Practical Prediction Theory

John Langford

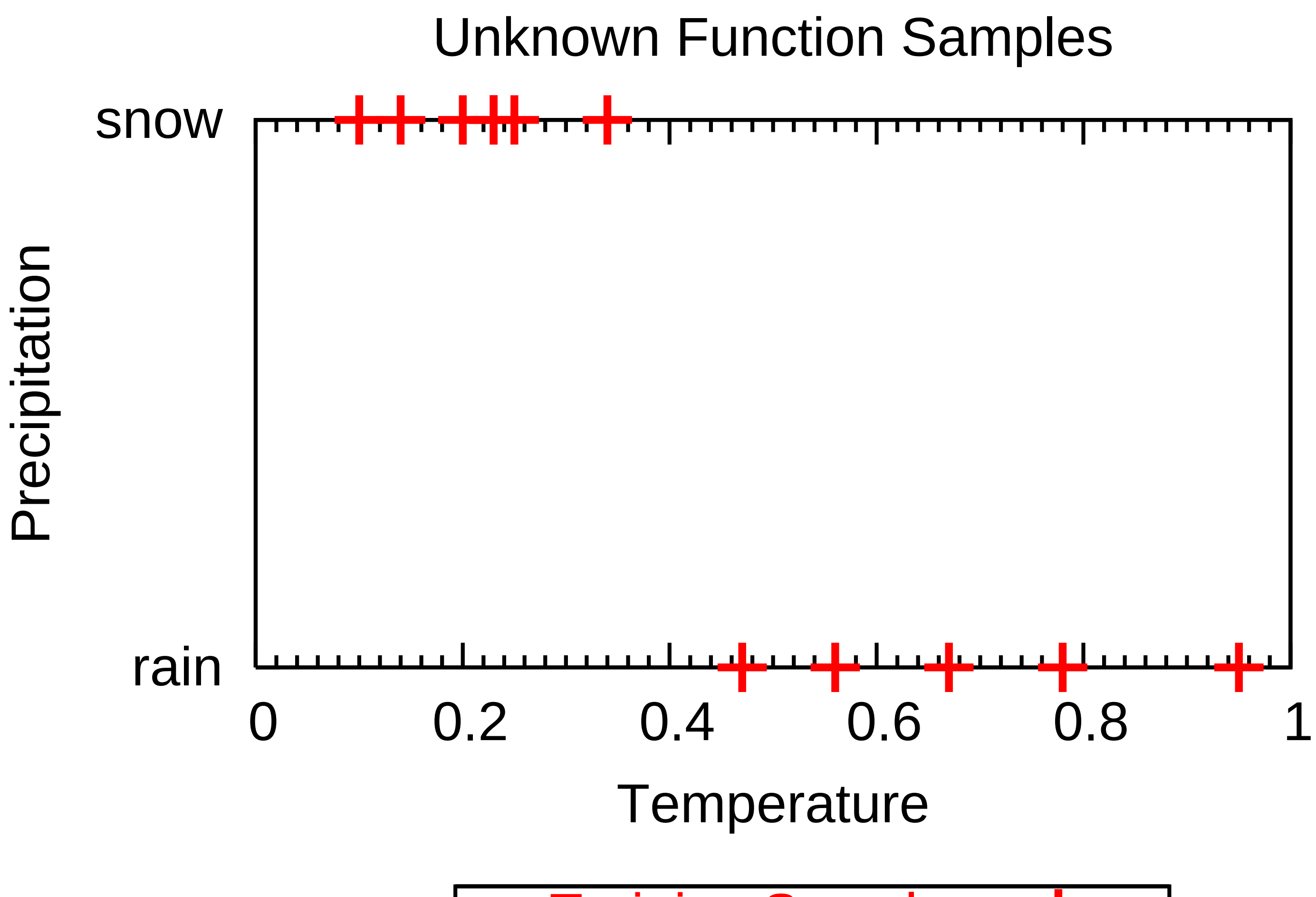
IBM Research

The Classification Problem:

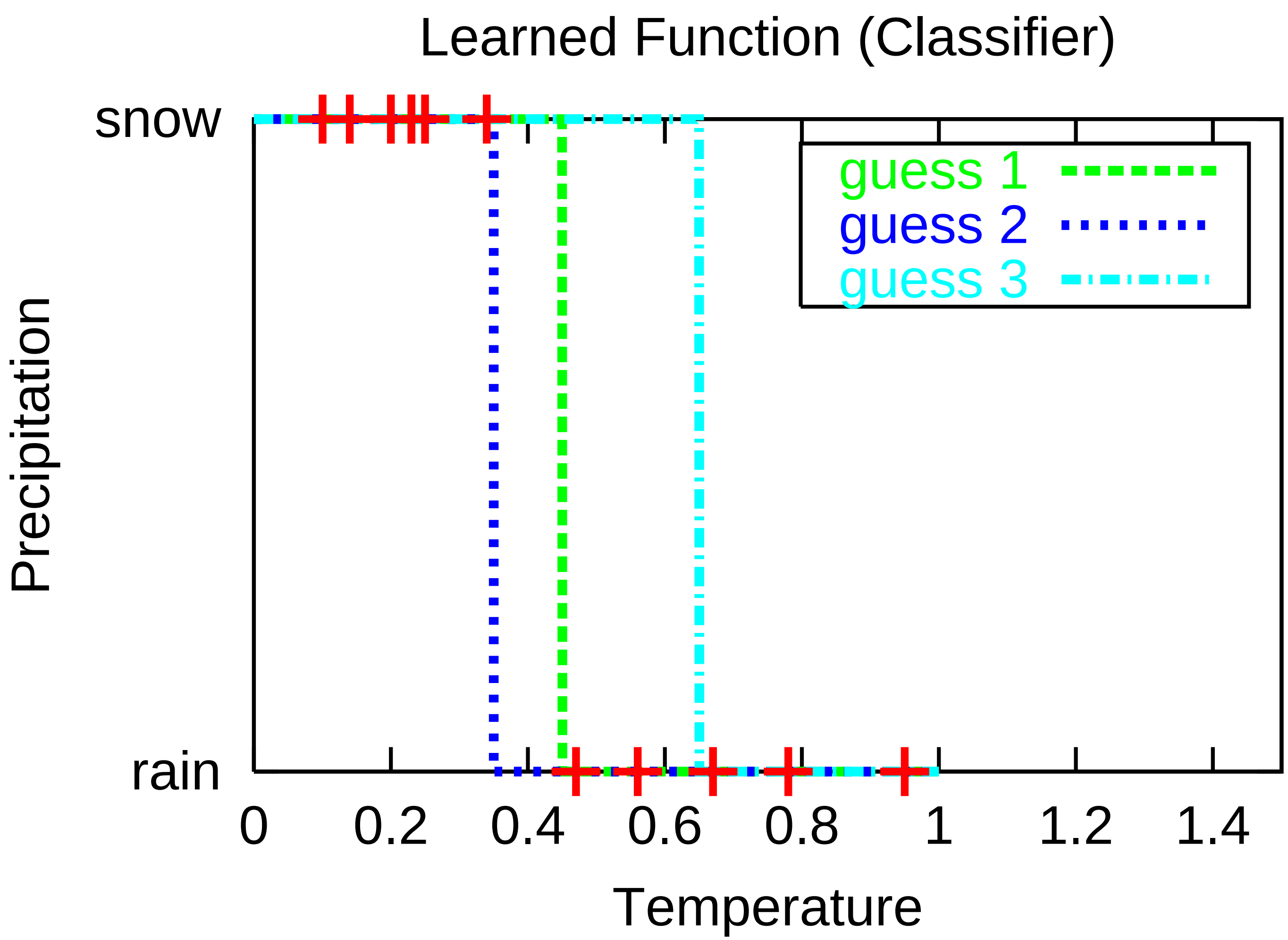
1. Get (input, class label) pairs from unknown function

2. Guess the unknown function

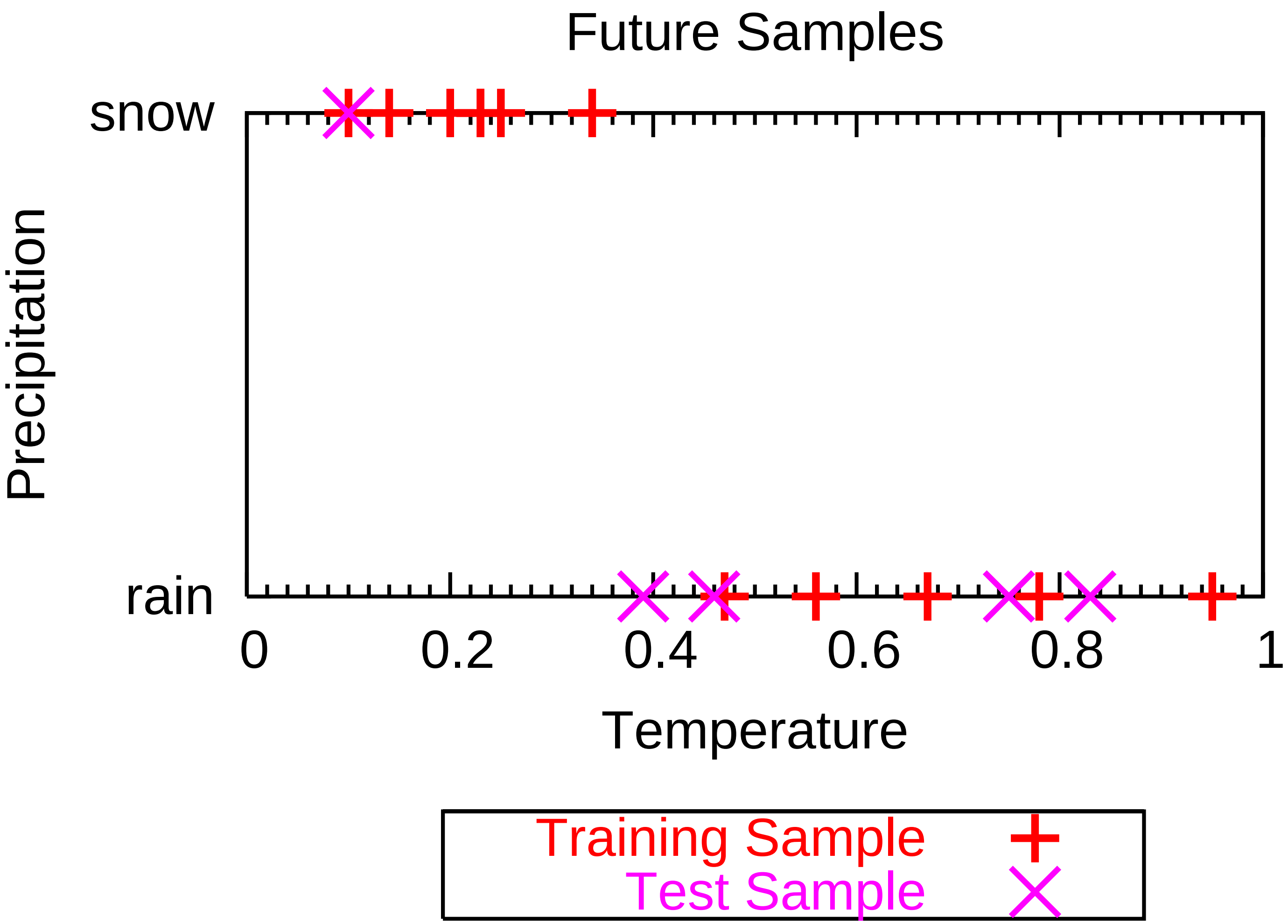
3. ... to minimize errors on future inputs



Unknown Function Samples
snow
rain
Precipitation
0
0.2
0.4
0.6
0.8
1
Temperature
Training Sample



Learned Function (Classifier)
snow
rain
Precipitation
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Temperature
guess 1
guess 2
guess 3



Future Samples
snow
rain
Precipitation
0
0.2
0.4
0.6
0.8
1
Temperature
Training Sample
Test Sample

Learning = Prediction ability

- We can't expect any prediction ability, in general.

