Combining Test and Train Set Bounds

John Langford

1. Test & Train set bounds

2. Combinations

3. A combination theorem

4. Technique and Method

The Model

X = (arbitrary) input space

$Y \in \{0, 1\}$ = prediction space

Assume: Examples (X, Y) are drawn independently from distribution D

Terminology

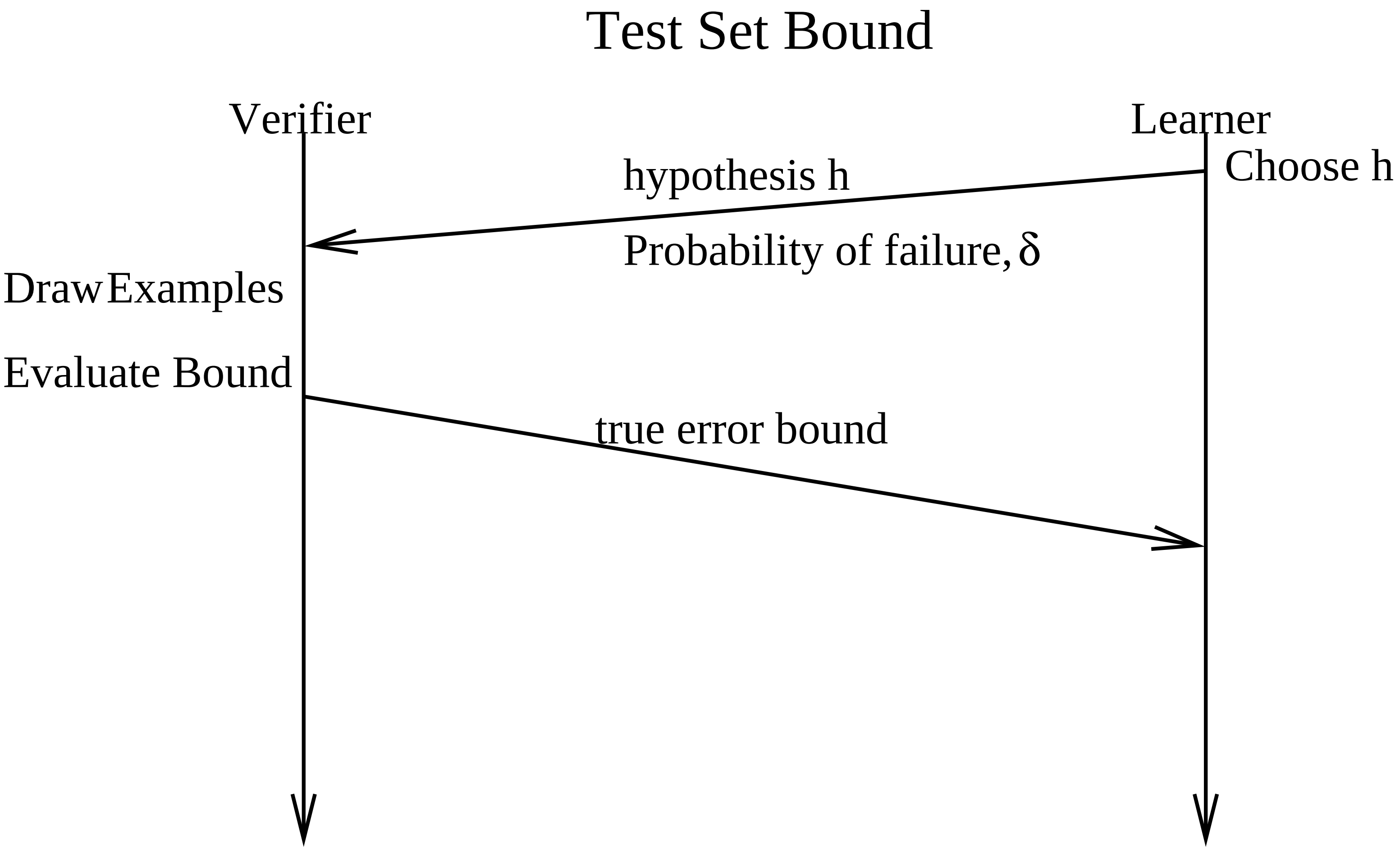
- Samples $= S$
- Learning algorithm $A: \; S \rightarrow h$
- True Error rate $e_D(h) = \mathrm{Pr}_D(h(x) \neq y)$
- Empirical Error rate $\widehat{e}_S(h) = \mathrm{Pr}_S(h(x) \neq y)$

