Overfitting and Sample Complexity

Columbia COMS4771, 2007

John Langford

Outline

1. The Basic Model
2. The Test Set Bound
3. Occam's Razor Bound

Model: Definitions

X = input space

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

$c : X \to Y$ = classifier

Model: Basic Assumption

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

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

Model: Derived quantities

The thing we want to know:

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

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:

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

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

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

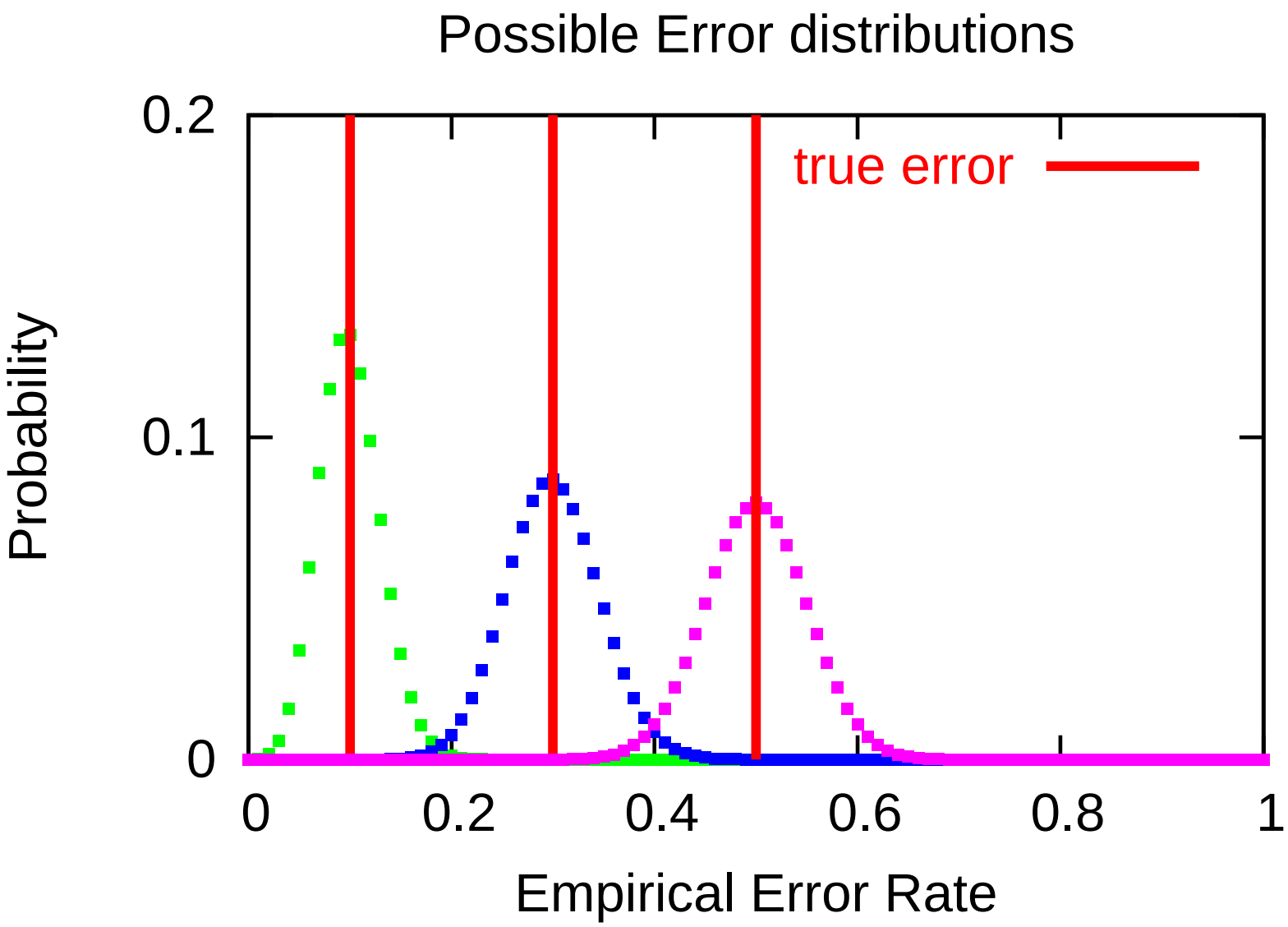
Model: Basic Observations

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

A: A Binomial.