- We can expect prediction ability, if examples come independently, sometimes.

Sample Complexity = the study of prediction ability, assuming indepedence.

Why Study Sample Complexity?

1. Methods for using more examples

2. To gain insight into learning

Methods for using more examples

Standard technique:

1. Divide samples into train and test set

2. train on train set

3. test on test set

Sample complexity approaches allow us to do better.

To gain insight into learning

1. Overfitting: sample complexity quantifies overfitting.

2. Learning algorithm design: What is a good pruning criteria? Why are large margin good? What other algorithms are likely to yield good results?

Outline

1. The Basic Model
2. The Test Set Bound
3. Occam's Razor Bound
4. PAC-Bayes Bound
5. Sample Compression Bound

Model: Definitions

X= input space

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

$c : X \to Y$ = classifier

Model: Basic Assumption

There exists a distribution $D(x, y)$ and all samples are drawn independently from D.

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

Model: Derived quantities

The thing we want to know:

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

The thing we have:

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

= "train error", "test error", or "observed error", depending on context.

(note: we overload the set S with the uniform distribution on S)

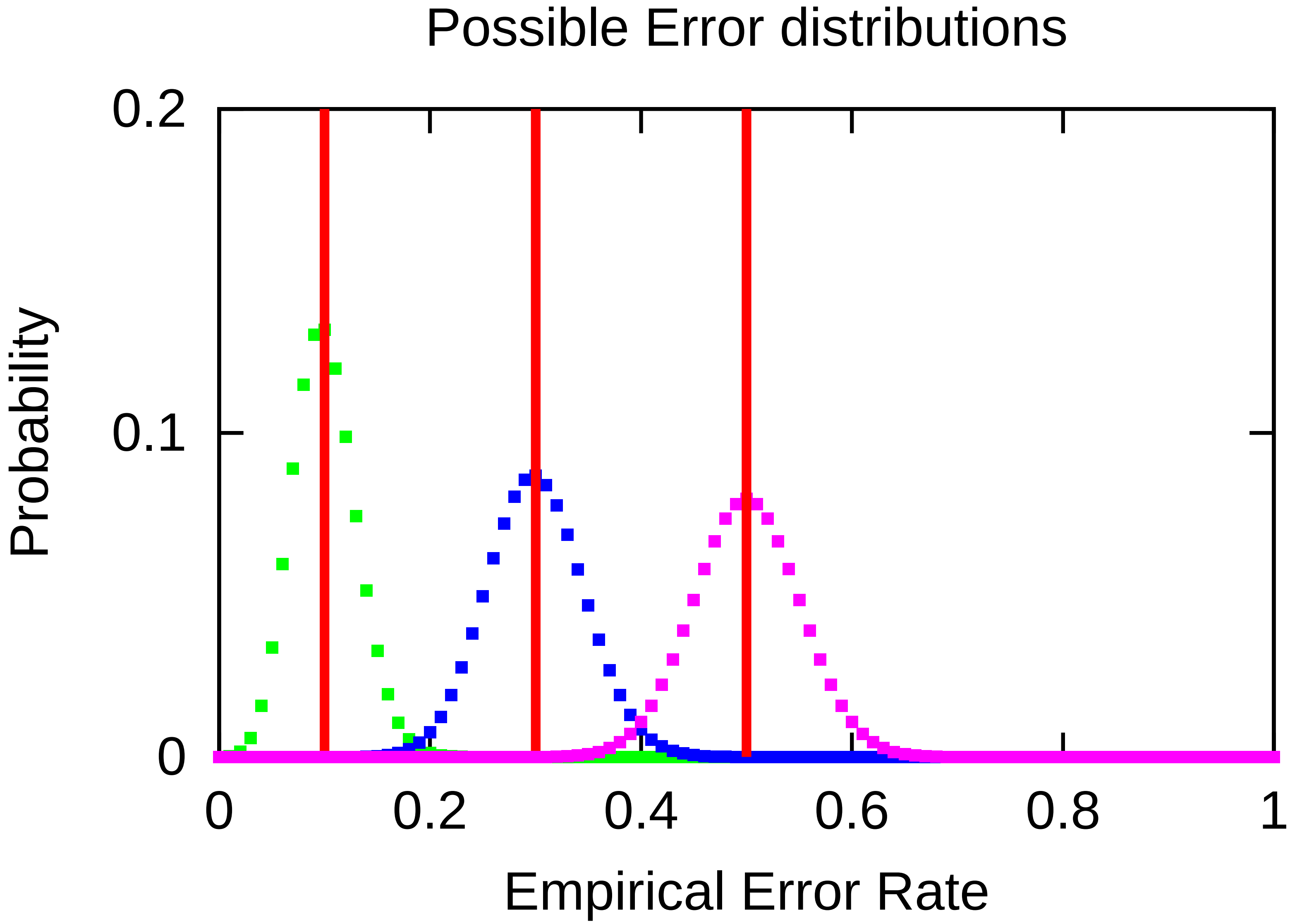
Model: Basic Observations

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

A: A Binomial.

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

$=$ probability of k heads in m flips of a coin with bias c_D.



Possible Error distributions
Probability
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
Empirical Error Rate

We use the cumulative:

$$\mathrm{Bin}\left(\frac{k}{m}, c_D\right) = \Pr_{S\sim D^m}\left(\hat{c}_S \leq \frac{k}{m} | c_D\right) = \sum_{i=0}^{k} \binom{m}{i} c_D^i (1 - c_D)^{m-i}$$

$=$ probability of observing k or fewer “heads” (errors) with m coins.

Actually, we use the pivot of the cumulative since we care about confidence intervals.

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

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

Outline

1. The Basic Model
2. The Test Set Bound
3. Occam's Razor Bound
4. PAC-Bayes Bound
5. Sample Compression Bound

Standard technique:

1. Cut the data into train set and test set

2. train on the train set

3. test on the test set

What does Sample Complexity say about this method?

Theorem: For all classifiers, c, for all $\delta \in (0,1]$

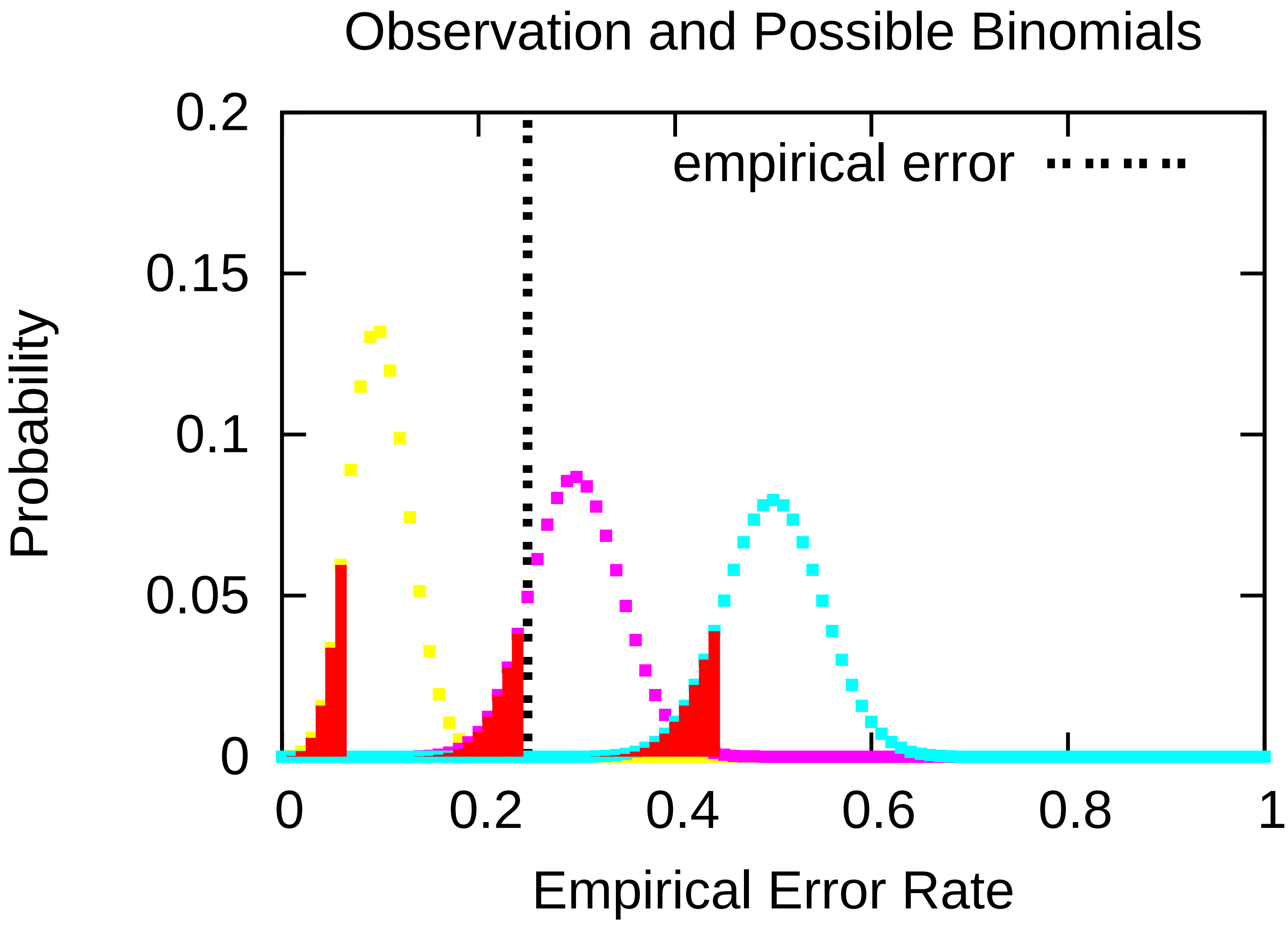
$$\Pr_{S \sim D^m}\left(c_D \leq \overline{\mathsf{Bin}}\left(\hat{c}_S, \delta\right)\right) \geq 1 - \delta$$

