Constructing and Using Tight Sample Complexity Bounds

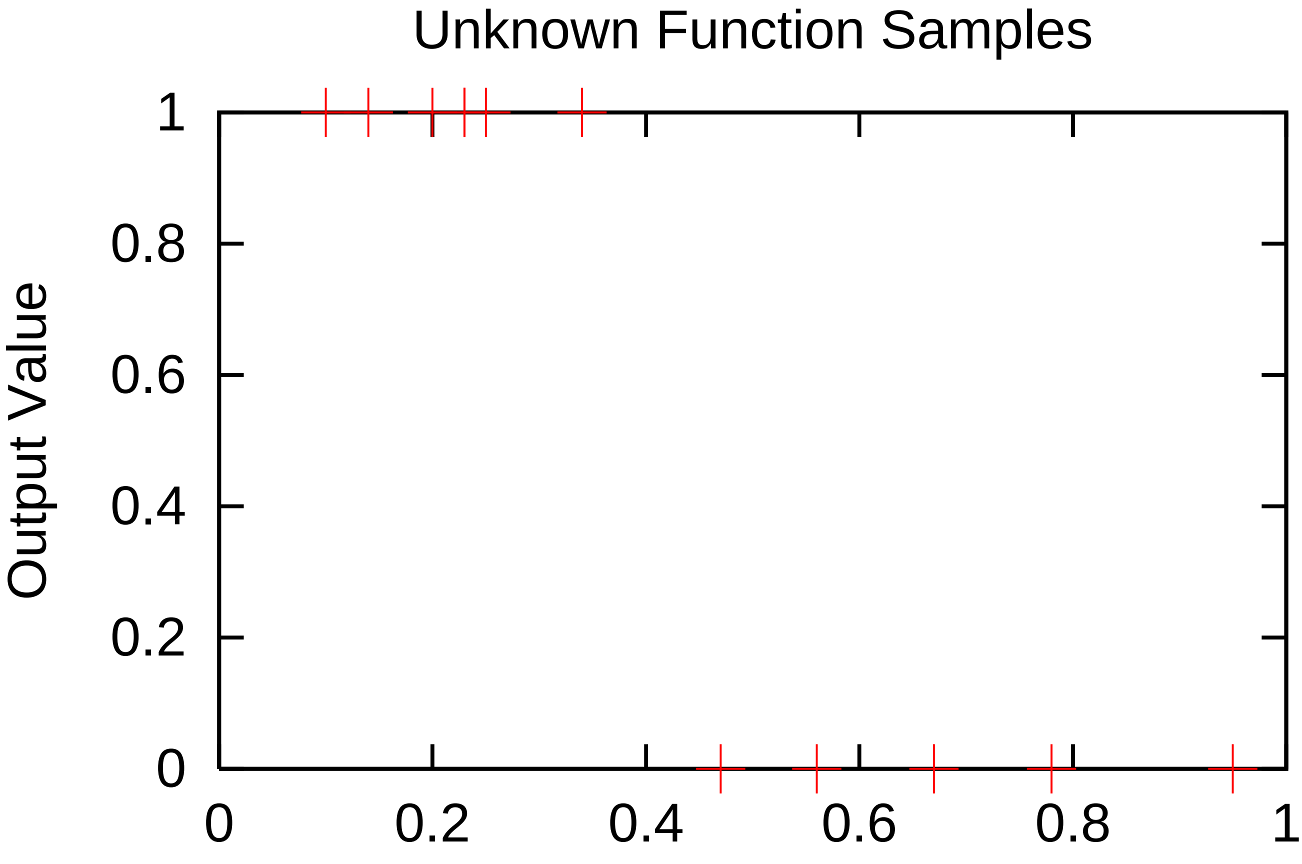
John Langford

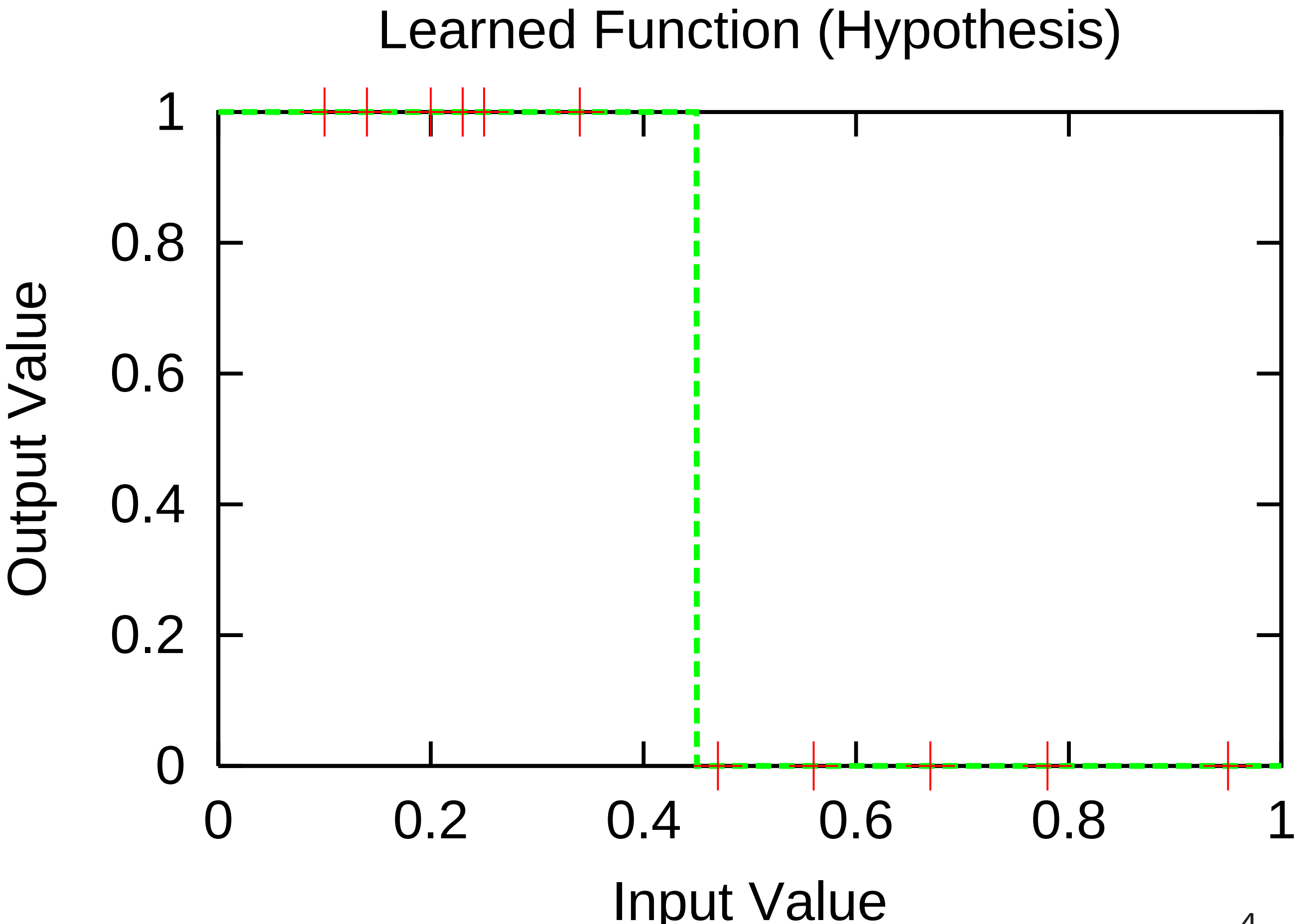
The Machine Learning Problem:

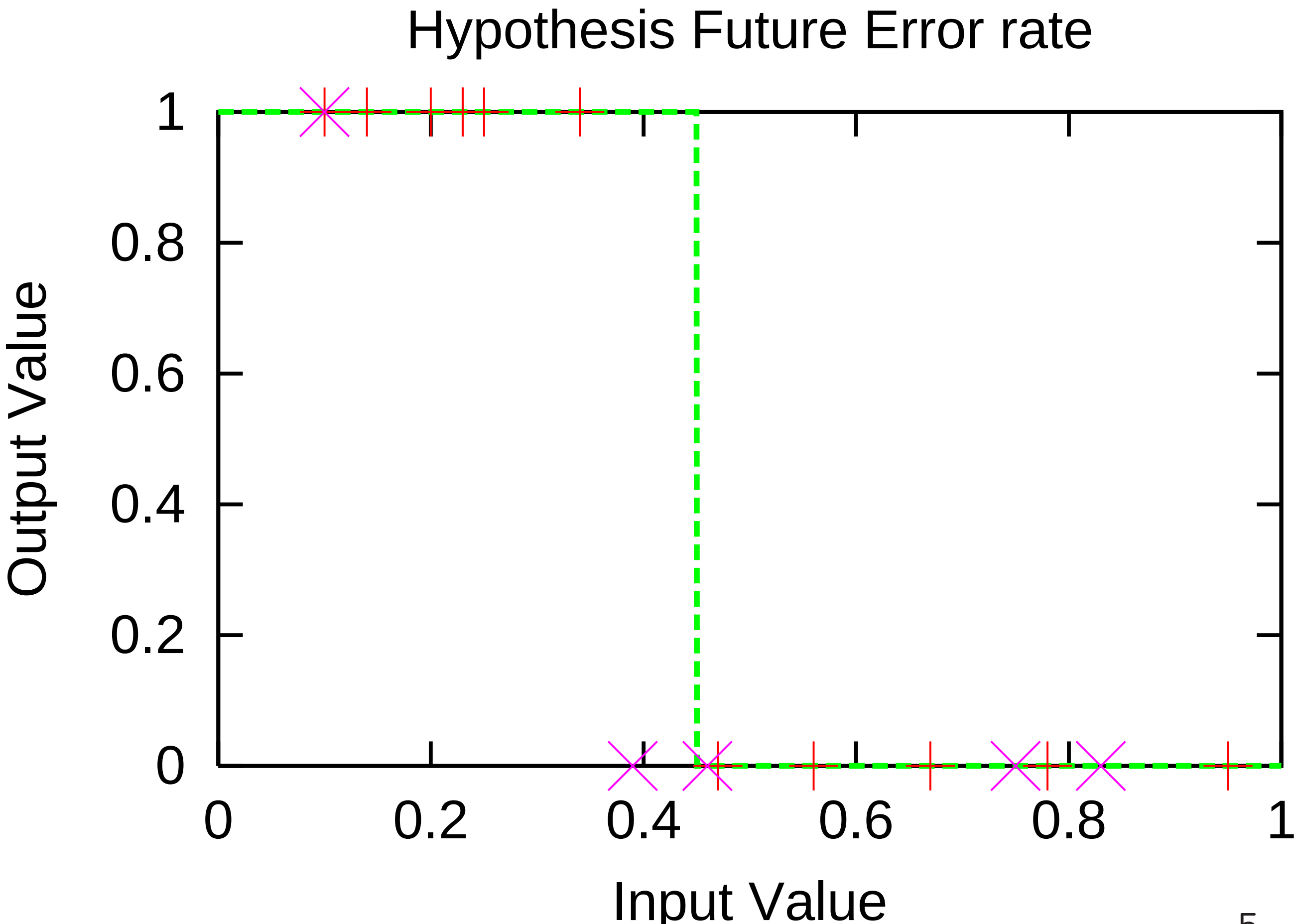
1. Get (input,output) pairs from unknown function

2. Guess the unknown function

3. ... to minimize errors on future inputs


Unknown Function Samples
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
Output Value
Input Value


Learned Function (Hypothesis)
Output Value
Input Value
0
0.2
0.4
0.6
0.8
1


Hypothesis Future Error rate
Output Value
Input Value
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

1. *What can you do with a future error bound?*

2. What is a Sample Complexity Bound?

3. How can you make one?

4. How well do they work?

What can you do with a future error bound?