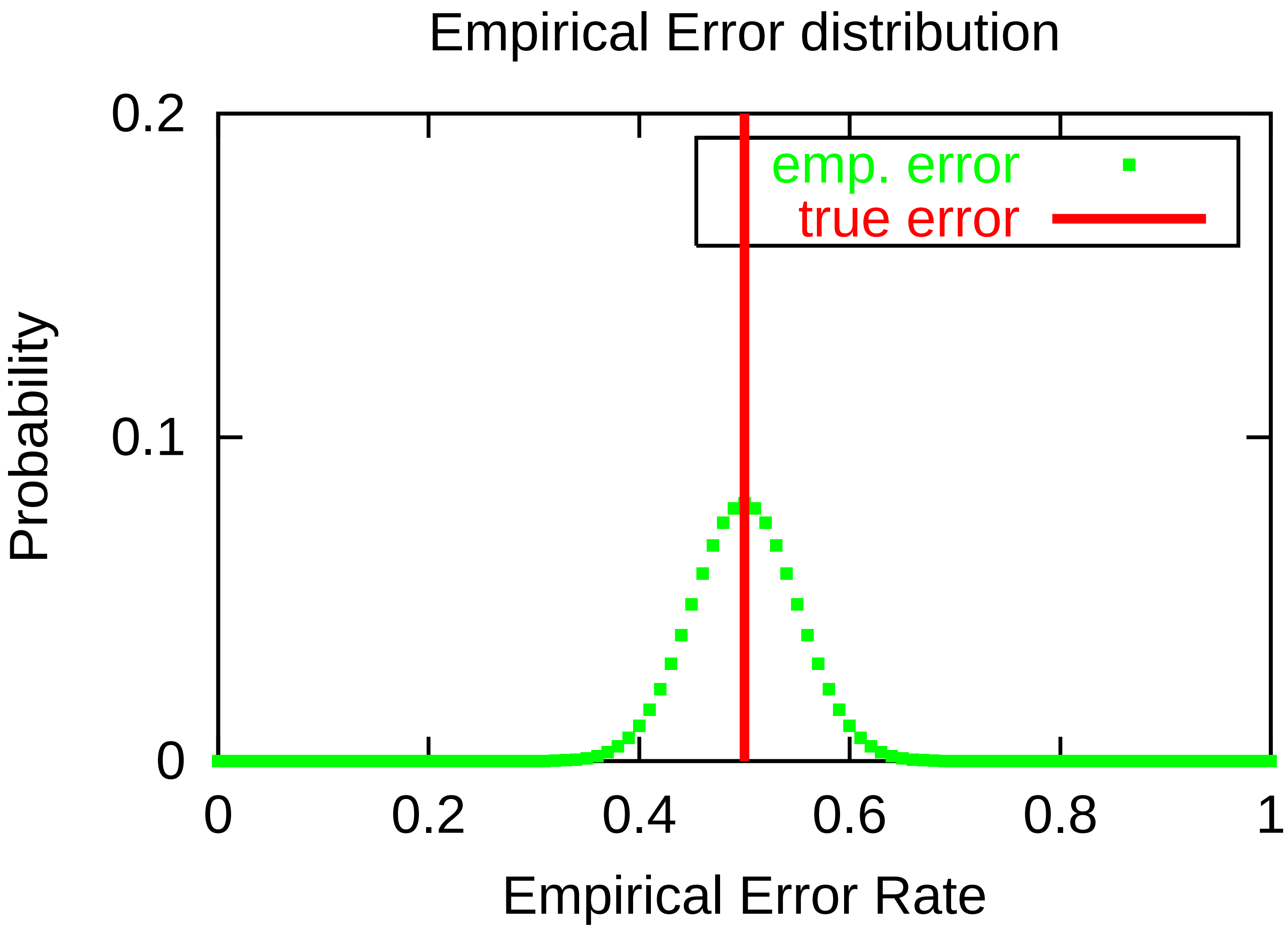
Proofs of Learning

1. Test set bound: Used in practice

