^d

$h(x) = \mathrm{sign}(\beta^T x)$

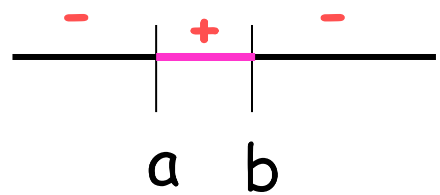


E.g., thresholds on the real line



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

E.g., intervals on the real line



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

An Effective Number of Hypotheses

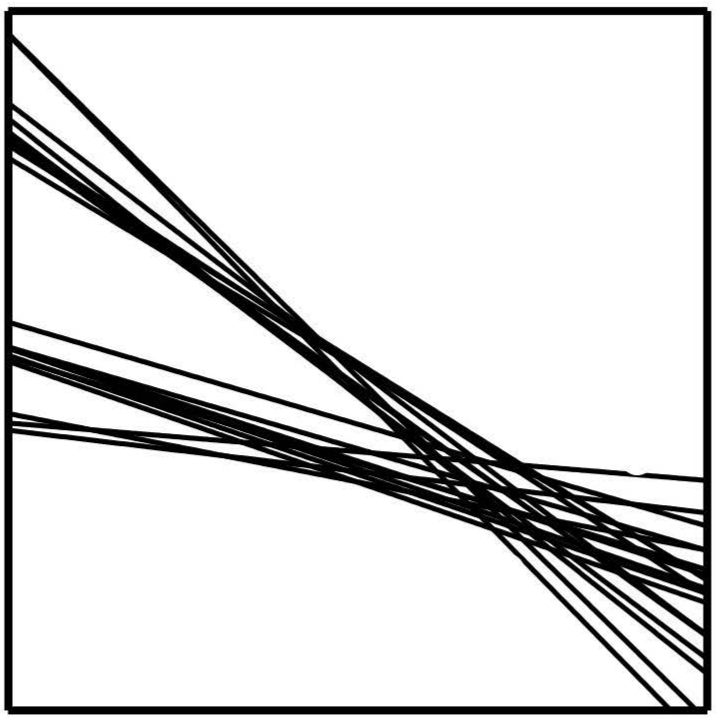
|H| only measures the maximum possible diversity of H