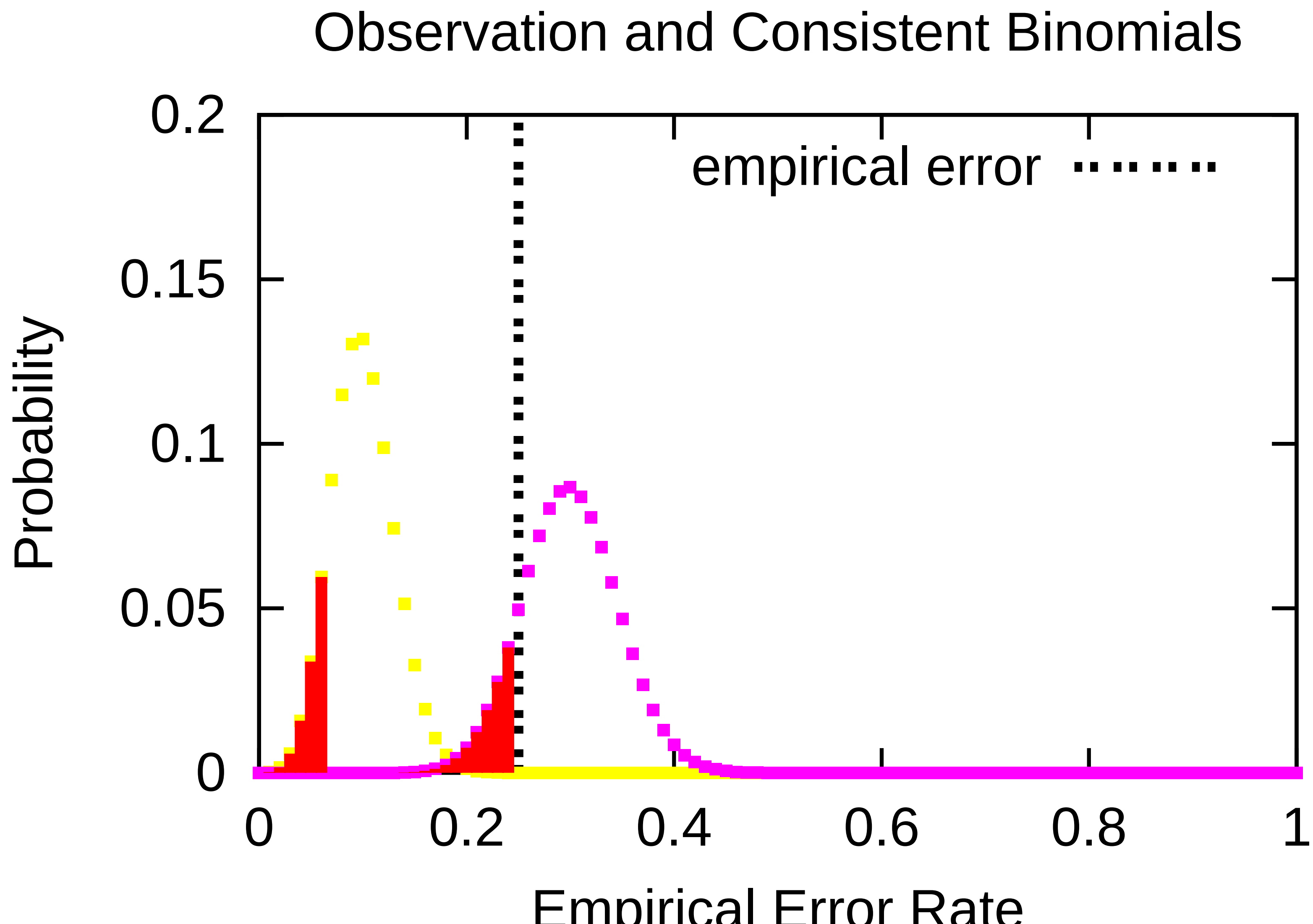
World's easiest proof: (by contradiction).

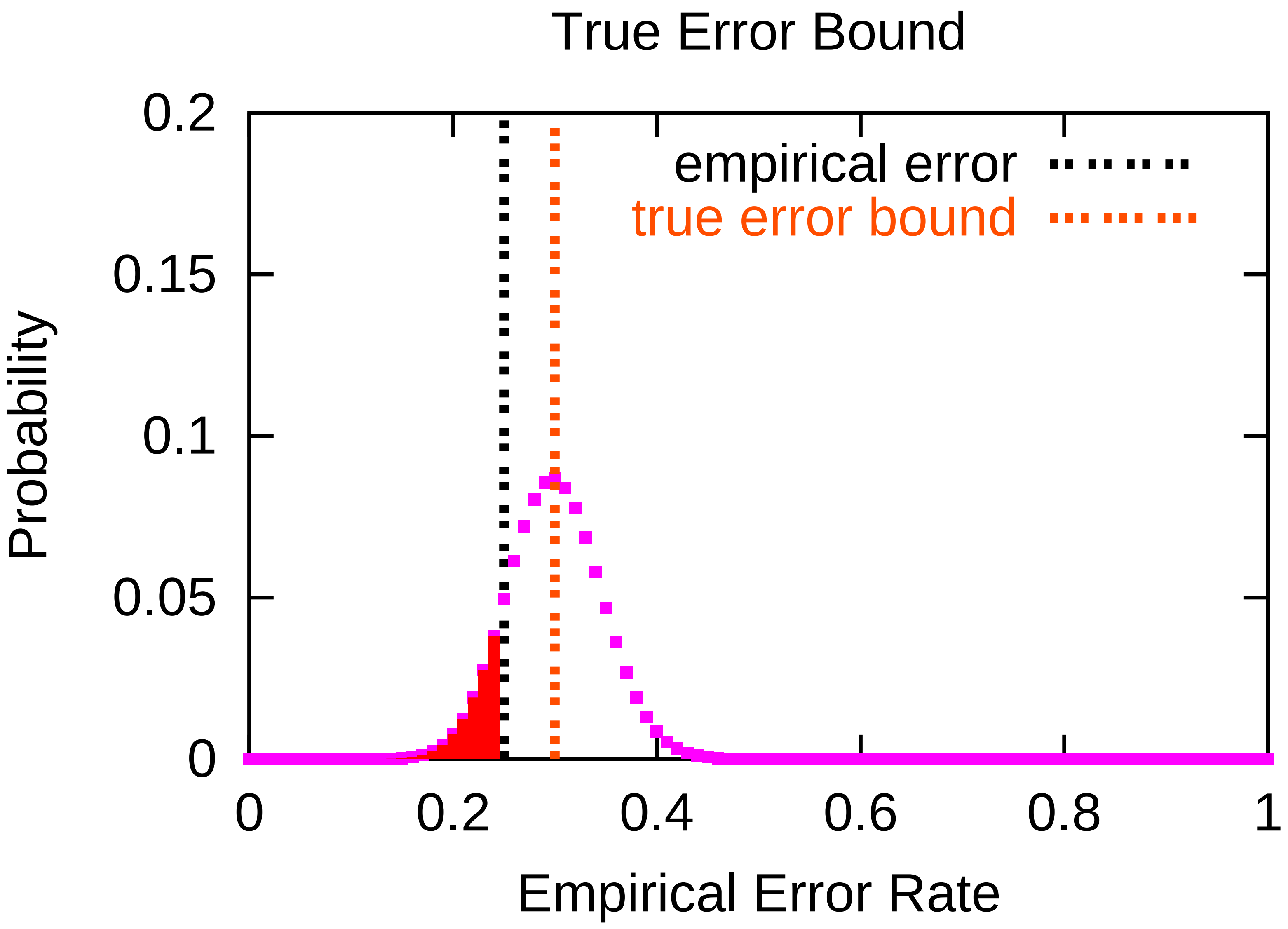
Assume $\mathsf{Bin}(\frac{k}{m}, c_D) \geq \delta$ (which is true with probability $1-\delta$).

Then, $\overline{\mathsf{Bin}}\left(\hat{c}_S, \delta\right) \geq c_D$

Next: proof by pictures


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
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1

Test set bound notes

Perfectly tight: there exist true error rates achieving the bound

Lower bound of the same form.

Primary use: verification of succesful learning

What does the bound mean?

Corollary: For all classifiers, c, for all $\delta \in (0,1]$

$$\Pr_{S \sim D^m}\left(c_D \leq \hat{c}_S + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}\right) \geq 1 - \delta$$

Proof: Use the Chernoff approximation,

$$\overline{\mathrm{Bin}}\left(\hat{c}_S, \delta\right) \leq \hat{c}_S + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

Note: NOT tight when $\hat{c}_S$ near 0 (our goal!)

