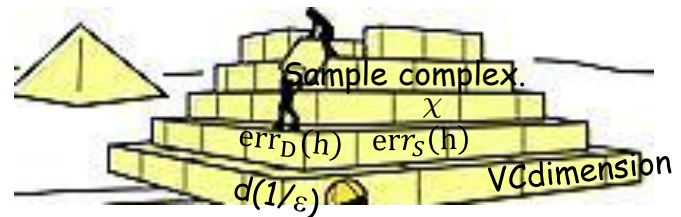# Machine Learning Theory II

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 $c^* \notin H$

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

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

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

$$err_D(h) \leq err_S(h) + \mathcal{O}\left(\sqrt{\frac{1}{m}\left(d + \ln\frac{1}{\delta}\right)}\right)$$
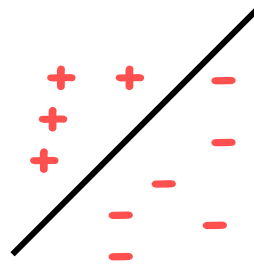
Some concepts:
1. Consistent learner
2. Version space (VS)
3. ε-exhausted VS
4. PAC-learnable
5. Hoeffding inequality
6. Dichotomy
7. Shattering
8. Shattering coefficient (growth function)
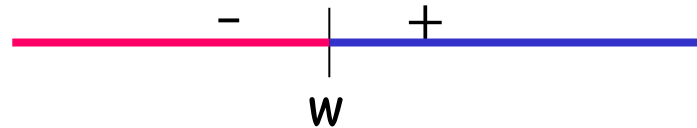9. VC-dimension
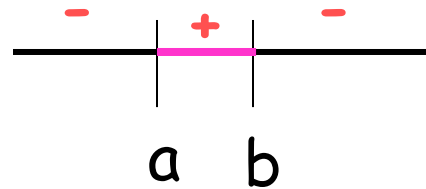10. Sauer's lemma

What if H is infinite?

E.g., linear separators in $\mathbb{R}^d$
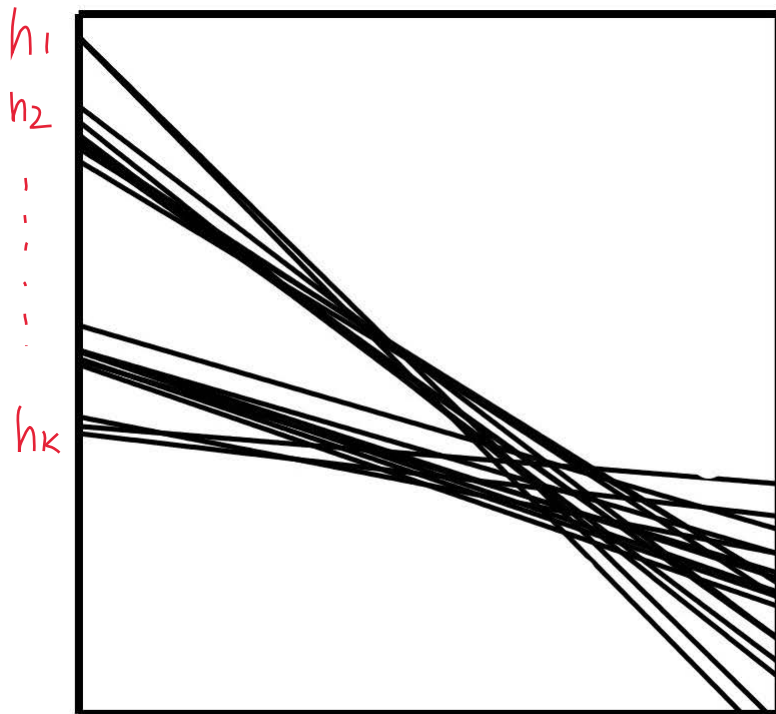
E.g., thresholds on the real line