h_1

h_2

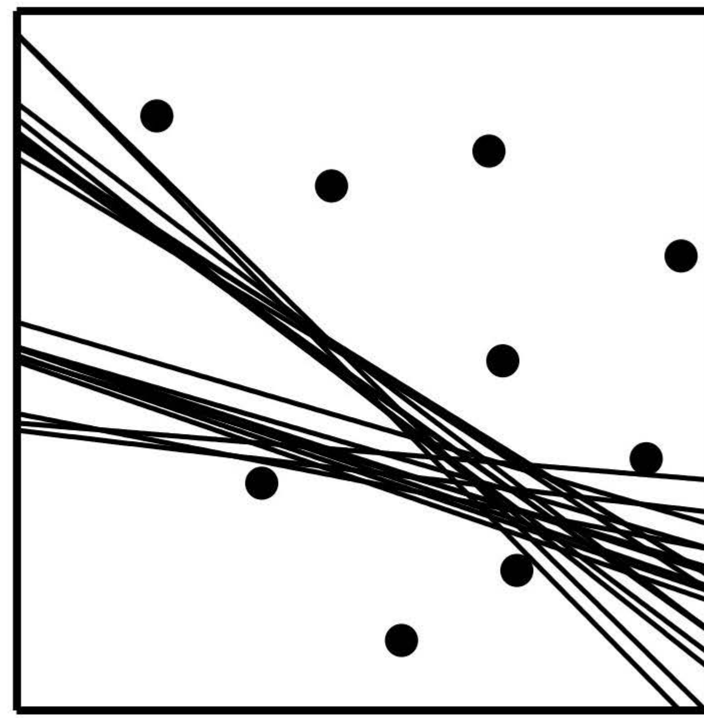
$\vdots$

h_K



$\mathcal{H}$

$|H| = \infty$

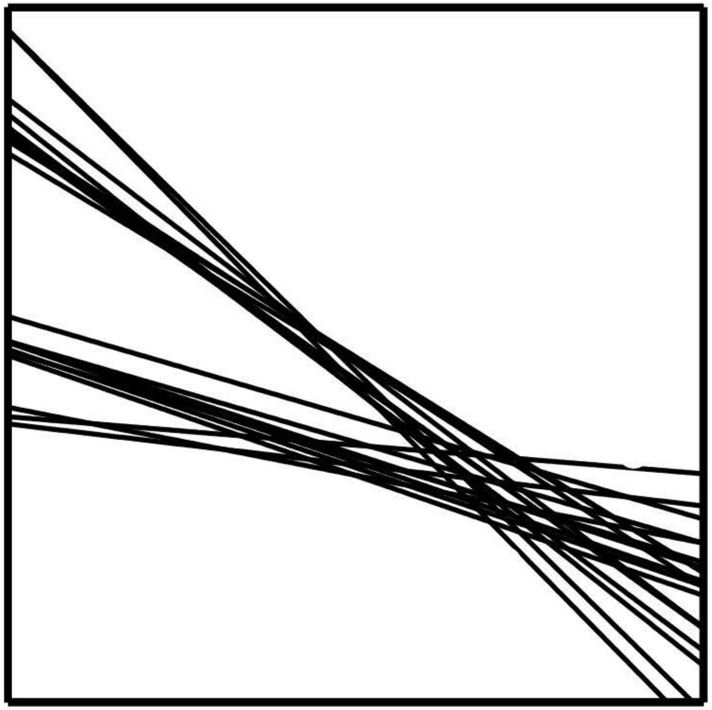


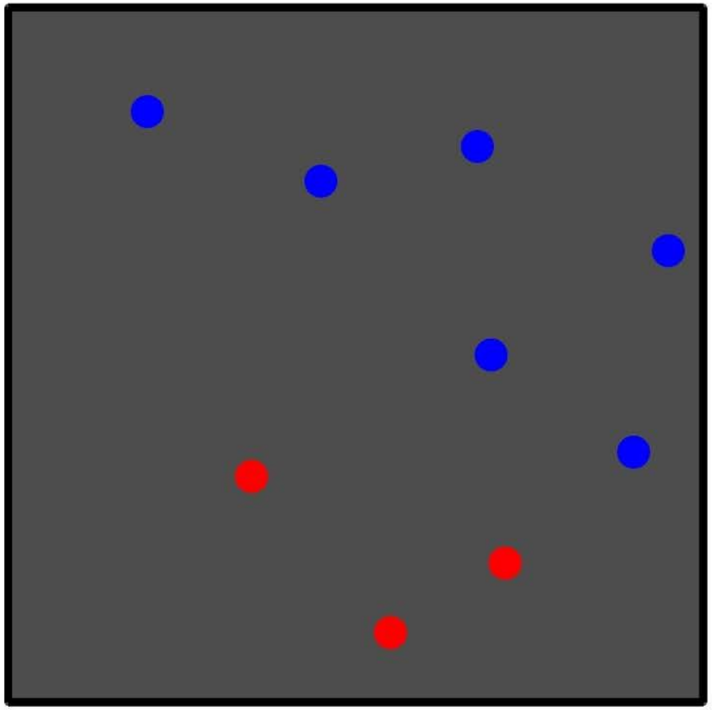
$\mathcal{H}$ through the eyes of the $\mathcal{D}$

$S = \{x_1, x_2, \ldots, x_9\}$

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

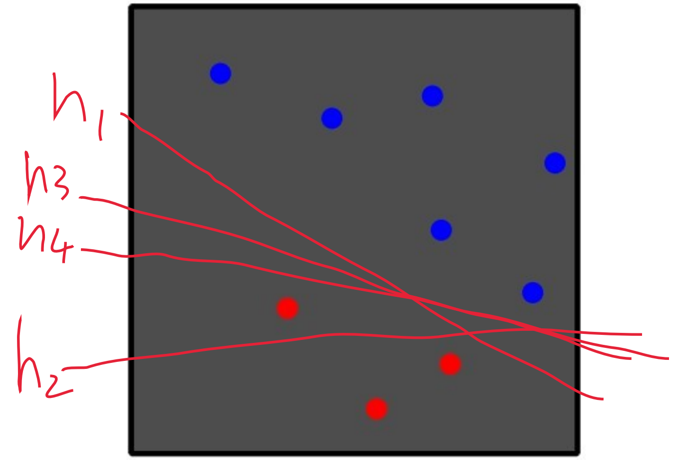
$|H|$ only measures the maximum possible diversity of H

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

$(h(x_1),\ldots,h(x_m))$ $h: X \to \{-1,+1\}$

A dichotomy of S

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



dichotomy

Growth Function.

*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. $H[m] = \max_{|S|=m} |H[S]|$, $H(S) = \{ h(x_1), \ldots, h(x_m) \mid h \in H \}$

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

$c^* \in H$

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

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

H: finite

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

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

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 $err_S(h) > 0$.