Comparsion: Empirical “confidence” intervals

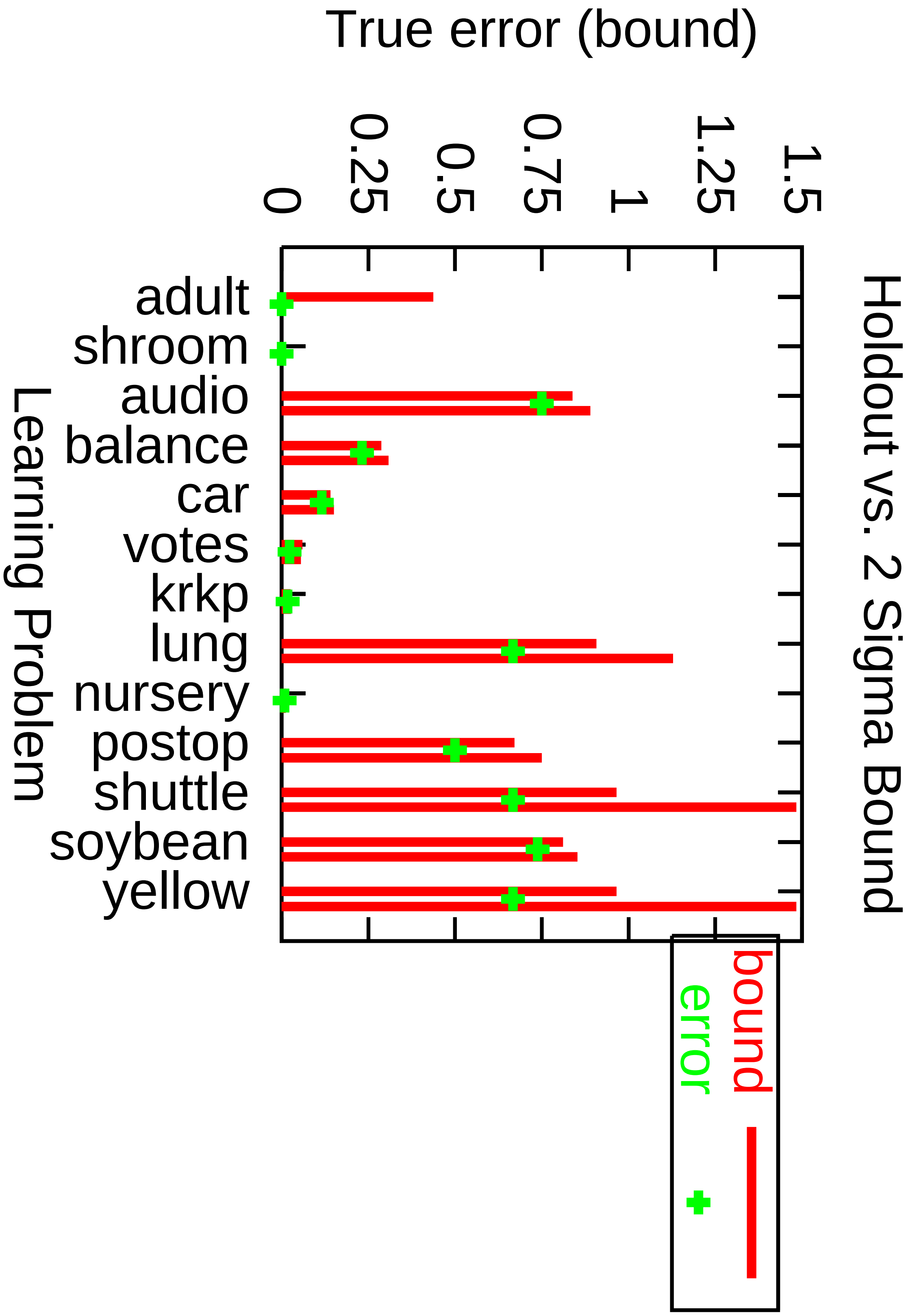
k= number of test errors, m = number of examples

$\mu = \frac{k}{m}$

$\sigma^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\mu - I(c(x_i) \neq y_i))^2$

pick bound $= \frac{k}{m} + 2\sigma$

How do they compare?



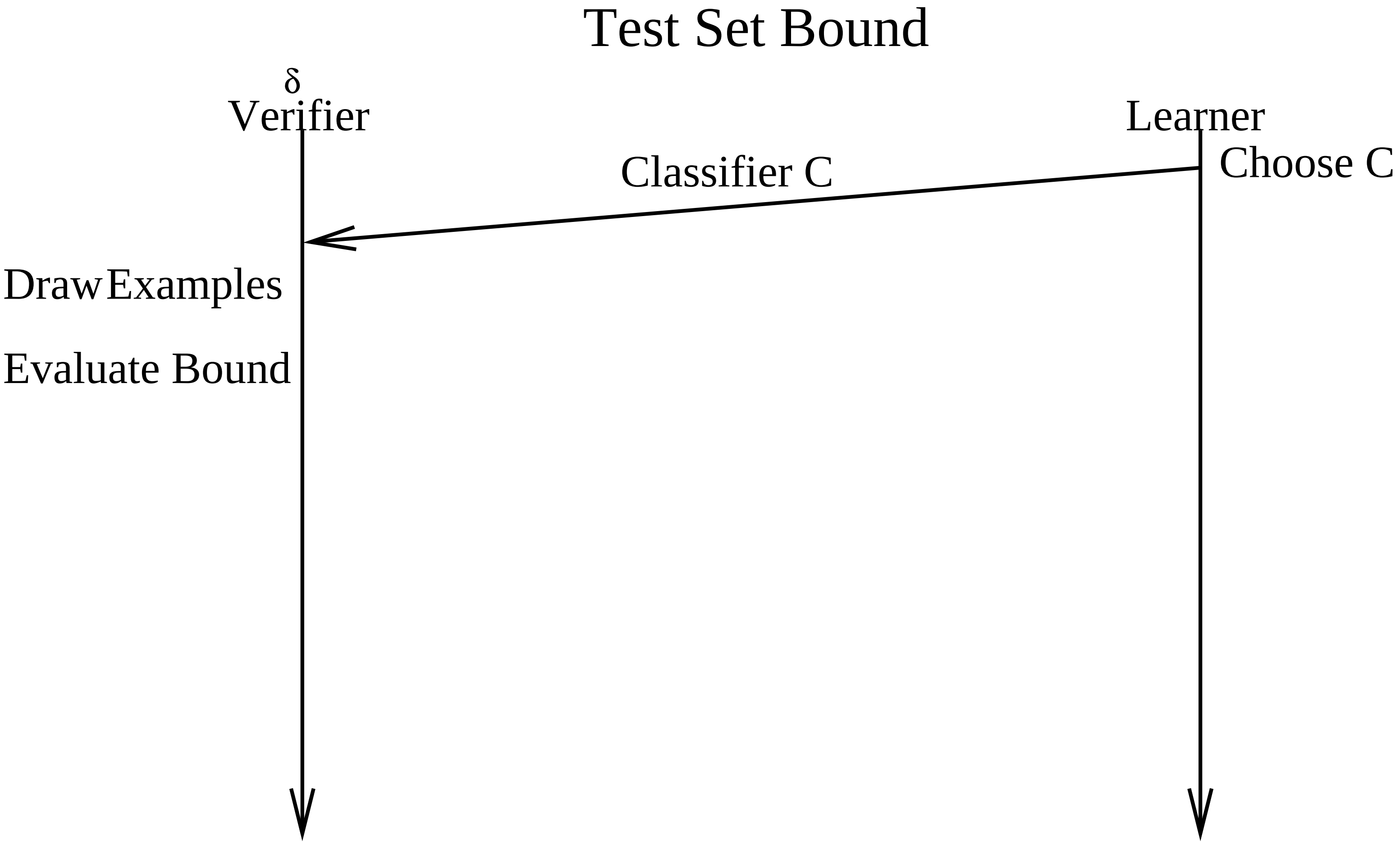
Holdout vs. 2 Sigma Bound
True error (bound)
0
0.25
0.5
0.75
1
1.25
1.5
Learning Problem
adult
shroom
audio
balance
car
votes
krkp
lung
nursery
postop
shuttle
soybean
yellow
bound
error

1. empirical confidence intervals are too pessimistic

2. empirical confidence intervals are optimistic

3. the test set bound always works

Interpretation: Interactive Proof of Learning



Outline

1. The Basic Model
2. The Test Set Bound
3. Occam's Razor Bound
4. PAC-Bayes Bound
5. Sample Compression Bound

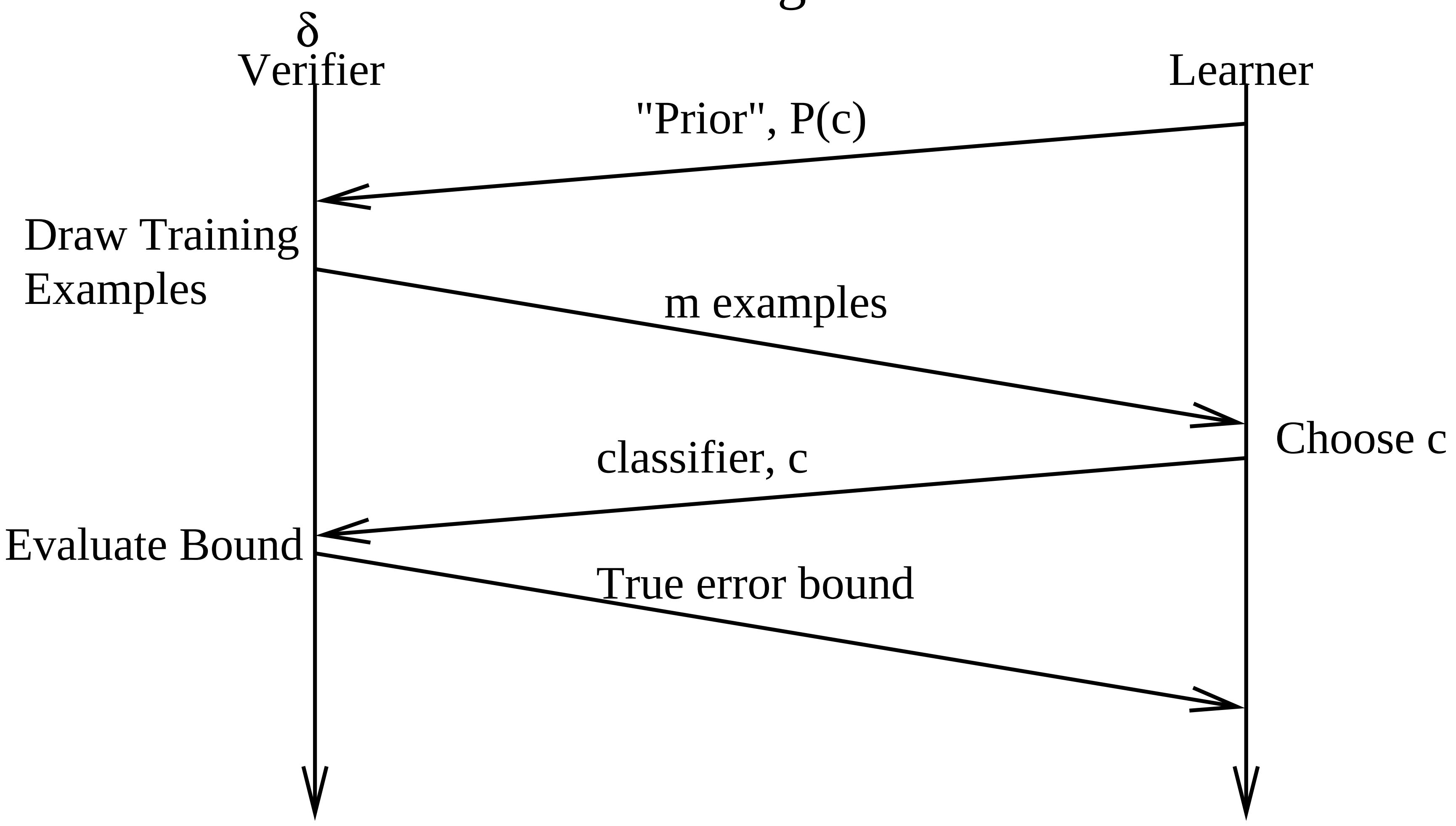
Occam's Razor bounds

- 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.

