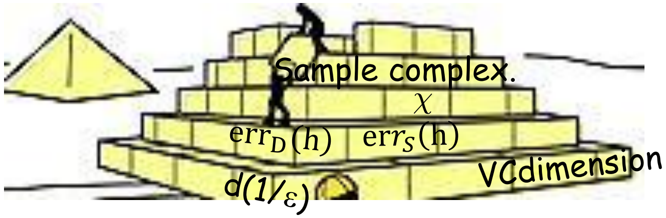
Machine Learning Theory II

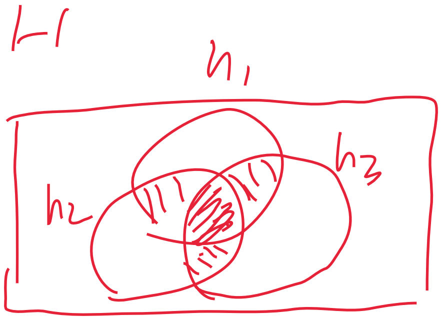
Maria-Florina (Nina) Balcan

February 11th, 2015

Today's focus

1. SLT for infinite H
2. Model selection





Machine Learning Theory

1. Binary classification
2. Noise-free labeling

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

?

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

Finite H

no free lunch

PAC-learn

VS.

Realizable $c^* \in H$

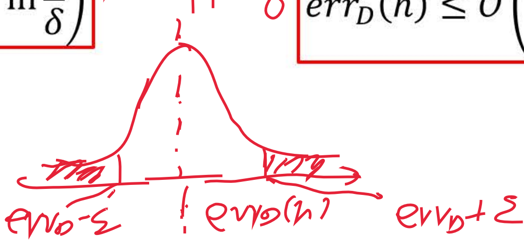
$$err_D(h) \le \frac{1}{m}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

over estimate

Agnostic $c^* \notin H$

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

Hoeffding



Infinite H

Realizable $c^* \in H$

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

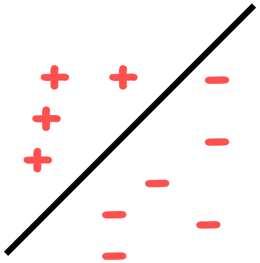
Agnostic $c^* \notin H$

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

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

What if H is infinite?