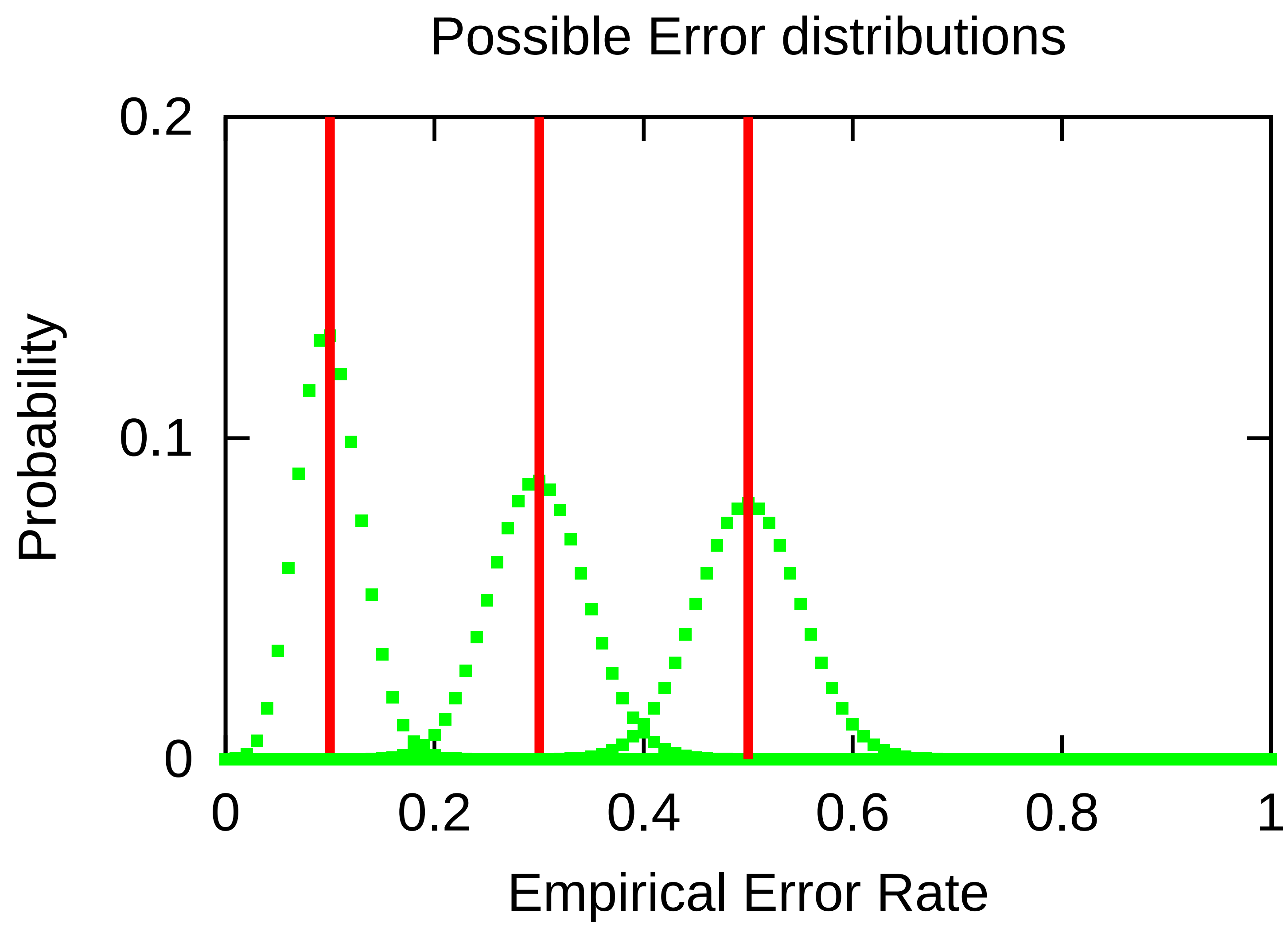
2. Train set bound: Not used, often loose

