Sample Complexity for Function Approximation. Model Selection.

Maria-Florina (Nina) Balcan

03/10/2018

Two Core Aspects of Machine Learning

Algorithm Design. How to optimize?

Computation

Automatically generate rules that do well on observed data.

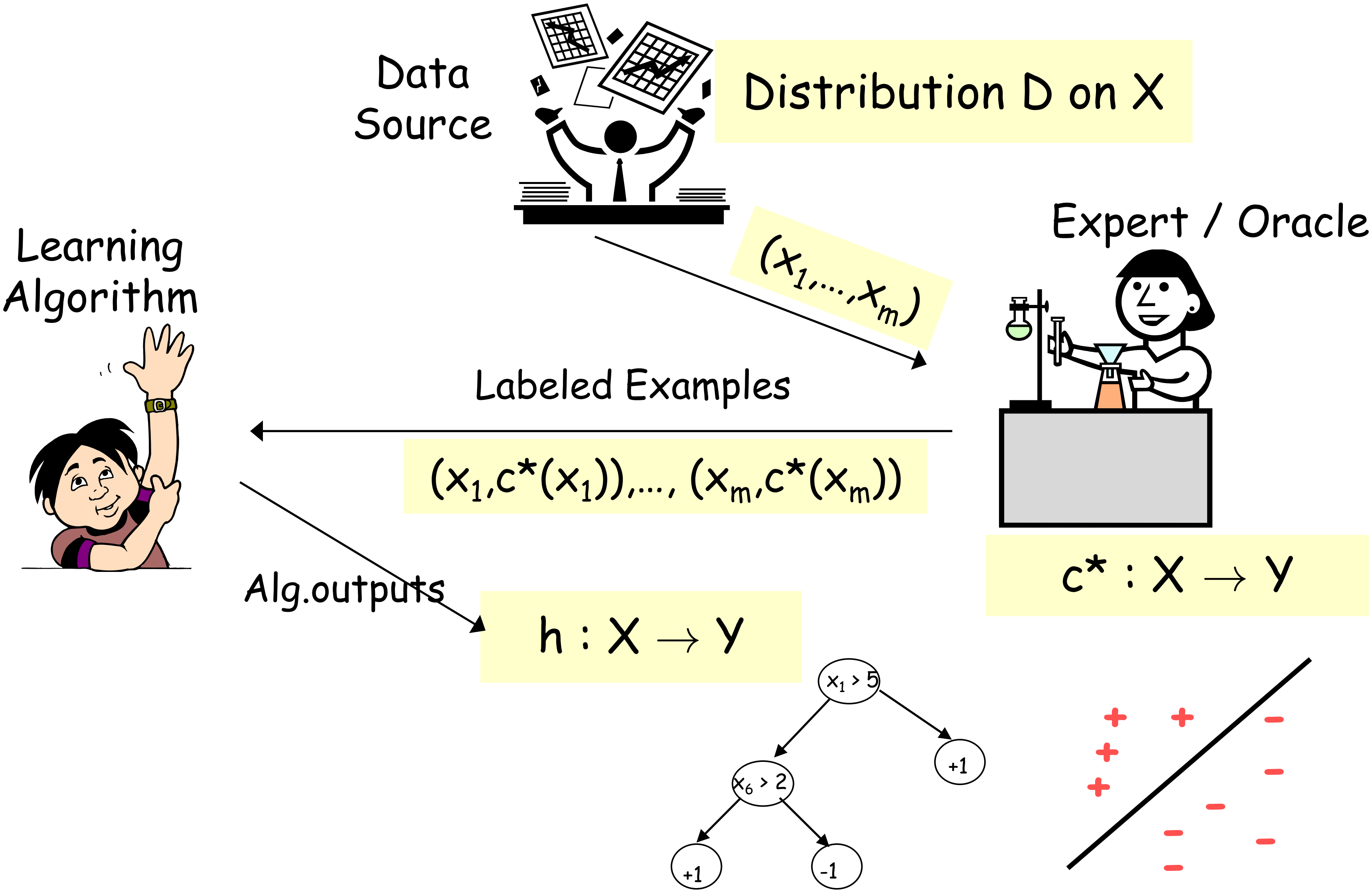
- E.g.: logistic regression, SVM, Adaboost, etc.

Confidence Bounds, Generalization

(Labeled) Data

Confidence for rule effectiveness on future data.

