Generalization, Overfitting, and Model Selection

Sample Complexity Results for Supervised Classification

Maria-Florina (Nina) Balcan

03/05/2018

Reminders

- Midterm Exam
  - Wed, March 7th
- Recitation
  - Tue, March 6th at 6:30-7:30pm
- Homework 3, due today at 5:00 PM

Midterm Exam

- **In-class exam on Wed, March 7th**
  - 4 problems
  - Format of questions:
    - Multiple choice
    - True / False (with justification)
    - Very short derivations
    - Short answers
    - Interpreting figures
  - No electronic devices
  - You are allowed to **bring** one $8\frac{1}{2}$ x 11 sheet of notes (front and back)

Midterm Exam

- **How to Prepare**
  - Attend the midterm recitation session: Thu, Oct. 6th at 6:00pm
  - Review this year's homework problems
  - Review prior year's exams and solutions (we'll post them)

Generalization, Overfitting, and Model Selection

Two Core Aspects of Machine Learning

Algorithm Design. How to optimize?

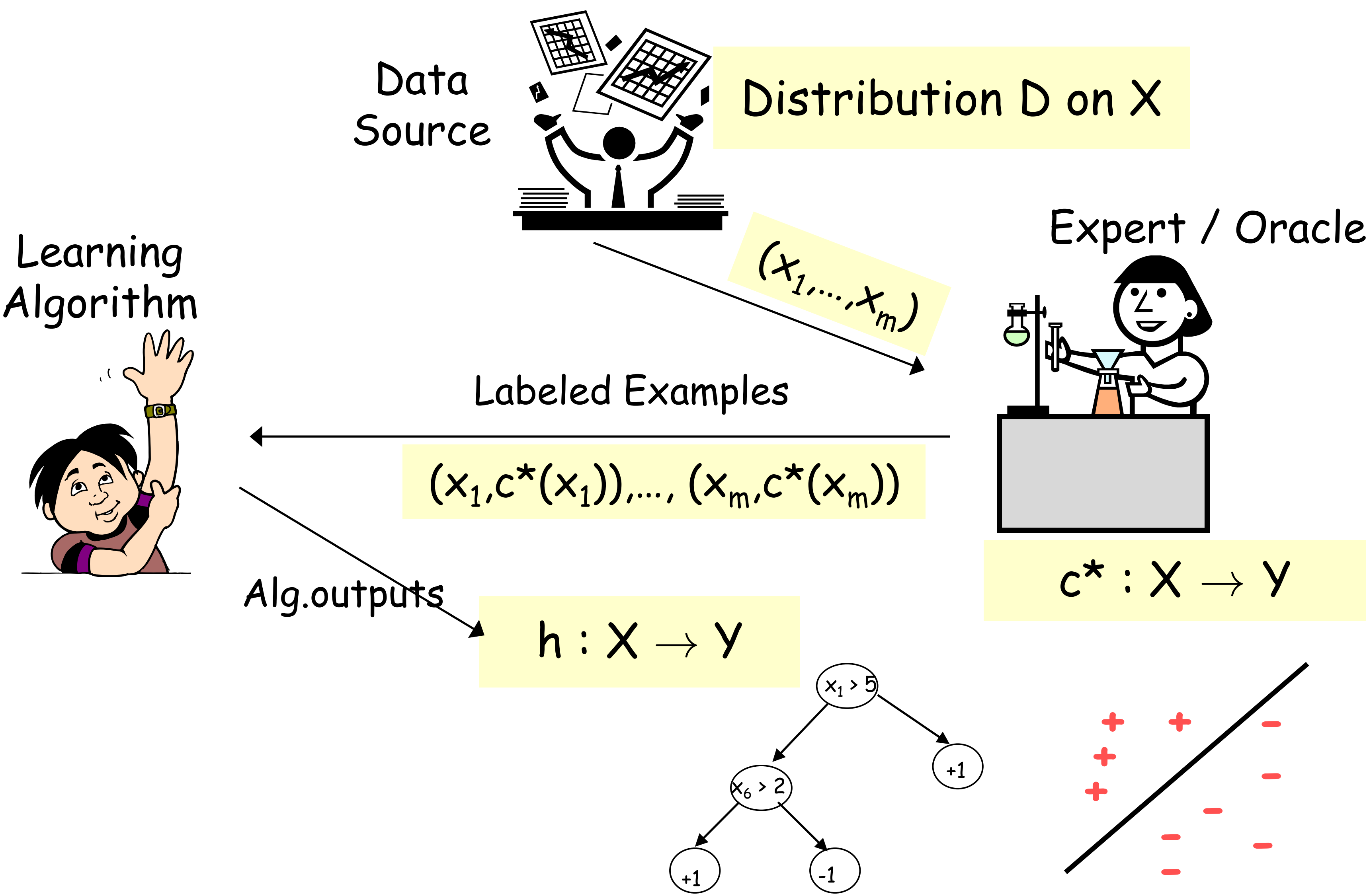
Automatically generate rules that do well on observed data.