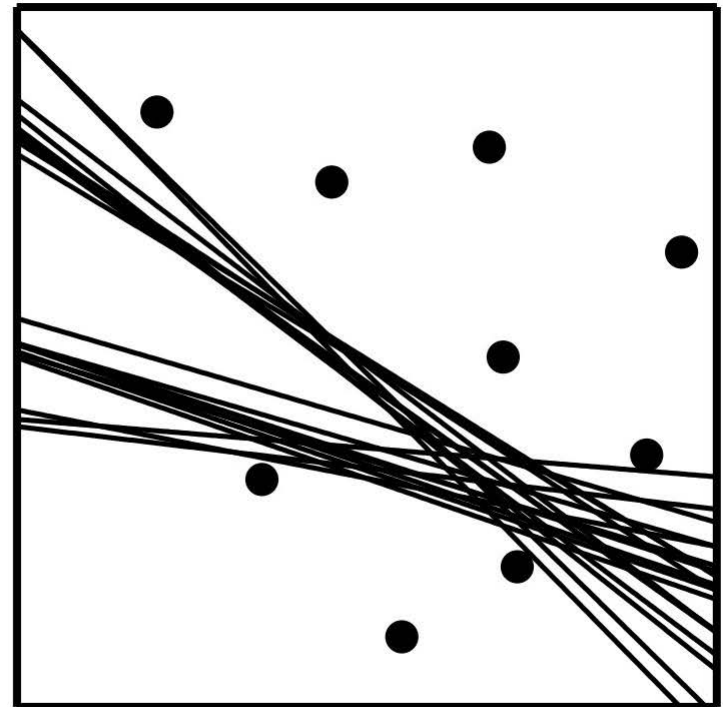E.g., intervals on the real line

An Effective Number of Hypotheses

|H| only measures the maximum possible diversity of H
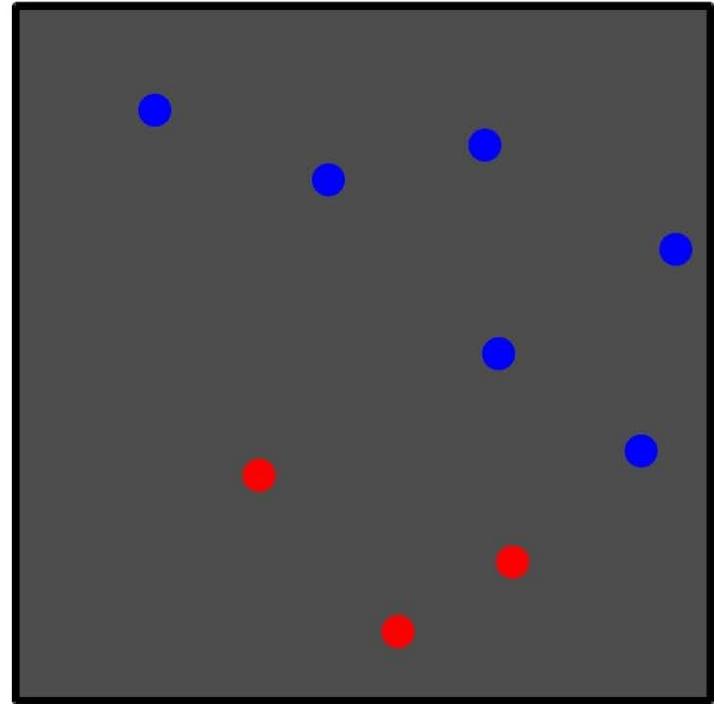