Holdout Bound Protocol

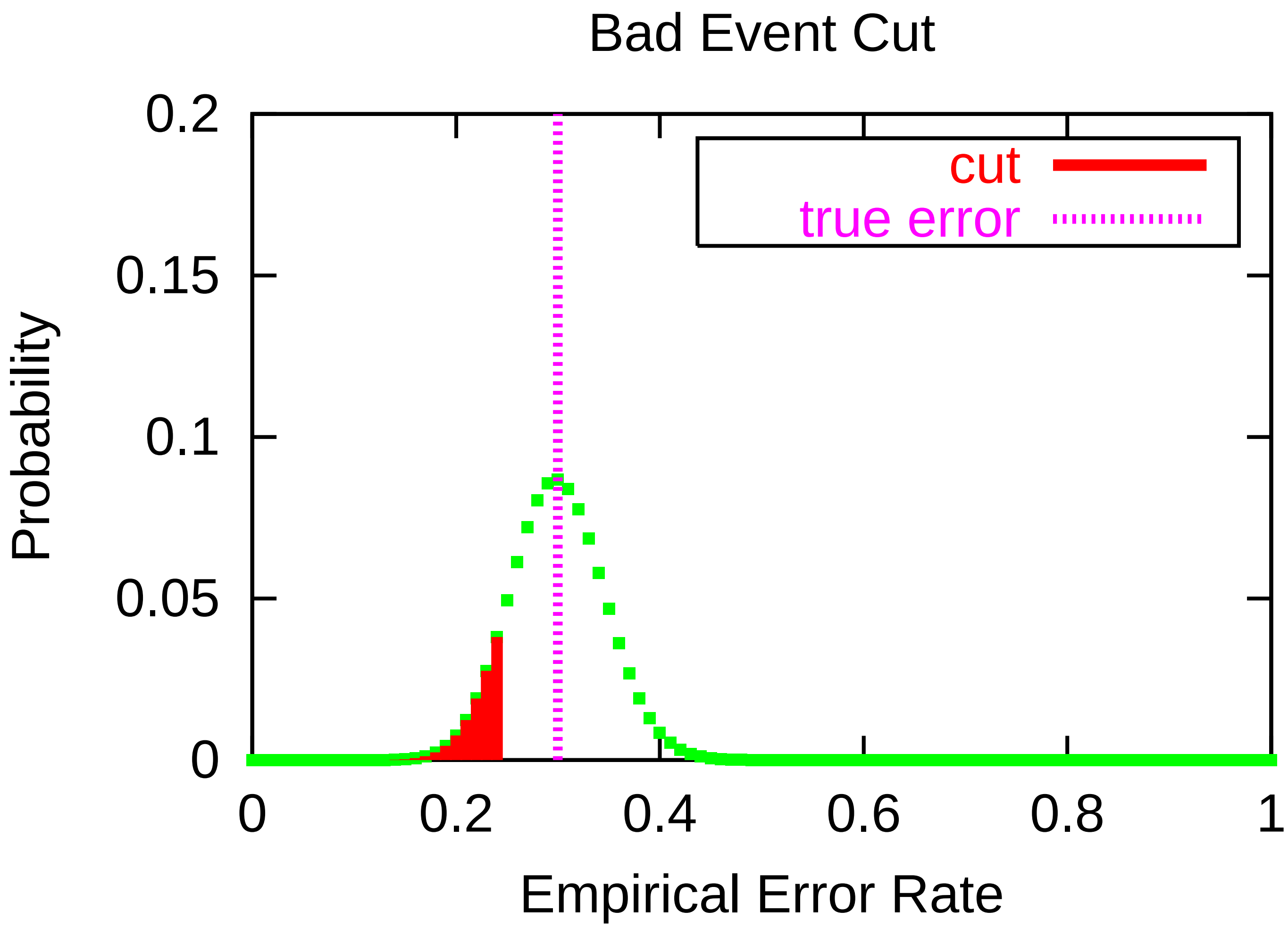


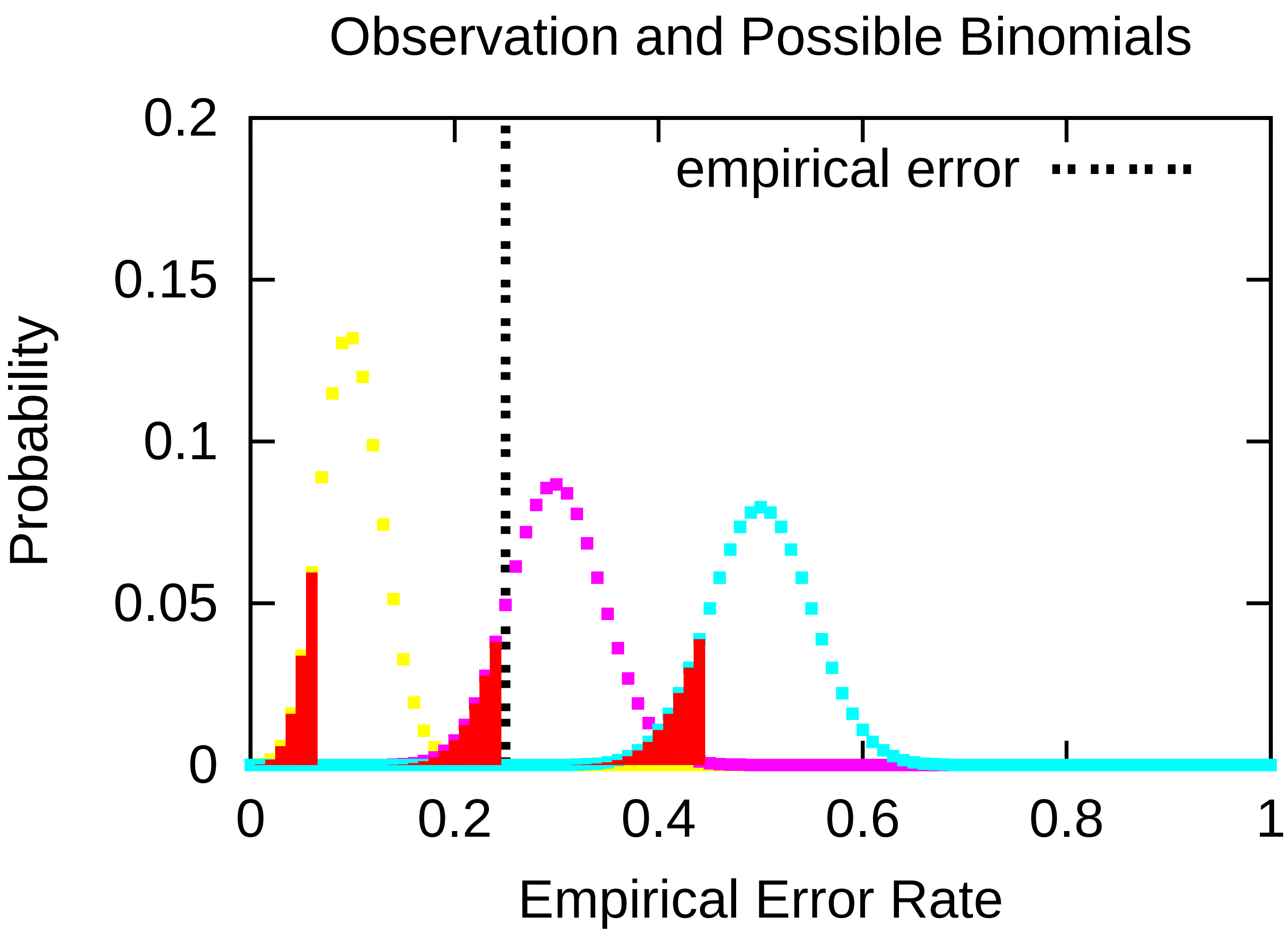


Empirical Error distribution
emp. error
true error
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
Probability
Empirical Error Rate