Confidence Bounds, Generalization

Confidence for rule effectiveness on future data.

- Our focus so far has been on Algorithm Design.
- In this module, Generalization Guarantees (not overfiting) – they apply to all algorithms we talk about throughout the course.

PAC/SLT models for Supervised Learning

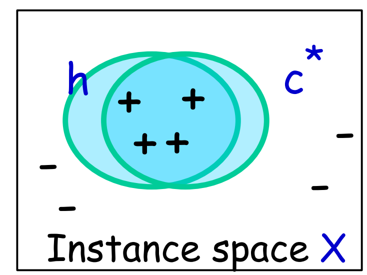


PAC/SLT models for Supervised Learning

- 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.

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

Bias: fix hypothesis space H [whose complexity is not too large]

- Realizable: $c^* \in H$.
- Agnostic: c^* "close to" H.

PAC/SLT models for Supervised Learning

- Algo sees training sample S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$, x_i i.i.d. from D
- Does optimization over S, find hypothesis $h \in H$.
- Goal: h has small error over D.

True error: $\mathrm{err}_D(h) = \Pr_{x \sim D}(h(x) \neq c^*(x))$

How often $h(x) \neq c^*(x)$ over future instances drawn at random from D

- But, can only measure:

Training error: $\mathrm{err}_S(h) = \frac{1}{m}\sum_i I\left(h(x_i) \neq c^*(x_i)\right)$

How often $h(x) \neq c^*(x)$ over training instances

Sample complexity: bound $err_D(h)$ in terms of $err_S(h)$

Sample Complexity: Finite Hypothesis Spaces

Realizable Case

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 $err_S(h) > 0$.

So, if $c^* \in H$ and can find consistent fns, then only need this many examples to get generalization error $\leq \epsilon$ with prob. $\geq 1-\delta$

Agnostic Case

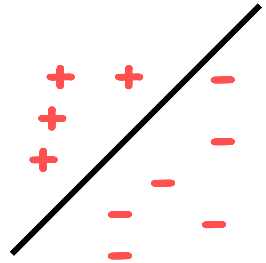
What if there is no perfect h?

Theorem After m examples, with probab. $\geq 1-\delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$, for

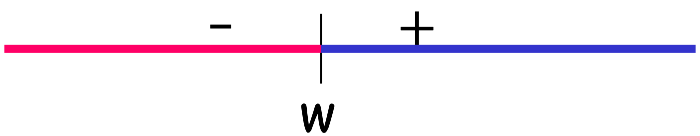
$$m \geq \frac{1}{2\varepsilon^2}\left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right]$$

What if H is infinite?

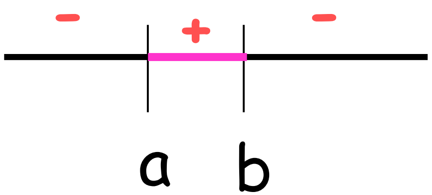
E.g., linear separators in R^d



E.g., thresholds on the real line



E.g., intervals on the real line



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 (labeled in all possible ways) by H.