$\mathcal{H}$ $\mathcal{H}$ through the eyes of the $\mathcal{D}$

An Effective Number of Hypotheses

|H| only measures the maximum possible diversity of H
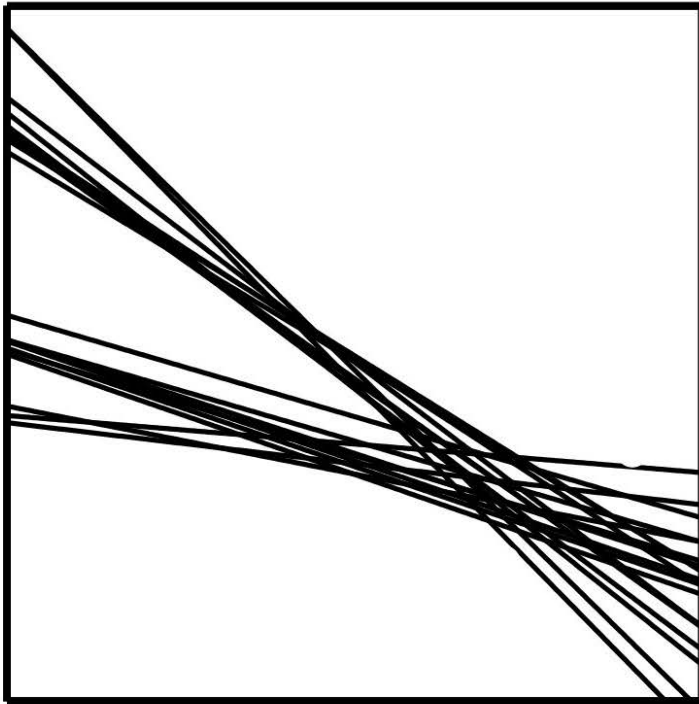
From the viewpoint of S, the entire H is just one dichotomy
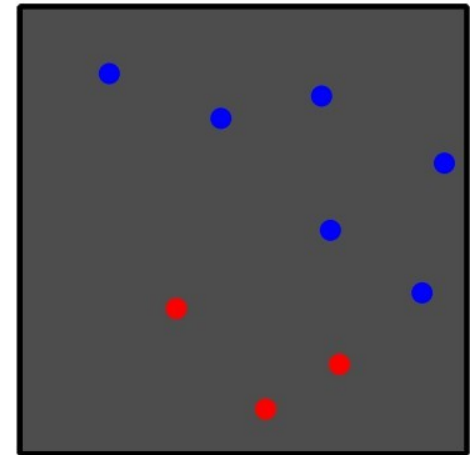
An Effective Number of Hypotheses

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

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

   $(h(x_1),...,h(x_m))$

   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

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

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

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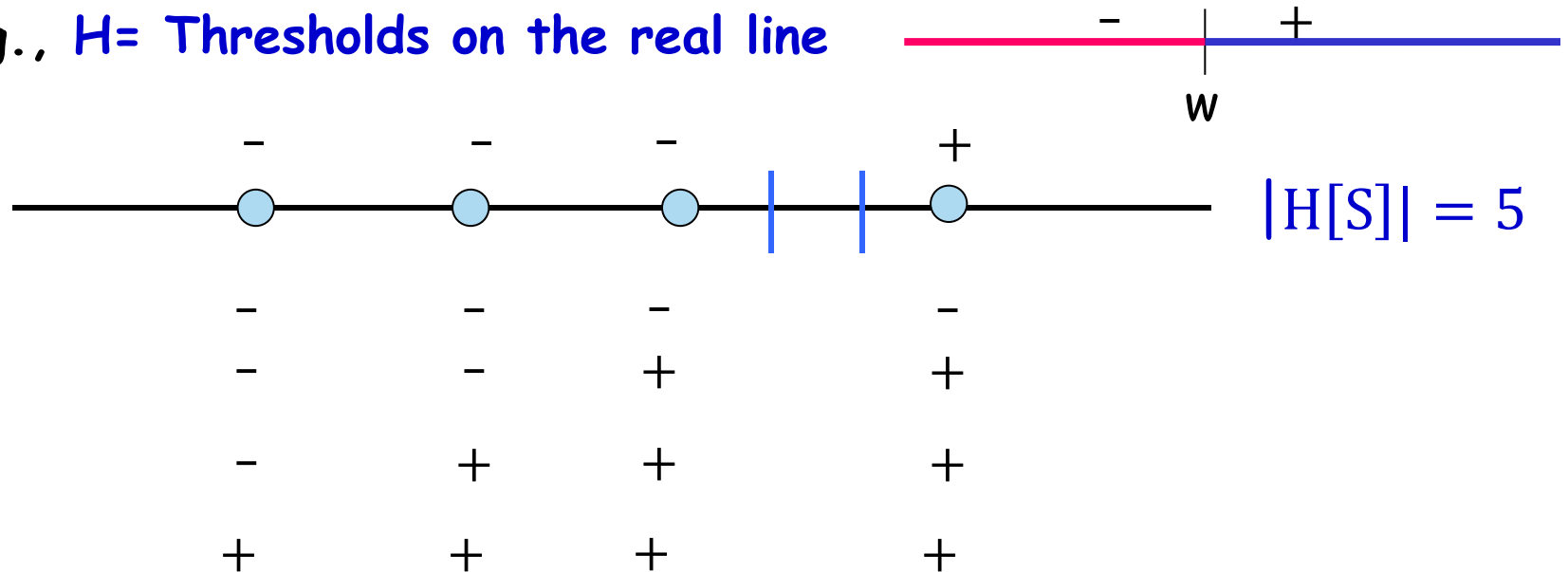
$$H[m] = \max_{|S|=m} |H[S]|$$