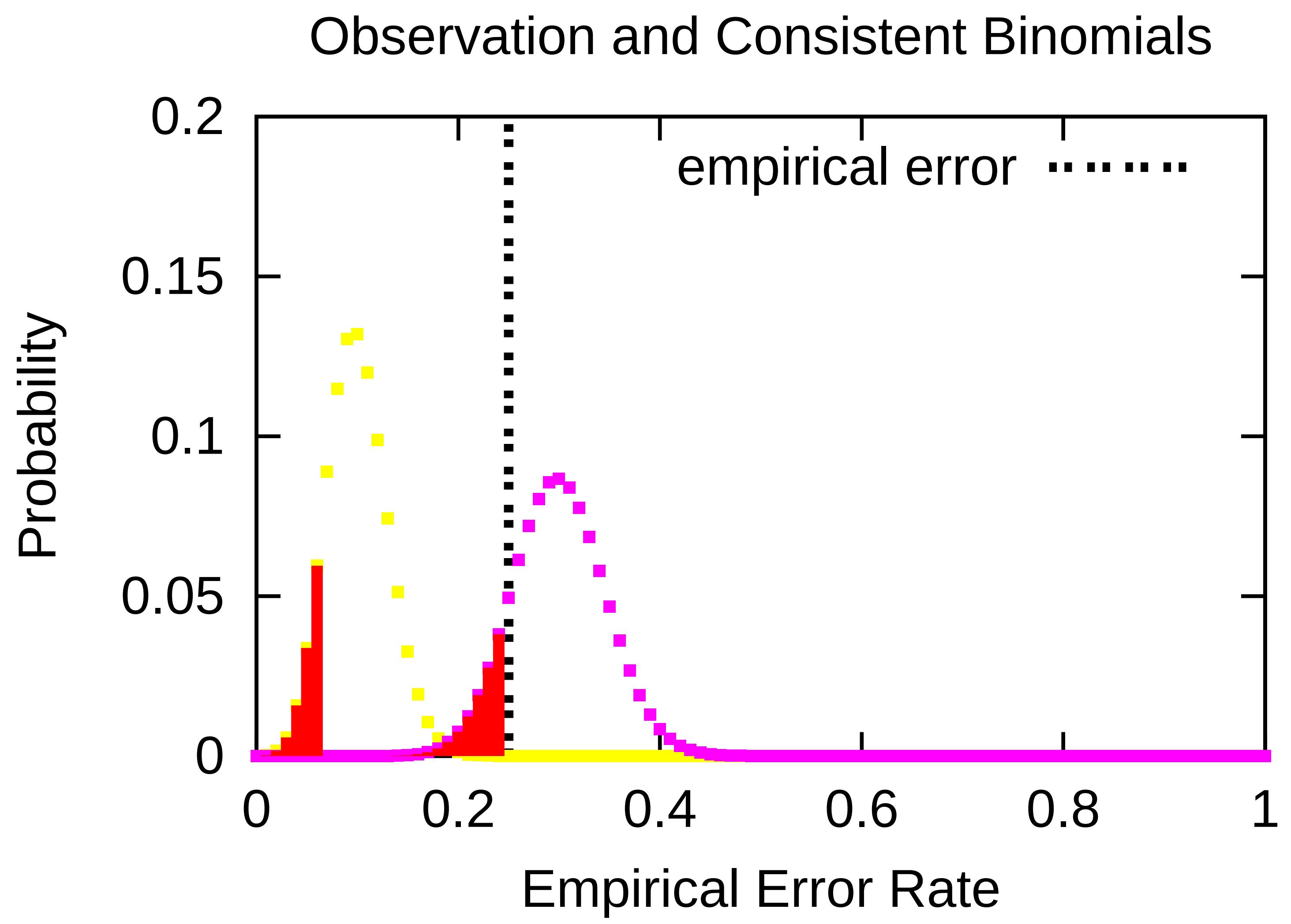

Possible Error distributions
Probability
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
Empirical Error Rate

How can you prove learning?