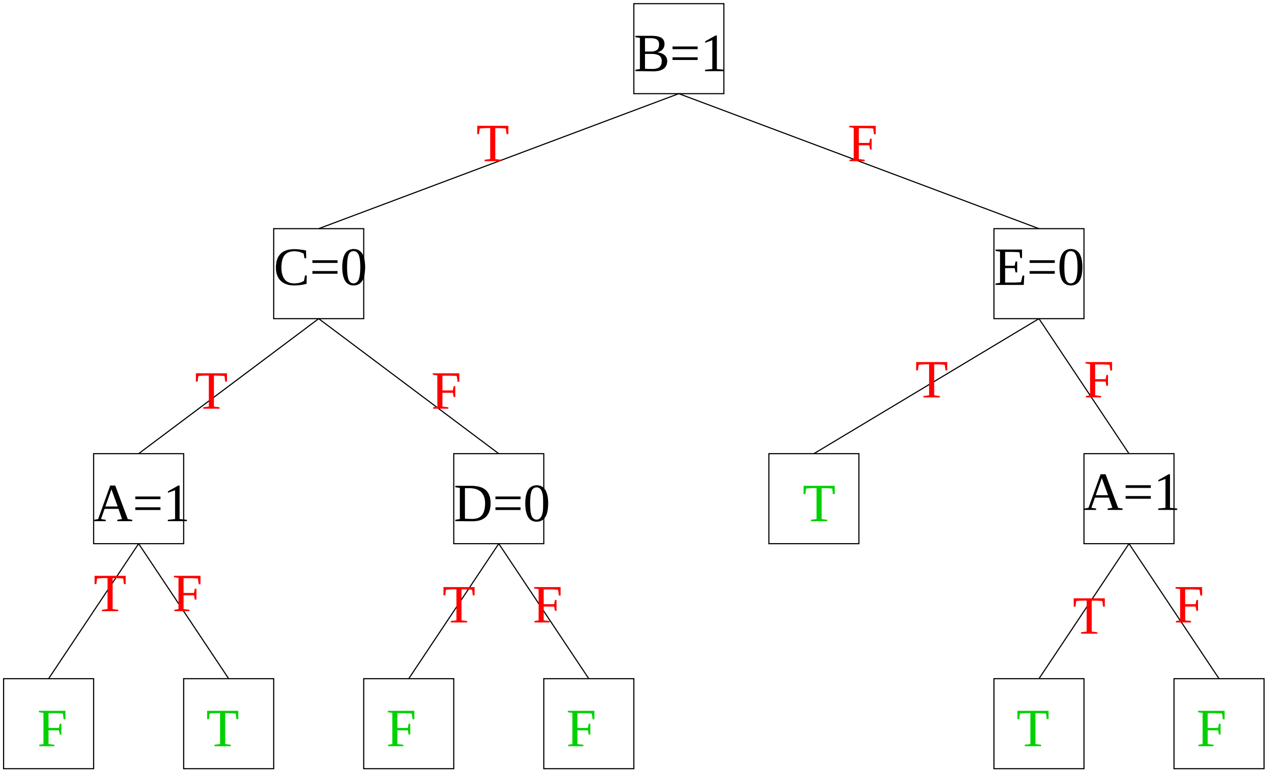
1. Learning algorithm design (done by many)

2. Choose the learning algorithm (sometimes done elsewhere)

3. Twiddle external knobs (done here)

4. Make external knobs internal (done here)

What are the knobs on a Decision Tree?



- When should I prune?

- What set of decisions should I use?

- How should I decide which test to use?

1. What can you do with a future error bound?

2. *What is a Sample Complexity Bound?*

3. How can you make one?

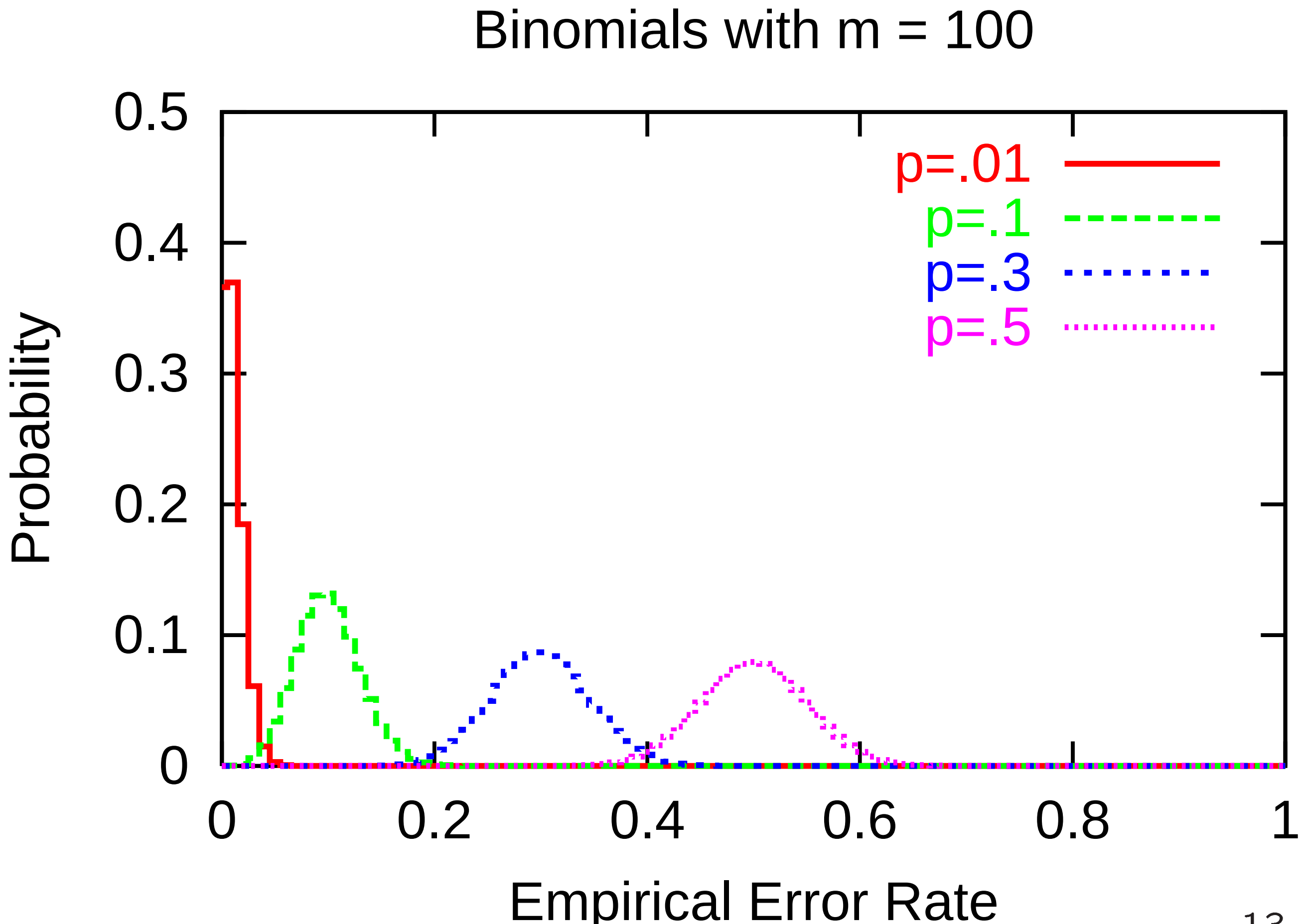
4. How well do they work?

Our Model

- (unknown) distribution D
- Draw m samples (x, y) independently from D

Terminology

- Hypothesis $= h$
- Samples $= S$
- True Error rate $e(h) = \mathrm{Pr}_D(h(x) \neq y)$
- Empirical Error rate $\widehat{e}(h) = \mathrm{Pr}_S(h(x) \neq y)$


Binomials with m = 100
p=.01
p=.1
p=.3
p=.5
Probability
Empirical Error Rate
0
0.1
0.2
0.3
0.4
0.5
0
0.2
0.4
0.6
0.8
1

The problem

1. Most hypotheses have true error near 0.5.

2. There are *many* hypotheses - like 2^{2^n}.

3. What is the probability that a bad one looks good?

We want to minimize the true error rate.

But this is sometimes inherently impossible.

Instead: minimize the gap between true error and empirical error.

Sample complexity bound:

How many samples are required to guarantee that the empirical error rate is near the true error rate?

1. What can you do with a future error bound?
2. What is a Sample Complexity Bound?
3. *How can you make one?*
4. How well do they work?

Sample Complexity Bounds

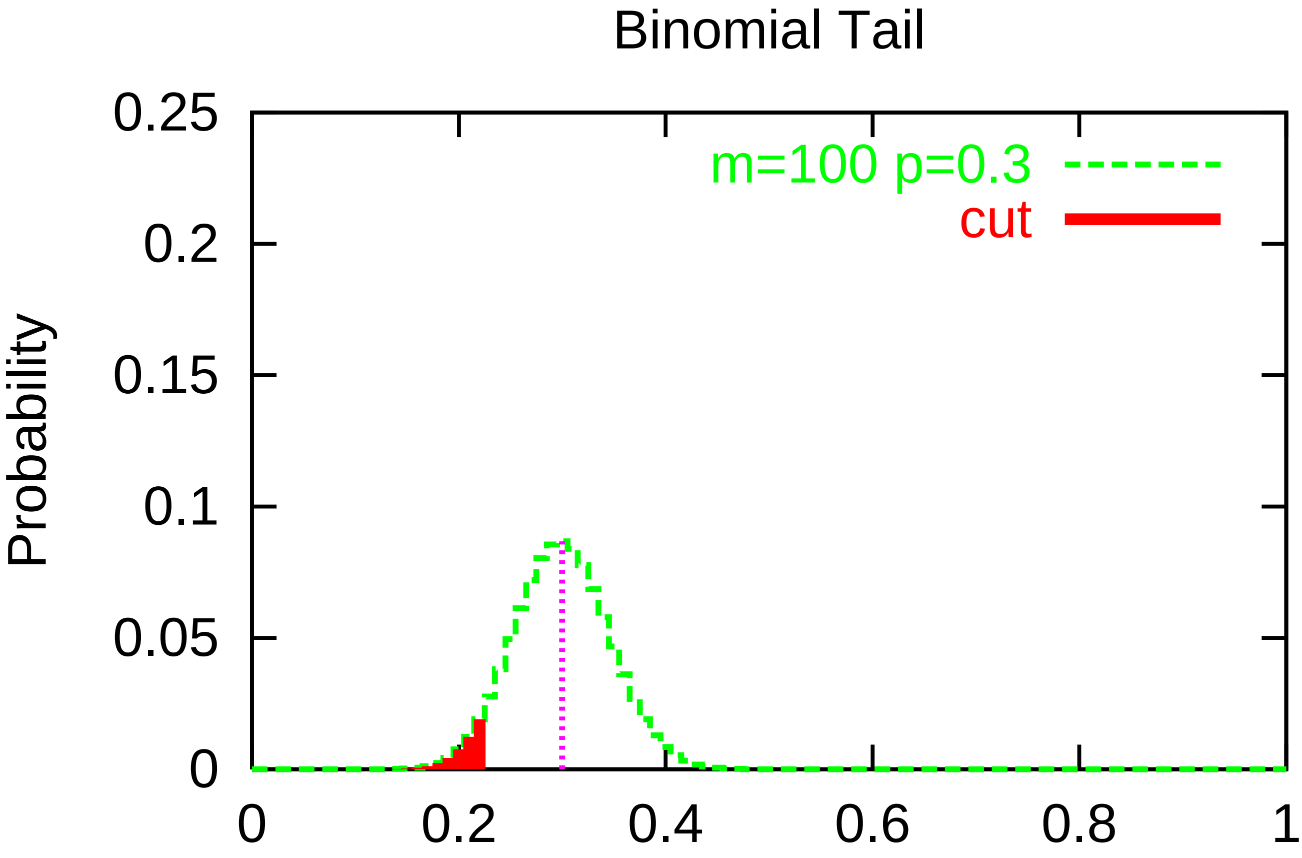
- Holdout: Used in practice (BKL99 for variants)
- Uniform: Way too loose
- Nonuniform: Occam's Razor (BEHW87), SRM (VC74,V95), Luckiness (SBWA96), Microchoice (LB99), etc...
- Double tail: Computable Shell (LM00)

How can you make a Bound?