PAC/SLT models for Supervised Classification

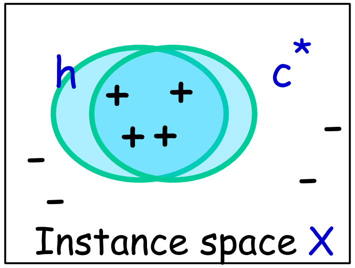


PAC/SLT models for Supervised Learning

- X – feature/instance space; distribution D over X

  e.g., $X = R^d$ or $X = \{0,1\}^d$

- Algo sees training sample S: $(x_1,c^*(x_1)),\ldots, (x_m,c^*(x_m))$, x_i i.i.d. from D
  - labeled examples - drawn i.i.d. from D and labeled by target c^*
  - labels $\in \{-1,1\}$ - binary classification



- Algo does optimization over S, find hypothesis h.
- Goal: h has small error over D.

$$err_D(h) = \Pr_{x\sim D}(h(x) \neq c^*(x))$$

- Fix hypothesis space H [whose complexity is not too large]
  - Realizable: $c^* \in H$.
  - Agnostic: c^* "close to" H.

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

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

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

Prob. over different samples of m training examples

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

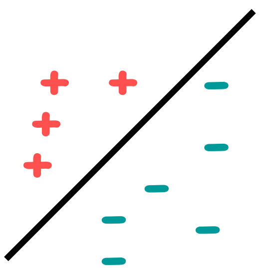
Theorem

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

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

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.

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

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

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], but get for something stronger.

2) Statistical Learning Theory style:

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

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

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

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

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).

Rademacher Complex: Binary classification

Fact: $H = \{h: X \to Y\}$ hyp. space (e.g., lin. sep) F= L(H), d=VCdim(H):

$$R_S(F) \le \sqrt{\frac{\ln(2|H[S]|)}{m}}$$

So, by Sauer's lemma, $R_S(F) \le \sqrt{\frac{2d\ln\left(\frac{em}{d}\right)}{m}}$

Theorem: For any H, any distr. D, w.h.p. $\ge 1-\delta$ all $h \in H$ satisfy:

$$err_D(h) \le err_S(h) + R_m(H) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

$$err_D(h) \le err_S(h) + \sqrt{\frac{2d\ln\left(\frac{em}{d}\right)}{m}} + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

generalization bound

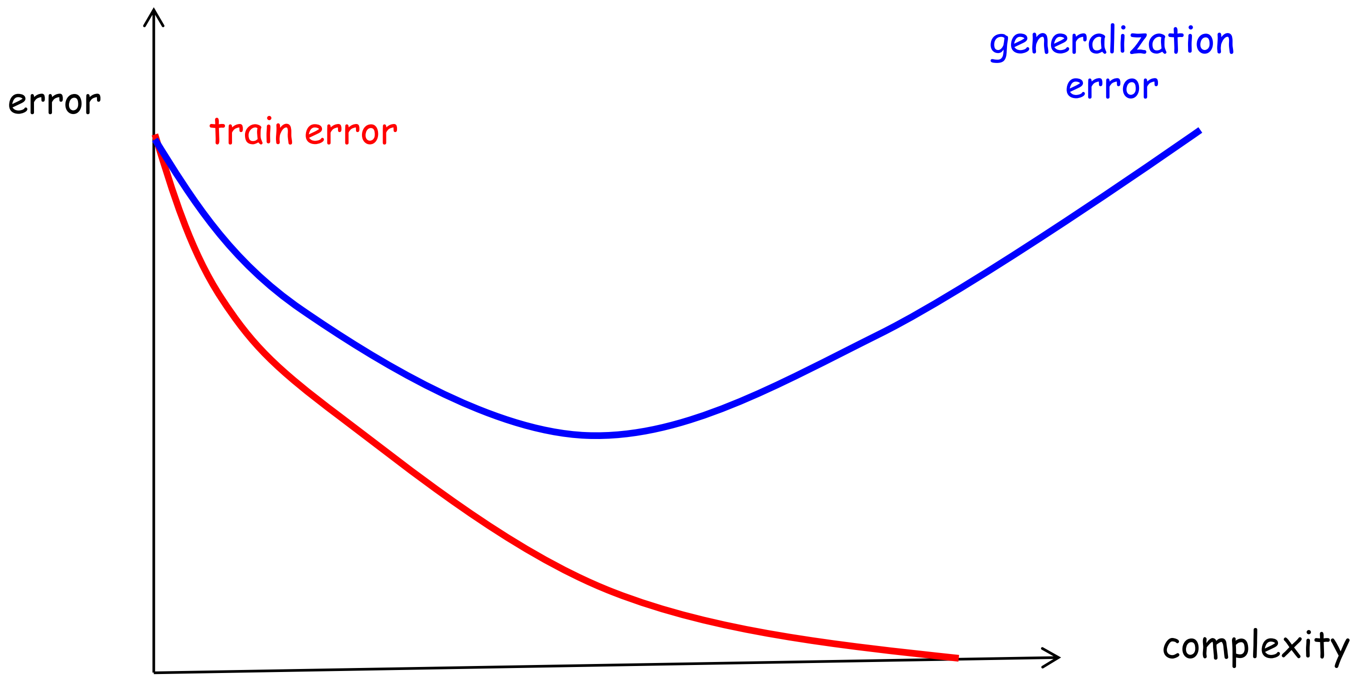
Many more uses!!! Margin bounds for SVM, boosting, regression bounds, etc.