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

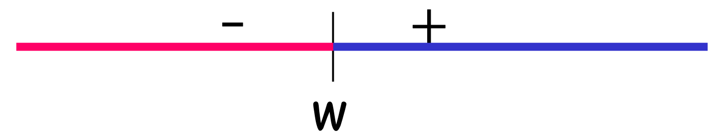
To show that VC-dimension is d:

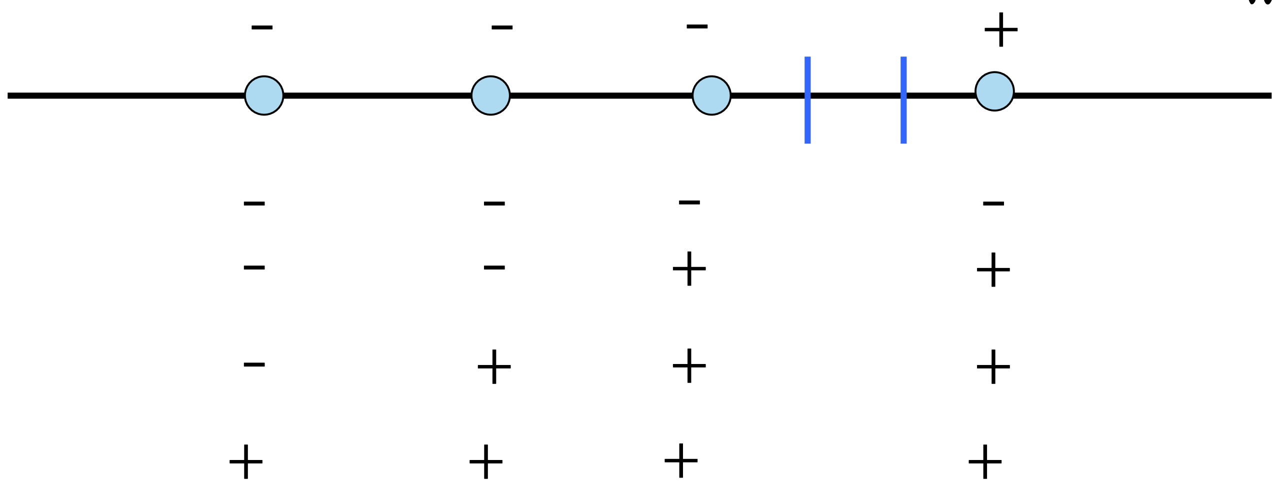
- there exists a set of d points that can be shattered
- there is no set of d+1 points that can be shattered.

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

True complexity of a hypothesis class

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





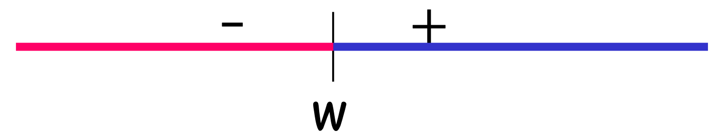
Can label 4 points with thresholds only in 5 ways

In general, can label m points with thresholds only in $m+1 \ll 2^m$

Shattering, VC-dimension

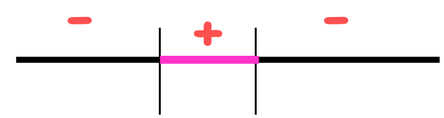
If the VC-dimension is d, that means there exists a set of d points that can be shattered, but there is no set of d+1 points that can be shattered.

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



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

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



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

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

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

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

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

Sample complexity linear in d

Interpretation: if double the number of features, thenwe only need roughly twice the number of samples to do well.