- $p =$ bias of a coin
- $\hat{p} =$ observed head rate in m flips

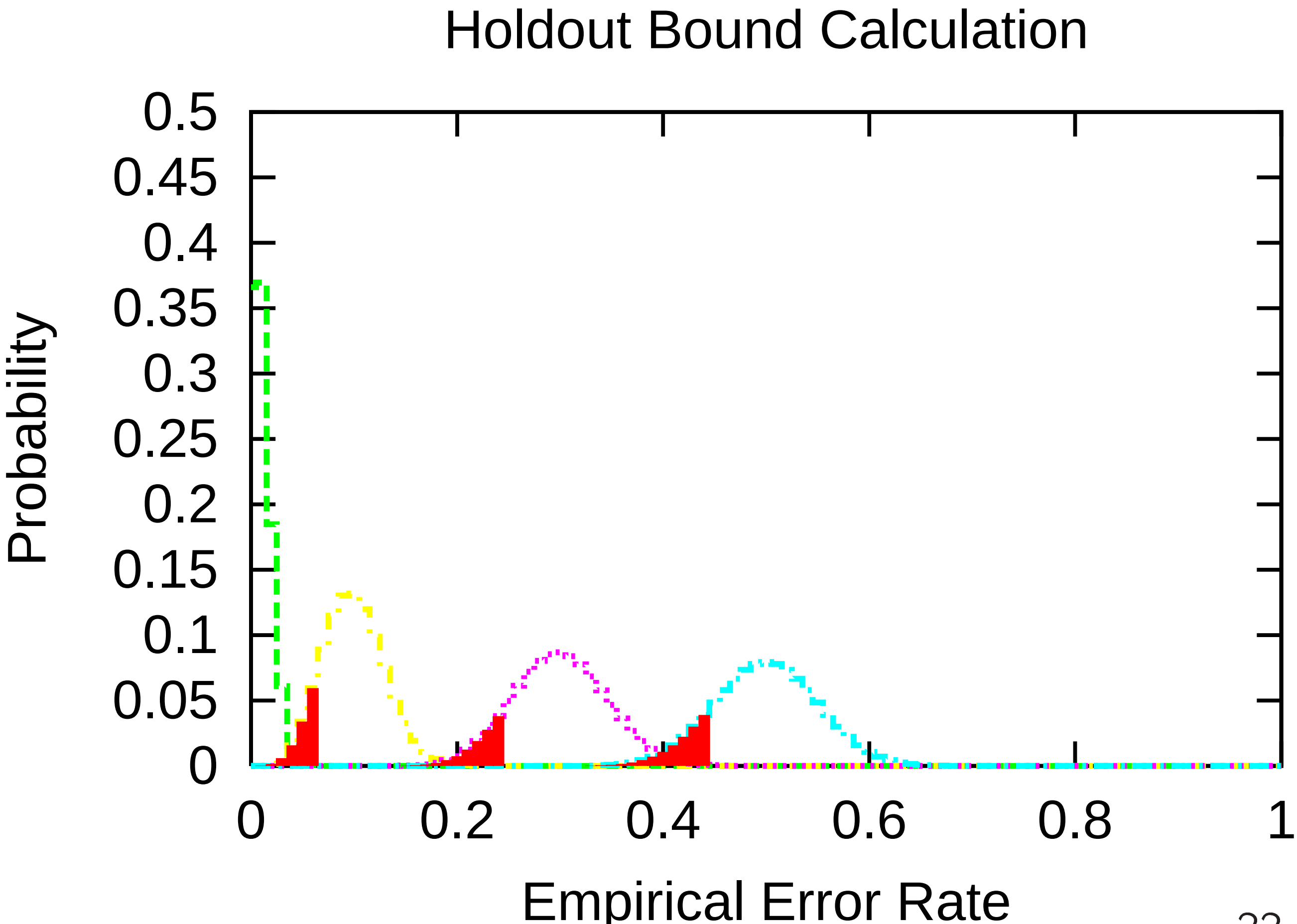
$$\Pr(p - \hat{p} > \epsilon) = \sum_{i=0}^{\lfloor m(p-\epsilon) \rfloor} \binom{m}{i} p^i (1-p)^{m-i}$$

$$\leq e^{-2m\epsilon^2}$$


Binomial Tail
m=100 p=0.3
cut
Probability
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
Empirical Error Rate

Q: Given the tail is cut, what is our bound?

A: The largest true error that did not cut the observation.


Holdout Bound Calculation
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Probability
0
0.2
0.4
0.6
0.8
1
Empirical Error Rate

We can verify if we have learned.

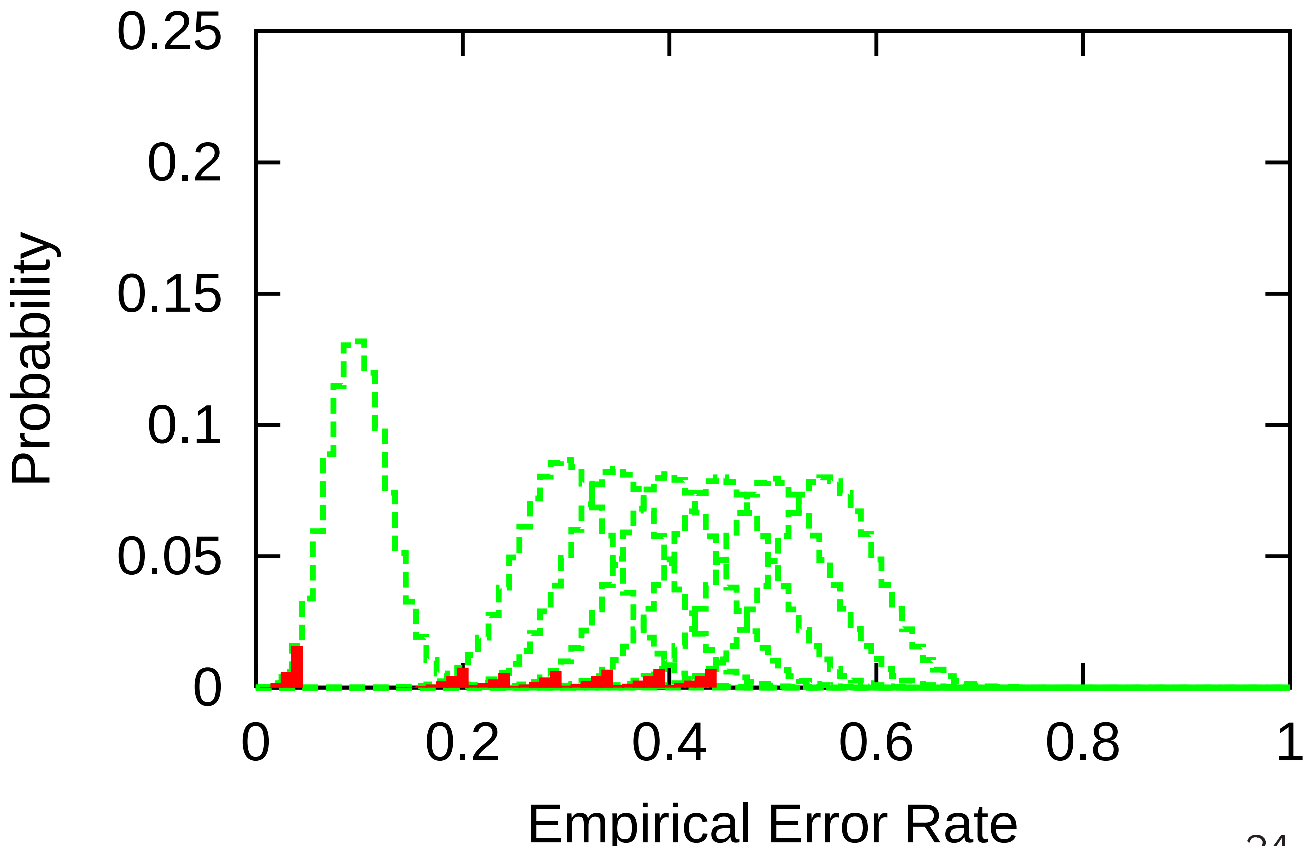
But: learning on one hypothesis is degenerate.