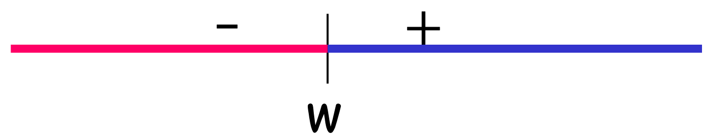
E.g., linear separators in R^d



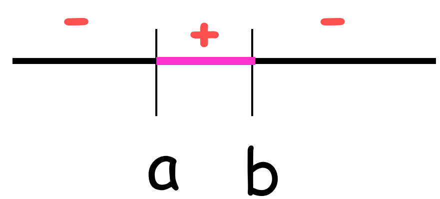
$h(x) = \text{sign}(w^T x)$

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

E.g., thresholds on the real line



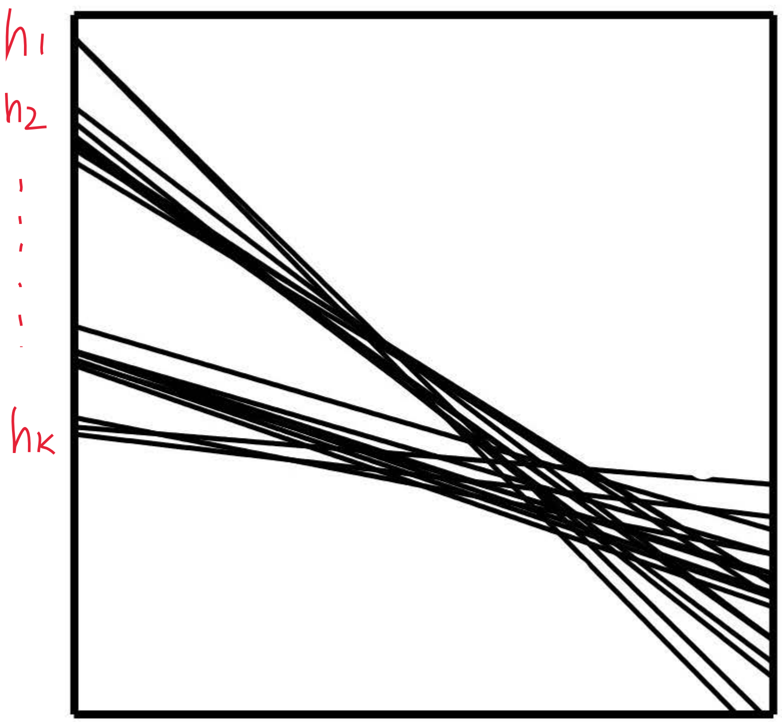
E.g., intervals on the real line



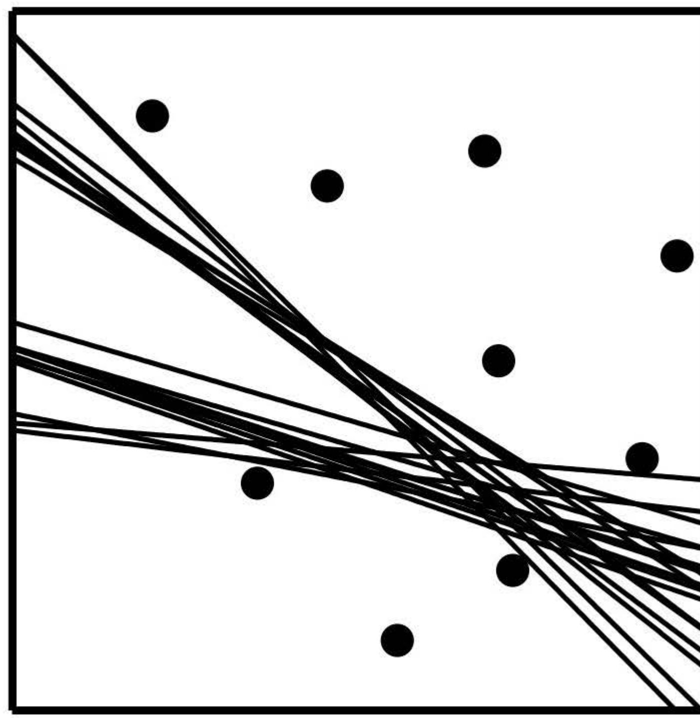
$h(x) = \begin{cases} 1, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$