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

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

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

$$|H[S]| = 5$$

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

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

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 (m choose 2) + 1 (for empty interval).
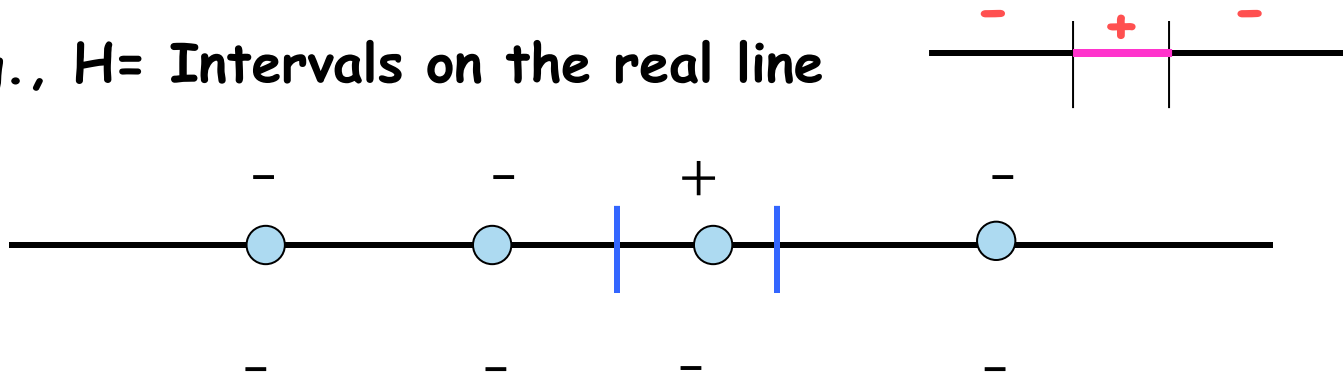
Effective number of hypotheses

- $H[S]$ – 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 $|H[S]| = 2^{|S|}$.

Sample Complexity: Infinite Hypothesis Spaces
Realizable Case

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 $err_S(h) > 0$.

- 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 $m \geq \frac{m}{\epsilon}(....)$ ☹

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

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 $err_S(h) > 0$.

Shattering, VC-dimension

Definition: 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, 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 $VCdim(H) = \infty$

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 $VCdim(H) = \infty$
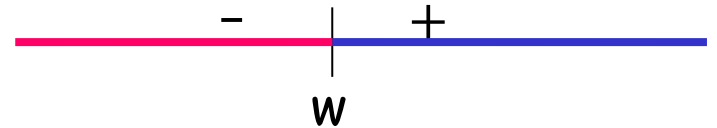
To show that VC-dimension is d:

- 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 $VCdim(H) \leq \log(|H|)$.

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

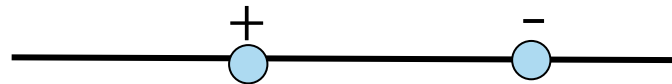$VCdim(H) = 1$

E.g., H= Intervals on the real line
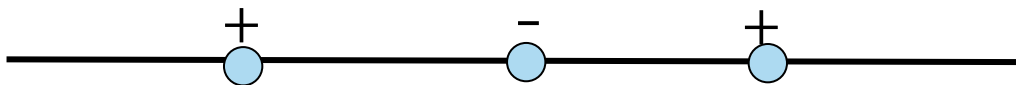
$VCdim(H) = 2$

