The Role of Bias in Learning Theory

John Langford

Outline

1. The Basic Model
2. The Test Set Bound
3. Occam's Razor Bound
4. PAC-MDL Bound

Model: Definitions

X = input space

$Y = \{0, 1\}$ = output space

$c : X \to Y$ = classifier

All samples are drawn independently from some unknown distribution $D(x, y)$.

$S = (x, y)^m \sim D^m$ is a sample set.

$c_D \equiv \Pr_{x,y \sim D}(c(x) \neq y) = \text{true error}$

$\hat{c}_S \equiv \Pr_{x,y \sim S}(c(x) \neq y) = \frac{1}{m} \sum_{i=1}^{m} I\left[c(x) \neq y\right]$

Model: Basic Observations

Q: What is the distribution of $\hat{c}_S$?

A: A Binomial.

$$\Pr_{S\sim D^m}\left(\hat{c}_S = \frac{k}{m}\middle| c_D\right) = \binom{m}{k} c_D^k (1-c_D)^{m-k}$$

= probability of k heads (errors) in m flips of a coin with bias c_D.

$$\begin{aligned} \mathrm{Bin}\left(\frac{k}{m}, c_D\right) &= \Pr_{S\sim D^m}\left(\hat{c}_S \le \frac{k}{m}\middle| c_D\right) \\ &= \textstyle\sum_{i=0}^{k} \binom{m}{i} c_D^i (1-c_D)^{m-i} \end{aligned}$$

= probability of observing k or fewer "heads" with m coins.

Model: basic quantities

Need confidence intervals $\Rightarrow$ use "pivot" of the cumulative

$$\overline{\mathsf{Bin}}\left(\frac{k}{m}, \delta\right) = \max\left\{p : \mathsf{Bin}\left(\frac{k}{m}, p\right) \geq \delta\right\}$$

= the largest true error such that the probability of observing k or fewer "heads" (errors) is at least δ.

Outline

1. The Basic Model
2. The Test Set Bound
3. Occam's Razor Bound
4. PAC-MDL Bound

Test Set Bound: Setting

Standard technique:

1. Cut the data into train set and test set
2. Train on the train set
3. Test on the test set

Test Set Bound: Theorem

Theorem: (Test Set Bound) For all classifiers c, for all D, for all $\delta \in (0,1]$ with probability $1-\delta$:

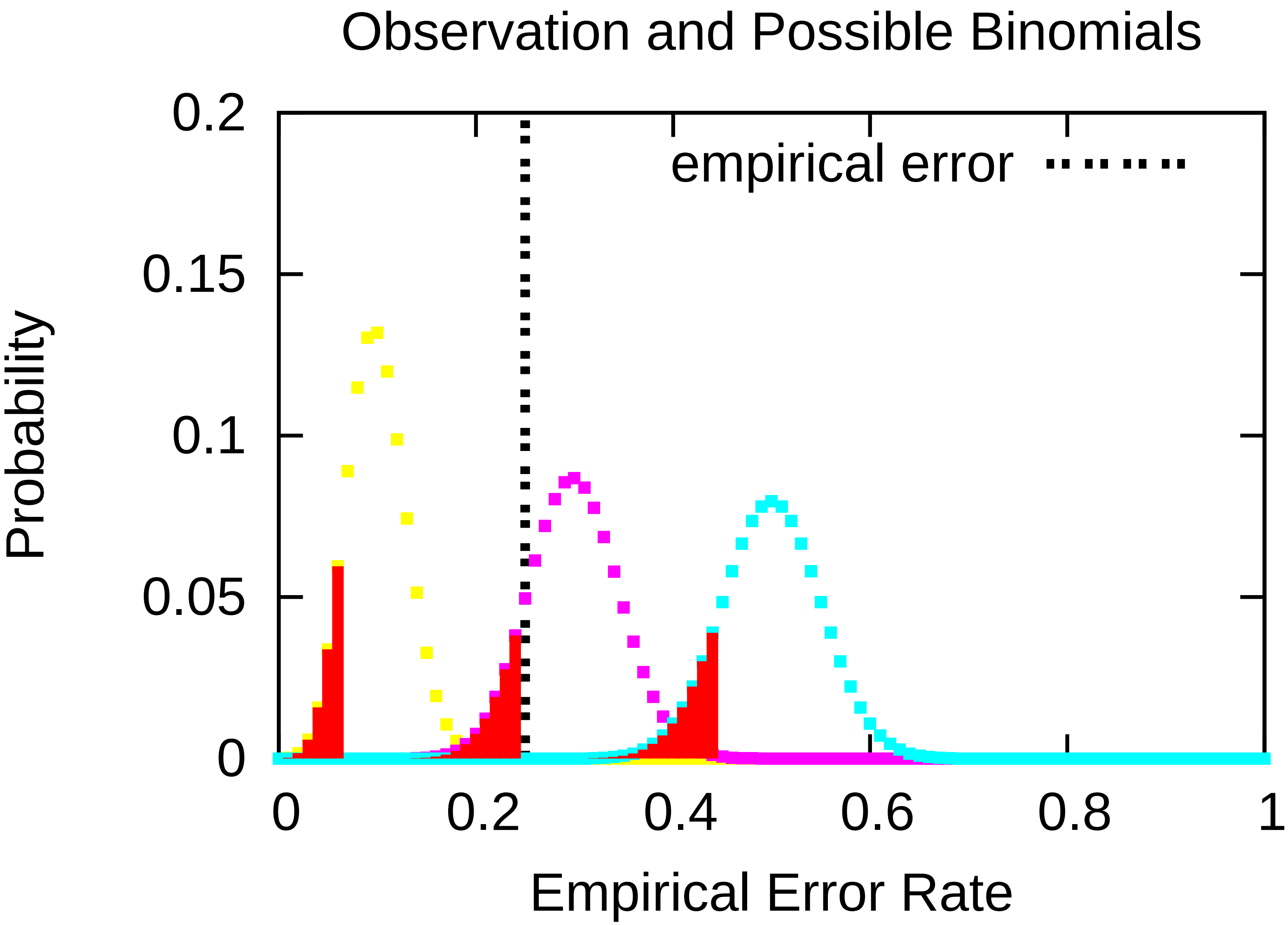
$$c_D \leq \overline{\mathrm{Bin}}\left(\hat{c}_S, \delta\right)$$

$$\Rightarrow \mathrm{KL}(\hat{c}_S || c_D) \leq \frac{\ln \frac{1}{\delta}}{m}$$