An Effective Number of Hypotheses

|H| only measures the maximum possible diversity of H



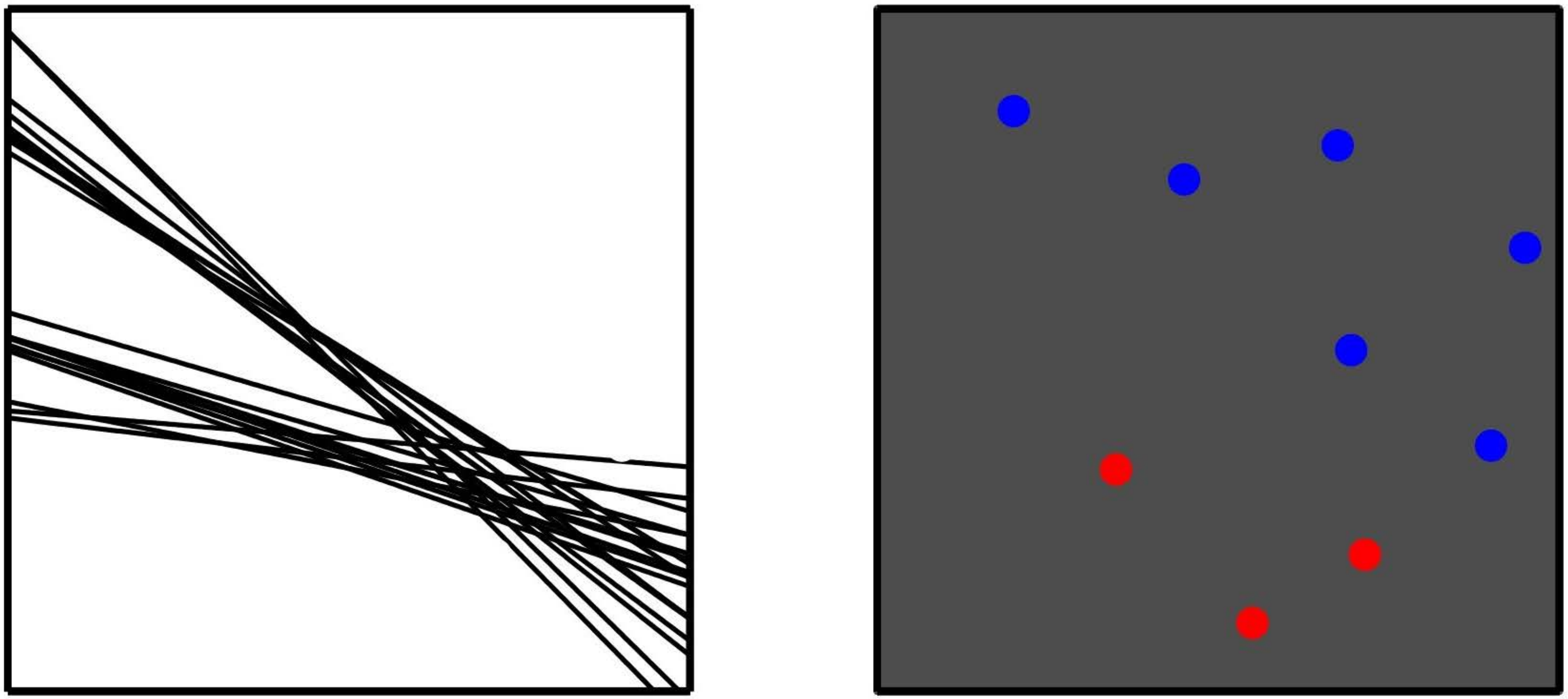
$\mathcal{H}$



$\mathcal{H}$ 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

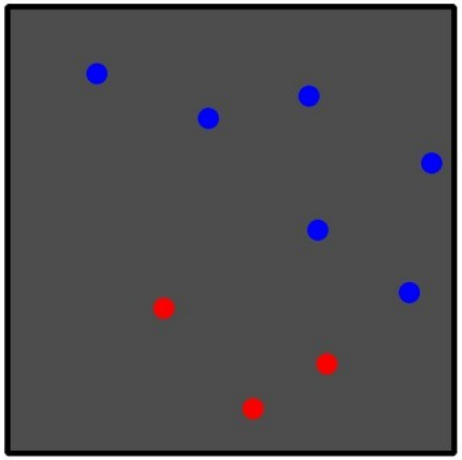
An Effective Number of Hypotheses

|H| only measures the maximum possible diversity of H

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

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

A dichotomy of S



dichotomy

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

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

$c^* \in H$

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

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

Theorem

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

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

$S = \{x_i\}_{i=1}^{m}$

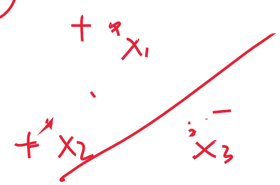
- $H[S]$ – the set of splittings of dataset S using concepts from H.
- $H[m]$ - max number of ways to split m points using concepts in H

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

$$H(S) = \{ \underbrace{(h(x_1), \ldots, h(x_m))}_{\text{dichotomy}} \mid h \in H \}, \quad |H(S)| \le 2^m < |H|$$

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

H : linear separator

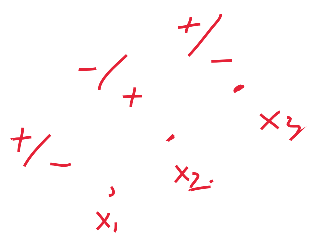
① 

$2^3 = 8$

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

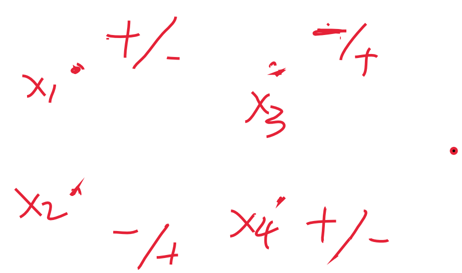
$|H(S)| = 2^3 = 8$

$H[3] = 8$



$|H(S)| = 8 - 2 = 6$

$H[3] = 8$

② 

$2^4 = 16$

$H(S) = \{ \underbrace{++++, +++-, \ldots, ----}_{16-2=14} \}$

$H[4] = 14$

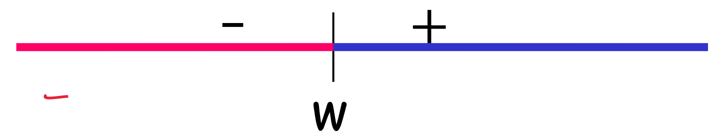
Effective number of hypotheses

- 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

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

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

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



−	+	+	−
−	−	−	+

$|H[S]| = 5$

$2^4 = 16$

$H(S) = \{\underbrace{(- \cdot \cdot), \ldots, (\cdot - -)}_{5}\}$

−	−	−	−
−	−	+	+
−	+	+	+
+	+	+	+

$\max_{S} |H(S)| = \binom{5}{1} = 5$

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

$H[m] = m+1$

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

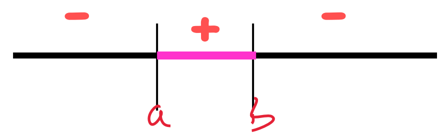
Effective number of hypotheses

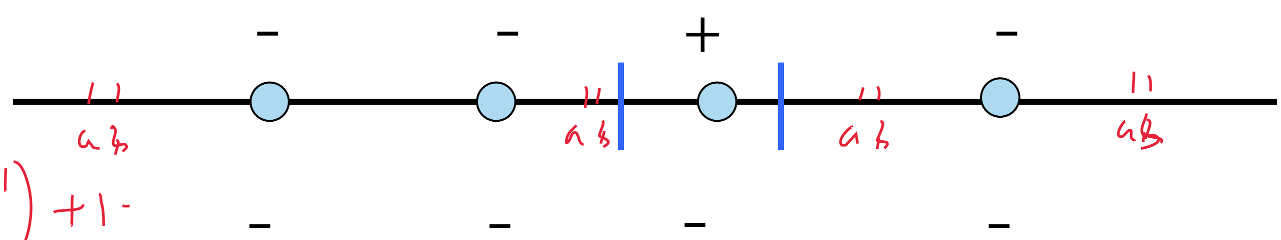
- H[S] - the set of splittings of dataset S using concepts from H.
- H[m] - max number of ways to split m points using concepts in H

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

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

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





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

$H[4] = 11$

$H[m] \le 2^m$

$\binom{m+1}{2} + 1 -$

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

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

$\forall S \subseteq X$

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

specific

Sample Complexity: Infinite Hypothesis Spaces

Realizable Case

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

probability $1-\delta$, $err_D(h) \leq \varepsilon$ approximately PAC-learn

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

$P(\text{bad}) < |H| e^{-\varepsilon m} \leq \delta$

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

$H[m] \leq 2^m$

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

then with probab. $1-\delta$, all $h \in H$ with $\overset{A}{err_D(h) \geq \varepsilon}$ $\Rightarrow$ have $\overset{B}{err_S(h) > 0}$.