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

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


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

Model: basic quantities

We use the cumulative:

$$\begin{aligned} \mathsf{Bin}\,(m,k,c_D) &= \mathsf{Pr}_{S\sim D^m}\,(\widehat{c}_S \le k|\, c_D) \\ &= \sum_{i=0}^{k} \binom{m}{i} c_D^i (1-c_D)^{m-i} \end{aligned}$$

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

Model: basic quantities

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

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

$=$ 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

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

What do sample complexity say about this method?

Test Set Bound: Theorem

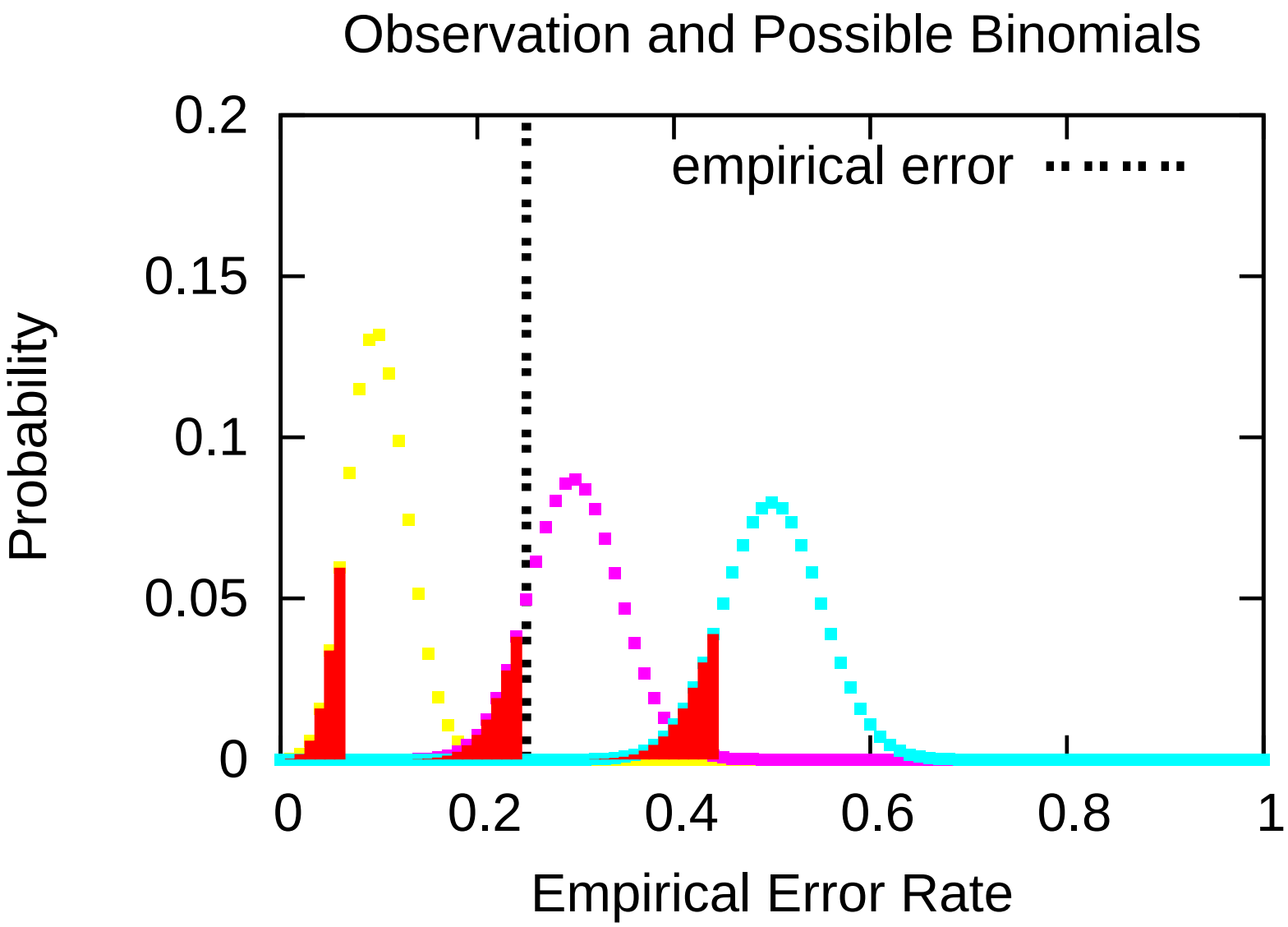
Theorem: (Test Set Bound) For all classifiers c, for all D, for all $\delta \in (0, 1]$:

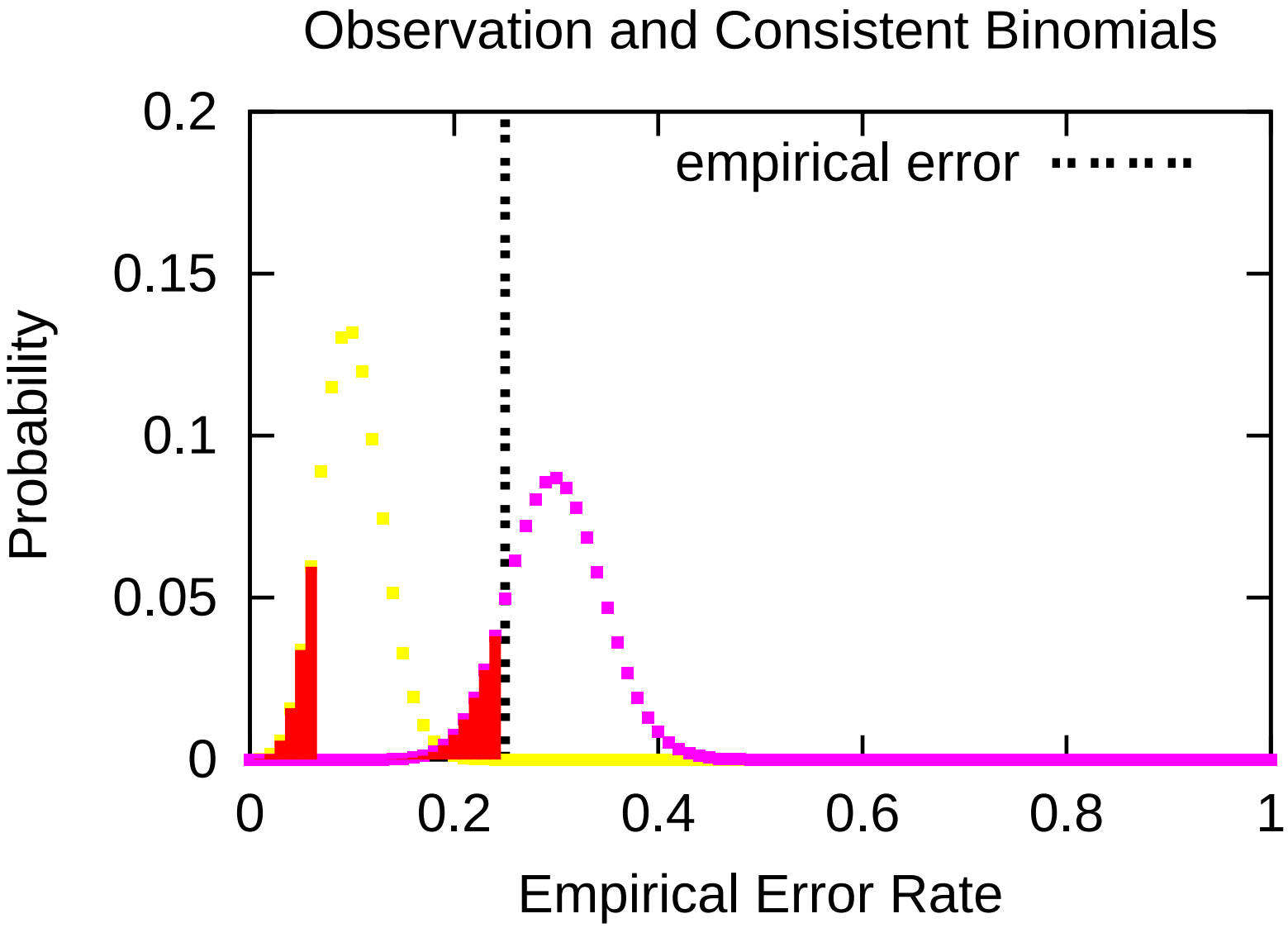
$$\Pr_{S \sim D^m} \left(c_D \leq \overline{\mathsf{Bin}} \left(m, \hat{c}_S, \delta \right) \right) \geq 1 - \delta$$

