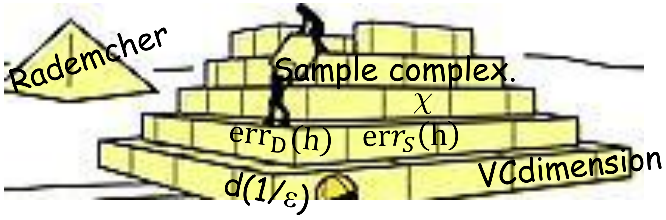
Sample Complexity for Function Approximation. Model Selection.

Maria-Florina (Nina) Balcan

February 16th, 2015



Two Core Aspects of Machine Learning

Algorithm Design. How to optimize? — Computation

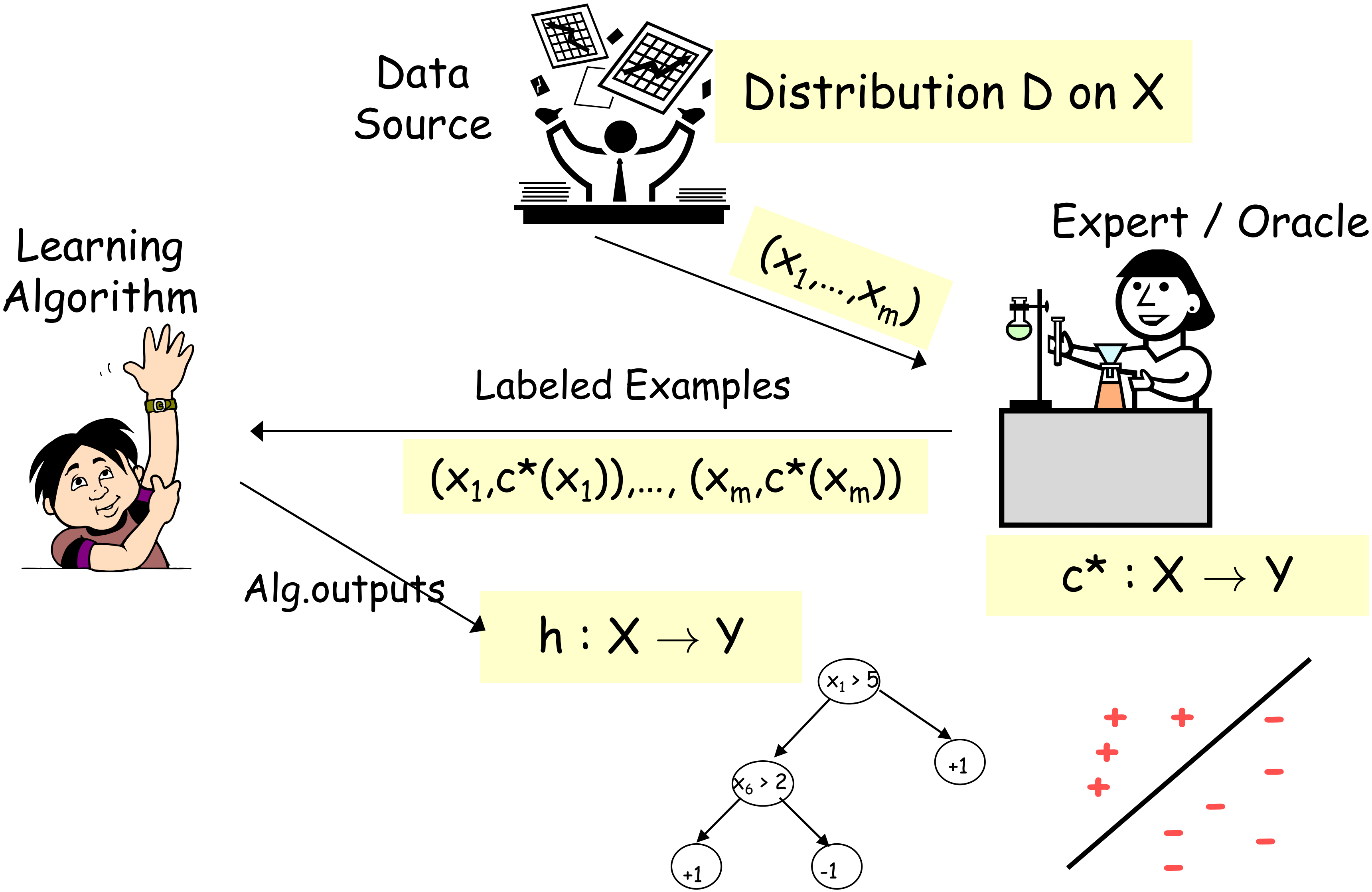
Automatically generate rules that do well on observed data.