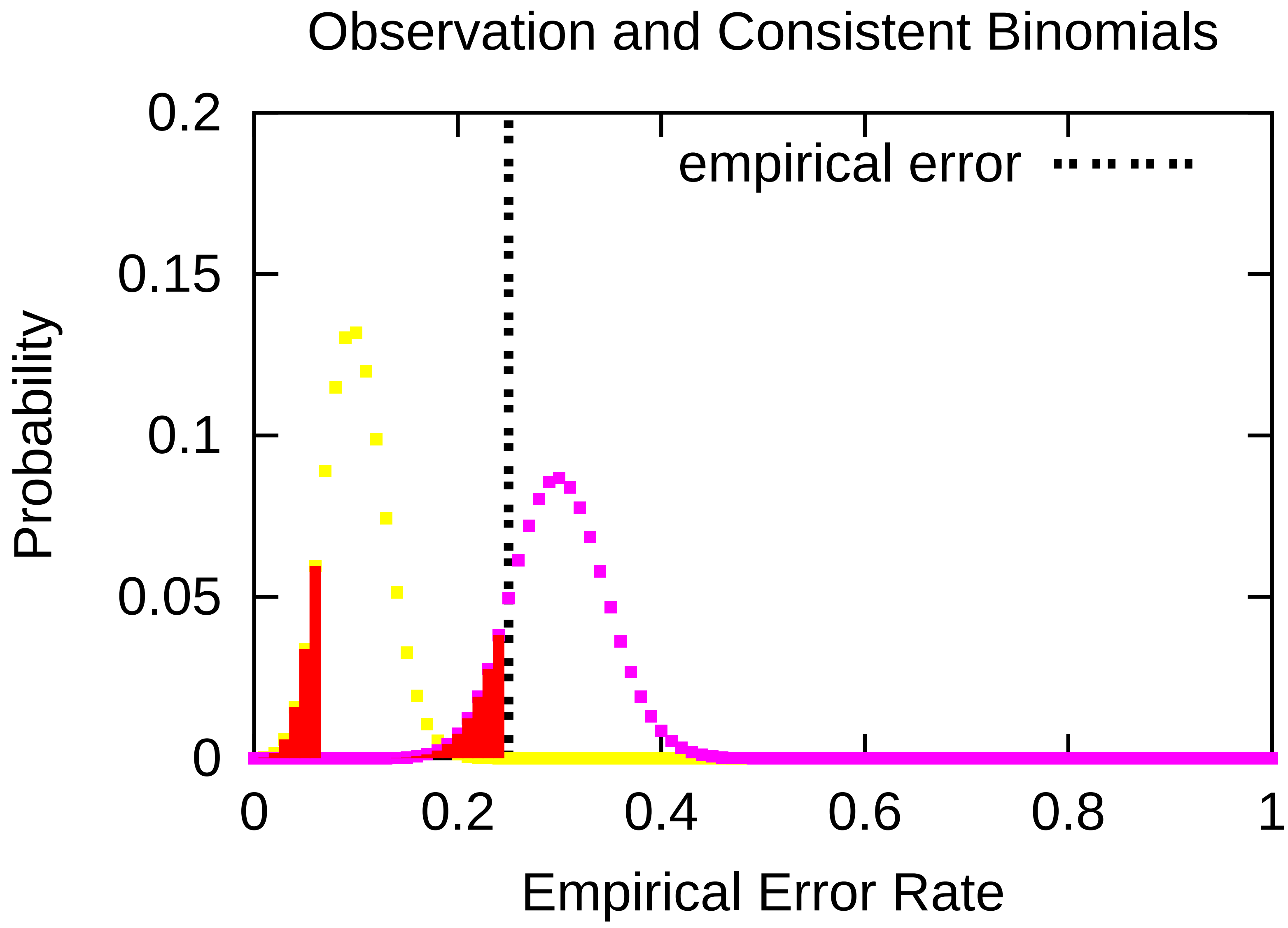
where $\mathrm{KL}(q||p) = q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p}$ for $q < p$

$$\Rightarrow c_D \leq \hat{c}_S + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