Training Set Bound



Occam's Razor Bound

Theorem: For all "priors" $P(c)$ over the classifiers, c, for all $\delta \in (0,1]$:

$$\Pr_{S \sim D^m}\left(\exists c: \; c_D \leq \overline{\mathsf{Bin}}\left(\hat{c}_S, \delta P(c)\right)\right) \geq 1 - \delta$$

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

Test set bound $\Rightarrow$

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

Negate to get:

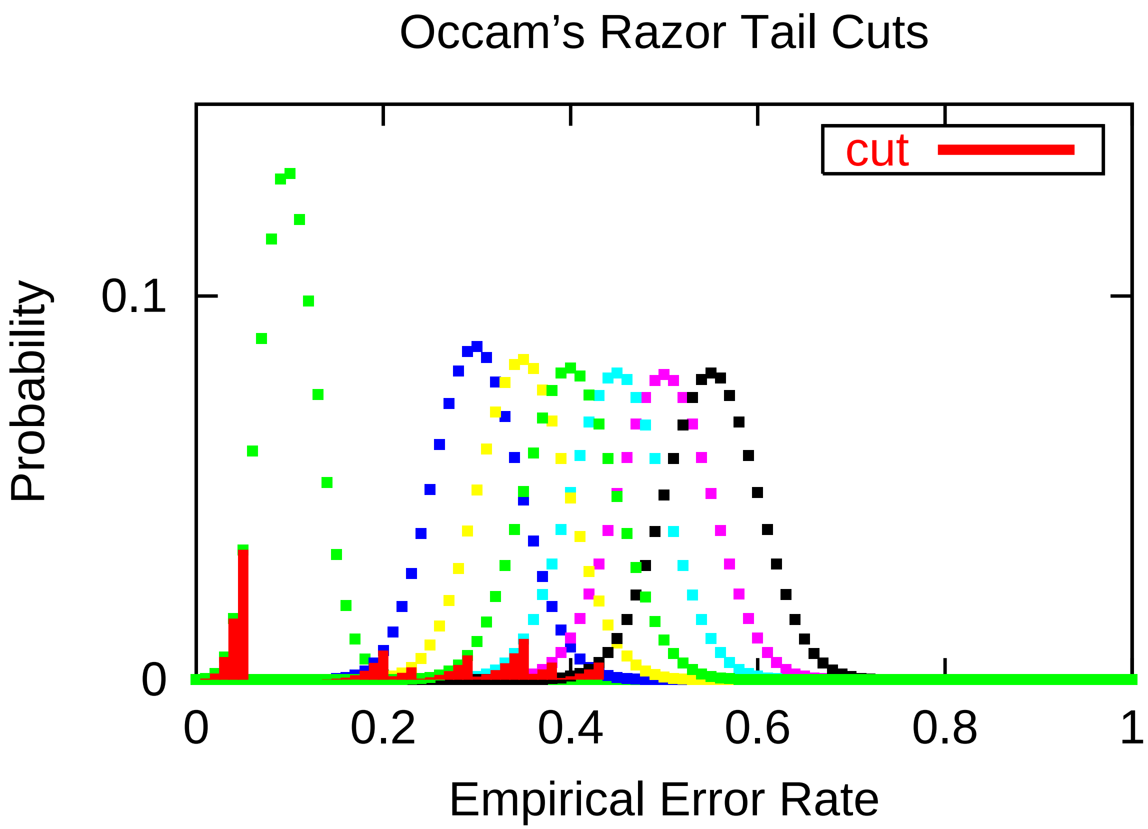
$$\forall c \quad \Pr_{S \sim D^m}\left(c_D > \overline{\mathsf{Bin}}\left(\hat{c}_S, \delta P(c)\right)\right) < \delta p(c)$$

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

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

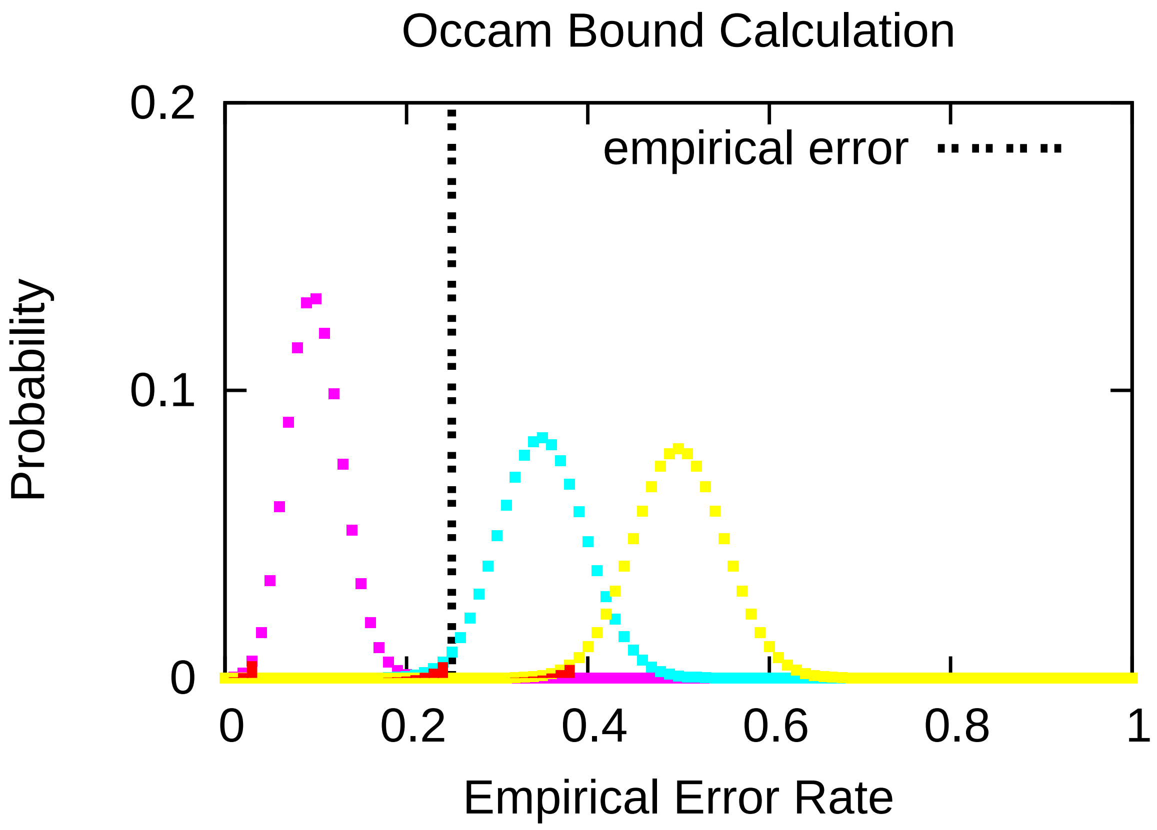
Negate again to get proof.

Next: Graphical proof

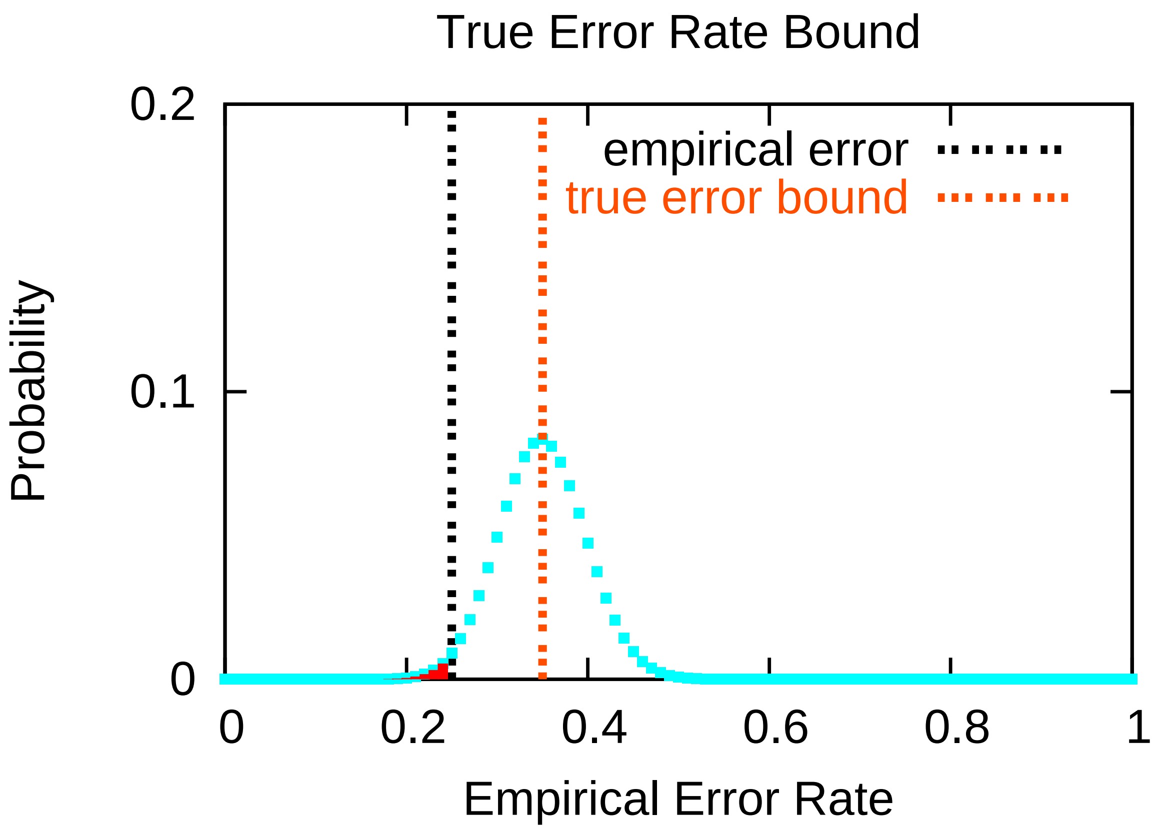


Each classifier is a Binomial with a different size tail cut.

With high probability no error falls in any tail.