- 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



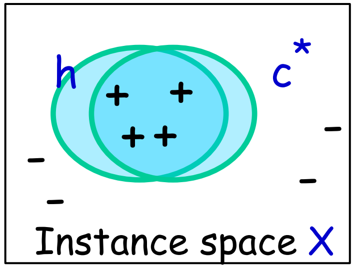
PAC/SLT models for Supervised Learning

- 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

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

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

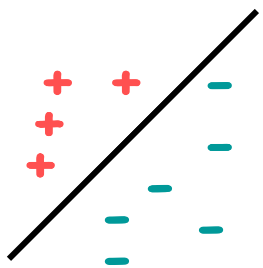
Theorem

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

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

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.

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 $\mathrm{err}_S(h)$

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 $|err_D(h) - err_S(h)| \leq \epsilon$.

Hoeffding bounds

Consider coin of bias p flipped m times.

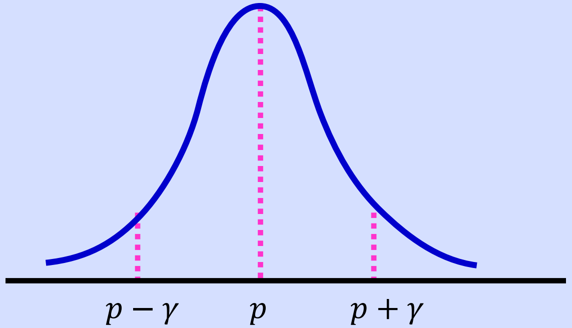
Let N be the observed # heads. Clearly $\mathrm{E}\left[\frac{\mathrm{N}}{\mathrm{m}}\right] = \mathrm{p}$.

[$\mathrm{N} = \mathrm{X}_1 + \mathrm{X}_2 + \ldots + \mathrm{X}_\mathrm{m}$, $\mathrm{X}_\mathrm{i} = 1$ with prob. p, 0 with prob 1-p.]

Hoeffding Inequality

Let $\gamma \in [0,1]$.

$$P\left[\left|\frac{N}{m} - p\right| \geq \gamma\right] \leq e^{-2m\gamma^2}$$



Exponentially decreasing tails

Tail inequality: bound probability mass in tail of distribution (how concentrated is a random variable around its expectation).

Sample Complexity: Finite Hypothesis Spaces Agnostic Case

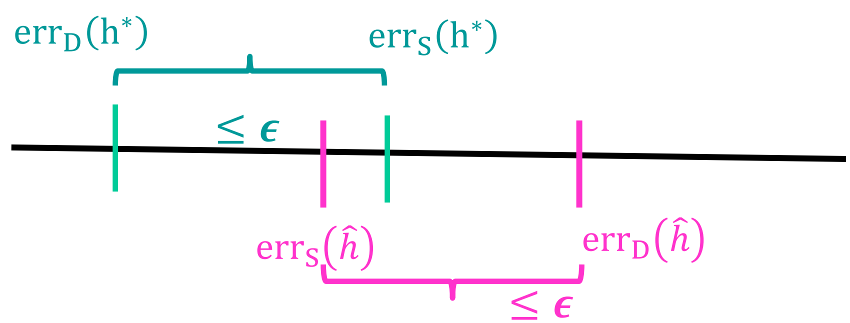
Theorem

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

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

Proof: Hoeffding & union bound.