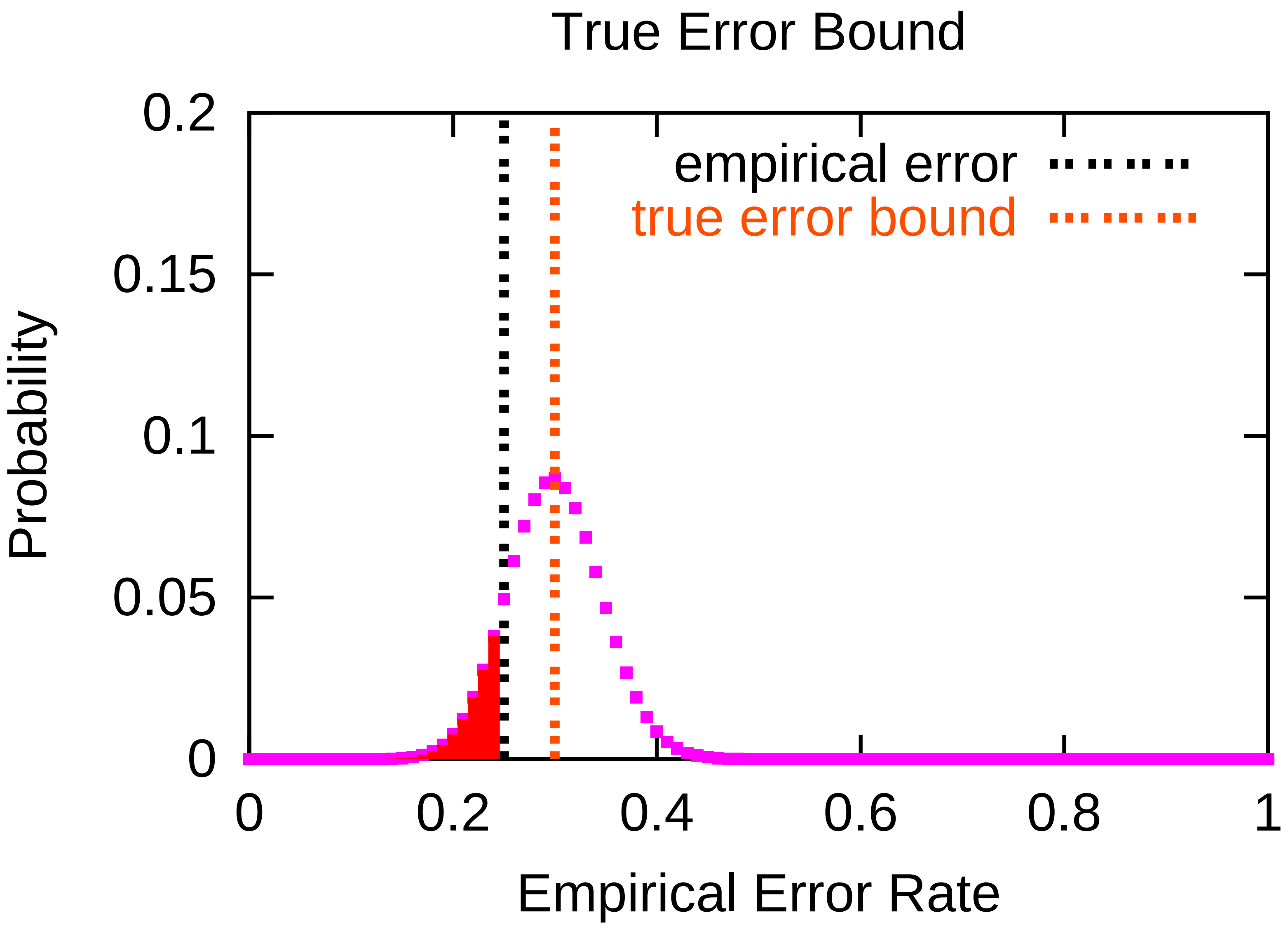
Proof: With probability $1-\delta$, observation not in δ tail.



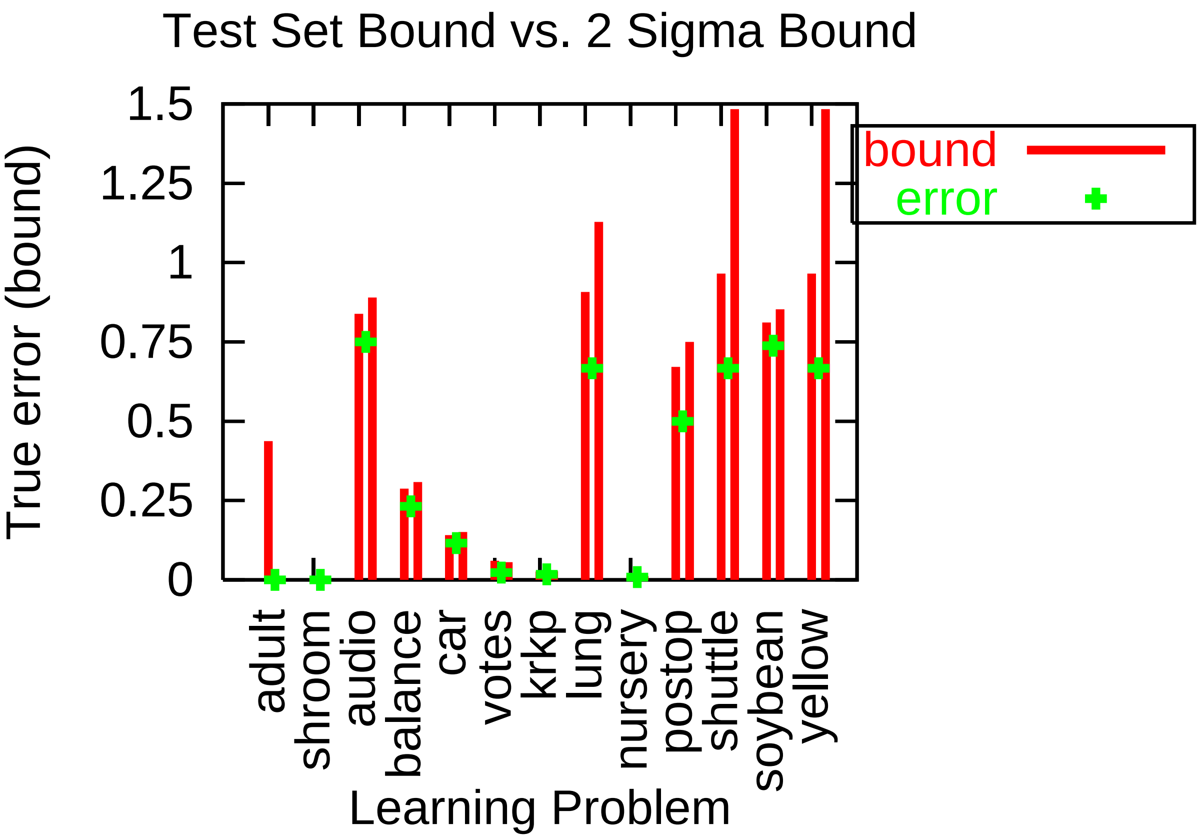
Observation and Possible Binomials
empirical error
Probability
Empirical Error Rate
0
0.05
0.1
0.15
0.2
0
0.2
0.4
0.6
0.8
1