$\neg B \Rightarrow \neg A$ $\quad$ $err_S(h)=0 \Rightarrow err_D(h) < \varepsilon$

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

- 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

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

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

$|H| \to H[m] \to VC$

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

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

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

$H[m] = O(m^d)$

Theorem

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

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

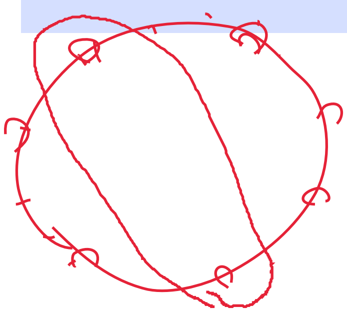
A set of points S is shattered by H is there are hypotheses in H that split S in all of the $2^{|S|}$ possible ways, all possible ways of classifying points in S are achievable using concepts in H.

Definition: VC-dimension (Vapnik-Chervonenkis dimension)

$|S| = m$

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 $\mathrm{VCdim}(\mathrm{H}) = \infty$ H: convex set



$$\mathrm{VCdim}(H) = \max\{|S| : H \text{ shatters } S\}$$
$$= \max\{m : H[m] = 2^m\}$$

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

To show that VC-dimension is d:

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

- there exists a set of d points that can be shattered

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

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

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

d

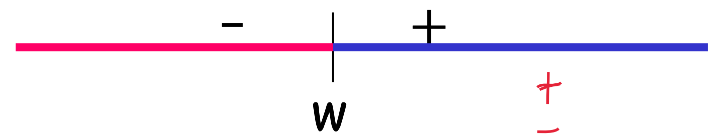
Fact: If H is finite, then $\mathrm{VCdim}(H) \leq \log(|H|)$.