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

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

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

- $H[S] = \{ h(x_1), \dots, h(x_m) \mid h \in H \}$
- $H[m] = \max_{|S|=m} |H[S]| = \max \{ |H[S]| \mid |S|=m, \forall S \subseteq D \}$

Linear separator $(X = R^2)$

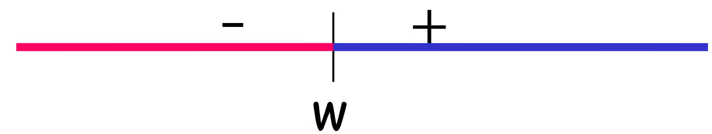
Effective number of hypotheses

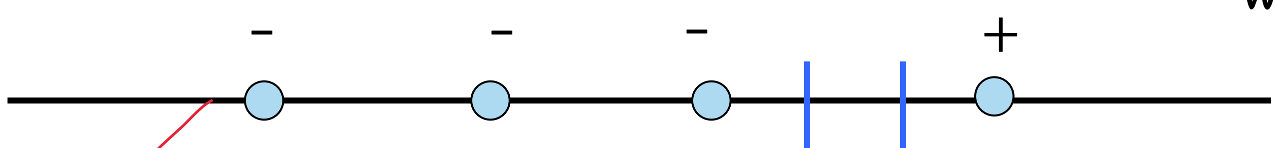
- H[S] – the set of splittings of dataset S using concepts from H.
- H[m] - max number of ways to split m points using concepts in H

$$H[m] = \max_{|S|=m} |H[S]| \qquad H[m] \leq 2^m$$

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

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





$|H[S]| = 5$

$H\{S\} =$

–	–	–	–
–	–	+	+
–	+	+	+
+	+	+	+

$H[4] = 5$

$H[m] = \binom{m+1}{1} = m+1$

$|H(S)| = \binom{m+1}{1}$

$= m+1 << 2^m$

VC-dim = 1

$\boxed{H\{m\} = 2^m, \quad m=1}$

$H\{m\} < 2^m, \quad m \geq 2$

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

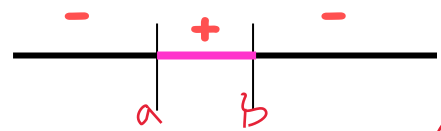
Effective number of hypotheses

- 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

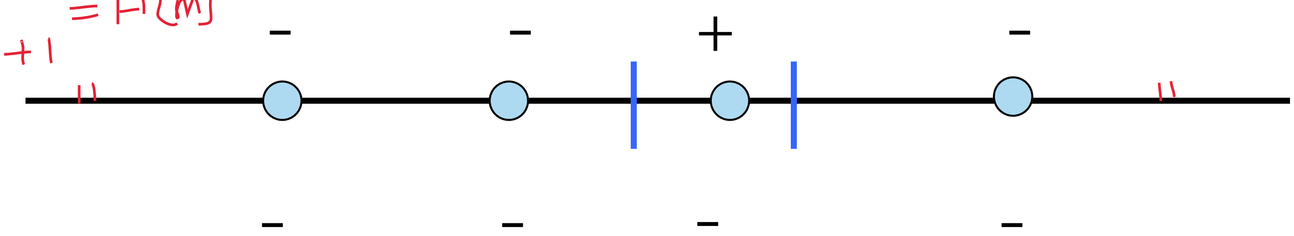
$$H[m] = \max_{|S|=m} |H[S]| \qquad H[m] \leq 2^m$$

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

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



$\binom{m+1}{2} + 1 = H\{m]$



$\binom{4+1}{2} + 1 = \frac{5 \times 4}{2} + 1 = 11$

$|H(S)| = 11$

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

VC-dim

interval $\begin{cases} H\{m] = 2^m, & m=1, \; m=2 \\ H\{m] < 2^m, & m \geq 3 \end{cases}$

There are m+1 possible options for the first part, m left for the second part, the order does not matter, so (m choose 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

$$H[m] = \max_{|S|=m} |H[S]| \qquad H[m] \leq 2^m$$

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

$H[m] =$
$|H(S)| = 2^m$

- From the viewpoint of S,

  H is the most powerful hypothesis space

Sample Complexity: Infinite Hypothesis Spaces
Realizable Case

($C^* \in H$, H. infinite).

$|H| \leftarrow H[m] \leftarrow$ VC-dim

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

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

$m \geq \frac{1}{\varepsilon}\left(\ln \boxed{|H|} + \ln \frac{1}{\delta} \right)$

(H finite.)