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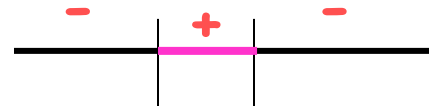
E.g., H= Union of k intervals on the real line $\text{VCdim(H)} = 2k$
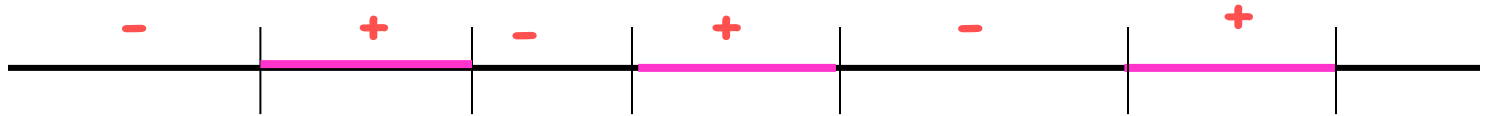
$\text{VCdim(H)} \geq 2k$

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

$\text{VCdim(H)} < 2k + 1$
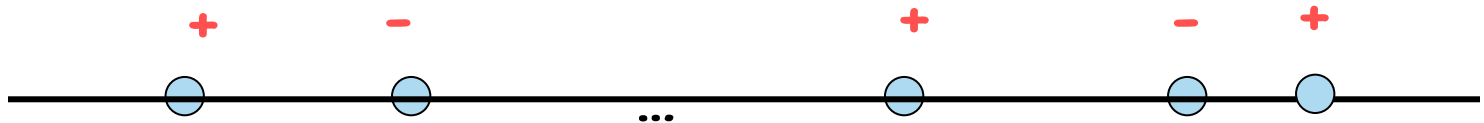
Shattering, VC-dimension

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

VCdim(H) $\geq$ 3
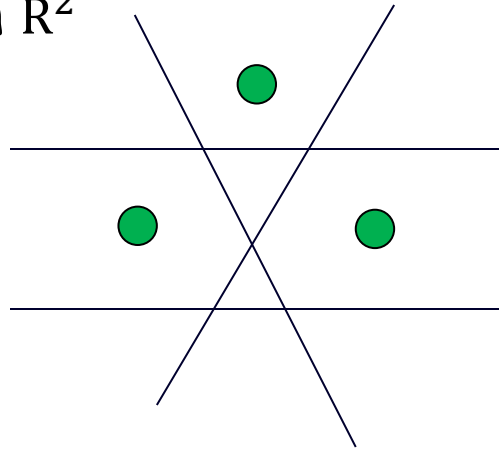
Shattering, VC-dimension

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

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

Case 1: one point inside the triangle formed by the others. Cannot label inside point as positive and outside points as negative.
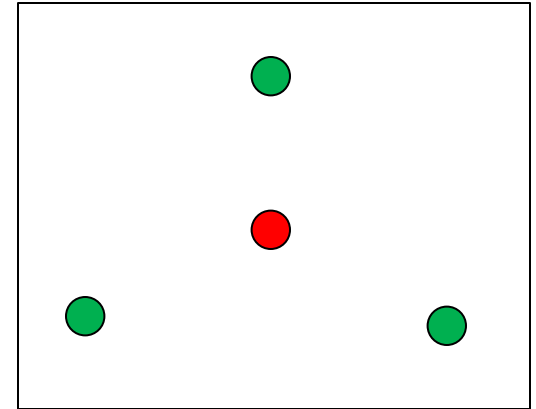
Case 2: all points on the boundary (convex hull). Cannot label two diagonally as positive and other two as negative.
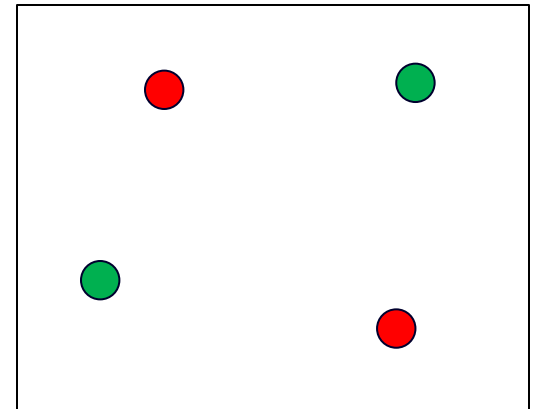
Fact: VCdim of linear separators in $\mathbb{R}^d$ is d+1