Sample Complexity: Infinite Hypothesis Spaces

Theorem (agnostic case)

$$m \geq \frac{c}{\epsilon^2}\left(VCdim(H) + \log\left(\frac{1}{\delta}\right)\right)$$

labeled examples are sufficient s.t. with probability at least $1-\delta$ for all h in H $|err_D(h) - err_S(h)| \leq \epsilon$

Statistical Learning Theory Style

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

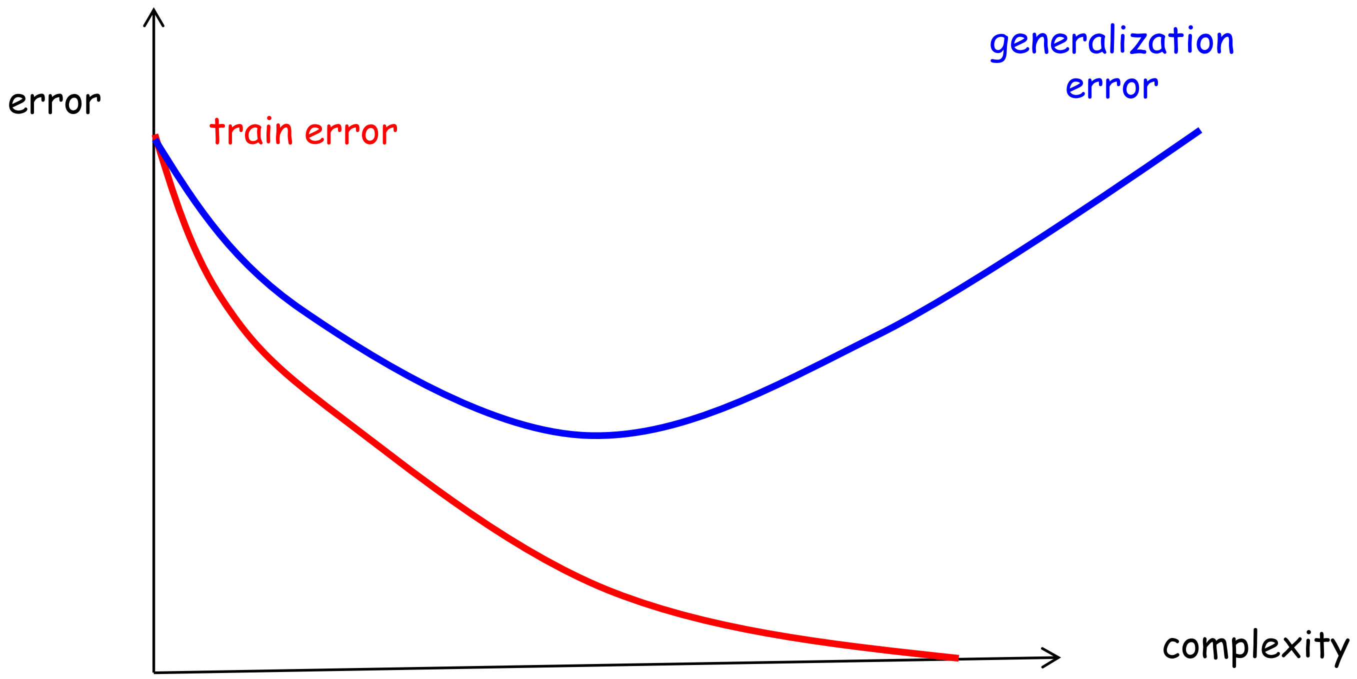
$$\mathrm{err}_{\mathrm{D}}(\mathrm{h}) \leq \mathrm{err}_{\mathrm{S}}(\mathrm{h}) + \sqrt{\frac{1}{2\mathrm{m}}\left(VCdim(H) + \ln\left(\frac{1}{\delta}\right)\right)}.$$

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.