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

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

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

- 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

$C^* \in H$

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: $H[m] = O\left(m^{VCdim(H)}\right)$ In practice → VC-dim ≃ # paras; → m ≃ 10 VC-dim

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 $err_S(h) > 0$.

($err_S(h)=0 \Rightarrow err_D(h)<\varepsilon$)

Shattering, VC-dimension

Definition: 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, all possible ways of classifying points in S are achievable using concepts in H.

Definition: VC-dimension (Vapnik-Chervonenkis dimension)

samples of S = m.

The VC-dimension of a hypothesis space H is the cardinality of the $\boxed{\text{largest}}$ set S that can be shattered by H.

$H\{m\} = 2^m$

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

$$\text{VC-dim}(H) = \max_{S \subset D} \{ |S| \mid H \text{ shatters } S \}$$

$$= \max_{S} \{ |S| \mid |H(S)| = 2^{|S|} \}$$

$$= \max_{m} \{ m \mid H\{m\} = 2^m \}$$

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

VC-dim = d.

To show that VC-dimension is d:

S • VC-dim(H) ≥ d

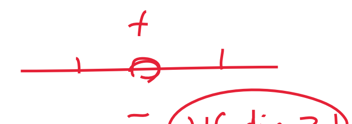
- there exists a set of d points that can be shattered
- there is no set of d+1 points that can be shattered.

• VC-dim(H) < d+1

Fact: If H is finite, then $\mathrm{VCdim}(H) \leq \log(|H|)$. $2^d < |H|$

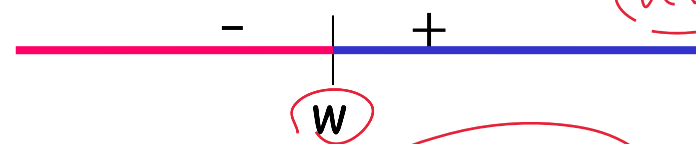
Shattering, VC-dimension

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.