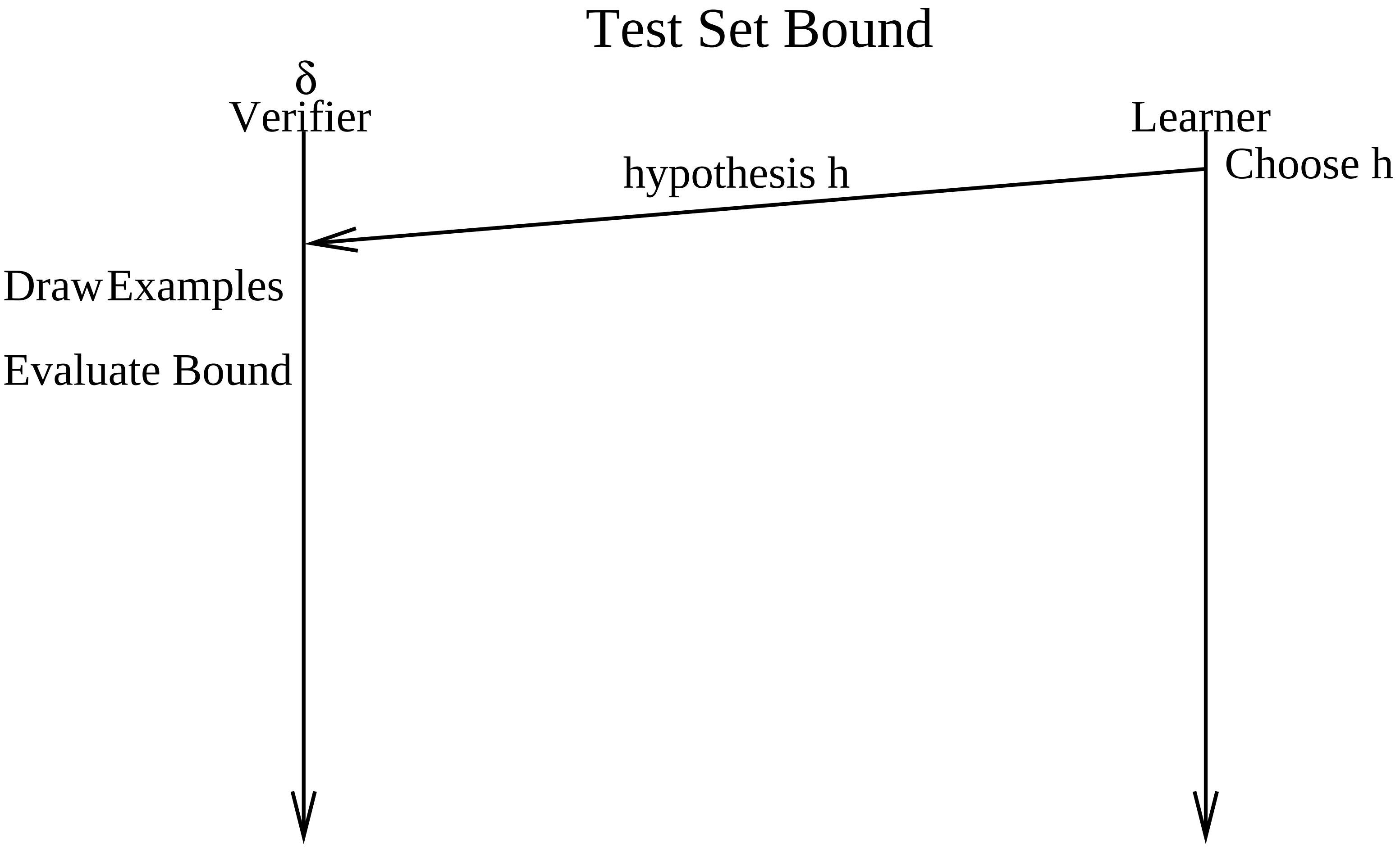
Observation and Consistent Binomials
empirical error
Probability
Empirical Error Rate
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1


True Error Bound
empirical error
true error bound
Probability
Empirical Error Rate
0
0.05
0.1
0.15
0.2
0
0.2
0.4
0.6
0.8
1