Can we use our bounds for model selection?

True Error, Training Error, Overfitting

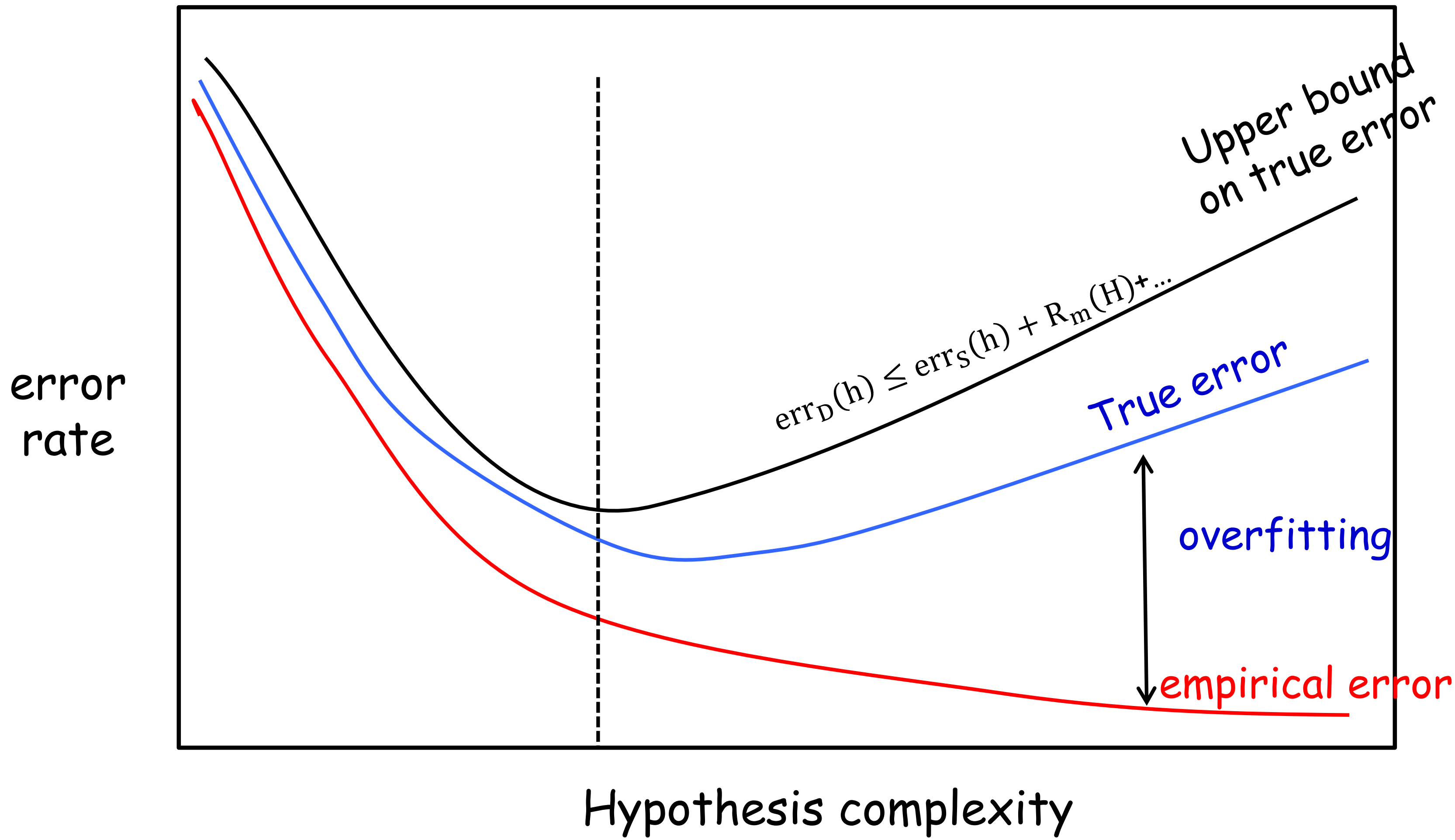
Model selection: trade-off between decreasing training error and keeping H simple.