- Fix h; by Hoeffding, prob. that $|\mathrm{err}_S(h) - \mathrm{err}_D(h)| \geq \epsilon$ is at most $2e^{-2m\epsilon^2}$
- By union bound over all $h \in H$, the prob. that $\exists h$ s.t. $|\mathrm{err}_S(h) - \mathrm{err}_D(h)| \geq \epsilon$ is at most $2|H|e^{-2m\epsilon^2}$. Set to δ. Solve.



Fact:

W.h.p. $\geq 1-\delta$, $\mathrm{err}_D(\hat{h}) \leq \mathrm{err}_D(h^*) + 2\epsilon$, $\hat{h}$ is ERM output, h^* is hyp. of smallest true error rate.

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 $h \in H$:

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

$\sqrt{\frac{1}{m}}$ 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 $h \in H$:

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

VCdimension Generalization Bounds

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

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).

Rademacher Complexity Bounds

[Koltchinskii&Panchenko 2002]

- Distribution/data dependent. Tighter for nice distributions.
- Apply to general classes of real valued functions & can be used to recover the VCbounds for supervised classification.
- Prominent technique for generalization bounds in last decade.

See "Introduction to Statistical Learning Theory"
O. Bousquet, S. Boucheron, and G. Lugosi.

Rademacher Complexity

Problem Setup

- A space Z and a distr. $D_{|Z}$
- F be a class of functions from Z to $[0,1]$
- $S = \{z_1, \ldots, z_m\}$ be i.i.d. from $D_{|Z}$

Want a high prob. uniform convergence bound, all $f \in F$ satisfy:

$$E_D[f(z)] \le E_S[f(z)] + \text{term}(\text{complexity of F, niceness of D/S})$$

What measure of complexity?

General discrete Y

E.g., $Z = X \times Y, Y = \{-1,1\}$, $H = \{h: X \to Y\}$ hyp. space (e.g., lin. sep)

$F = L(H) = \{l_h: X \to Y\}$, where $l_h(z = (x,y)) = 1_{\{h(x) \neq y\}}$ [Loss fnc induced by h and 0/1 loss]

Then $E_{z \sim D}[l_h(z)] = \text{err}_D(h)$ and $E_S[l_h(z)] = \text{err}_S(h)$.

$$\text{err}_D[h] \le \text{err}_S[h] + \text{term}(\text{complexity of H, niceness of D/S})$$

Rademacher Complexity

Space Z and a distr. $D_{|Z}$; F be a class of functions from Z to $[0,1]$

Let $S = \{z_1, \ldots, z_m\}$ be i.i.d from $D_{|Z}$.

The empirical Rademacher complexity of F is:

$$\hat{R}_m(F) = E_{\sigma_1,\ldots,\sigma_m}\left[\sup_{f \in F} \frac{1}{m} \sum_i \sigma_i f(z_i)\right]$$

where σ_i are i.i.d. Rademacher variables chosen uniformly from $\{-1,1\}$.

The Rademacher complexity of F is: $R_m(F) = E_S[\hat{R}_m(F)]$

sup measures for any given set S and Rademacher vector σ, the max correlation between $f(z_i)$ and σ_i for all $f \in F$

So, taking the expectation over σ this measures the ability of class F to fit random noise.

Rademacher Complexity

Space Z and a distr. $D_{|Z}$; F be a class of functions from Z to $[0,1]$

Let $S = \{z_1, \dots, z_m\}$ be i.i.d from $D_{|Z}$.

The empirical Rademacher complexity of F is:

$$\hat{R}_m(F) = E_{\sigma_1,\dots,\sigma_m}\left[\sup_{f \in F} \frac{1}{m} \sum_i \sigma_i f(z_i)\right]$$

where σ_i are i.i.d. Rademacher variables chosen uniformly from $\{-1,1\}$.

The Rademacher complexity of F is: $R_m(F) = E_S[\hat{R}_m(F)]$

Theorem: Whp all $f \in F$ satisfy:

Useful if it decays with m.

$$E_D[f(z)] \le E_S[f(z)] + 2R_m(F) + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

$$E_D[f(z)] \le E_S[f(z)] + 2\,\hat{R}_m(F) + 3\sqrt{\frac{\ln(1/\delta)}{m}}$$

Rademacher Complexity

Space Z and a distr. $D_{|Z}$; F be a class of functions from Z to [0,1]

Let $S = \{z_1, \ldots, z_m\}$ be i.i.d from $D_{|Z}$.