Test set bounds in action



Interpretation: Interactive Proof of Learning



Outline

1. The Basic Model
2. The Test Set Bound
3. Occam's Razor Bound
4. PAC-MDL Bound

Training Set Bounds in General

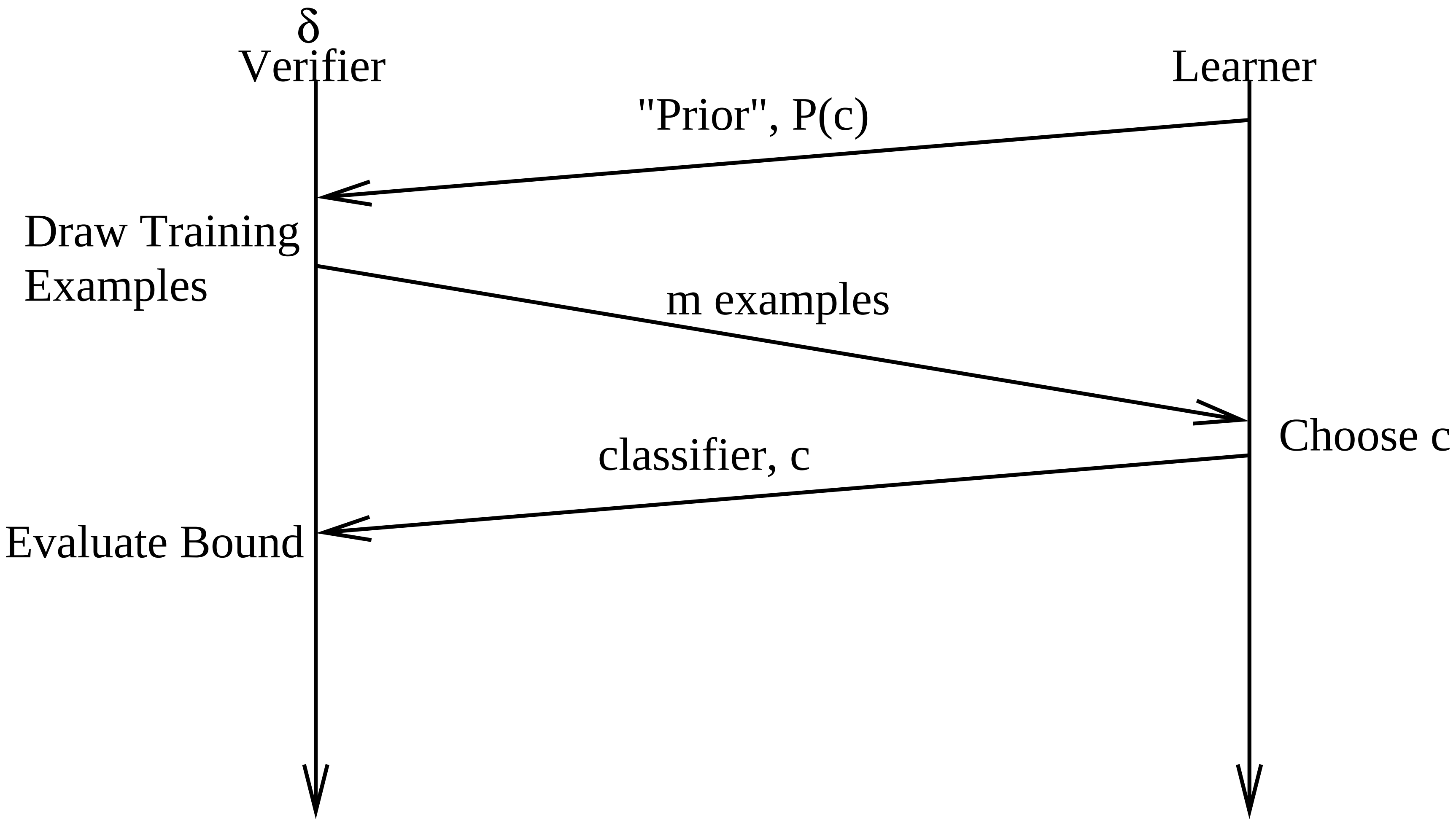
- Sometimes the holdout set is *critical* for learning.

- Sometimes we want bounds to guide learning

$\Rightarrow$Train set bounds

Occam's Razor bound is the simplest train set bound.

Occam's Razor Bound Protocol



Occam's Razor Bound

Theorem: (Occam's Razor Bound) For all "priors" $P(c)$ over the classifiers c, for all D, for all $\delta \in (0,1]$ with probability at least $1-\delta$:

$$\forall c: \; c_D \leq \overline{\mathrm{Bin}}\left(\hat{c}_S, \delta P\left(c\right)\right)$$

$$\Rightarrow \mathrm{KL}(\hat{c}_S || c_D) \leq \frac{\ln \frac{1}{P(c)} + \ln \frac{1}{\delta}}{m}$$

$$\Rightarrow c_D \leq \hat{c}_S + \sqrt{\frac{\ln \frac{1}{P(c)} + \ln \frac{1}{\delta}}{2m}}$$

Compare with test set bound: $\delta \rightarrow \delta P(c)$.

Occam's Razor Bound: Proof

First note: $P(c) = 0 \Rightarrow \overline{\mathrm{Bin}}\left(\hat{c}_S, 0\right) = 1$.

Test set bound $\Rightarrow$

$$\forall c \quad \Pr_{S \sim D^m}\left(c_D \leq \overline{\mathrm{Bin}}\left(\hat{c}_S, \delta P(c)\right)\right) \geq 1 - \delta P(c)$$

Negate to get:

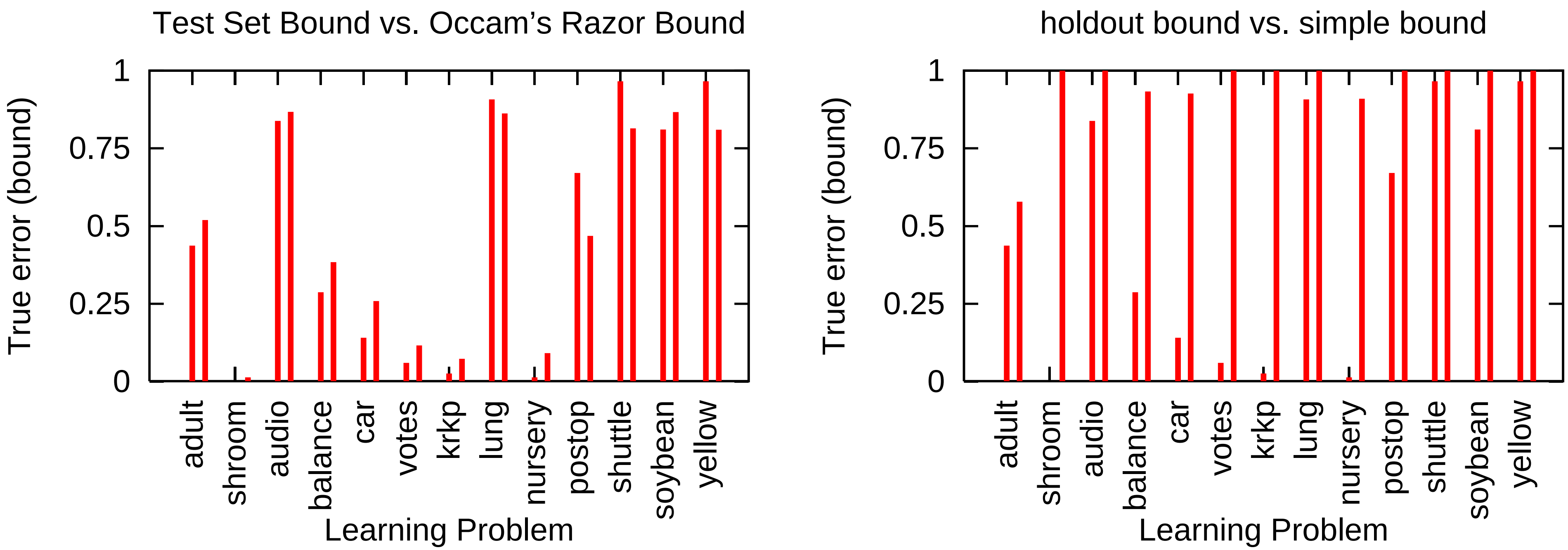
$$\forall c \quad \Pr_{S \sim D^m}\left(c_D > \overline{\mathrm{Bin}}\left(\hat{c}_S, \delta P(c)\right)\right) < \delta P(c)$$