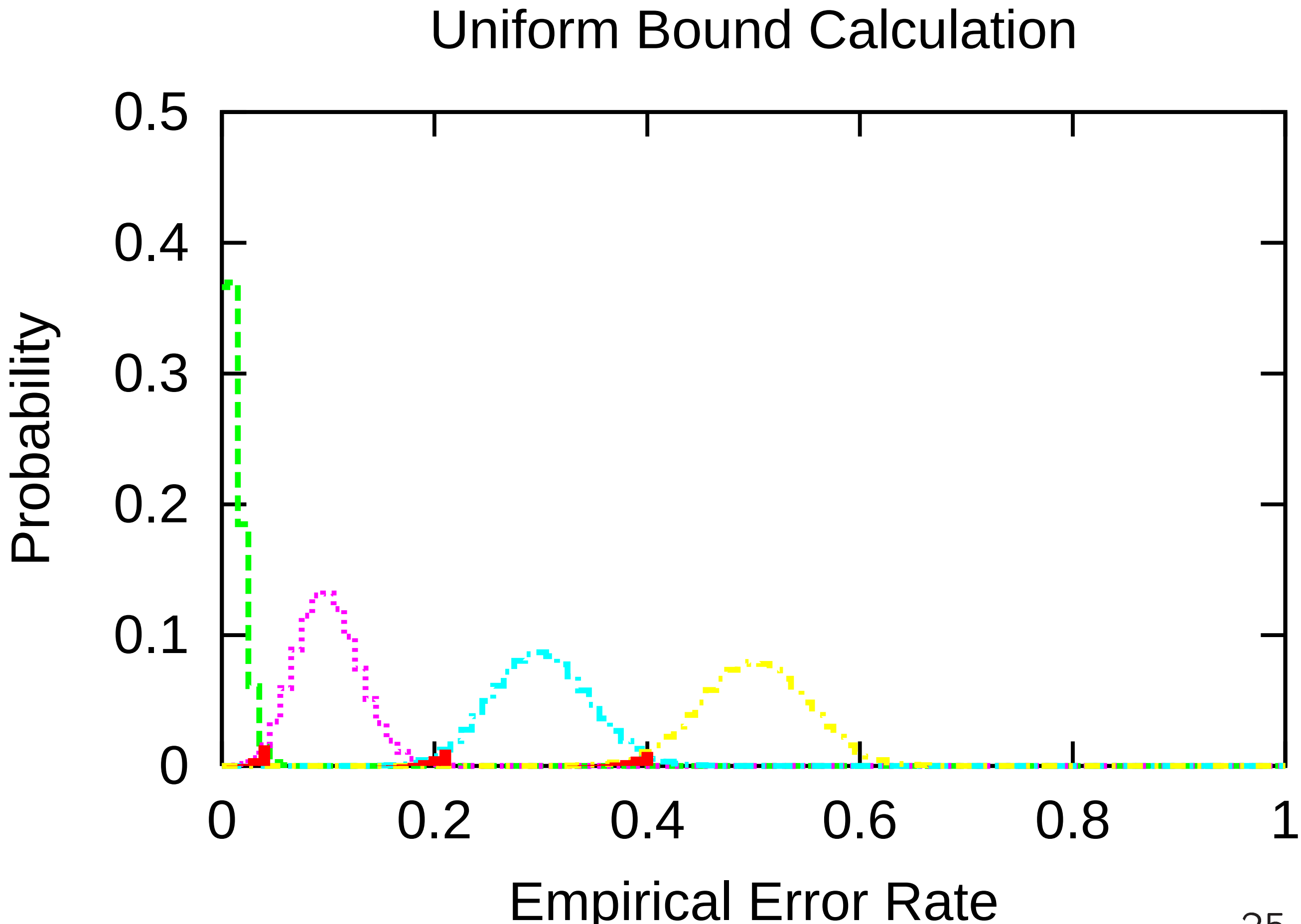
Q: What if there are multiple hypotheses?

A: Cut all tails and add up the probability of error

$$\Pr(\exists h \in H : e(h) - \widehat{e}(h) > \epsilon) \leq |H| e^{-2m\epsilon^2}$$

Uniform Tail Cuts




Uniform Bound Calculation
Probability
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
Empirical Error Rate

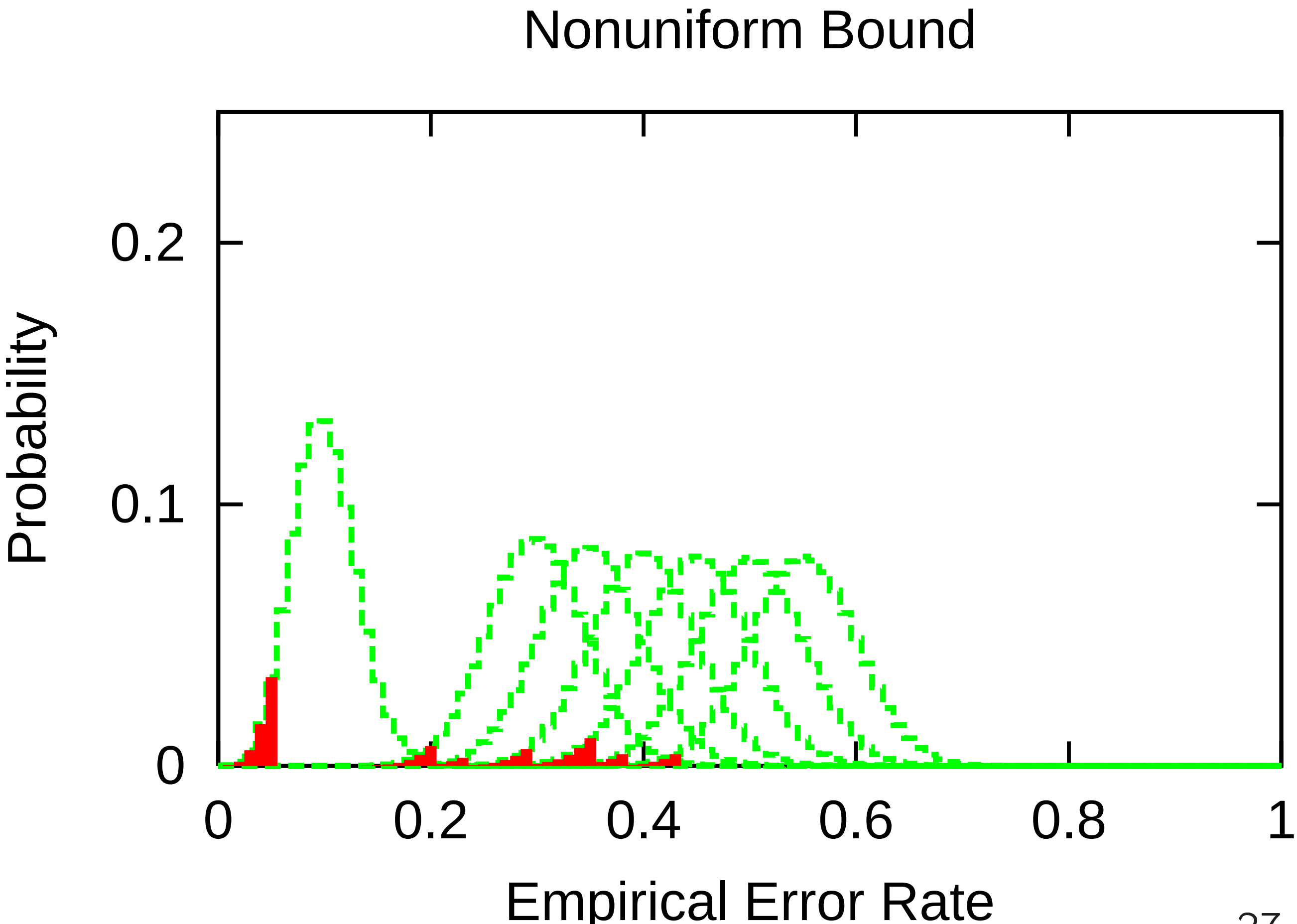
We can learn confidently on finite hypothesis spaces.

But: The bound is loose. Can we make it tighter?

Q: What if we have some *a priori* reason to like one hypothesis?

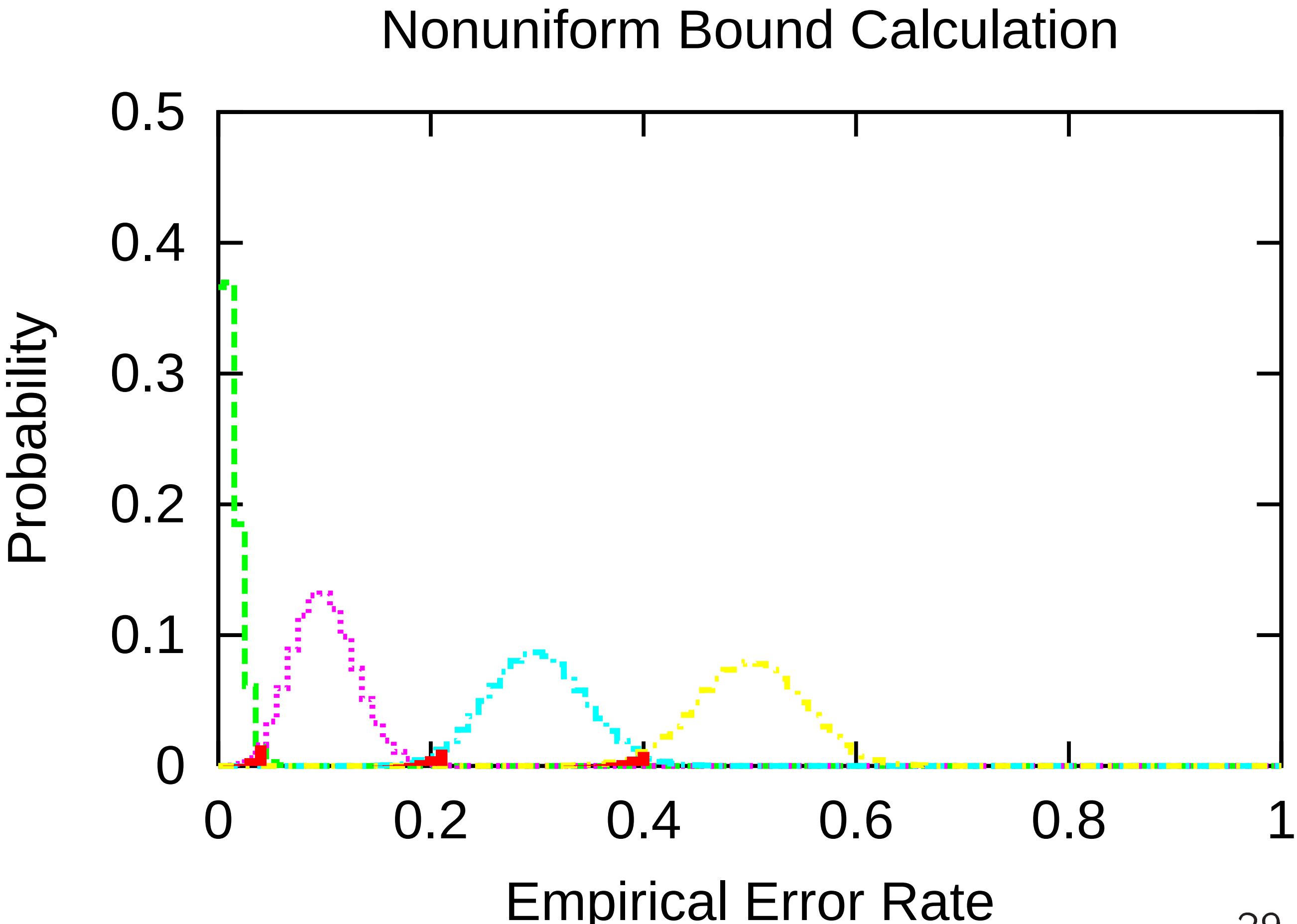
A: Then we can cut the tails in a nonuniform way.

$$\Pr(\exists h \in H : e(h) - \hat{e}(h) > \sqrt{\frac{\ln \frac{1}{p(h)} + \ln \frac{1}{\delta}}{2m}}) \leq \delta$$


Nonuniform Bound
Probability
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
Empirical Error Rate

Q: How do we calculate a bound?

A: Given h we know $p(h)$ so calculate the worst hypothesis consistent with a binomial tail cut at mass $\delta p(h)$.