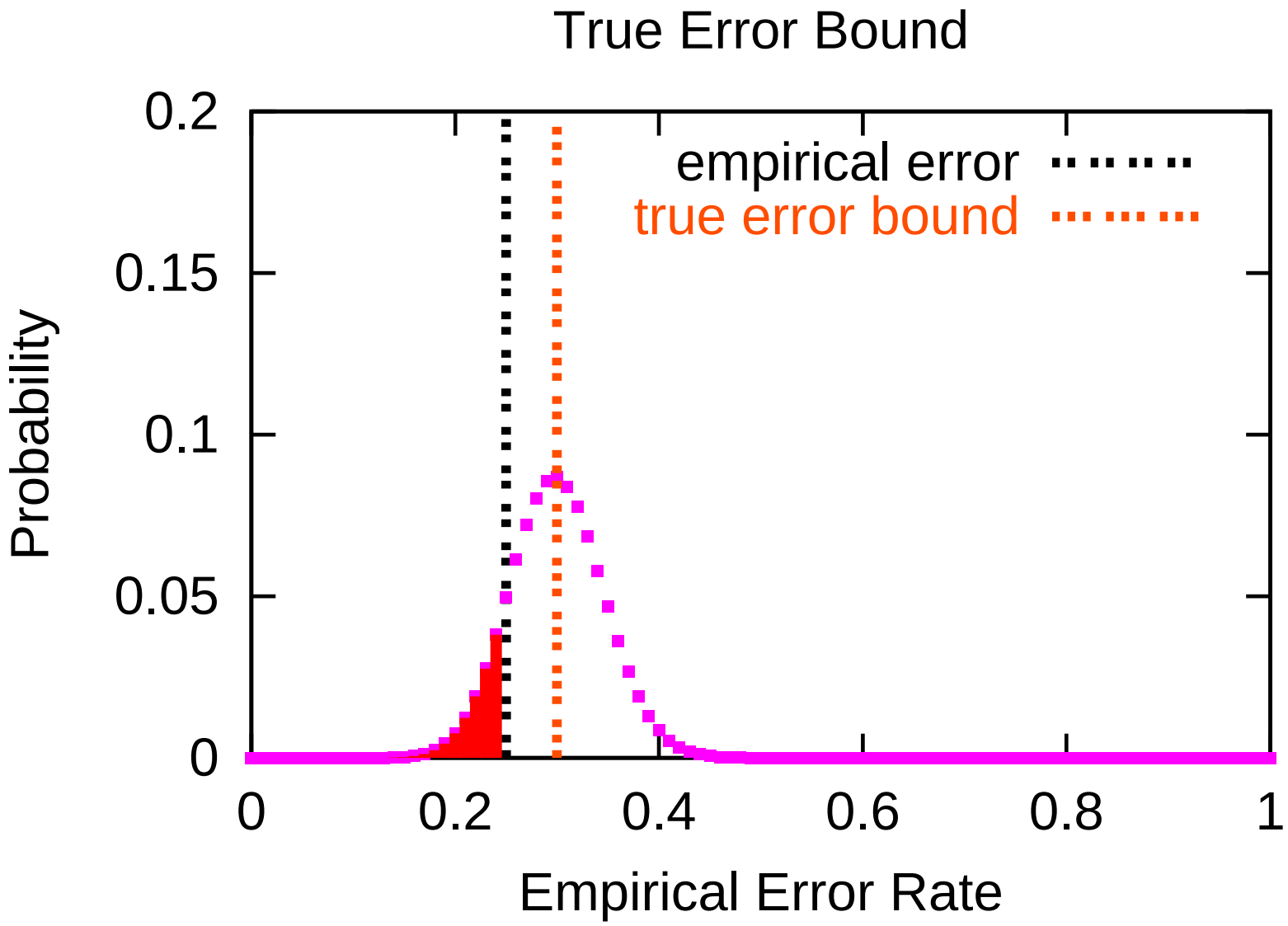
World's easiest proof: (by contradiction).