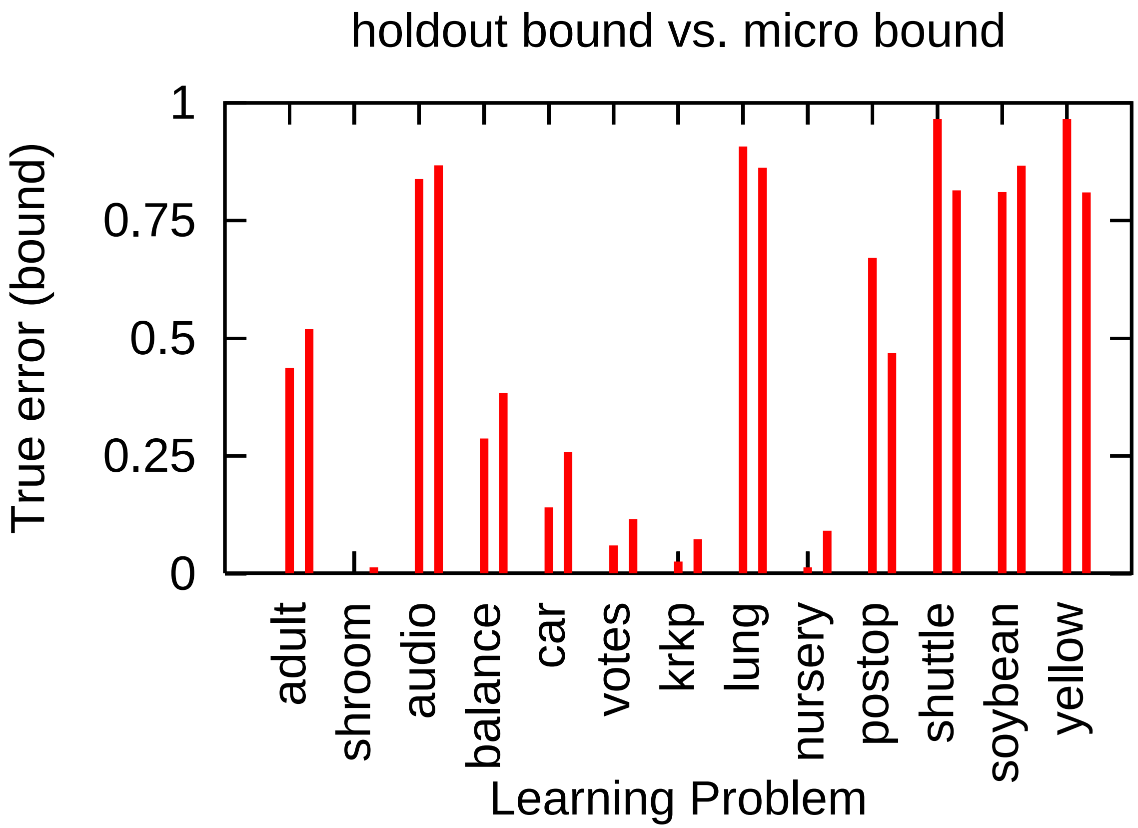
The chosen classifier has an unknown true error rate.



Bound = the largest true error rate for which the observation is not in the tail.

Results: Decision Trees

- ID3 decision tree + pruning
- probability of failure $= \delta = 0.1$
- Discrete problems from UCI database of Machine Learning problems.
- 100% of data used for training set bounds
- 80%/20% Train/Test split for test set bounds
- Minimal selection bias



Left bar = holdout bound, right bar = Occam's Razor Bound

Outline

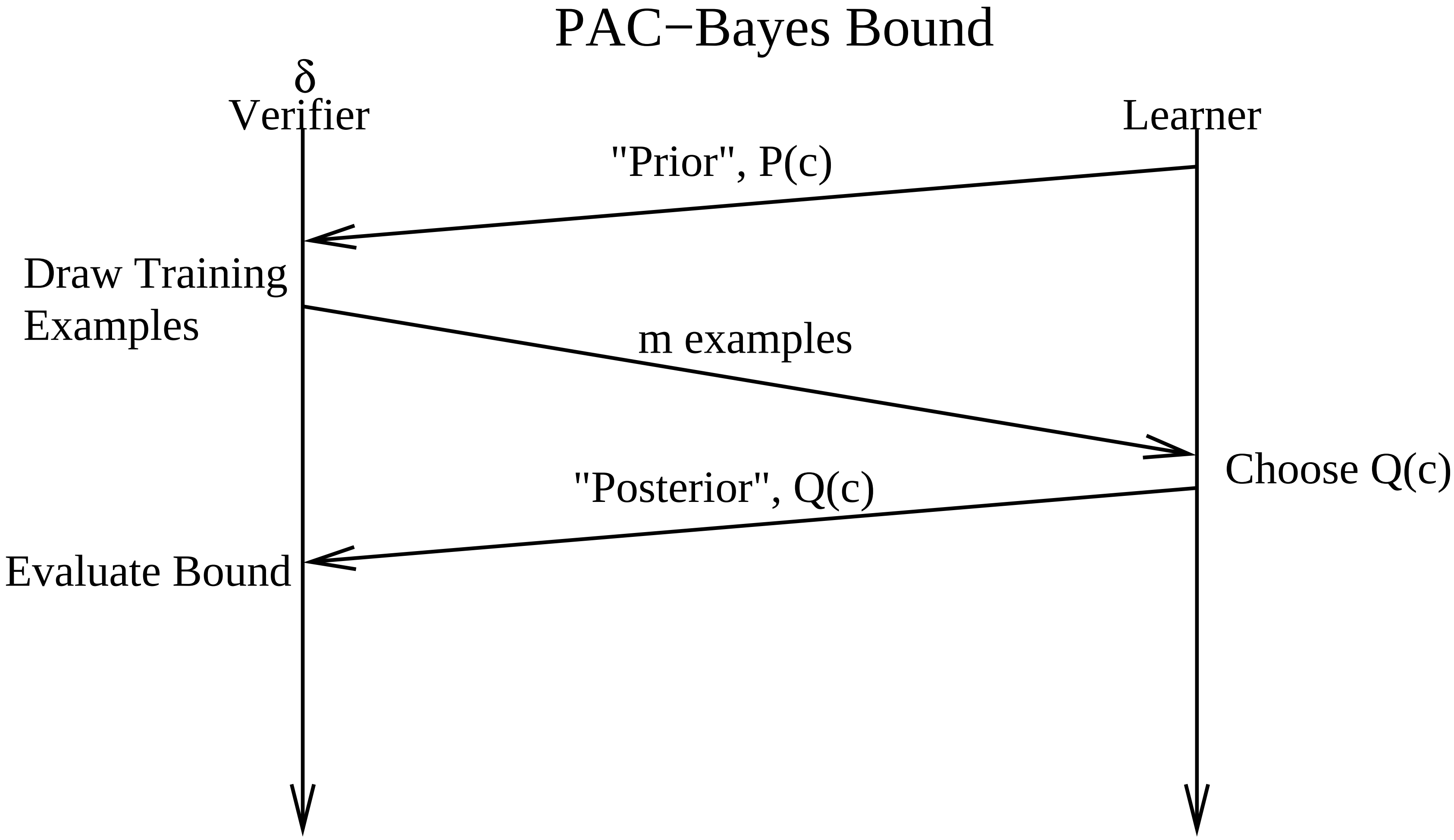
1. The Basic Model

2. The Test Set Bound

3. Occam's Razor Bound

4. PAC-Bayes Bound

5. Sample Compression Bound


PAC−Bayes Bound
δ
Verifier
Learner
"Prior", P(c)
Draw Training
Examples
m examples
Choose Q(c)
"Posterior", Q(c)
Evaluate Bound

average true error:

$$Q_D \equiv E_{c\sim Q} c_D$$

average train error:

$$\hat{Q}_S \equiv E_{c\sim Q} \hat{c}_S$$

Theorem: (PAC-Bayes Bound) For all "priors" $P(c)$ over the classifiers, c, for all $\delta \in (0,1]$:

$$\Pr_{S\sim D^m}\left(\forall Q(c):\ \mathsf{KL}\left(\hat{Q}_S \| Q_D\right) \leq \frac{\mathsf{KL}(Q\|P) + \ln \frac{m+1}{\delta}}{m}\right) \geq 1 - \delta$$

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

$\mathsf{KL}(Q\|P) = E_{c\sim Q} \ln \frac{Q(c)}{P(c)}$

Lemma 1: For all $P(c)$, For all $\delta \in (0, 1]$:

$$\Pr_{S \sim D^m}\left(E_{c \sim P} \frac{1}{\Pr_{S \sim D^m}(\hat{c}_S)} \leq \frac{m+1}{\delta}\right) \geq 1 - \delta$$

Proof:

$$\forall c \ \ E_{S \sim D^m} \frac{1}{\Pr_{S \sim D^m}(\hat{c}_S)} = \sum_{\frac{k}{m}} \Pr_{S \sim D^m}\left(\hat{c}_S = \frac{k}{m}\right) \frac{1}{\Pr_{S \sim D^m}\left(\hat{c}_S = \frac{k}{m}\right)} = m+1$$

$$\Rightarrow E_{S \sim D^m} E_{c \sim P} \frac{1}{\Pr_{S \sim D^m}(\hat{c}_S)} = m + 1$$

Use the Markov inequality ($X \geq 0$, $EX = \mu$, $\Rightarrow \Pr(X > \frac{\mu}{\delta}) \leq \delta$):

$$\forall P \ \ \Pr_{S \sim D^m}\left(E_{c \sim P} \frac{1}{\Pr_{S \sim D^m}(\hat{c}_S)} \leq \frac{m+1}{\delta}\right) \leq \delta$$

Lemma 2: For all $Q(c)$: $\frac{E_{c\sim Q} \ln \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)}}{m} \geq \mathrm{KL}(\hat{Q}_S || Q_D)$

Proof: $\frac{1}{m} E_{c\sim Q} \ln \frac{1}{\binom{m}{m\hat{c}_S} c_D^{m\hat{c}_S}(1-c_D)^{m(1-\hat{c}_S)}}$

$$= E_{c\sim Q}\left[\hat{c}_S \ln \frac{1}{c_D} + (1-\hat{c}_S)\ln\frac{1}{1-c_D}\right] - \frac{E_{c\sim Q} \ln \binom{m}{m\hat{c}_S}}{m}$$

Jensen's inequality (f concave $\Rightarrow Ef(X) \geq f(EX)$):

$$E_{c\sim Q}\left[\hat{c}_S \ln \frac{1}{c_D} + (1-\hat{c}_S)\ln\frac{1}{1-c_D}\right] \geq \hat{Q}_S \ln \frac{1}{Q_D} + \left(1-\hat{Q}_S\right)\ln\frac{1}{1-Q_D}$$

$$\frac{E_{c\sim Q} \ln \binom{m}{m\hat{c}_S}}{m} \leq \frac{E_{c\sim Q} \ln e^{mH(\hat{c}_S)}}{m} = \frac{E_{c\sim Q} mH(\hat{c}_S)}{m} \leq H\left(\hat{Q}_S\right)$$

$\Rightarrow$lemma

Proof (of PAC-Bayes bound)

Let

$$P_G(c) = \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)\, E_{c\sim P}\frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)}} P(c)$$

$$\Rightarrow 0 \le \mathsf{KL}(Q||P_G) = E_{c\sim Q} \ln \frac{Q(c)}{P(c)} \Pr_{S\sim D^m}(\hat{c}_S)\, E_{c\sim P} \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)}$$

$$= \mathsf{KL}(Q||P) - E_{c\sim Q} \ln \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)} + \ln E_{c\sim P} \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)}$$

$$\Rightarrow E_{c\sim Q} \ln \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)} \le \mathsf{KL}(Q||P) + \ln E_{c\sim P} \frac{1}{\Pr_{S\sim D^m}(\hat{c}_S)}$$

$+$ Lemma 1&2 $\Rightarrow$proof

Is the PAC-Bayes bound tight enough to be useful? YES

Application: true error bounds for Support Vector Machines.