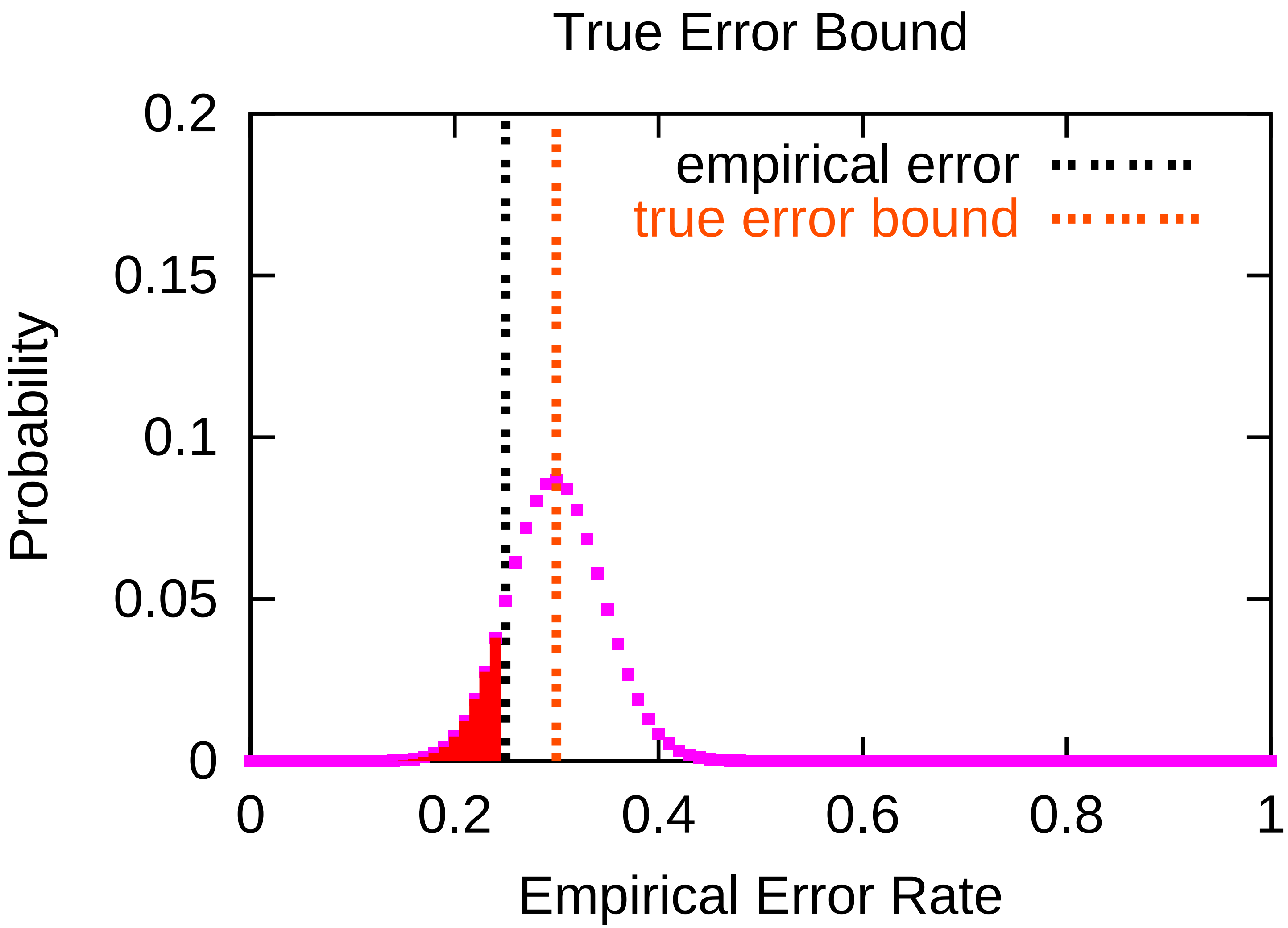




Observation and Possible Binomials
empirical error
Probability
Empirical Error Rate
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1