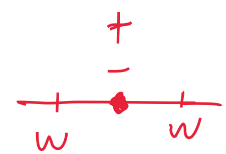
$2^d \leq |H|$

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



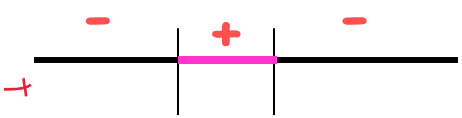


① VCdim(H) ≥ 1

VCdim(H) = 1

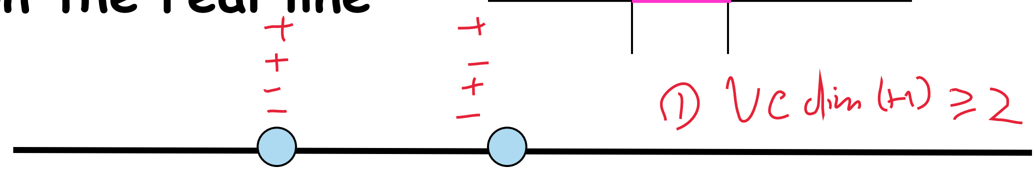
② VCdim(H) < 2

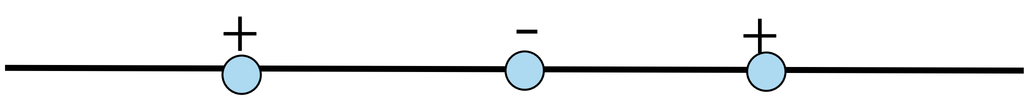
E.g., H= Intervals on the real line



① VCdim(H) ≥ 2

VCdim(H) = 2



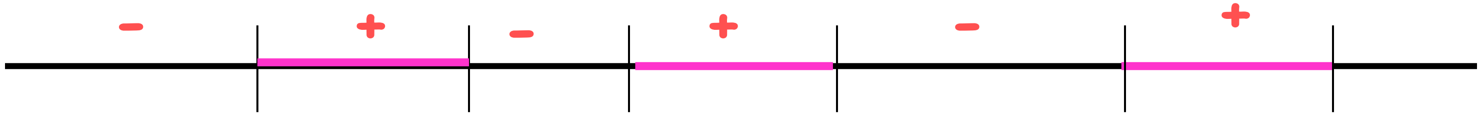


② VCdim(H) < 3

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



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

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

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

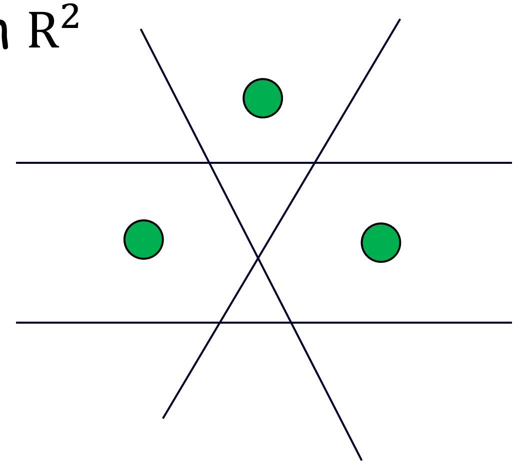
Shattering, VC-dimension

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

① $\mathrm{VCdim(H)} \geq 3$

② $\mathrm{VC\ dim}(H) < 4$

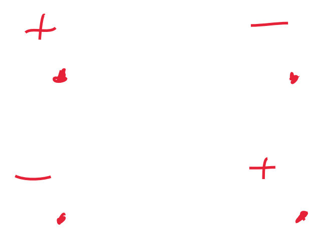
$\Rightarrow \mathrm{VC\ dim}(H) = 3$



$2^3 = 8$

+ + +
+ + −
⋮
− − −

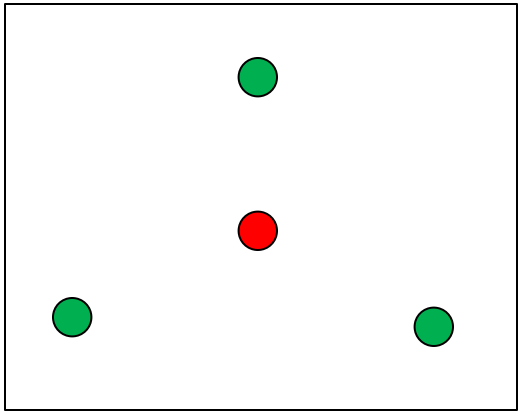
8 dichotomy



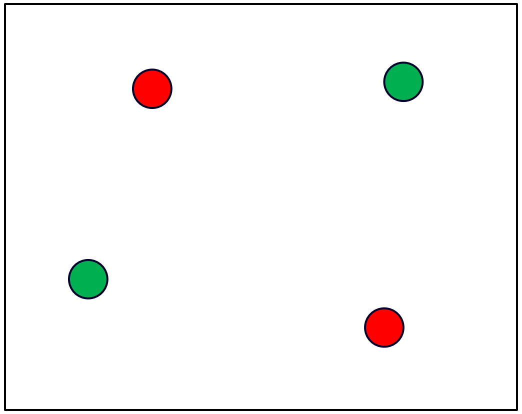
Shattering, VC-dimension

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

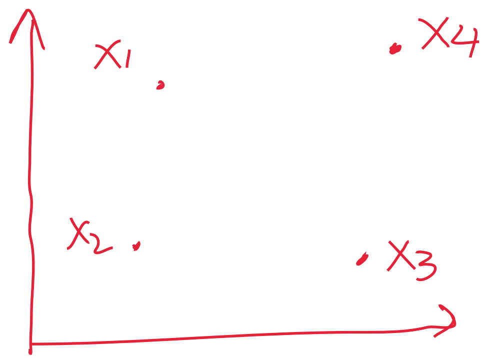
$X^T = (x_1, x_2)$

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

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

$\boxed{H: \text{quadratic separator}}$

(#params = 6)



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

① $H(S) \overset{?}{=}$

② $H[m] \overset{?}{=}$

③ $VCdim(H) \geq 4$?

Sauer's Lemma

Sauer's Lemma:

Let d = VCdim(H)

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

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

Hint: try proving it for intervals…

Sample Complexity: Infinite Hypothesis Spaces

Realizable Case

$c^* \in H$

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

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

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

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

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

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 $err_S(h) > 0$.

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

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 R^d n

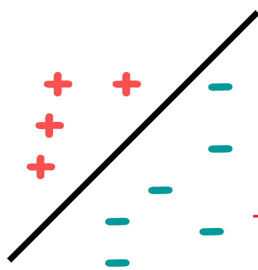
$$m = O\left(\frac{1}{\varepsilon}\left[d\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

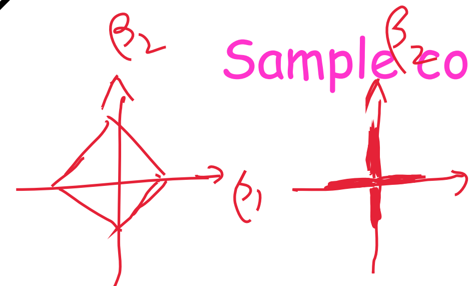
VCdim(H)= n d+1

Sample complexity linear in d

$f(x) = x^T\beta + \beta_0$

$(x \in \mathbb{R}^n)$





$\|\beta\|_0 = \sum_{j=1}^{n} \mathbb{I}(\beta_j \neq 0)$

$= \sum_{j=1}^{n} |\beta_j|$

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

ridge: $\|\beta\|_2^2$, lasso: $\|\beta\|_1$

Practical rule of thumb: VCdim(H) ~ #free parameters of H

What if $c^* \notin H$?

Sample Complexity: Uniform Convergence

Agnostic Case

$|H| \to H[m] \to VCdim(H)$

Empirical Risk Minimization (ERM)

- Input: S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$
- Output: Find h in H with smallest $\mathrm{err}_S(h)$

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).

$1/\epsilon^2$ 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 $|err_D(h) - err_S(h)| < \varepsilon$.