Classifier form:

$$c(x) = \mathsf{sign}\left(\vec{w} \cdot \vec{x}\right)$$

PAC-Bayes Margin bound

$\bar{F}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ = cumulative distribution of a Gaussian

$Q(\vec{w}, \mu) = N(\mu, 1) \times N(0, 1)^{n-1}$ where first direction parallel to $\vec{w}$

$\gamma(\vec{x}, y) = \frac{y\vec{w}\cdot\vec{x}}{||\vec{w}||||\vec{x}||}$ = normalized margin

$\hat{Q}(\vec{w}, \mu)_S = E_{\vec{x},y\sim S} \bar{F}\left(\mu\gamma(\vec{x}, y)\right)$ = stochastic error rate

Corollary: (PAC-Bayes Margin Bound) For all distributions D, for all $\delta \in (0, 1]$, we have:

$$\Pr_{S\sim D^m}\left(\forall \vec{w}, \mu :\ \mathrm{KL}\left(\hat{Q}(\vec{w}, \mu)_S || Q(\vec{w}, \mu)_D\right) \leq \frac{\frac{\mu^2}{2} + \ln\frac{m+1}{\delta}}{m}\right) \geq 1 - \delta$$

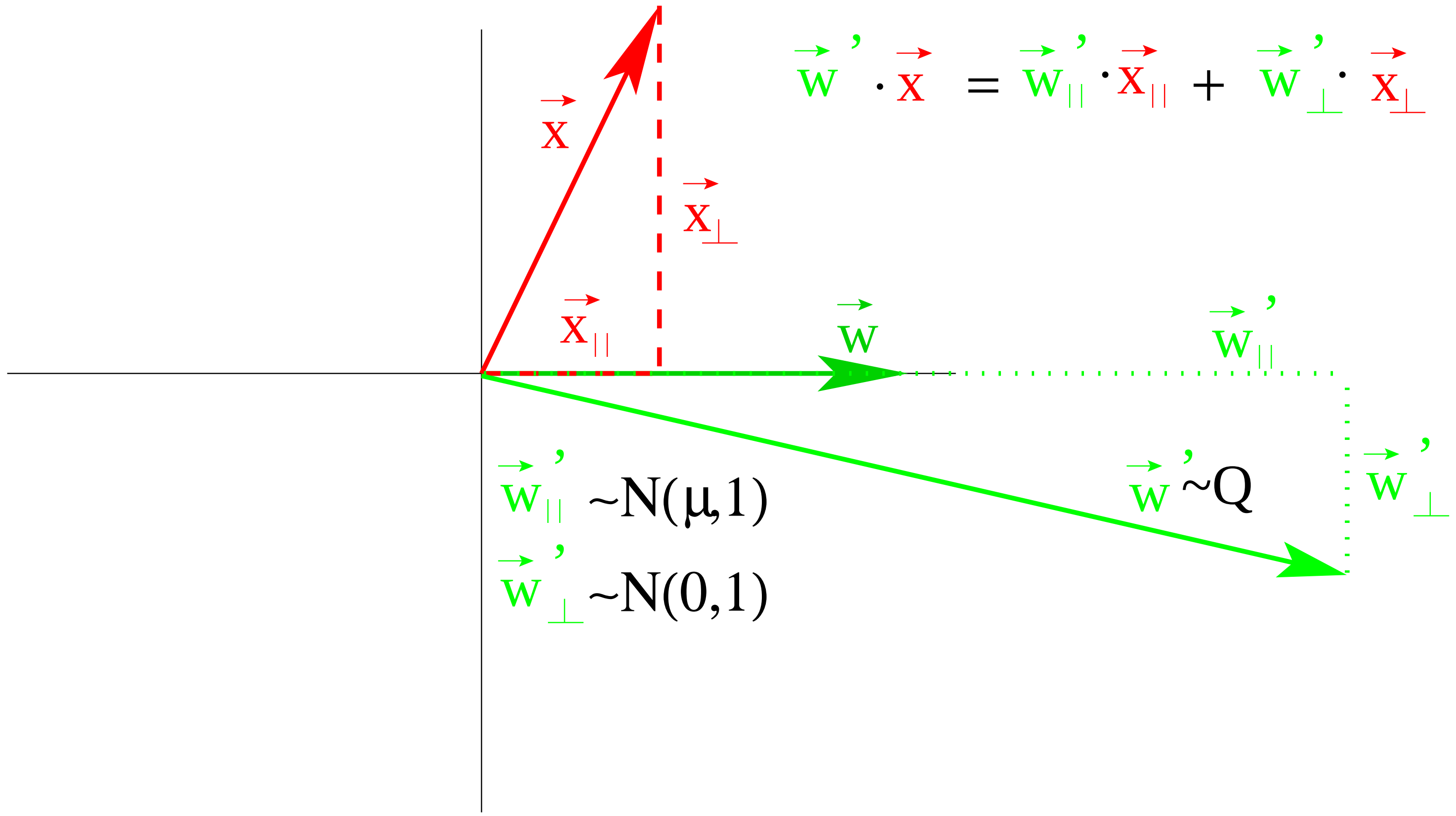
Proof: Set $P = N(0,1)^n$

$Q(\vec{w}, \mu) = N(\mu, 1) \times N(0,1)^{n-1}$ with first direction parallal to $\vec{w}$

Gaussian $\Rightarrow$ coordinate system reorientable

$$\Rightarrow \mathrm{KL}(Q||P) = \mathrm{KL}(N(0,1)^{n-1}||N(0,1)^{n-1}) + \mathrm{KL}(N(\mu,1)||N(0,1))$$

$$= \frac{\mu^2}{2}$$



$$\hat{Q}(\vec{w},\mu)_S = E_{\vec{x},y\sim S,\vec{w}'\sim Q(\vec{w},\mu)} I\left(y \neq \text{sign}\left(\vec{w}' \cdot \vec{x}\right)\right)$$

$$= E_{\vec{x},y\sim S} E_{w'_{||}\sim N(\mu,1)} E_{w'_{\perp}\sim N(0,1)} I\left(y(w'_{||}x_{||} + w'_{\perp}x_{\perp}) \leq 0\right)$$

PAC-Bayes Margin proof: the end

$$= E_{\vec{x},y\sim S} E_{z'\sim N(0,1)} E_{w'_\perp \sim N(0,1)} I\left(y\mu \leq -yz' - yw'_\perp \frac{x_\perp}{x_{||}} \right)$$

The sum of two Gaussians is a Gaussian $\Rightarrow$

$$= E_{\vec{x},y\sim S} E_{v\sim N\left(0,1+\frac{x_\perp^2}{x_{||}^2}\right)} I\left(y\mu \leq -yv \right)$$

$$= E_{\vec{x},y\sim S} E_{v\sim N\left(0,\frac{1}{\gamma(\vec{x},y)^2}\right)} I\left(y\mu \leq -yv \right)$$

$$= E_{\vec{x},y\sim S} \bar{F}\left(\mu\gamma(\vec{x},y) \right)$$

Application to SVM

SVM classifier:

$$c(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i k(x_i, x)\right)$$

k is a kernel $\Rightarrow \exists \vec{\Phi}: \quad k(x_i, x) = \vec{\Phi}(x_i) \cdot \vec{\Phi}(x)$ so:

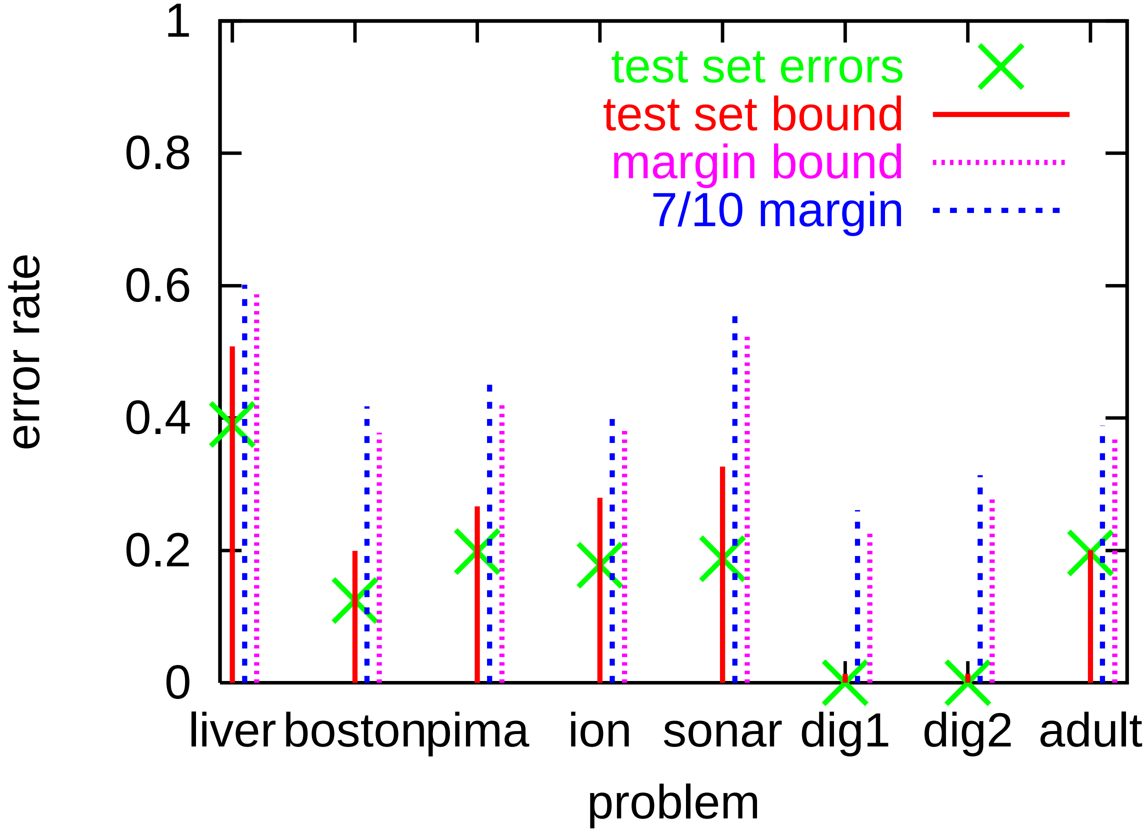
$\vec{w} \cdot \vec{x} = \sum_{i=1}^{m} \alpha_i k(x_i, x)$