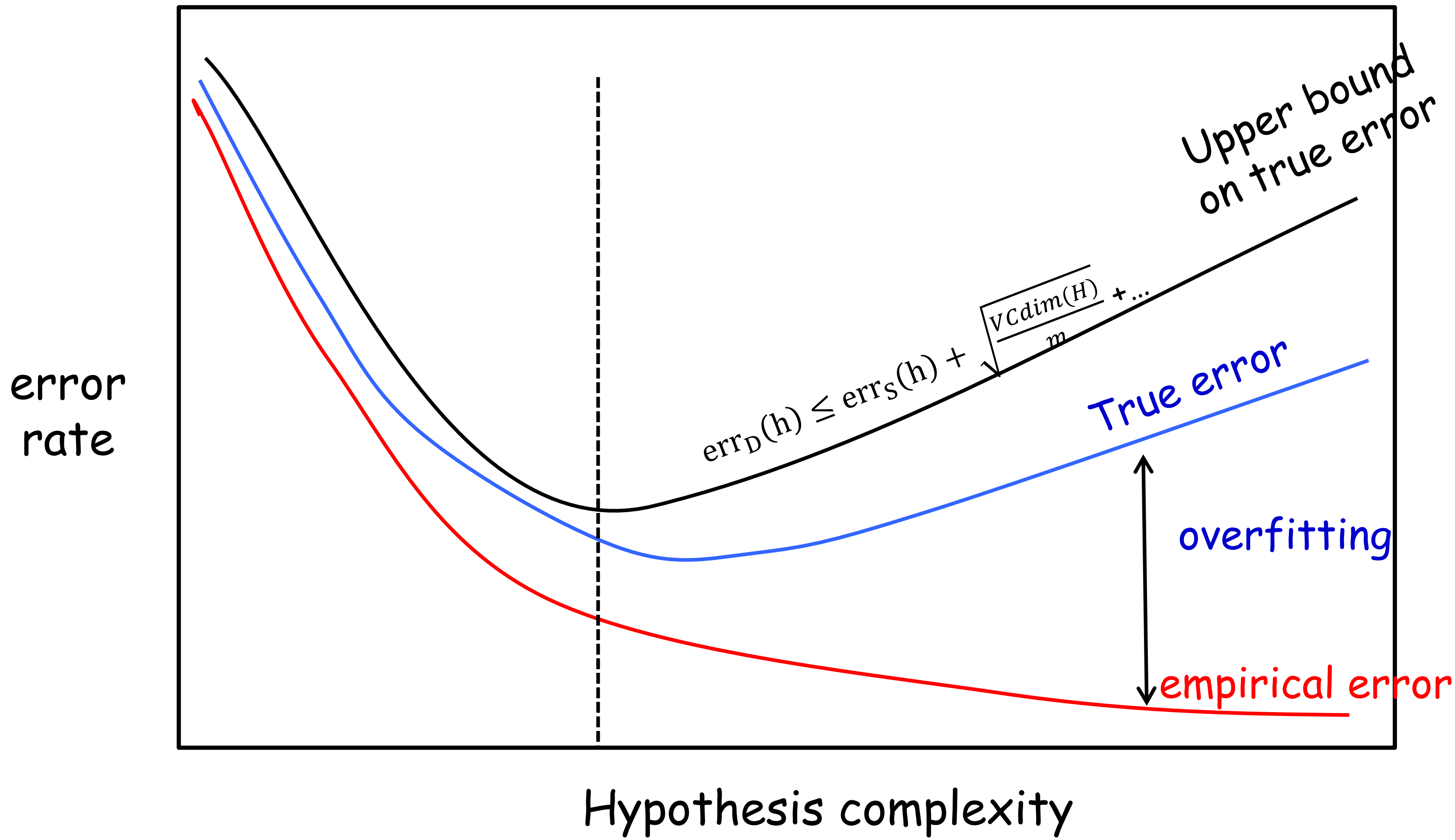
$$\mathrm{err}_D(\mathrm{h}) \leq \mathrm{err}_S(\mathrm{h}) + \sqrt{\frac{VCdim(H)}{m}} + \ldots$$



Structural Risk Minimization (SRM)

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

(E.g., H_i= decision trees of depth i)