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

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


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


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


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

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 Test Set Bound mean?

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

$$\Pr_{S \sim D^m} \left(\mathsf{KL}\left(\frac{\widehat{c}_S}{m} || c_D \right) \leq \frac{\ln \frac{1}{\delta}}{m} \right) \geq 1 - \delta$$

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

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

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

Proof: Use the Chernoff approximation. Full details in the notes.

Test Set Bound: Example

Suppose $\delta = 0.1$

Suppose $m = 100$

Suppose $\widehat{c}_S = 2$

Square root Chernoff bound: $\Rightarrow c_D \in [-0.102, 0.142]$

Exact calculation $\Rightarrow c_D \in [0.0045, 0.0616]$

Outline

1. The Basic Model

2. The Test Set Bound

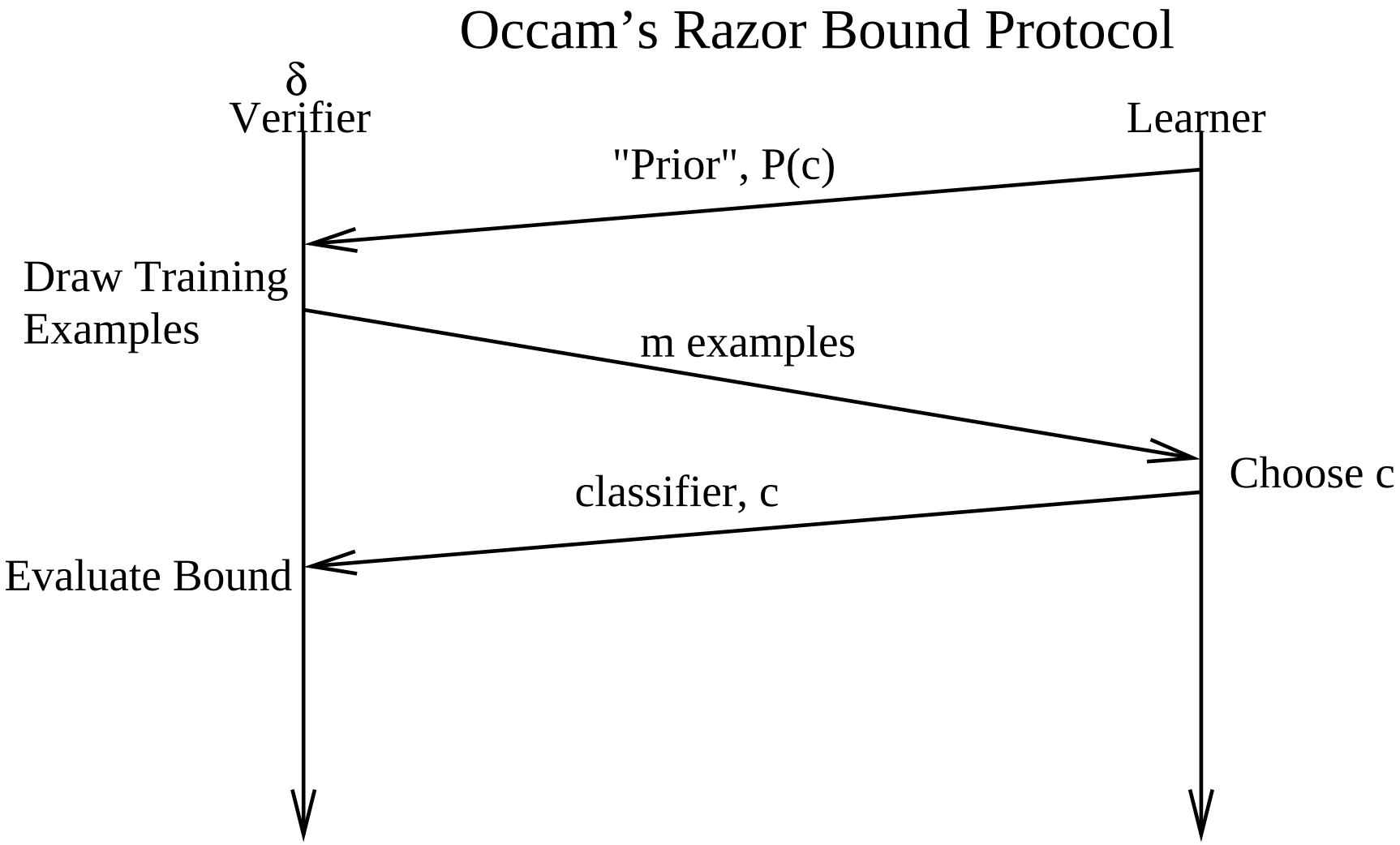
3. Occam's Razor Bound

Training Set Bounds in General

- Sometimes a 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
δ
Verifier
Learner
"Prior", P(c)
Draw Training