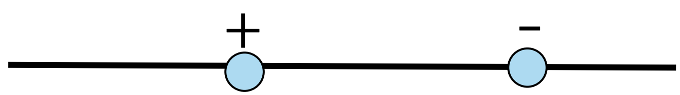
VC-dim ≥ 1

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

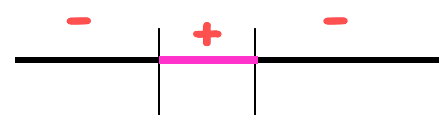


VC-dim < 2

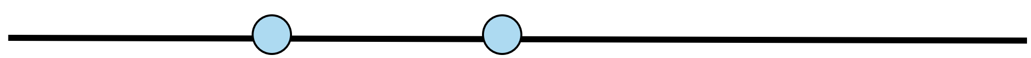
$\text{VCdim(H)} = 1$

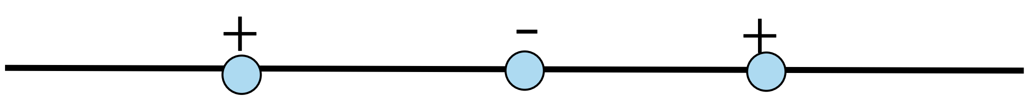


E.g., H= Intervals on the real line



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

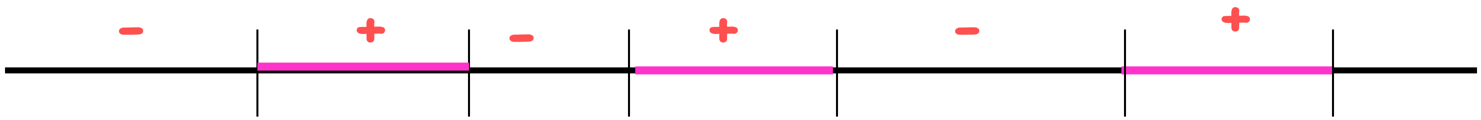




Shattering, VC-dimension

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



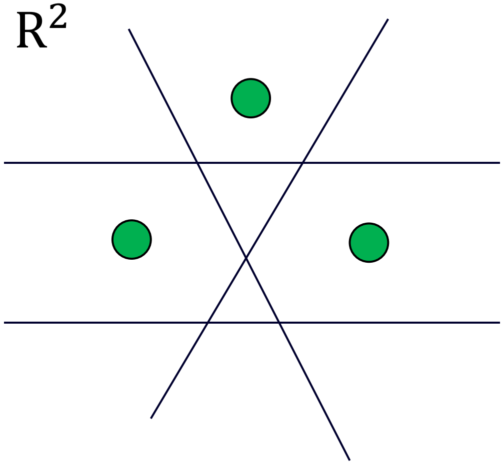
$\mathrm{VCdim(H)} \geq 2\mathrm{k}$

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

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

Shattering, VC-dimension

E.g., **H=** linear separators in R^2



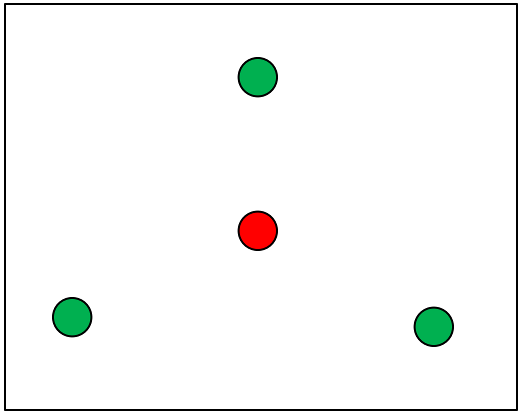
$\mathrm{VCdim(H)} \geq 3$

Shattering, VC-dimension

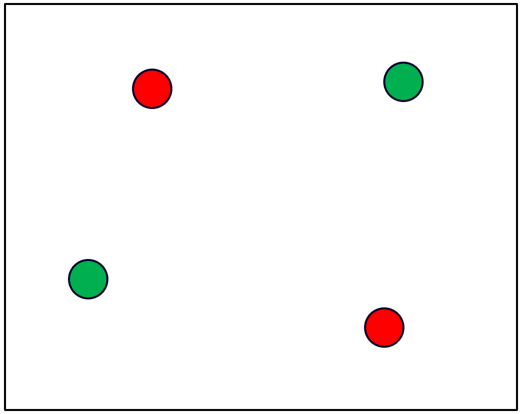
E.g., **H=** linear separators in R^2

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

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 R^d is d+1

Today's Quiz

Sauer's Lemma

Sauer's Lemma:

Let d = VCdim(H)

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

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: $\mathrm{H[m] = O\left(m^{VCdim(H)}\right)}$

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

Theorem

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

Prob. over different samples of m training examples

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

Linear in $1/\epsilon$

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: Infinite Hypothesis Spaces
Realizable Case

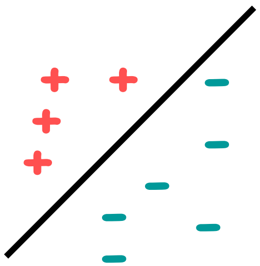
Theorem

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

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) ~ #free parameters of H*

What if $c^* \notin H$?

Sample Complexity: Uniform Convergence
Agnostic Case

Empirical Risk Minimization (ERM)

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

Theorem

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

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

$1/\epsilon^2$ dependence [as opposed to1/ϵ for realizable]

Theorem

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

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).

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

Theorem

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

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

2) Statistical Learning Theory style:

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

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

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

Sample Complexity: Infinite Hypothesis Spaces
Agnostic Case

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

Theorem

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

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 $\mathrm{h} \in \mathrm{H}$:

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

VCdimension Generalization Bounds

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

VC bounds: distribution independent bounds

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

  [nearly tight in the WC over choice of D]