The empirical Rademacher complexity of F is:

$$\hat{R}_m(F) = E_{\sigma_1,\ldots,\sigma_m}\left[\sup_{f \in F} \frac{1}{m} \sum_i \sigma_i f(z_i)\right]$$

where σ_i are i.i.d. Rademacher variables chosen uniformly from $\{-1,1\}$.

The Rademacher complexity of F is: $R_m(F) = E_S[\hat{R}_m(F)]$

E.g.,:

1) F={f}, then $\hat{R}_m(F) = 0$

[Linearity of expectation: each $\sigma_i f(z_i)$ individually has expectation 0.]

2) F={all 0/1 fnc}, then $\hat{R}_m(F) = 1/2$

[To maximize set $f(z_i) = 1$ when $\sigma_i = 1$ and $f(z_i) = 0$ when $\sigma_i = -1$. Then quantity inside expectation is $\#1's \in \sigma$, which is m/2 by linearity of expectation.]

Rademacher Complexity

Space Z and a distr. $D_{|Z}$; F be a class of functions from Z to [0,1]

Let $S = \{z_1, \dots, z_m\}$ be i.i.d from $D_{|Z}$.

The empirical Rademacher complexity of F is:

$$\hat{R}_m(F) = E_{\sigma_1,\dots,\sigma_m}\left[\sup_{f\in F} \frac{1}{m}\sum_{o} \sigma_i f(z_i)\right]$$

where σ_i are i.i.d. Rademacher variables chosen uniformly from $\{-1,1\}$.

The Rademacher complexity of F is: $R_m(F) = E_S[\hat{R}_m(F)]$

E.g.,:

1) F={f}, then $\hat{R}_m(F) = 0$

2) F={all 0/1 fnc}, then $\hat{R}_m(F) = 1/2$

3) F=L(H), H=binary classifiers then: $R_S(F) \le \sqrt{\frac{\ln(2|H[S]|)}{m}}$

H finite: $R_S(F) \le \sqrt{\frac{\ln(2|H|)}{m}}$

Rademacher Complexity Bounds

Space Z and a distr. $D_{|Z}$; F be a class of functions from Z to [0,1]

Let $S = \{z_1, \dots, z_m\}$ be i.i.d from $D_{|Z}$.

The empirical Rademacher complexity of F is:

$$\widehat{R}_m(F) = E_{\sigma_1,\dots,\sigma_m}\left[\sup_{f\in F} \frac{1}{m}\sum_o \sigma_i f(z_i)\right]$$

where σ_i are i.i.d. Rademacher variables chosen uniformly from $\{-1,1\}$.

The Rademacher complexity of F is: $R_m(F) = E_S[\widehat{R}_m(F)]$

Theorem: Whp all $f \in F$ satisfy: Data dependent bound!

$$E_D[f(z)] \le E_S[f(z)] + 2R_m(F) + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

$$E_D[f(z)] \le E_S[f(z)] + 2\,\widehat{R}_m(F) + 3\sqrt{\frac{\ln(1/\delta)}{m}}$$

Bound expectation of each f in terms of its empirical average & the RC of F

Proof uses Symmetrization and Ghost Sample Tricks! (same as for VC bound)