Examples
m examples
Choose c
classifier, c
Evaluate Bound

Occam's Razor Bound

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

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

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

Corollary: For all $P(c)$, for all D, for all $\delta \in (0, 1]$:

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

Occam's Razor Bound: Proof

Test set bound $\Rightarrow$

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

Occam's Razor Bound: Proof

Test set bound $\Rightarrow$

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

Negate to get:

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

Occam's Razor Bound: Proof

Test set bound $\Rightarrow$

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

Negate to get:

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

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

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

Occam's Razor Bound: Proof

Test set bound $\Rightarrow$

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

Negate to get:

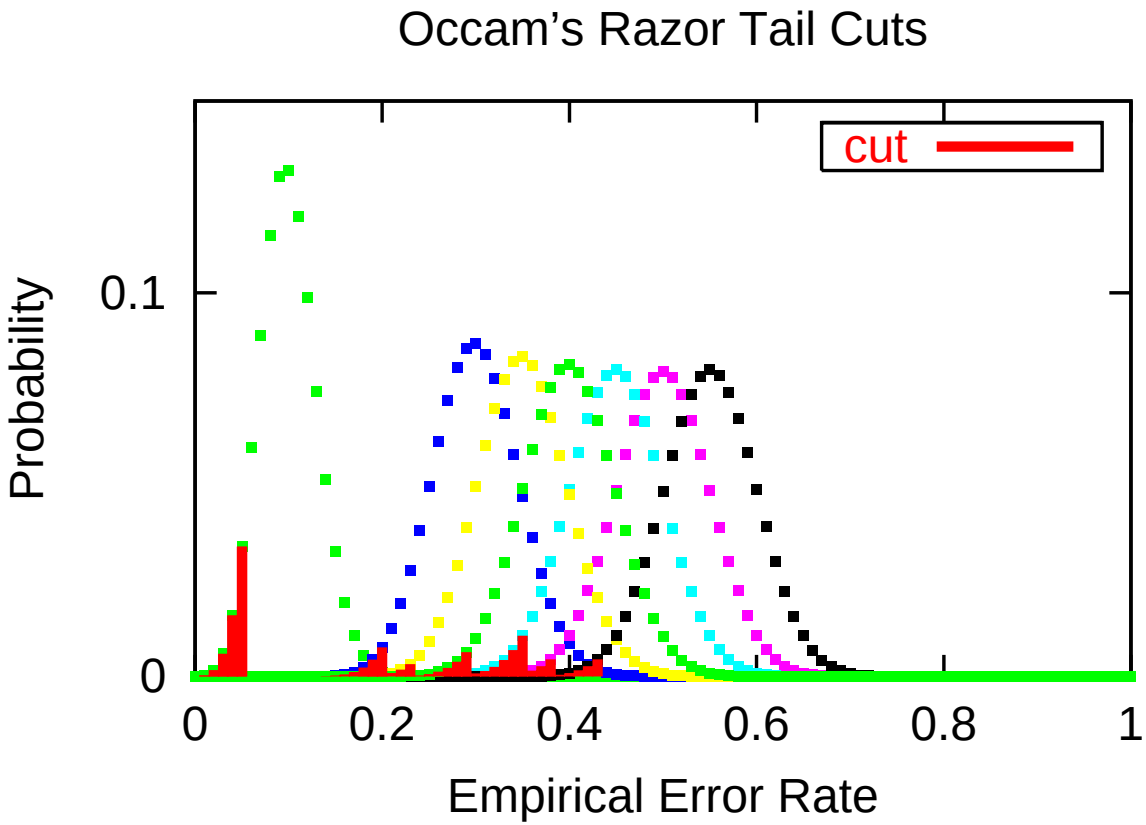
$$\forall c \quad \Pr_{S \sim D^m} \left(c_D > \overline{\mathsf{Bin}}\left(m, \hat{c}_S, \delta P(c)\right) \right) < \delta P(c)$$