Nonuniform Bound Calculation
Probability
Empirical Error Rate
0
0.1
0.2
0.3
0.4
0.5
0
0.2
0.4
0.6
0.8
1

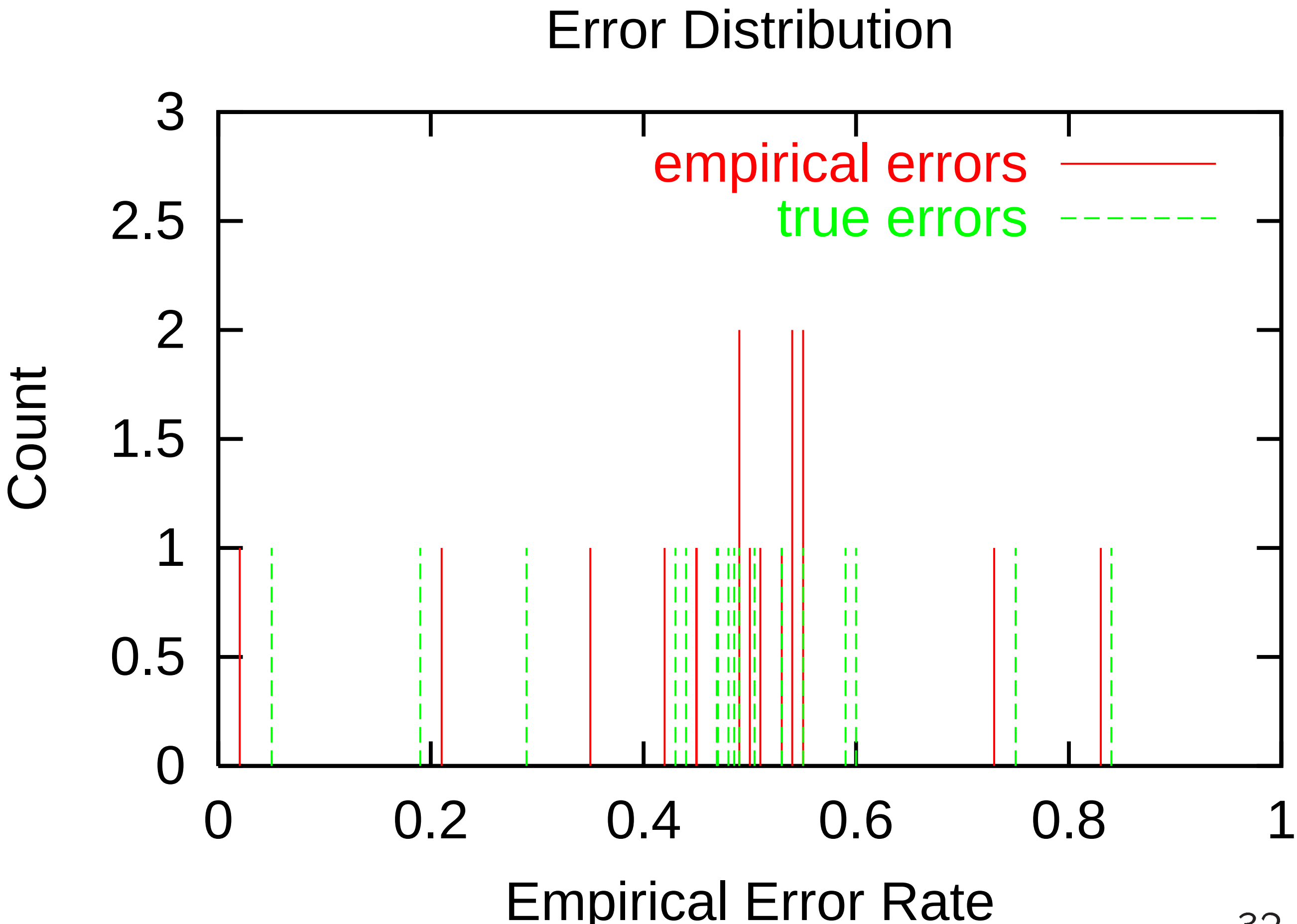
- Nonuniform bound is *tight*.

- Binomial coefficient calculations are *tight*.

Can't do better without making extra assumptions.

I lied

We can do better - but it requires considerably more computation.

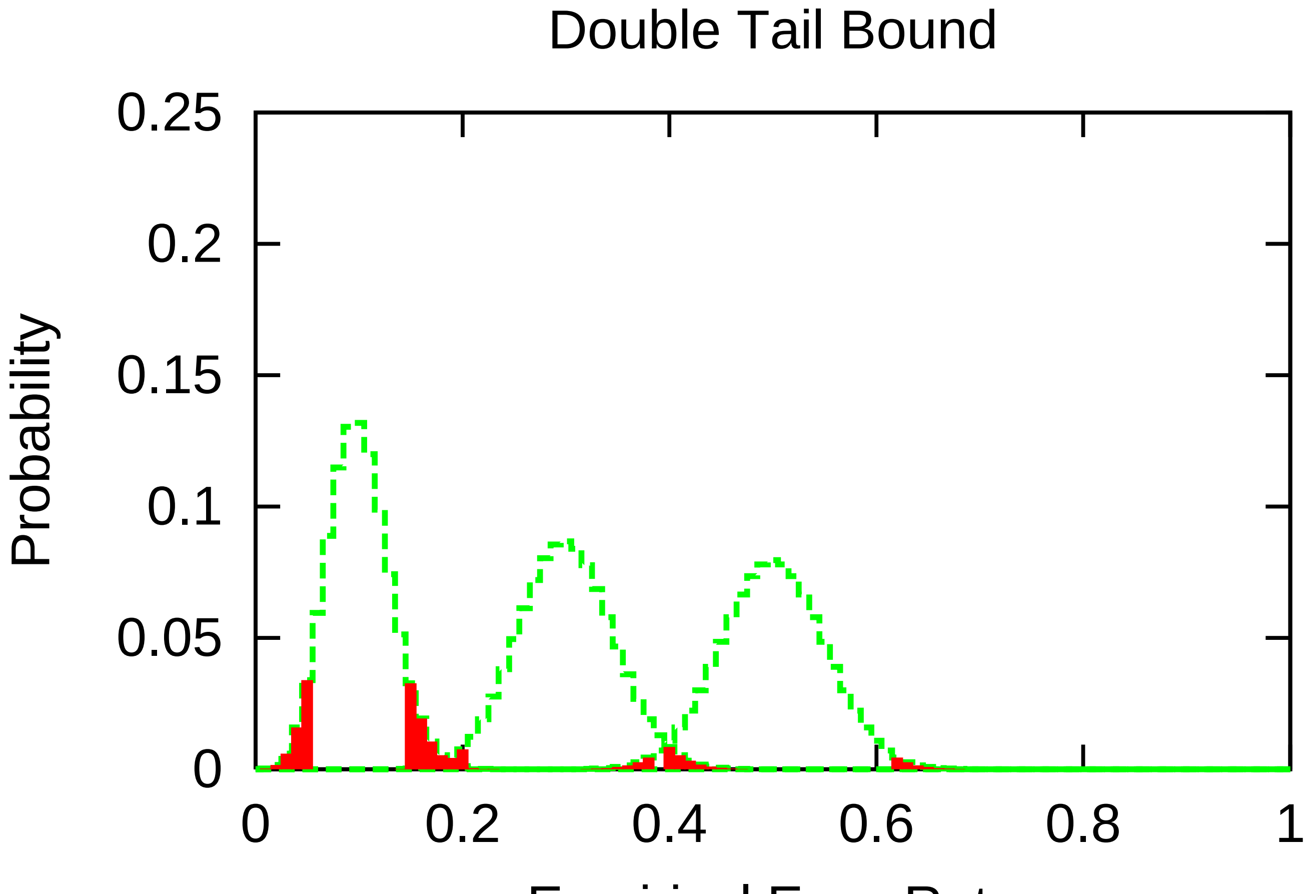

Error Distribution
empirical errors
true errors
Count
Empirical Error Rate
0
0.5
1
1.5
2
2.5
3
0
0.2
0.4
0.6
0.8
1

We will:

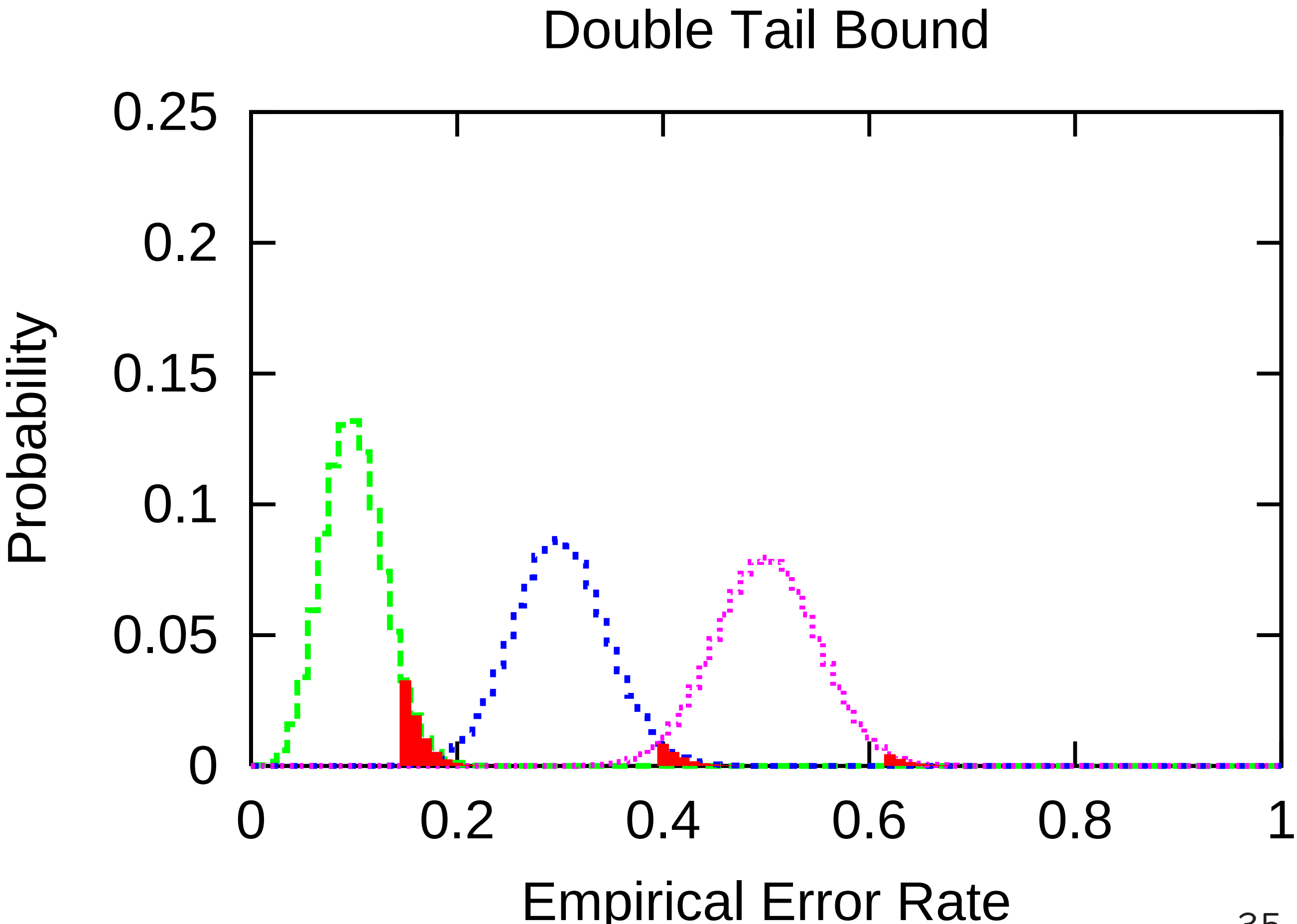
1. Cut the lower and *upper* tails.