Today's Quiz

Sauer's Lemma

Sauer's Lemma:

Let d = VCdim(H)

- $m \leq d$, then $H[m] = 2^m$

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

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

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

Prob. over different samples of m training examples

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

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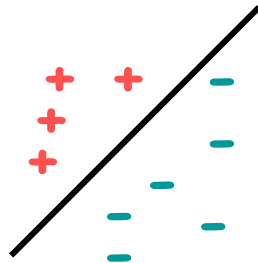
$$m = O\left(\frac{1}{\varepsilon}\left[d\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

VCdim(H)= d+1

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 to $1/\epsilon$ 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).

Theorem

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

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

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 h $\in$ 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 $h \in H$:

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

VCdimension Generalization Bounds

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

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).
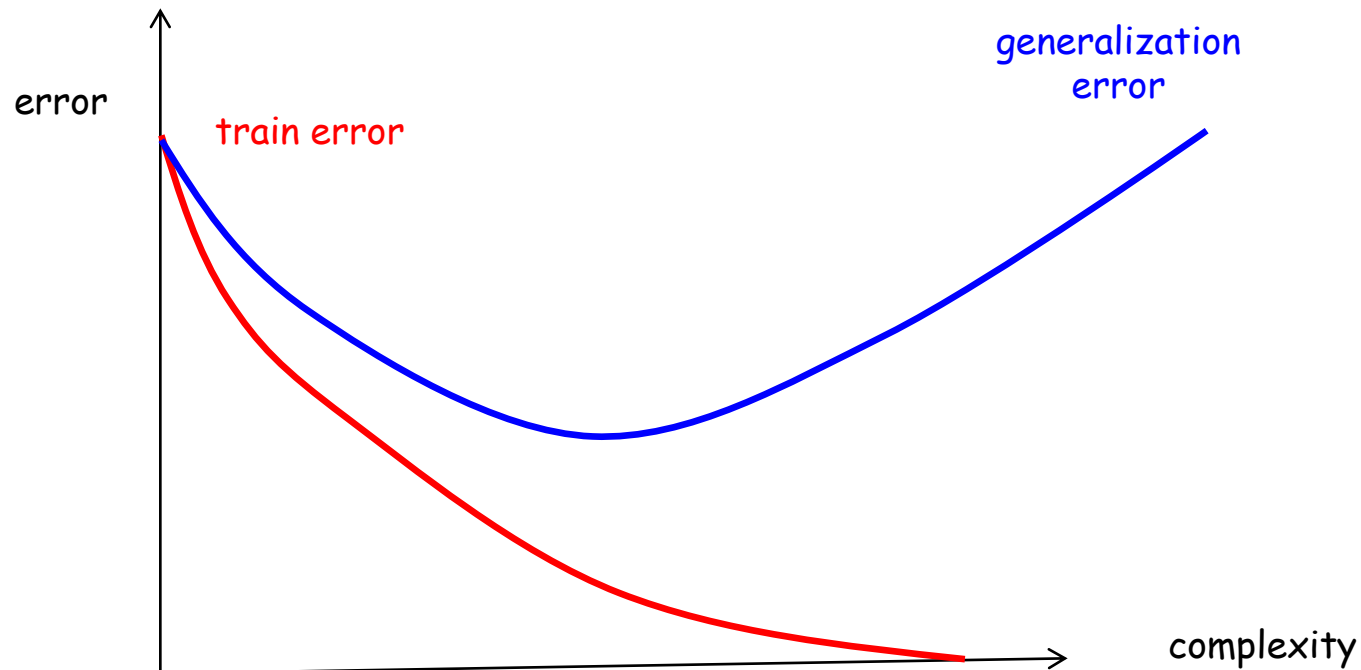
Can we use our bounds for model selection?