$err_S(h) - \varepsilon \leq \boxed{err_D(h) \leq err_S(h) + \varepsilon}$

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 $h \in H$:

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

Sample Complexity: Infinite Hypothesis Spaces

Agnostic Case

$H : \forall h$

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

Theorem

$m \gtrsim 10 \; VCdim(H)$

$\simeq$ # paras of H

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

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

H: linear separator in $\mathbb{R}^n$

$n=2$: $\quad n = 2000$

$m \gtrsim 30 \quad m \gtrsim 20000$

$VCdim(H) = d$

2) Statistical Learning Theory style:

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

Vapnik

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

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

$\frac{d}{m}\uparrow$

$\frac{d}{m}\downarrow$

VCdimension Generalization Bounds

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

$\forall h \in H$

VC bounds: distribution independent bounds

C D

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

  [nearly tight in the WC over choice of D]

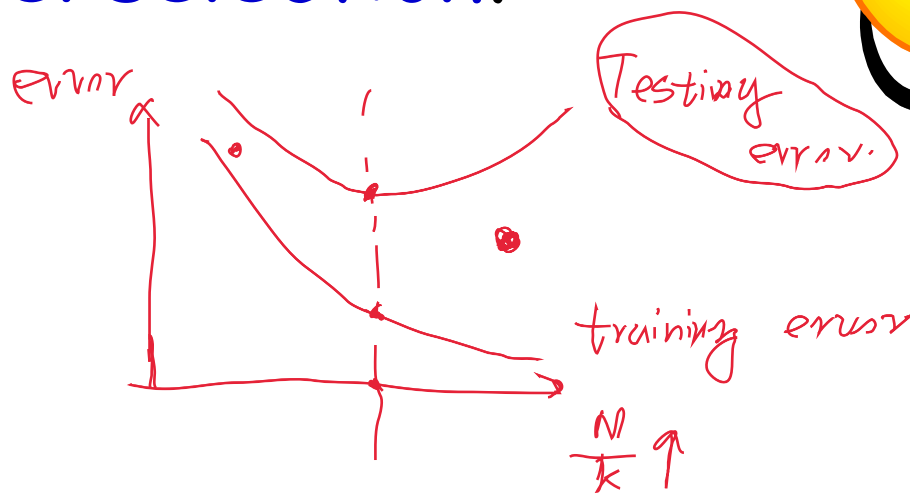
- Might be very loose specific distr. that are more benign than the worst case....

- Hold only for binary classification; we want bounds for fns approximation in general (e.g., multiclass classification and regression).

$|\mathrm{err}_D(h) - \mathrm{err}_S(h)| < \varepsilon$

Can we use our bounds for **model selection**?