$\vec{x} \cdot \vec{x} = k(x, x)$

$\vec{w} \cdot \vec{w} = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$

$$\Rightarrow \gamma(x, y) = \frac{\sum_{i=1}^{m} \alpha_i k(x_i, x)}{\sqrt{k(x, x) \sum_{i,j=1,1}^{m,m} \alpha_i \alpha_j k(x_i, x_j)}}$$

PAC-Bayes Margin Bound Results



Outline

1. The Basic Model

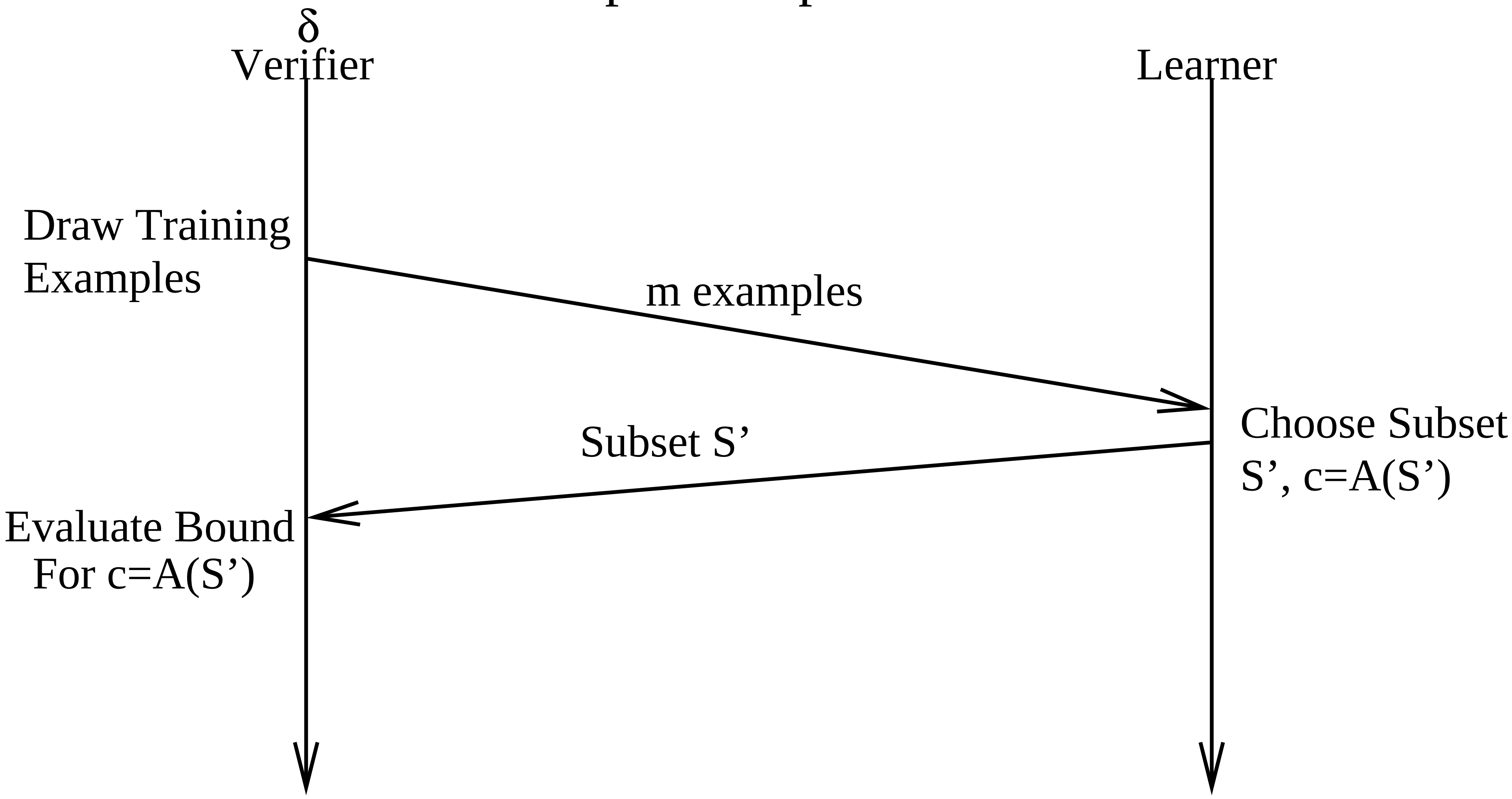
2. The Test Set Bound

3. Occam's Razor Bound

4. PAC-Bayes Bound

5. Sample Compression Bound

Sample Compression Bound



$A =$ learning algorithm

$S' \subseteq S =$ subset of training set.

The idea: if we knew S' in advance that $S - S'$ would act as a test set. We don't know S' in advance so the bound is looser.

Theorem: (Sample Compression Bound) For all $\delta \in (0,1]$, D, A:

$$\Pr_{S \sim D^m}\left(\forall S' \subseteq S \text{ with } c = A(S') :\ \ c_D \leq \overline{\mathrm{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S'|}}\right)\right) \geq 1-\delta$$

Proof: (Note that S' is unambiguous before S is drawn)

$$\forall S' \subseteq S \text{ with } c = A(S') : \Pr_{S \sim D^m}\left(c_D \leq \overline{\mathrm{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S-S'|}}\right)\right) \geq 1 - \frac{\delta}{m\binom{m}{|S-S'|}}$$

Negate to get:

$$\forall S' \subseteq S \text{ with } c = A(S') : \Pr_{S \sim D^m}\left(c_D > \overline{\mathrm{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S-S'|}}\right)\right) < \frac{\delta}{m\binom{m}{|S-S'|}}$$

Use union bound ($\Pr(A \text{ or } B) \leq \Pr(A) + \Pr(B)$) over each S'

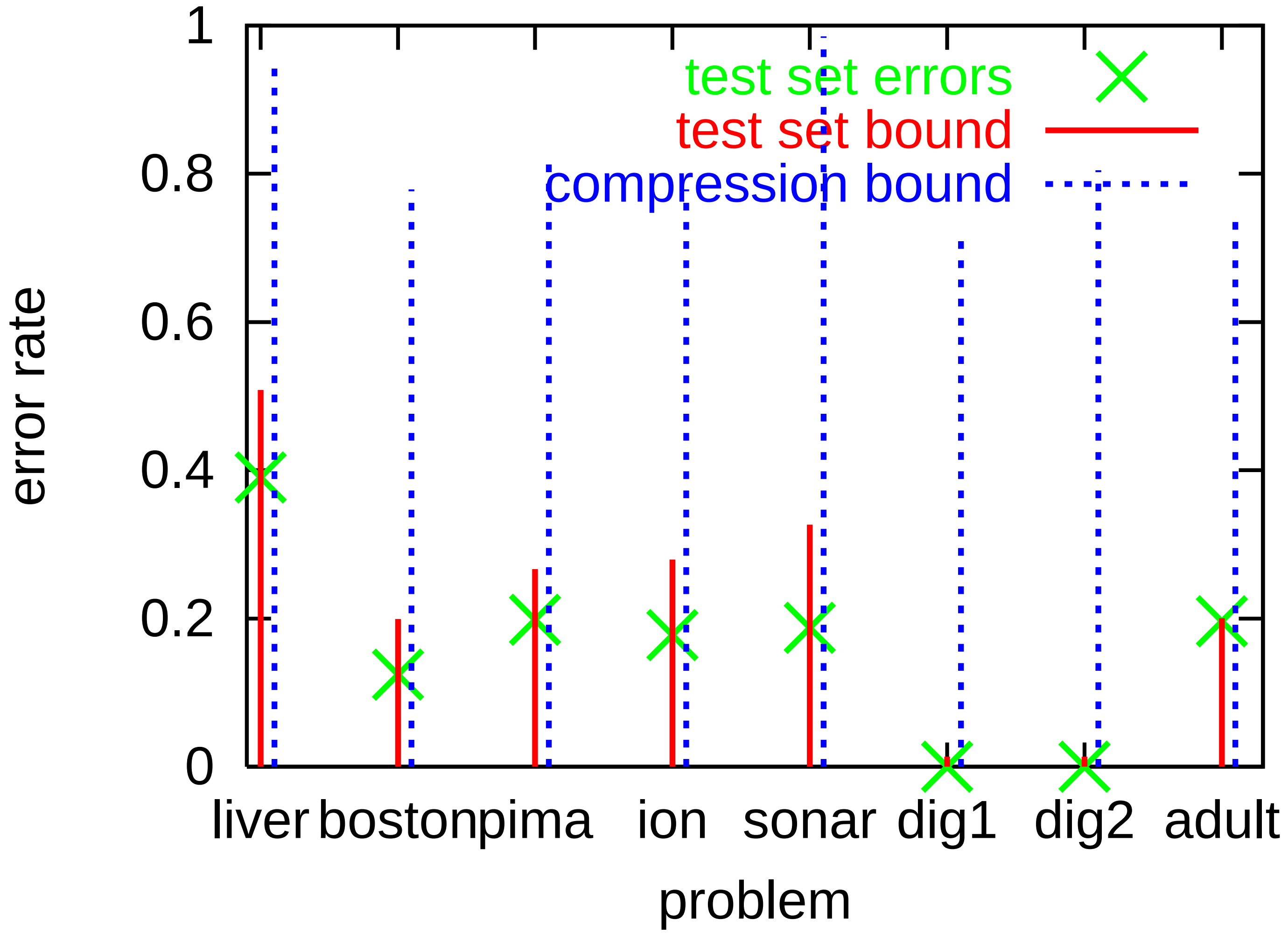
$$\Rightarrow \Pr_{S \sim D^m}\left(\exists S' \subseteq S \text{ with } c = A(S') :\ c_D > \overline{\mathrm{Bin}}\left(\hat{c}_{S-S'}, \frac{\delta}{m\binom{m}{|S-S'|}}\right)\right) < \delta$$

Negate again $\Rightarrow$ theorem

Application: Support Vector Machines

If S'= set of support vectors than $A(S') = A(S)$.

How well does the sample compression bound work with a support vector machine?


test set errors
test set bound
compression bound
1
0.8
0.6
0.4
0.2
0
error rate
liver
boston
pima
ion
sonar
dig1
dig2
adult
problem

Answer: Not so well.

Why not?

The SVM learning algorithm achieves 'incidental' sparsity rather than optimizing for it explicitly (in contrast to the margin).

Note work by Mario Marchand and John Shawe-Taylor using Sample Compression bound variant $\Rightarrow$ good quantitative results

Conclusion

Use a confidence interval best for comparison.

Amongst estimation methods, test set or K-fold cross validation are reasonable.

Some code at:

http://www-2.cs.cmu.edu/~jcl/programs/bound/bound.html