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

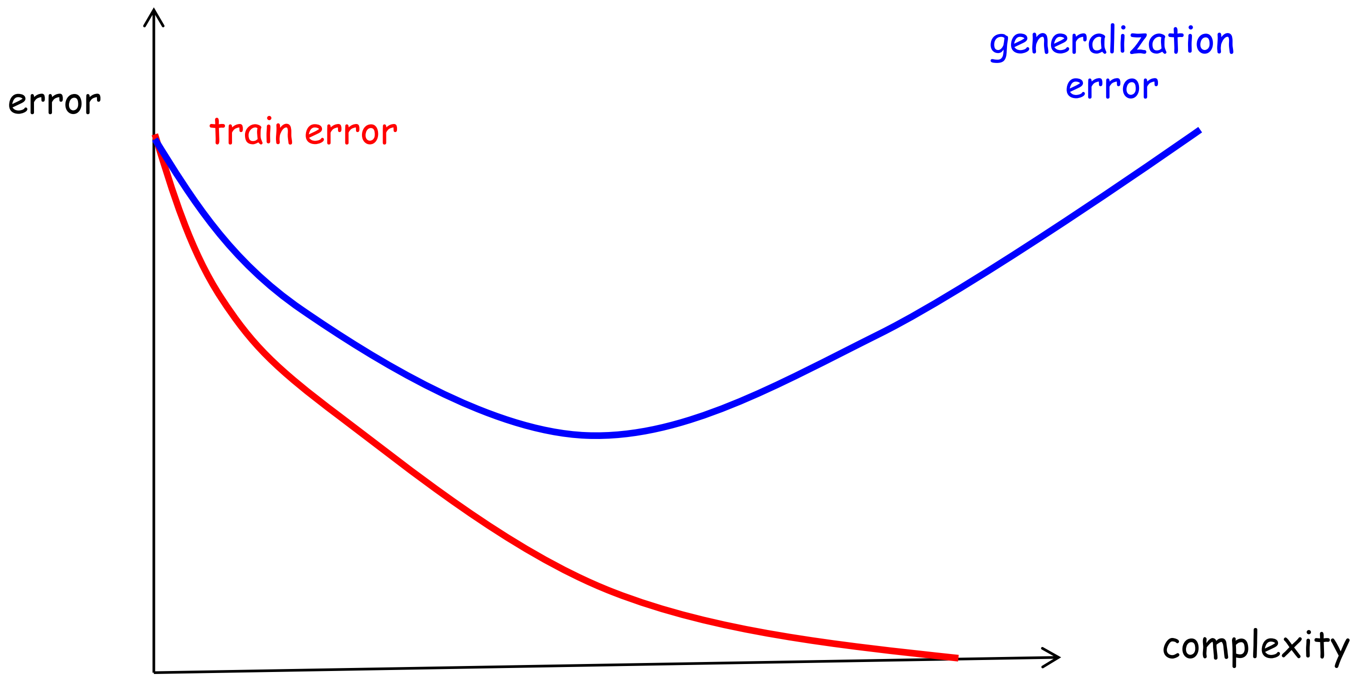
Many more uses!!! Margin bounds for SVM, boosting, regression bounds, etc.

Can we use our bounds for model selection?