knn



$err_D(h)$

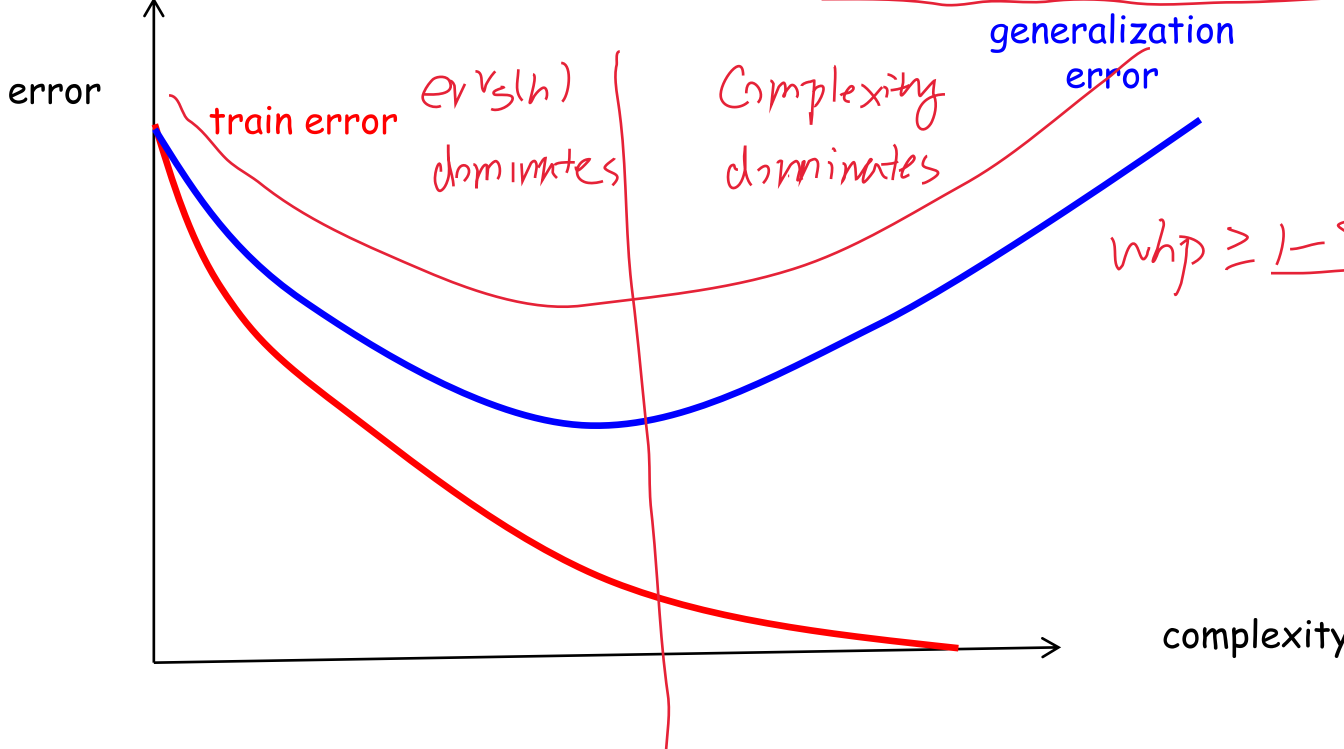
True Error, Training Error, Overfitting

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

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

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

model complexity penalty



$\frac{d}{m} \uparrow$

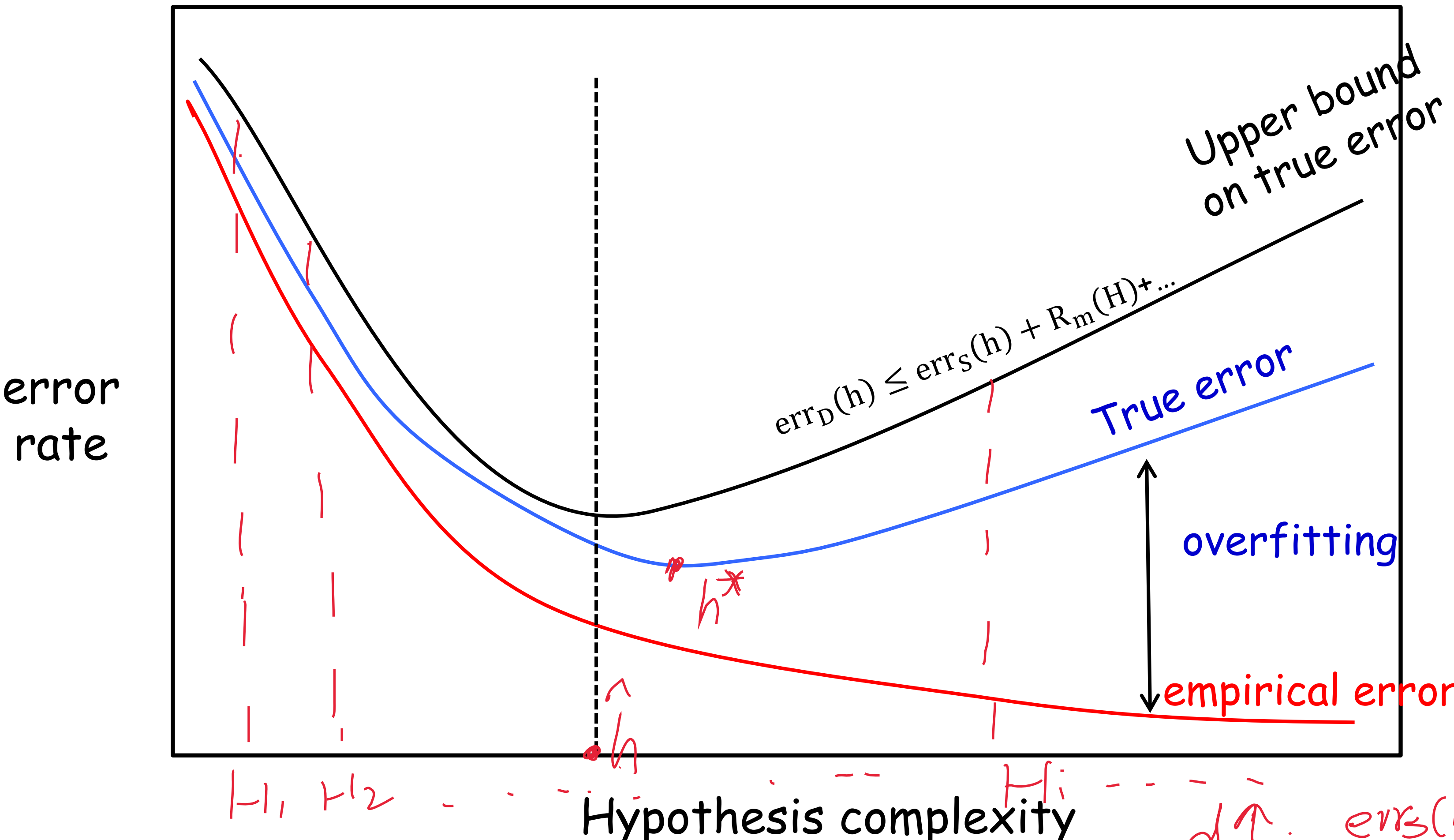
whp $\ge 1-\delta$

Structural Risk Minimization (SRM)

$d_1 < d_2 < d_3 < \cdots < d_i < \cdots$

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

$c^* \notin H$



$d\uparrow$: $err_S(h) \approx 0$

$d\downarrow$: $err_D(h) \approx err_S(h)$

What happens if we increase m?