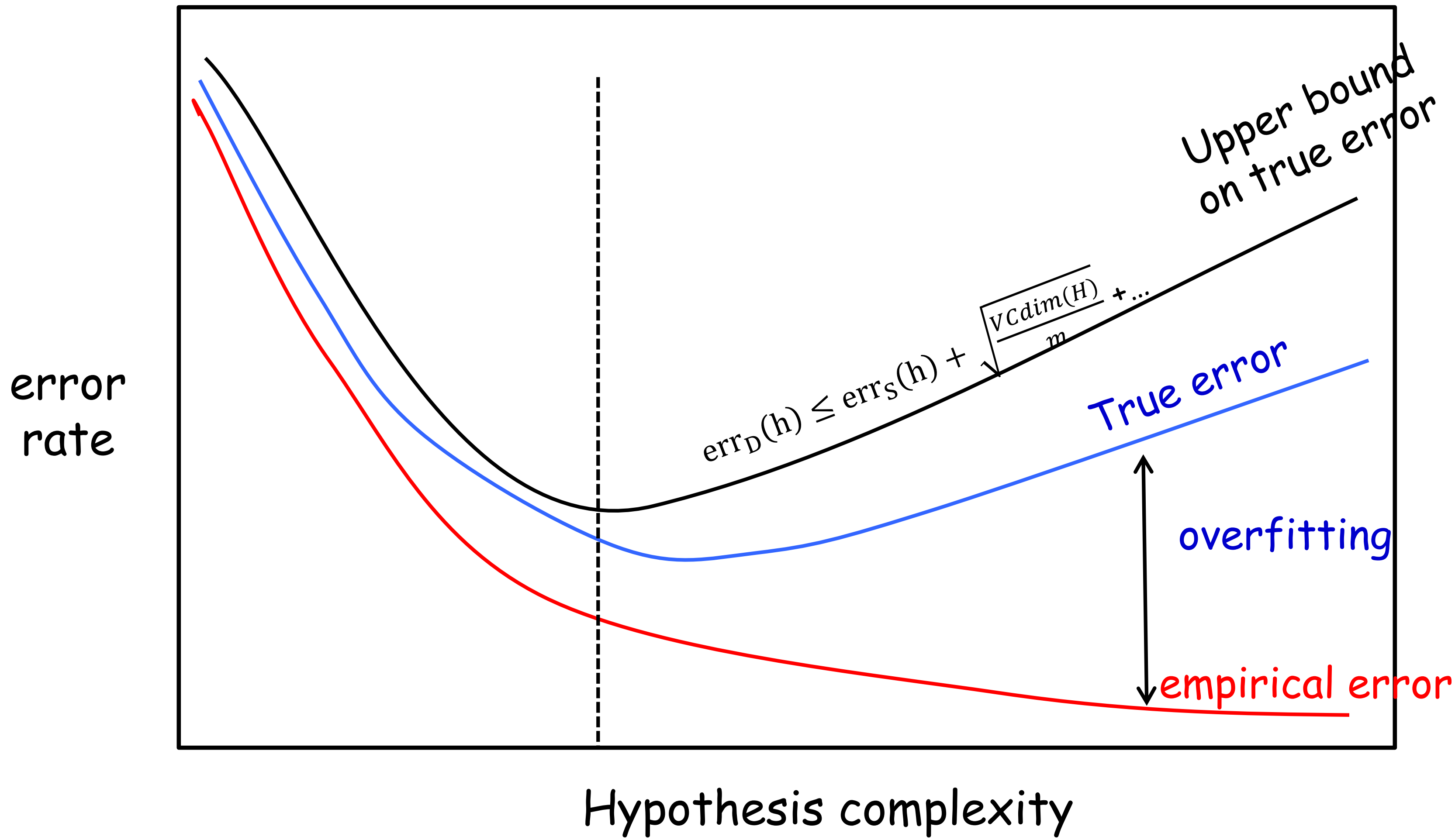


What happens if we increase m?

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)

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

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

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

- **Cross Validation:**
  - Hold out part of the training data and use it as a proxy for the generalization error

What you should know

- The importance of sample complexity in Machine Learning.

- Understand meaning of PAC bounds (what PAC stands for, meaning of parameters ϵ and δ).

- Shattering, VC dimension as measure of complexity, form of the VC bounds.

- Model Selection, Structural Risk Minimization.