True Error, Training Error, Overfitting

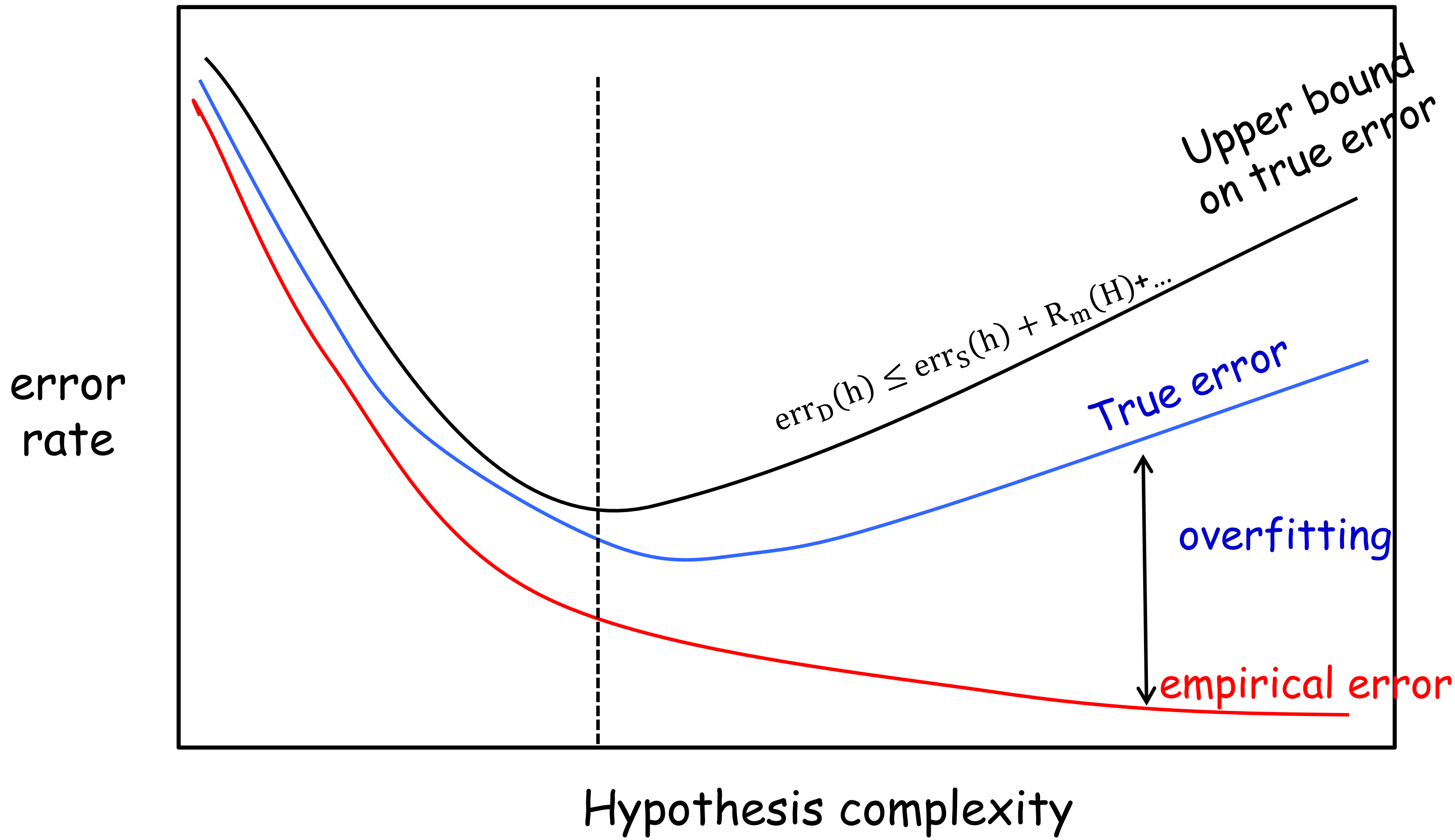
Model selection: trade-off between decreasing training error and keeping H simple.

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



Structural Risk Minimization (SRM)