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

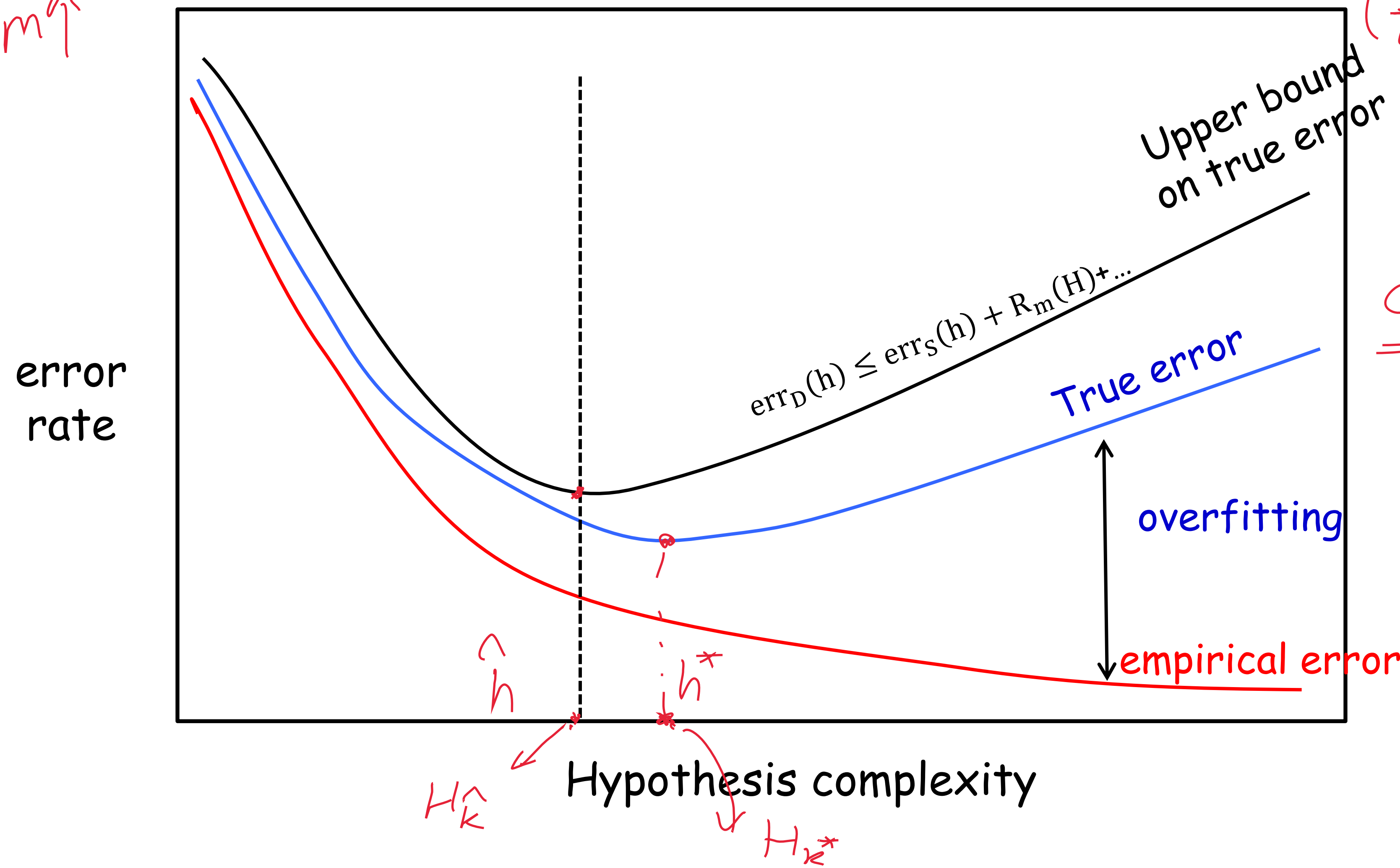
Structural Risk Minimization (SRM)

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

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

$\left(\frac{\ln m}{m}\right)$

$m\uparrow$



$c^* \notin H$

Structural Risk Minimization (SRM)

$d_1 < d_2 < d_3 < \cdots < d_i < \cdots$

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

$\min_{h \in H} err_D(h)$: expected risk minimization

$\min_{h \in H} err_S(h)$: empirical risk

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

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

$\min_{h \in H} err_S(h) + Complexity(H)$

- $\hat{k} = \mathrm{argmin}_{k \geq 1}\{\mathrm{err}_S(\hat{h}_k) + \mathrm{complexity}(H_k)\}$

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

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

Proof:

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

  $|err_D(h) - err_S(h)| < \varepsilon$

- Whp, $\mathrm{err}_S(h^*) \leq \mathrm{err}_D(h^*) + \mathrm{complexity}(H_{k^*})$.

  $err_S(h) - \varepsilon < err_D(h) < err_S(h) + \varepsilon$

Techniques to Handle Overfitting

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

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

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

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

- Feature selection.

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.