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)

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

error rate

Upper bound on true error

$err_D(h) \leq err_S(h) + R_m(H) + \ldots$

True error

overfitting

empirical error

Hypothesis complexity
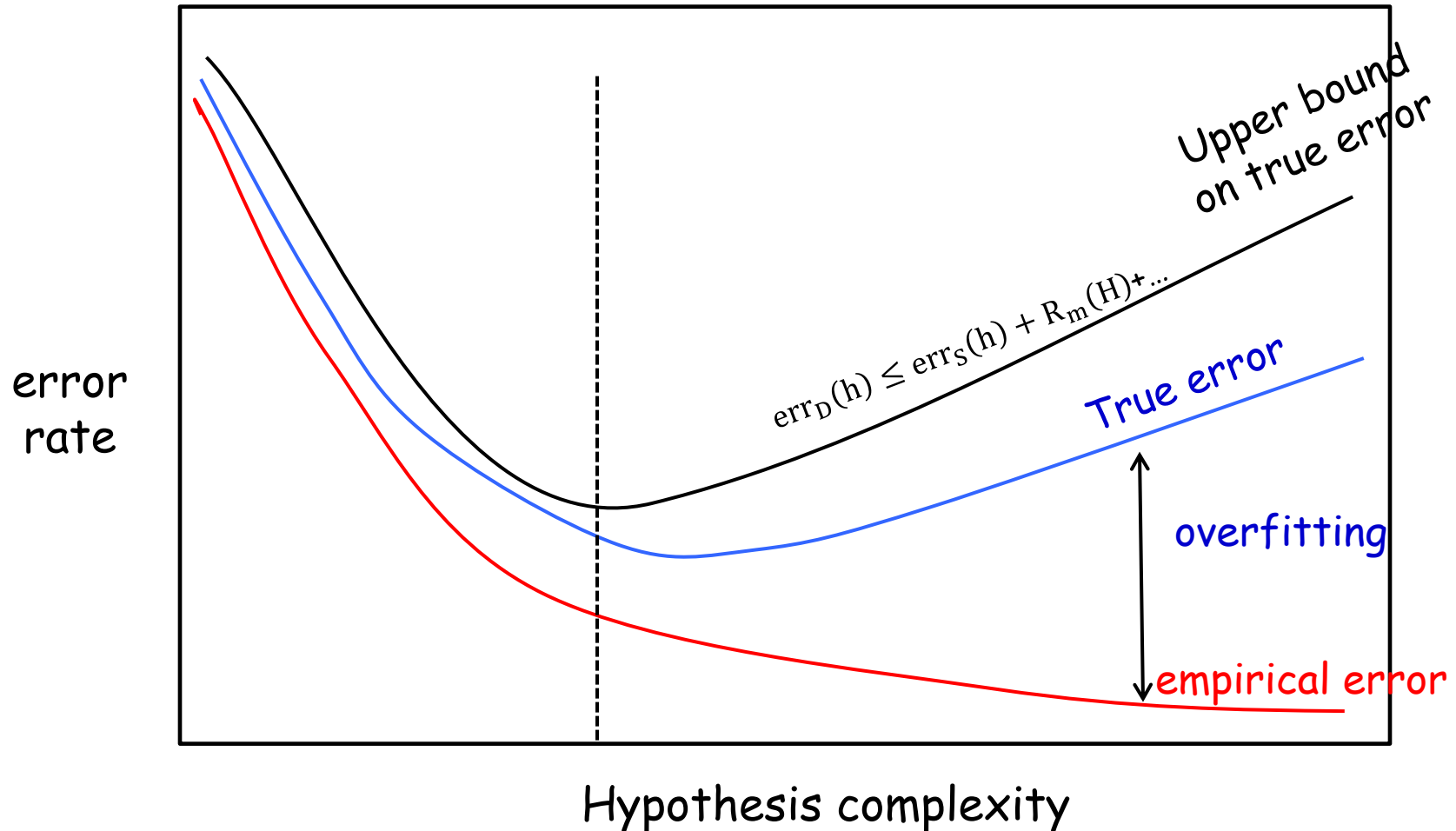
What happens if we increase m?