Apply union bound: $\Pr(A \text{ or } B) \leq \Pr(A) + \Pr(B)$ repeatedly.

$$\Pr_{S \sim D^m}\left(\exists c: \;\; c_D > \overline{\mathrm{Bin}}\left(\hat{c}_S, \delta P(c)\right)\right) < \sum_{c: P(c) \neq 0} \delta P(c) = \delta$$

Negate again to prove.

Results for Decision Trees, 80/20 train/test split



Left bar = test set bound, right bar = Occam's Razor or Uniform Bound

Outline

1. The Basic Model
2. The Test Set Bound
3. Occam's Razor Bound
4. PAC-MDL Bound

Transduction Definitions

$S = (x, y)^m \sim D^m$ = train set

$S' = (x, y)^n \sim D^n$ = test set

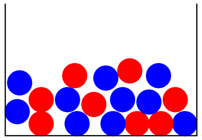
$\sigma : X^{|S|+|S'|} \rightarrow Y^{|S|+|S'|}$ = transductive classifier

$\hat{\sigma}_S \equiv \sum_{x,y \in S} I(\sigma(S, S')(x) \neq y)$ = train error

$\hat{\sigma}_{S'} \equiv \sum_{x,y \in S'} I(\sigma(S, S')(x) \neq y)$ = test error

Exact Bound

Bucket(m, n, a, b) = Probability of getting at least b blue balls in $a + b$ draws from a bucket with m red balls and n blue balls



$= \sum_{t=b}^{a+b} \frac{\binom{m}{a+b-t}\binom{n}{t}}{\binom{m+n}{a+b}}$

$b_{\mathrm{max}}(m, n, a, \delta) = \max\{b : \mathrm{Bucket}(m, n, a, b) \geq \delta\}$

Theorem: (PAC-MDL Bound) For all measures $P(\sigma)$, with probability $1-\delta$ over draws of $S, S' \sim D^{m+n}$:

$$\forall \sigma : \ \hat{\sigma}_{S'} \leq b\mathrm{max}(m, n, \hat{\sigma}_S, P(\sigma)\delta)$$

$$\Rightarrow \forall \sigma : \ \hat{\sigma}_S > 0 \text{ or } \hat{\sigma}_{S'} \leq \frac{|\sigma| + \log_2 \frac{1}{\delta}}{\log_2 \left(1 + \frac{m}{n}\right)}$$