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



Structural Risk Minimization (SRM)

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

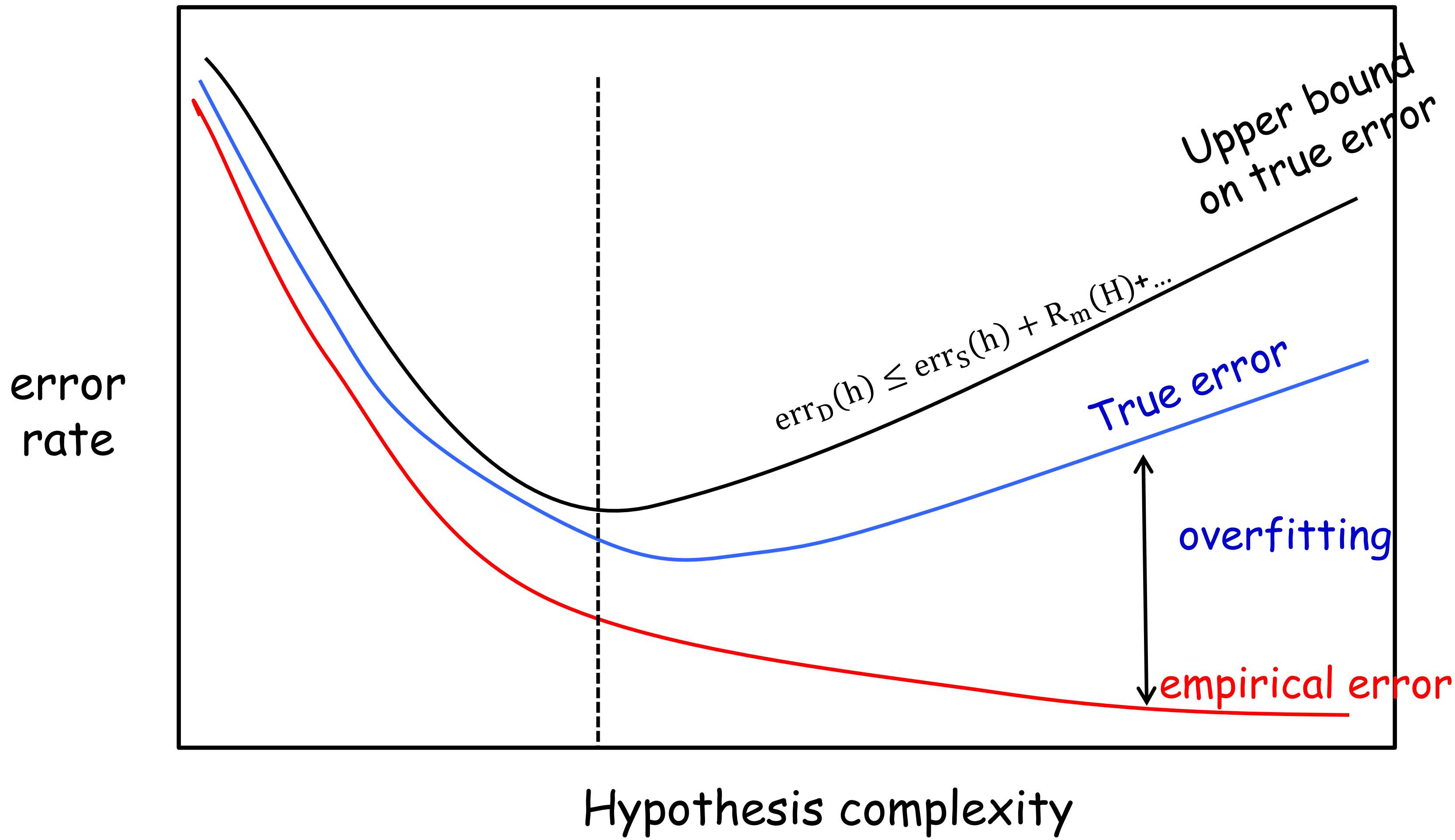


What happens if we increase m?

Black curve will stay close to the red curve for longer, everything shift 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)

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

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

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

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

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

Proof:

- We chose $\hat{\mathrm{h}}$ s.t. $\mathrm{err_s}(\hat{\mathrm{h}}) + \mathrm{complexity}(\mathrm{H}_{\hat{\mathrm{k}}}) \leq \mathrm{err_S(h^*)} + \mathrm{complexity(H_{k^*})}$.
- Whp, $\mathrm{err_D}(\hat{\mathrm{h}}) \leq \mathrm{err_s}(\hat{\mathrm{h}}) + \mathrm{complexity}(\mathrm{H}_{\hat{\mathrm{k}}})$.
- Whp, $\mathrm{err_S(h^*)} \leq \mathrm{err_D(h^*)} + \mathrm{complexity(H_{k^*})}$.

Techniques to Handle Overfitting

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

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

  - 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

What you should know

- Notion of sample complexity.
- Understand reasoning behind the simple sample complexity bound for finite H [exam question!].
- Shattering, VC dimension as measure of complexity, Sauer's lemma, form of the VC bounds (upper and lower bounds).
- Rademacher Complexity.
- Model Selection, Structural Risk Minimization.