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

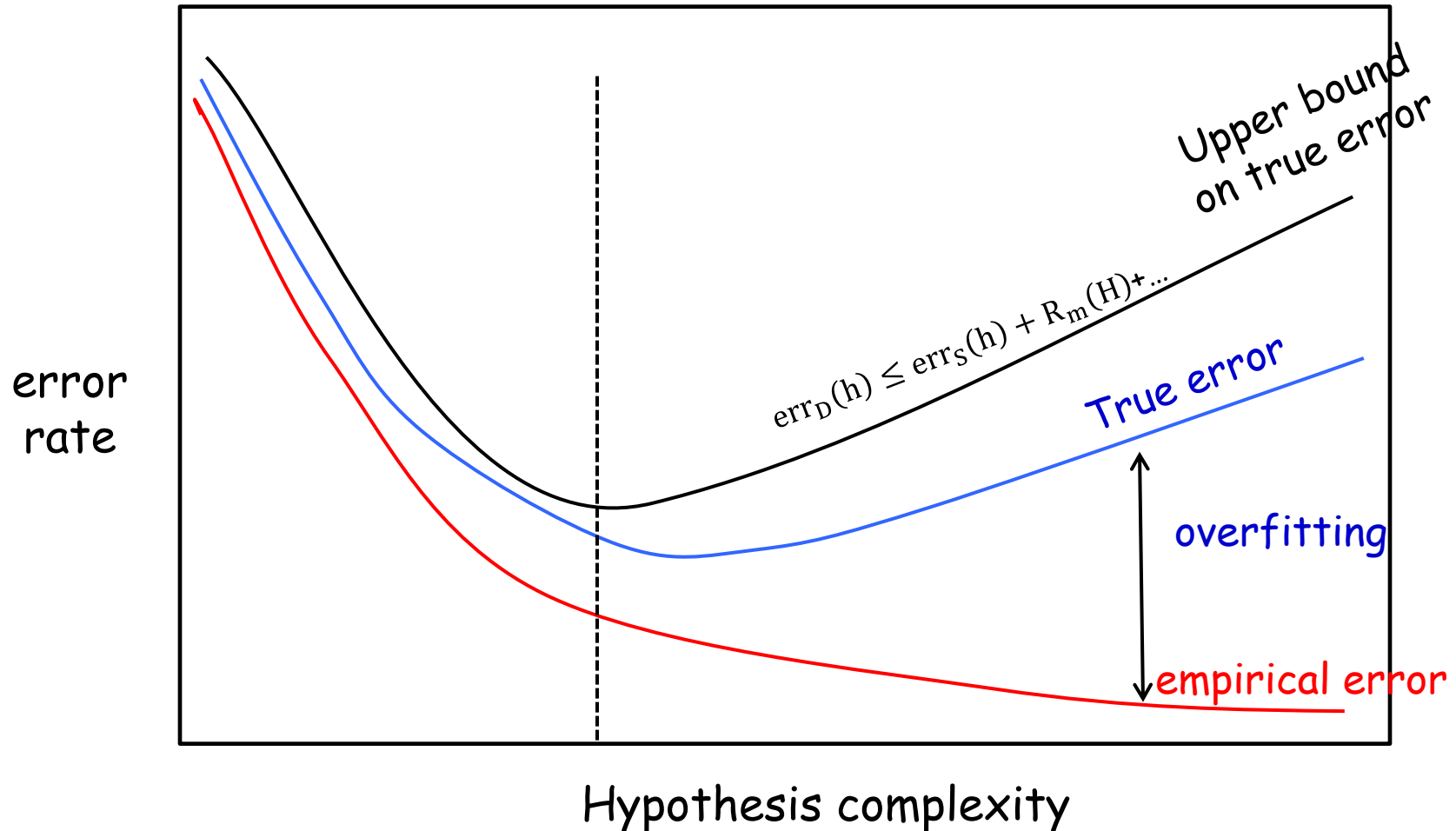# Structural Risk Minimization (SRM)

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

error rate

Upper bound on true error

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

True error

overfitting

empirical error

Hypothesis complexity

Structural Risk Minimization (SRM)

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

- $\hat{h}_k = \operatorname{argmin}_{h \in H_k}\{\operatorname{err}_S(h)\}$

  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).** $H_1 \subseteq H_2 \subseteq \cdots \subseteq H_i \subseteq \ldots$

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

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

    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.