Clustering application

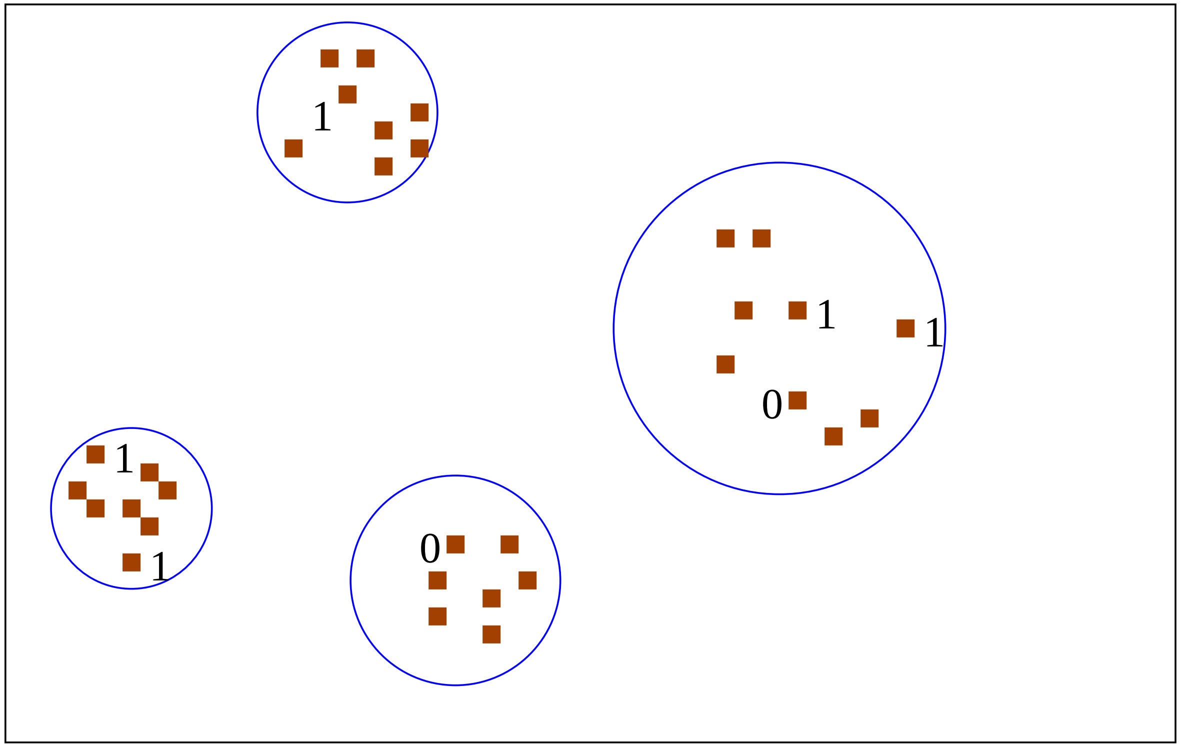
Suppose $z^t = (x, y)^m, x^n$

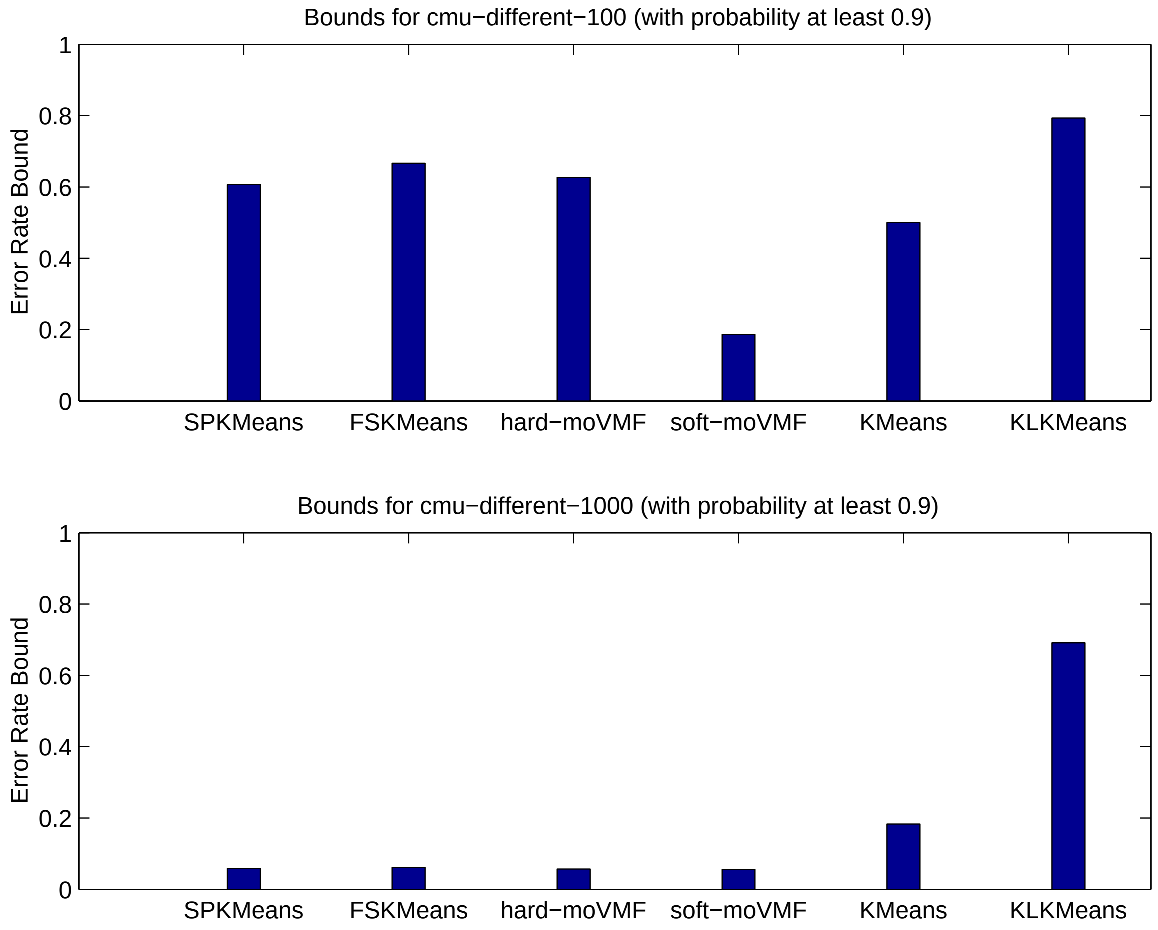
1. Cluster on x^{m+n}

2. Choose best constant label y for each cluster using y^m.

3. For each x, predict the label of it's cluster.

A Set of Clustered Points, With Labels





$$P(\sigma) = \frac{2^{-k}}{k(k+1)r}$$

k = number of clusters, r = number of random initializations

Transduction $\rightarrow$ Induction: Infinite samples

$$\lim_{|S'|\rightarrow\infty} \frac{\widehat{\sigma}_{S'}}{|S'|} = c_D$$

$\Rightarrow$ Occam's Razor Bound

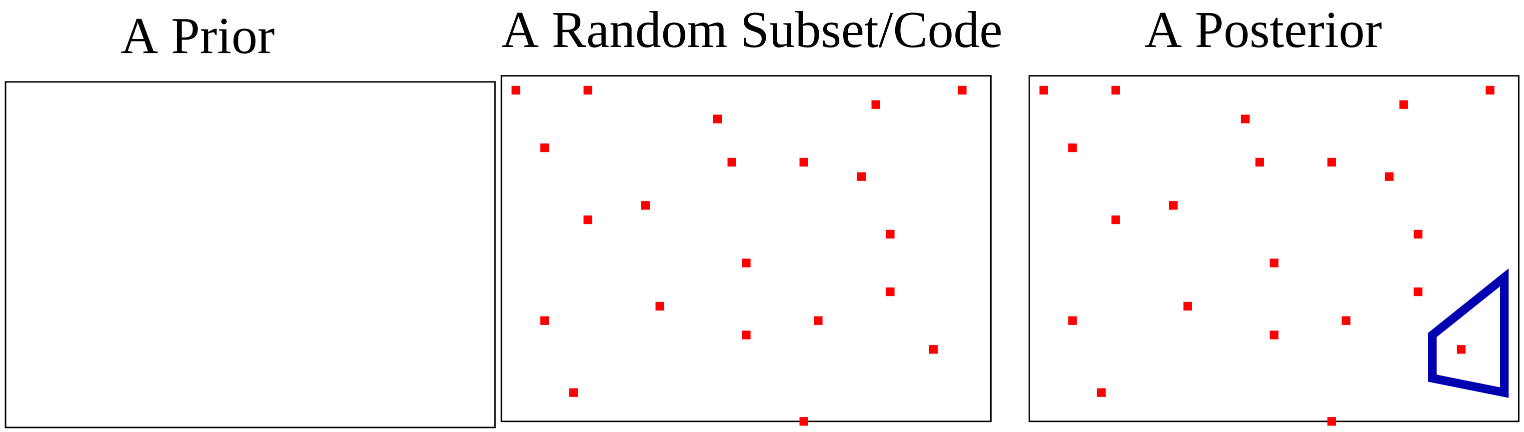
Transduction $\rightarrow$ Induction: Finite samples