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

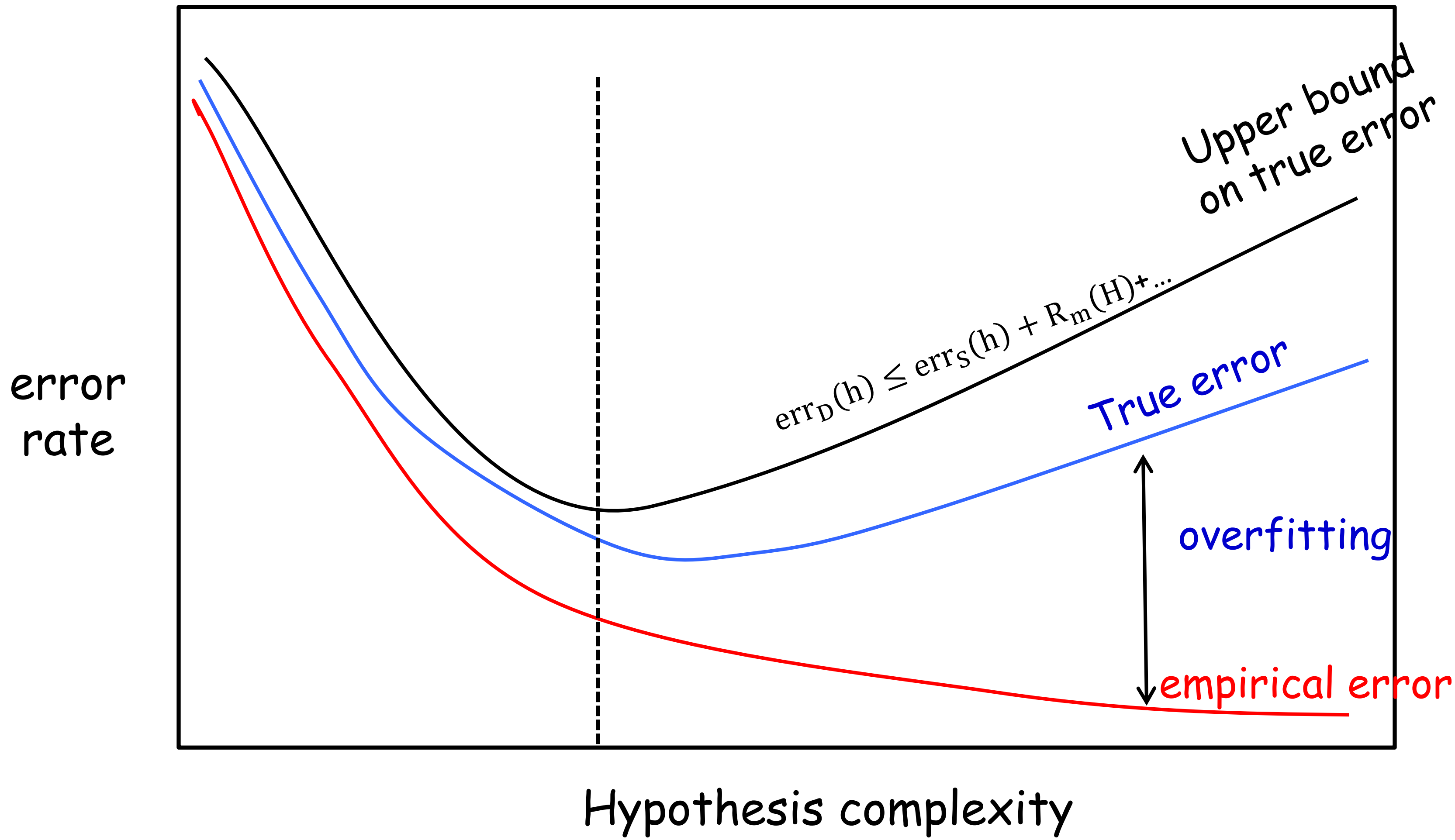


What happens if we increase m?

Black curve will stay close to the red curve for longer, everything 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: $\hat{h} = \mathrm{argmin}_{k\geq 1}\{\mathrm{err}_S(\hat{h}_k) + \mathrm{complexity}(H_k)\}$

  - 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

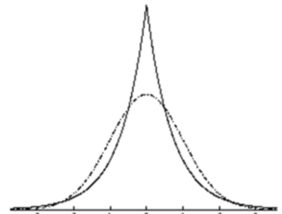
What you should know

- 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.

L2 vs. L1 Regularization

$$W = \arg\max_W \ \ln P(W) + \sum_l \ln(P(Y^l|X^l;W))$$

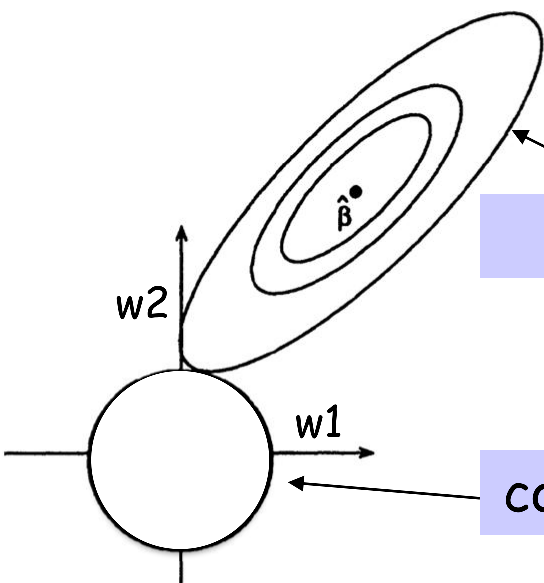
Gaussian P(W)
→ L2 regularization



Laplace P(W)
→ L1 regularization

$$\ln P(W) \propto \sum_i w_i^2$$

$$\ln P(W) \propto \sum_i |w_i|$$



constant P(Data|W)

constant P(W)