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

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

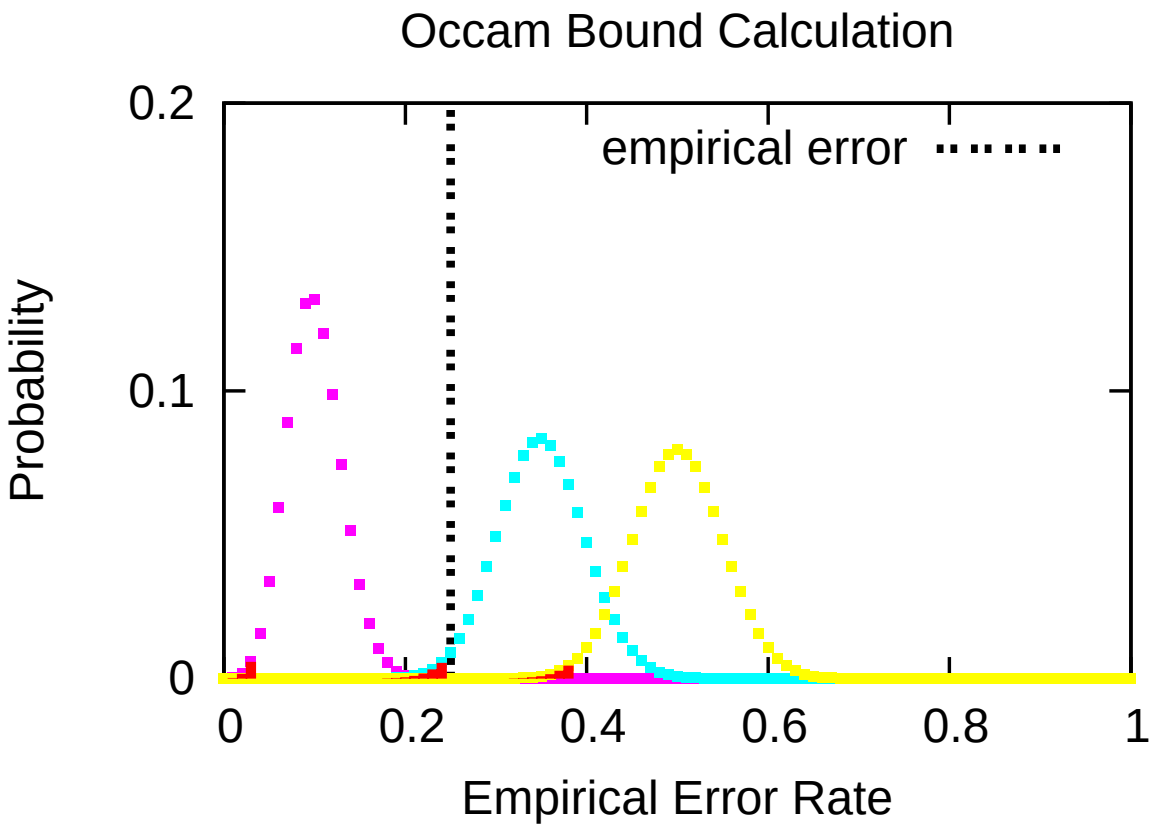
Negate again to get proof.