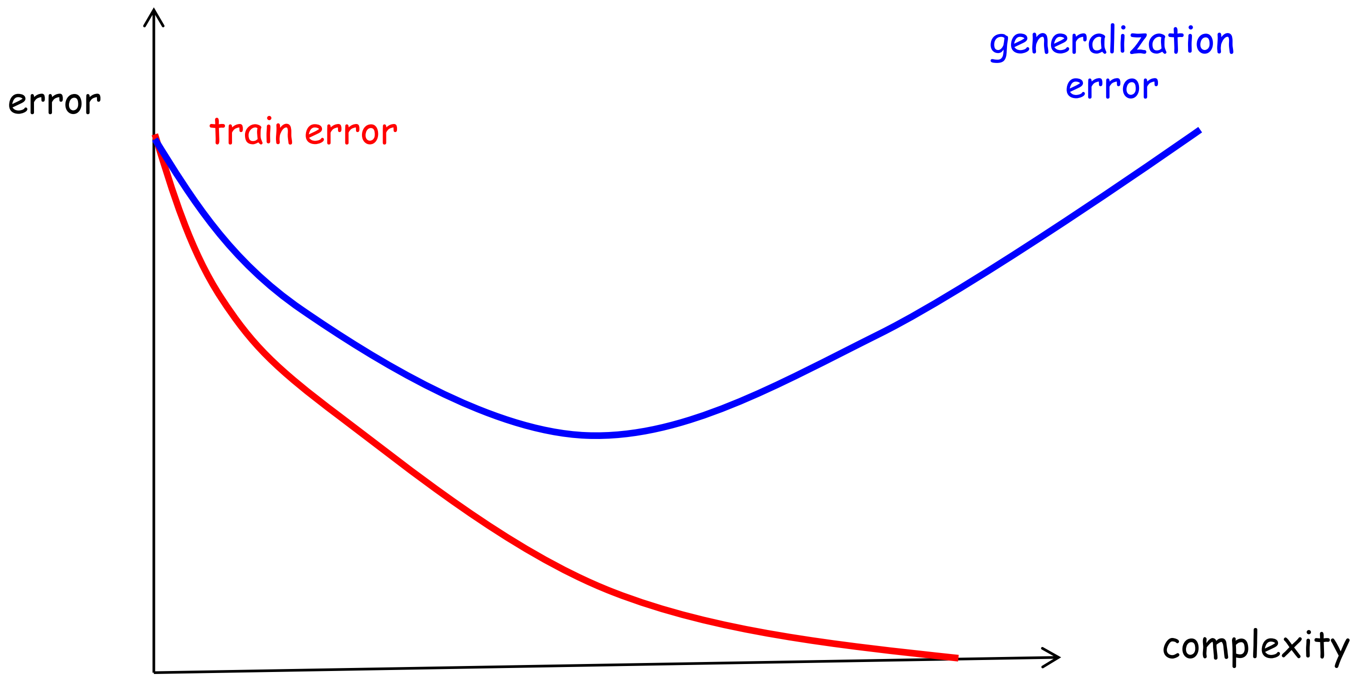
- Might be very loose specific distr. that are more benign than the worst case....
- Hold only for binary classification; we want bounds for fns approximation in general (e.g., multiclass classification and regression).

Can we use our bounds for model selection?

True Error, Training Error, Overfitting

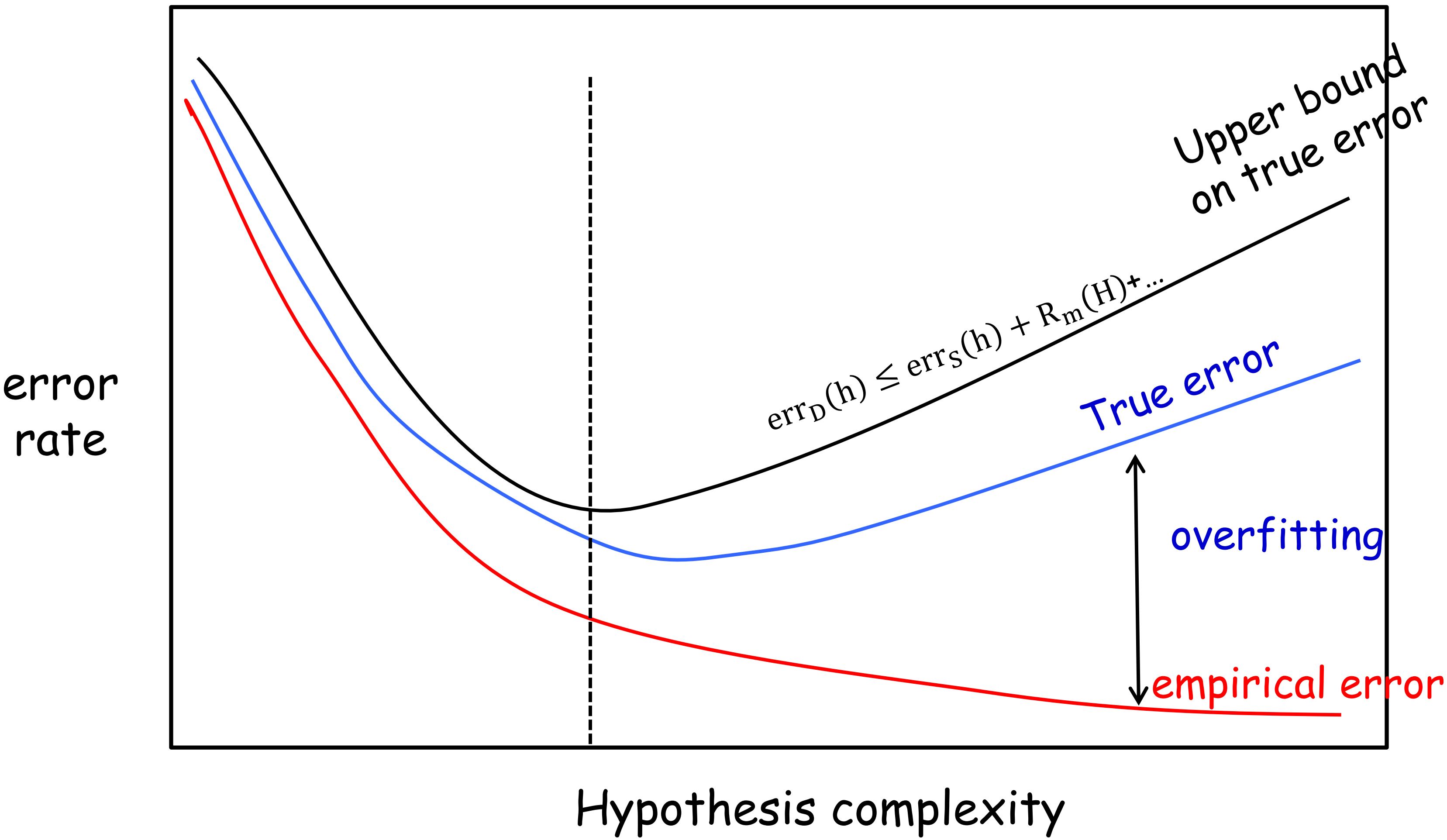
Model selection: trade-off between decreasing training error and keeping H simple.