2. Calculate worst case true error given the empirical error and upper tail cut.

3. Calculate the probability of a bad event given worst case true errors.



Double Tail Bound
Probability
Empirical Error Rate
0
0.05
0.1
0.15
0.2
0.25
0
0.2
0.4
0.6
0.8
1

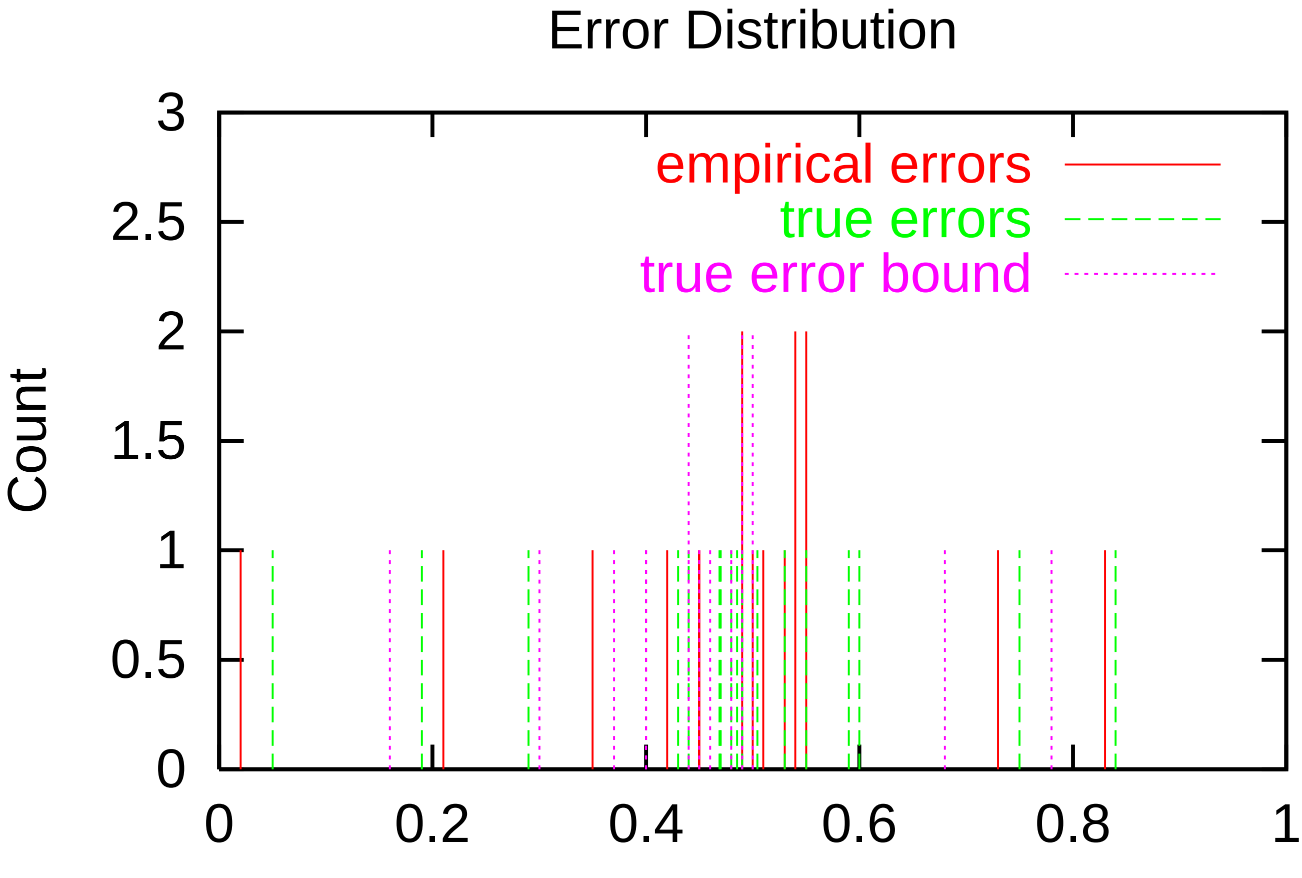


Double Tail Bound
Probability
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
Empirical Error Rate

We know how to bound the probability of a bad event for one hypothesis.

But there are *many* hypotheses.

So look at the distribution of empirical errors across *all* hypotheses.


Error Distribution
empirical errors
true errors
true error bound
Count
Empirical Error Rate
0
0.5
1
1.5
2
2.5
3
0
0.2
0.4
0.6
0.8
1

1. What can you do with a future error bound?

2. What is a Sample Complexity Bound?

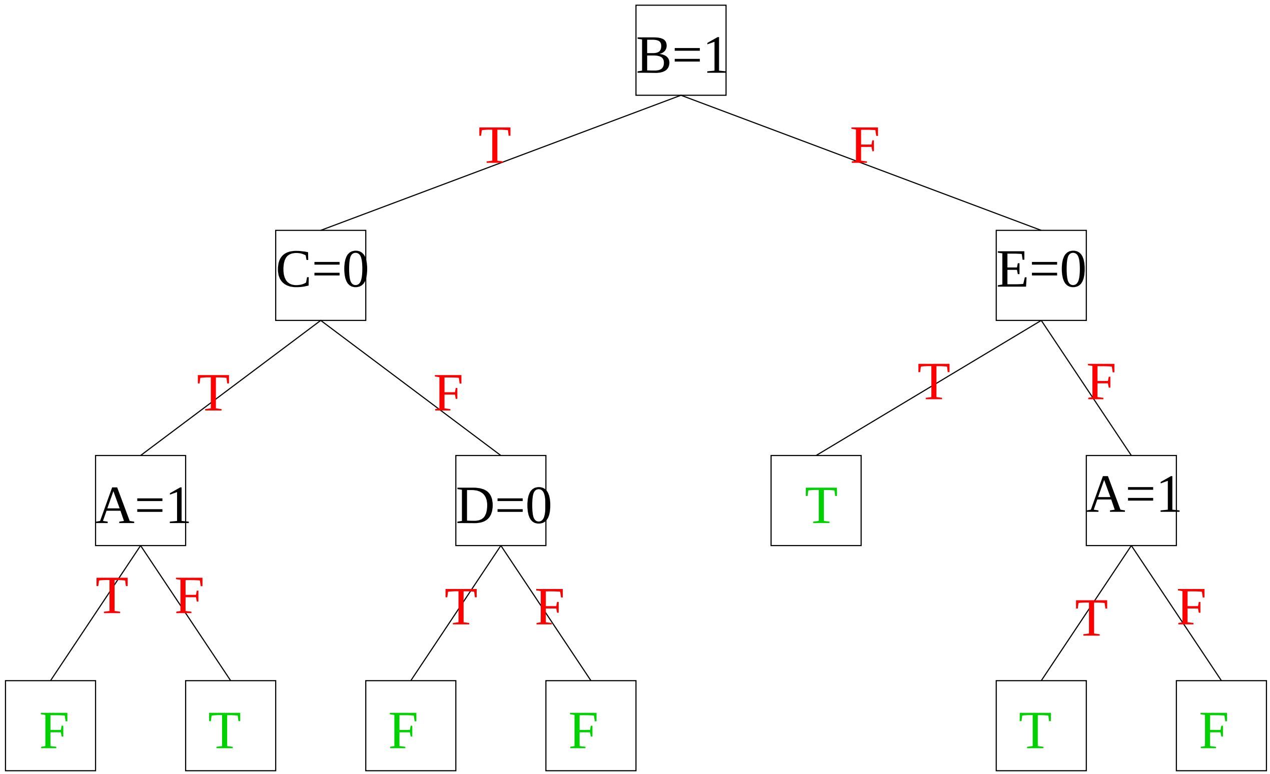
3. How can you make one?

4. *How well do they work?*

How well do the Sample Complexity Bounds work?

We will test on decision trees from the UCI database.

Example Decision Tree



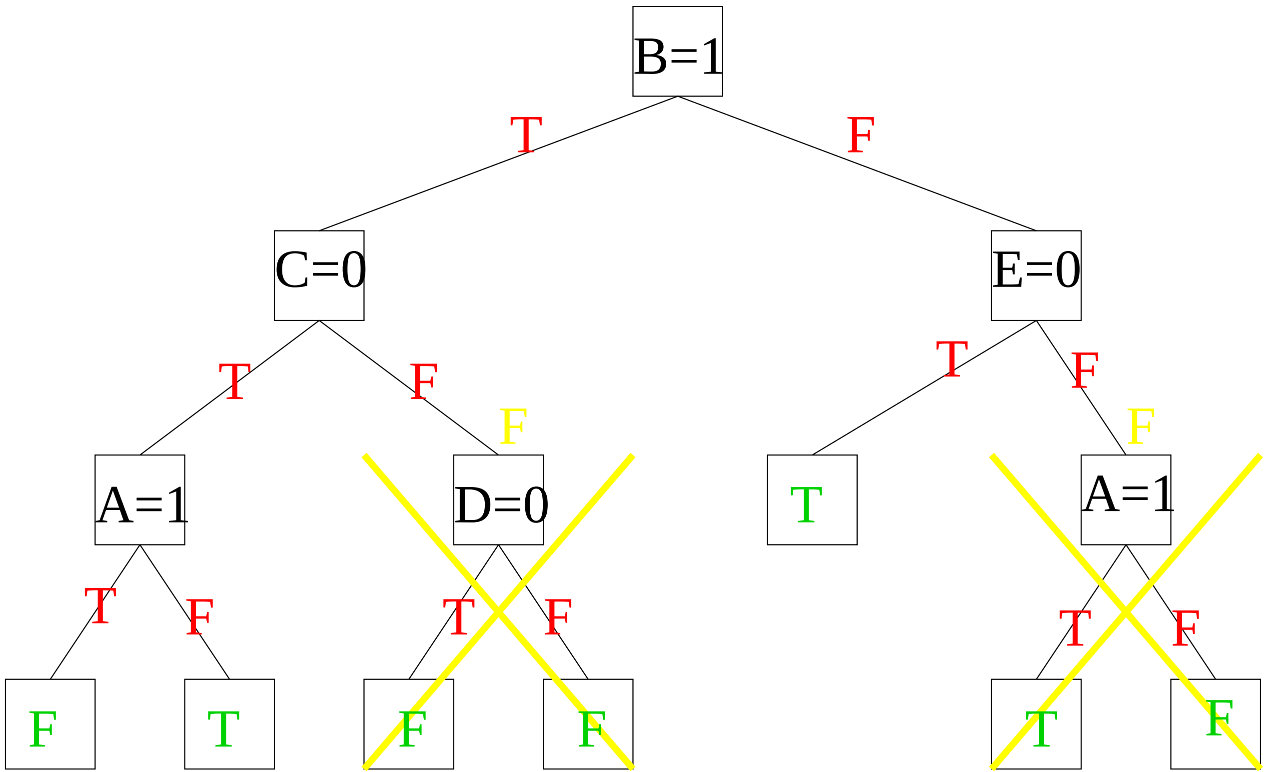
Optimizations:

- Pruning: To find a preferred hypothesis

- Sampling: Make Double Tail feasible

- Fast Sampling: Make Double Tail feasible

Pruning: Smaller decision tree ⇒ tighter bounds



Sampling: Trading a looser bound for less computation

1. Given a uniform samples from the hypothesis space, what is the worst case empirical error distribution?

2. Given the worst case empirical error distribution, what is the worst case true error distribution?

3. Given the worst case true error distribution, what is the worst case probability of an error?

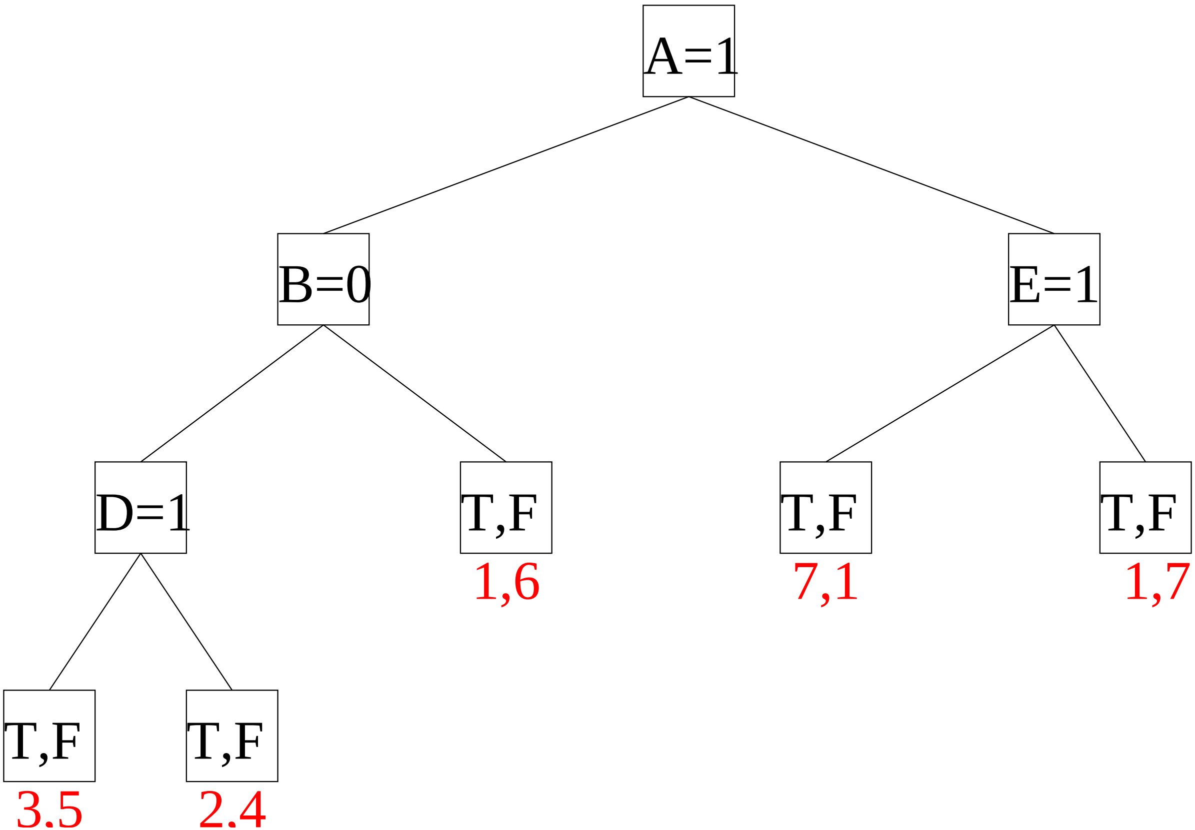
We can calculate a Double tail bound variant.

But the approximation only improves logarithmically in the number of samples.

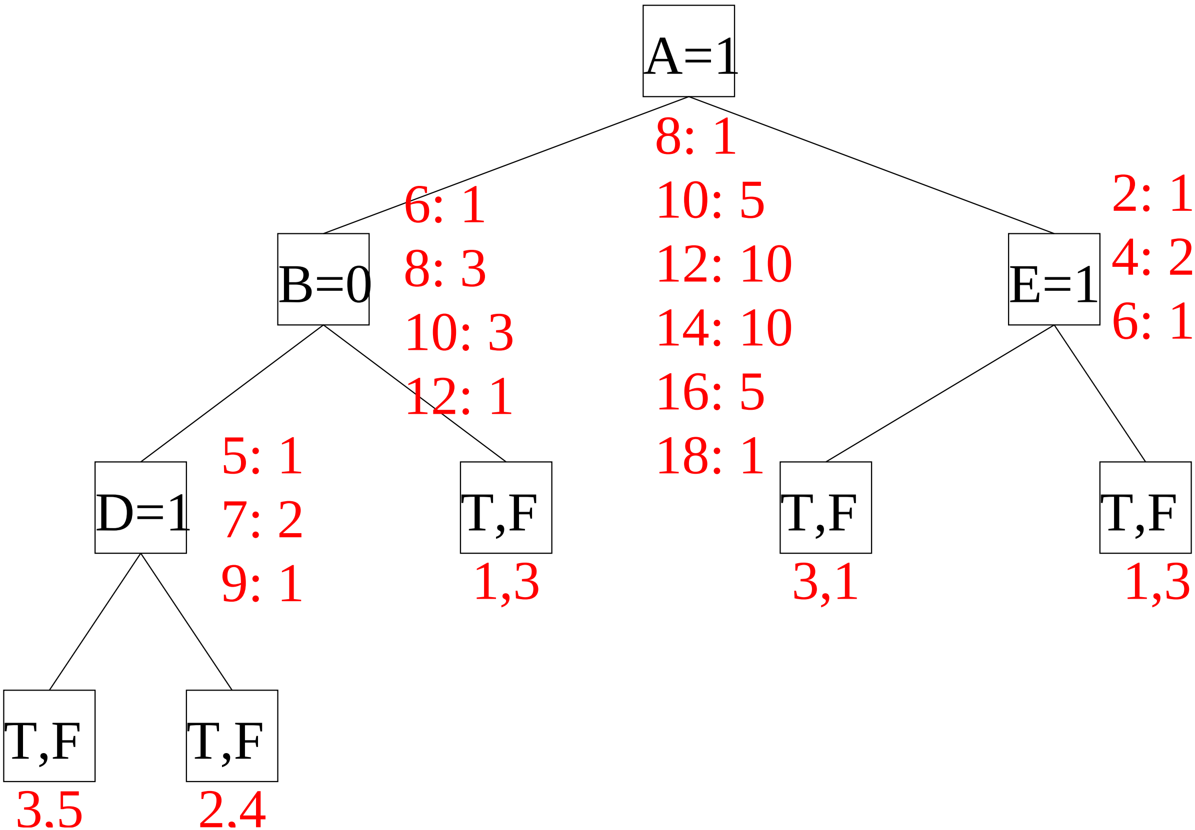
Q: How can we do better?

A: Sample *much* faster

Fast Sampling: Cheating



Fast Sample example



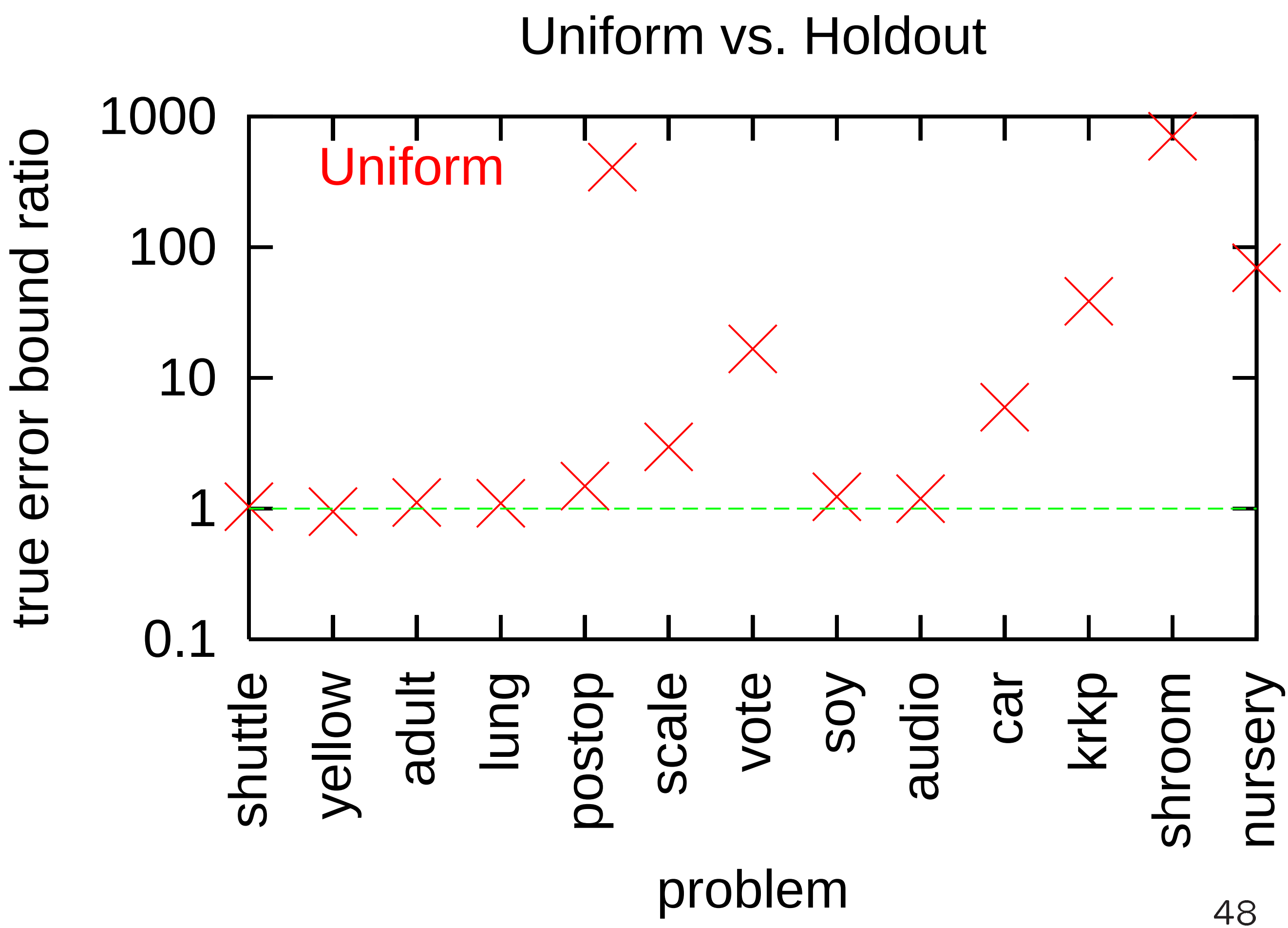
Now, how tight are the various bounds? We have:

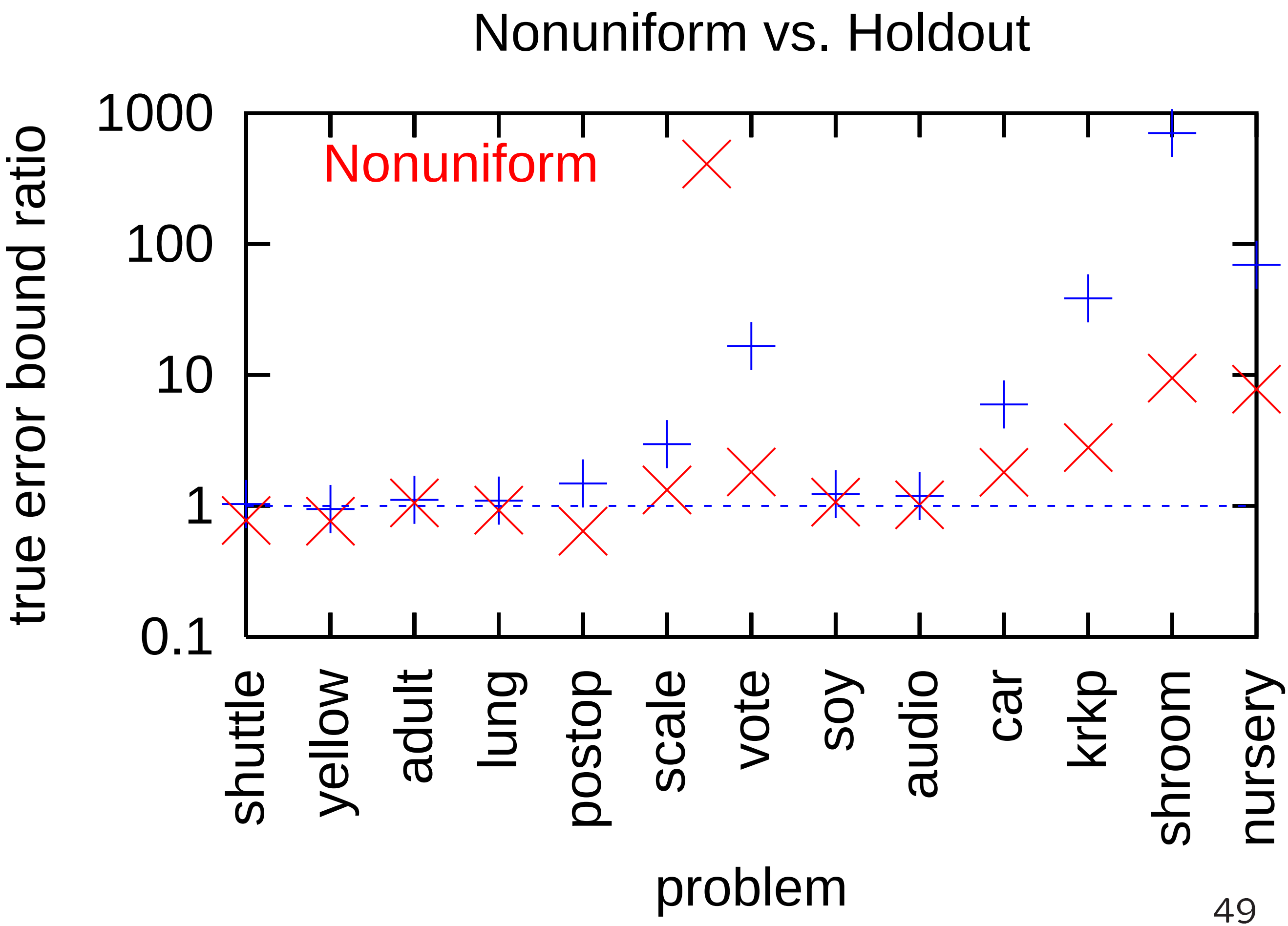
- Holdout

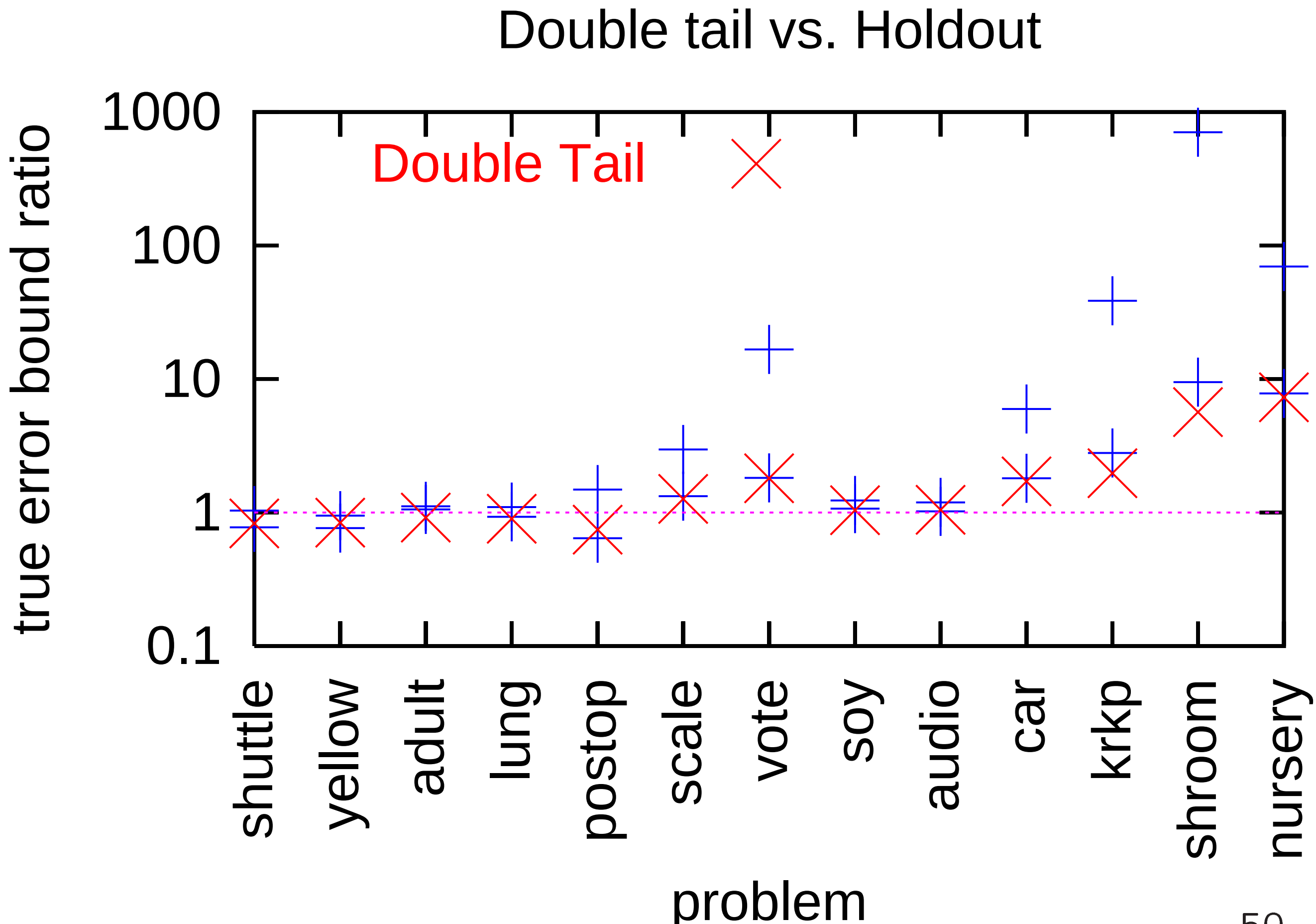
- Uniform

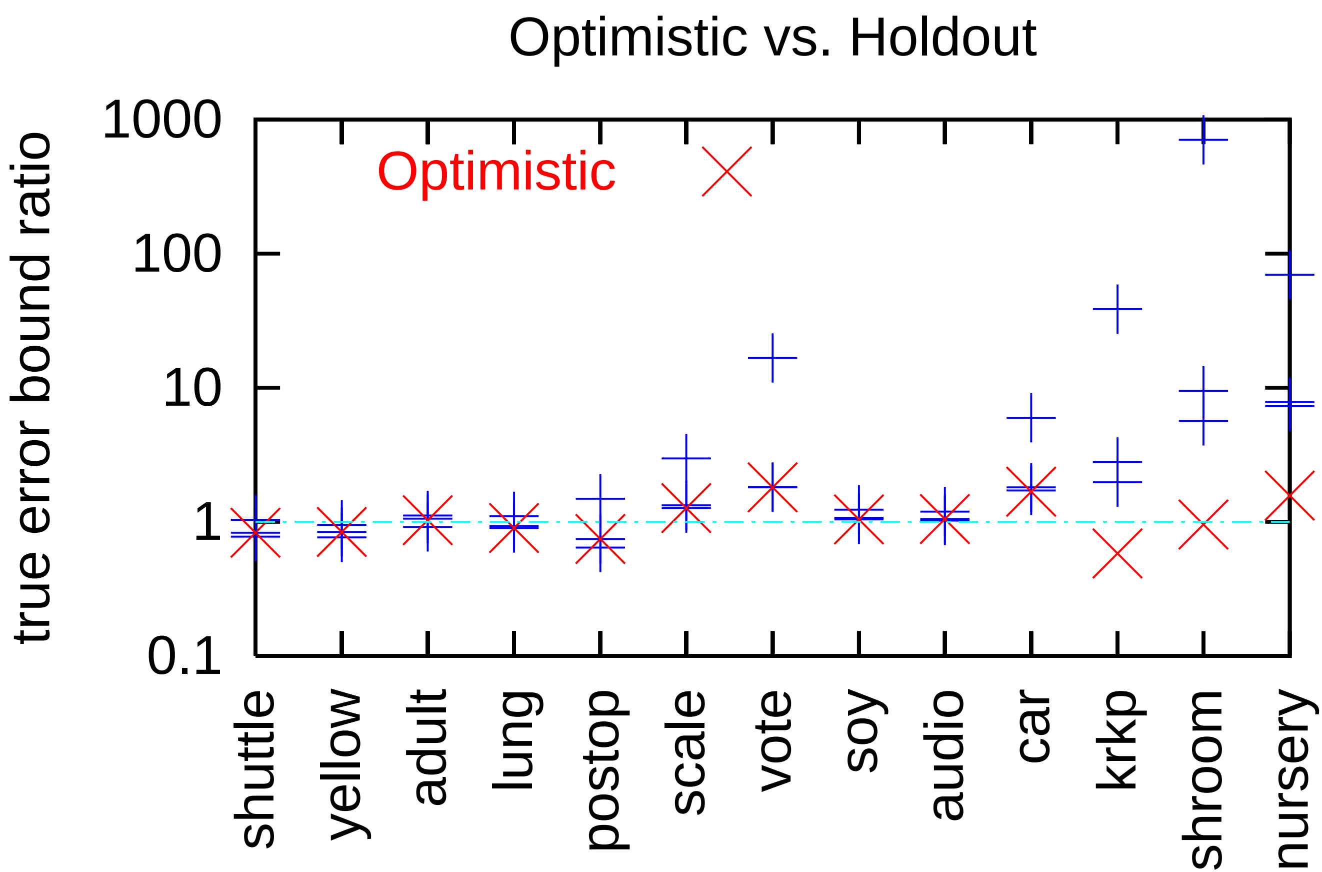
- Nonuniform

- Double tail


Uniform vs. Holdout
true error bound ratio
1000
100
10
1
0.1
Uniform
shuttle
yellow
adult
lung
postop
scale
vote
soy
audio
car
krkp
shroom
nursery
problem


Nonuniform vs. Holdout
Nonuniform
true error bound ratio
1000
100
10
1
0.1
shuttle
yellow
adult
lung
postop
scale
vote
soy
audio
car
krkp
shroom
nursery
problem


Double tail vs. Holdout
Double Tail
true error bound ratio
1000
100
10
1
0.1
shuttle
yellow
adult
lung
postop
scale
vote
soy
audio
car
krkp
shroom
nursery
problem



Optimistic vs. Holdout
true error bound ratio
1000
100
10
1
0.1
Optimistic
shuttle
yellow
adult
lung
postop
scale
vote
soy
audio
car
krkp
shroom
nursery
problem

Conclusion

1. Other bounds beat Holdout with less then 100 examples.

2. Holdout wins for more then 1000 examples.

3. Further counting optimizations could be useful.

Technical Details at:

http://www.cs.cmu.edu/~jcl/papers/papers.html