Next: Graphical proof

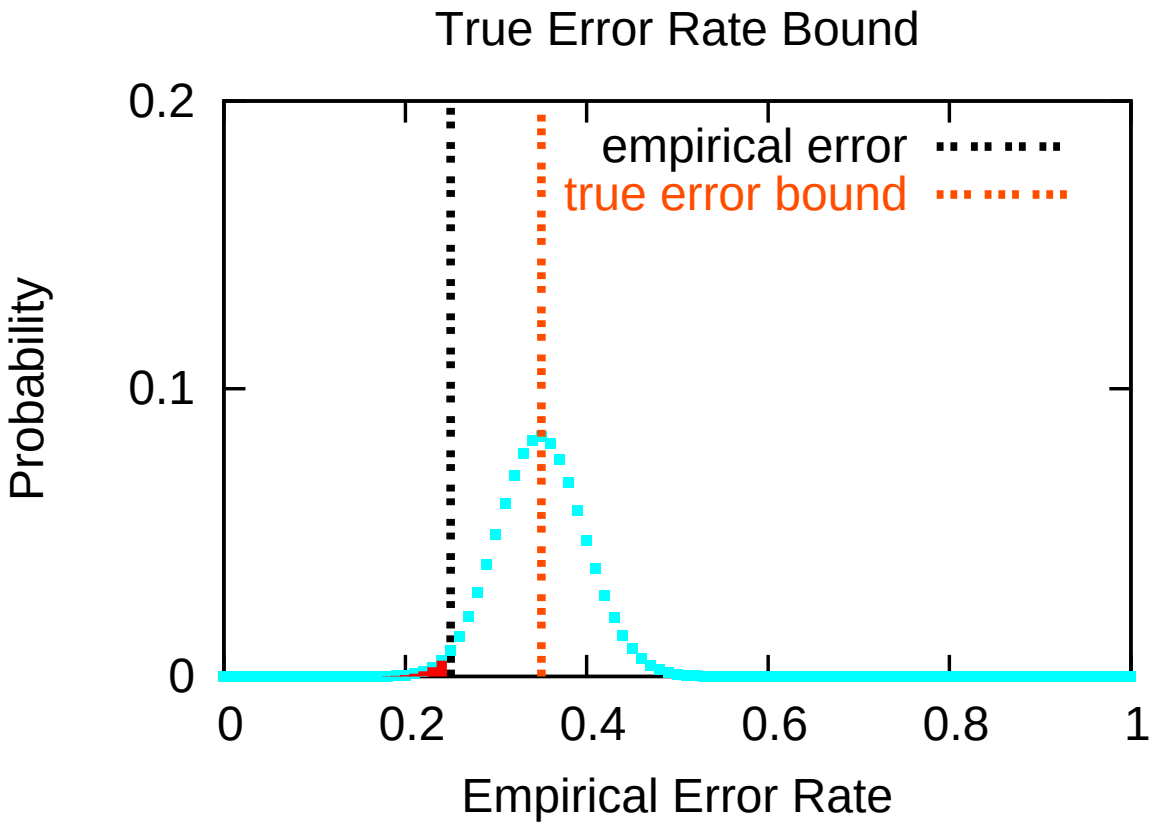


Each classifier is a Binomial with a different size tail cut.

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.

Occam's Razor Bound: Example

Suppose $\delta = 0.1$

Suppose $m = 100$

Suppose $P(c) = 0.1$

Suppose $\widehat{c}_S = 2$

Square root Chernoff $\Rightarrow c_D \in [-0.143, 0.183]$

Exact calculation $\Rightarrow c_D \in [0.001, 0.089]$

Conclusion

1. A real confidence interval to compare classifiers is good.
2. Test set bound very simple.
3. Train set bounds tell you something about how to design an algorithm, but are somewhat loose also.

Code for bound calculation at:

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

Midterm Thursday!