$$\mathrm{err}_D(h) \leq \mathrm{err}_S(h) + R_m(H) + \ldots$$



Structural Risk Minimization (SRM)

$H_1 \subseteq H_2 \subseteq H_3 \subseteq \cdots \subseteq H_i \subseteq \ldots$

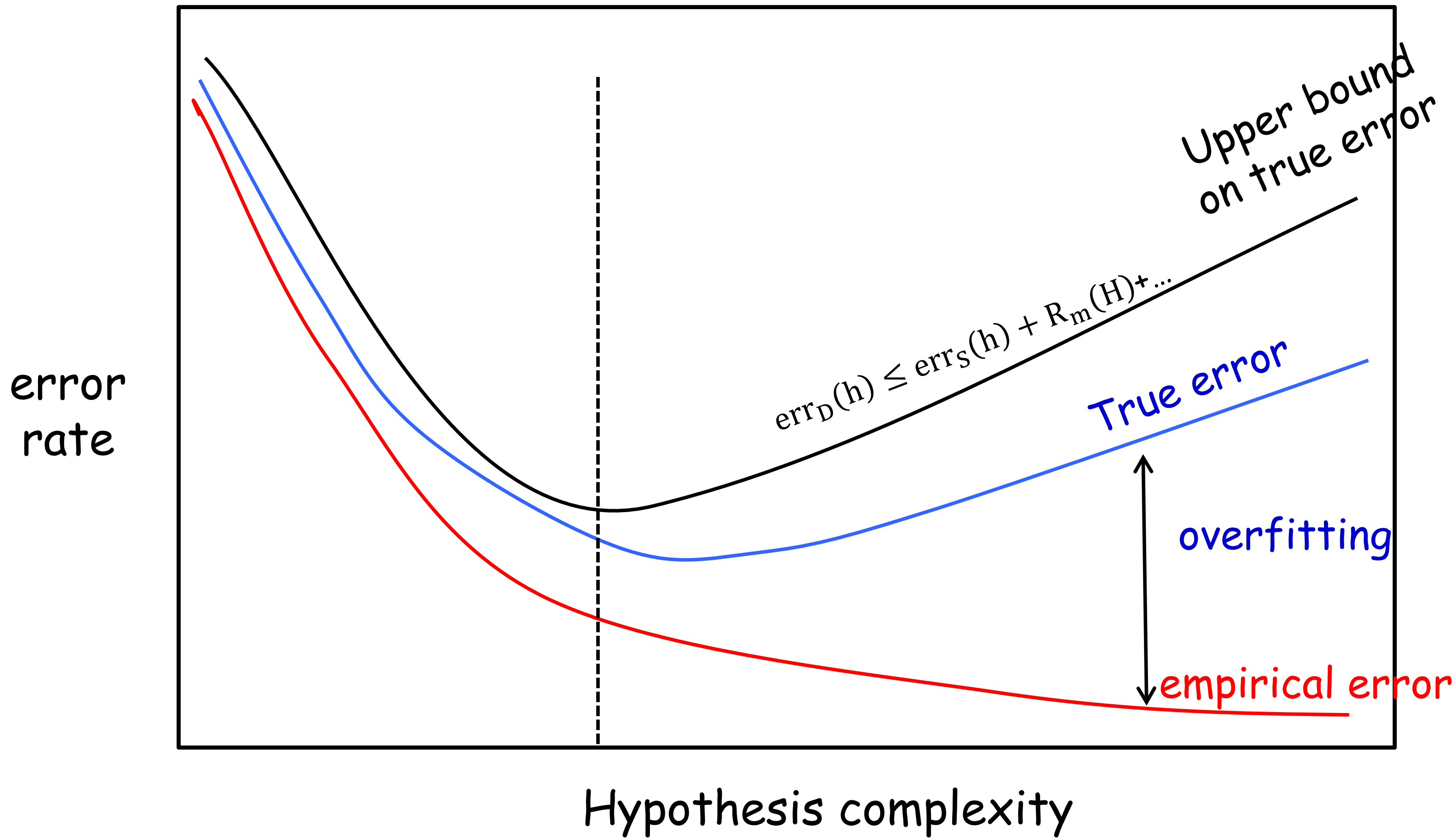


What happens if we increase m?

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

Structural Risk Minimization (SRM)

$H_1 \subseteq H_2 \subseteq H_3 \subseteq \cdots \subseteq H_i \subseteq \ldots$



Structural Risk Minimization (SRM)

- $H_1 \subseteq H_2 \subseteq H_3 \subseteq \cdots \subseteq H_i \subseteq \ldots$
- $\hat{\mathrm{h}}_{\mathrm{k}} = \mathrm{argmin}_{\mathrm{h} \in \mathrm{H}_{\mathrm{k}}} \{\mathrm{err}_{\mathrm{S}}(\mathrm{h})\}$

  As k increases, $\mathrm{err}_{\mathrm{S}}(\hat{\mathrm{h}}_{\mathrm{k}})$ goes down but complex. term goes up.

- $\hat{k} = \mathrm{argmin}_{\mathrm{k} \geq 1} \{\mathrm{err}_{\mathrm{S}}(\hat{\mathrm{h}}_{\mathrm{k}}) + \mathrm{complexity}(\mathrm{H}_{\mathrm{k}})\}$

  Output $\hat{h} = \hat{h}_{\hat{k}}$

Claim: W.h.p., $\mathrm{err}_{\mathrm{D}}(\hat{\mathrm{h}}) \leq \min_{\mathrm{k}^*} \min_{\mathrm{h}^* \in \mathrm{H}_{\mathrm{k}^*}} [\mathrm{err}_{\mathrm{D}}(\mathrm{h}^*) + 2\mathrm{complexity}(\mathrm{H}_{\mathrm{k}^*})]$