Theorem: For all measures $P(\sigma)$, with probability $1-\delta$ over draws of $S \sim D^m$:

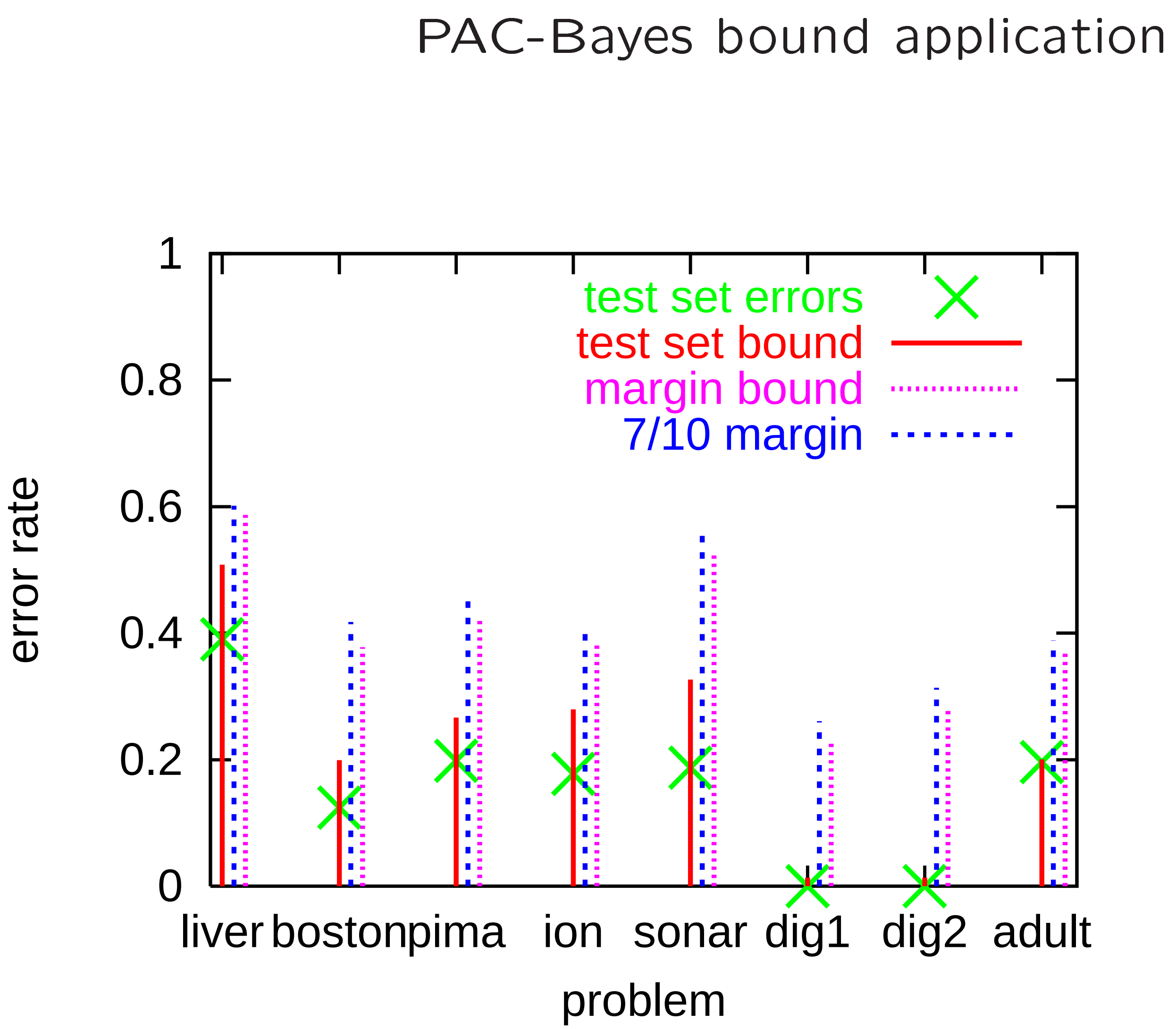
$$\forall c \;\; c_D \le \overline{\mathrm{Bin}}\left(m, \frac{\max_{\sigma \text{ consistent with } c} |\sigma| + \log_2 \frac{1}{\delta}}{\log_2\left(1+\frac{m}{n}\right)}, \frac{1}{2}\right)$$

$\Rightarrow$ VC bound, (almost) Sample Compression Bound, Rademacher bound (?).

Transduction $\rightarrow$ Induction: The PAC-Bayes Case



Communication language = random classifier language drawn from P


PAC-Bayes bound application
test set errors
test set bound
margin bound
7/10 margin
error rate
1
0.8
0.6
0.4
0.2
0
liver
boston
pima
ion
sonar
dig1
dig2
adult
problem

End of Story? No.

-Some slack in arguments.

-Does not account for “shell bound” or “localized rademacher” bounds.

For many details, code & papers, see:

http://hunch.net/~jl/projects/prediction_bounds/prediction_bounds.html