Proof:

- We chose $\hat{\mathrm{h}}$ s.t. $\mathrm{err}_{\mathrm{s}}(\hat{\mathrm{h}}) + \mathrm{complexity}(\mathrm{H}_{\hat{\mathrm{k}}}) \leq \mathrm{err}_{\mathrm{S}}(\mathrm{h}^*) + \mathrm{complexity}(\mathrm{H}_{\mathrm{k}^*})$.
- Whp, $\mathrm{err}_{\mathrm{D}}(\hat{\mathrm{h}}) \leq \mathrm{err}_{\mathrm{s}}(\hat{\mathrm{h}}) + \mathrm{complexity}(\mathrm{H}_{\hat{\mathrm{k}}})$.
- Whp, $\mathrm{err}_{\mathrm{S}}(\mathrm{h}^*) \leq \mathrm{err}_{\mathrm{D}}(\mathrm{h}^*) + \mathrm{complexity}(\mathrm{H}_{\mathrm{k}^*})$.

Techniques to Handle Overfitting

- **Structural Risk Minimization (SRM).** $H_1 \subseteq H_2 \subseteq \cdots \subseteq H_i \subseteq \ldots$

  Minimize gener. bound: $\hat{h} = \mathrm{argmin}_{k \geq 1}\{\mathrm{err}_S(\hat{\mathrm{h}}_k) + \mathrm{complexity}(\mathrm{H}_k)\}$

  - 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: $\mathrm{err}_S(\mathrm{h}) + \lambda||\mathrm{h}||^2$

Some norm when H is a vector space; e.g., L_2 norm

Picked through cross validation

- **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.