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

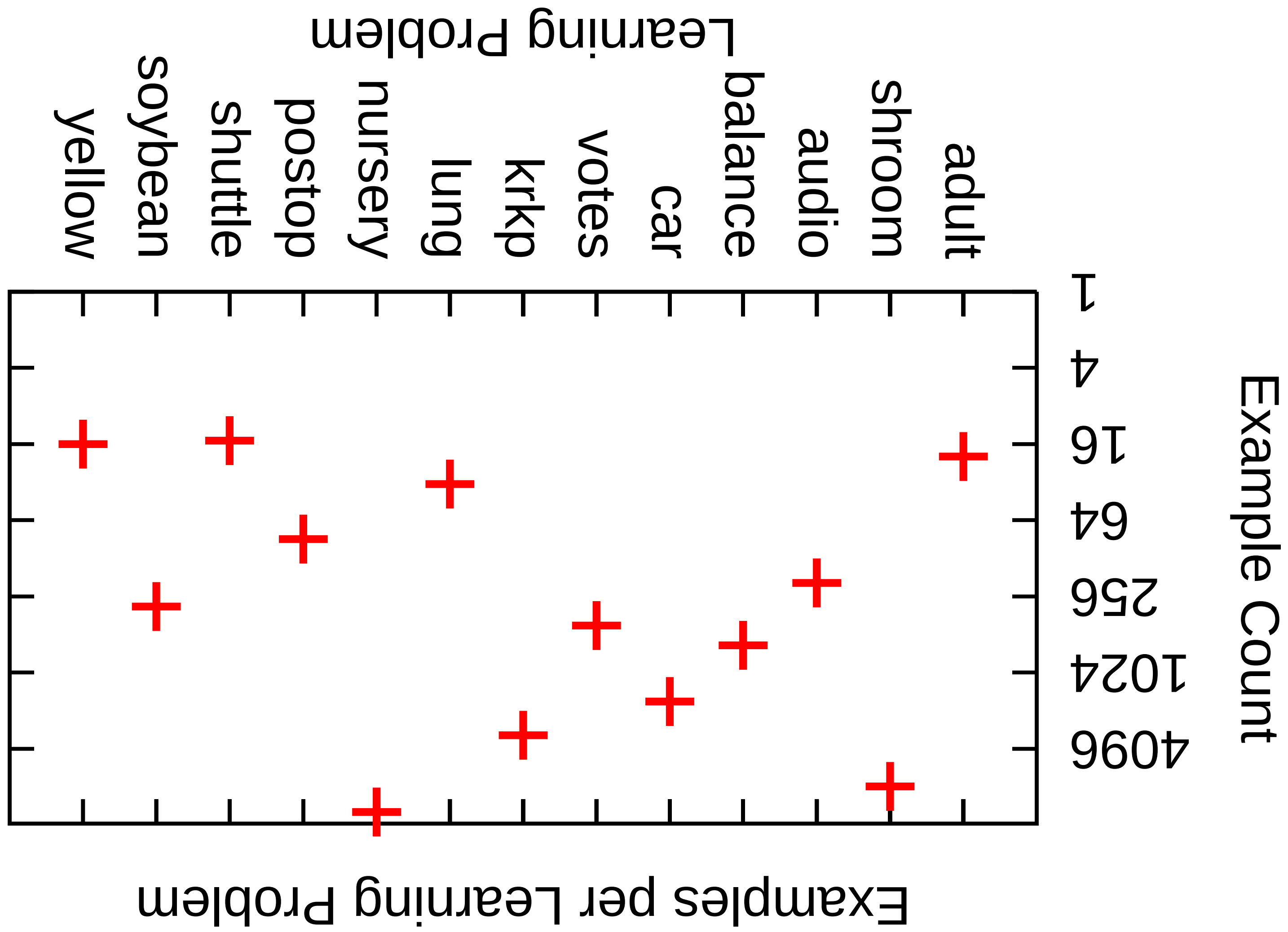
Let $\mathrm{Bin}_{p,m}(k) = \sum_{i=0}^{k} \left(\begin{array}{c} m \\ i \end{array} \right) p^i (1-p)^{m-i}$ = Binomial cumulative distribution

Theorem:

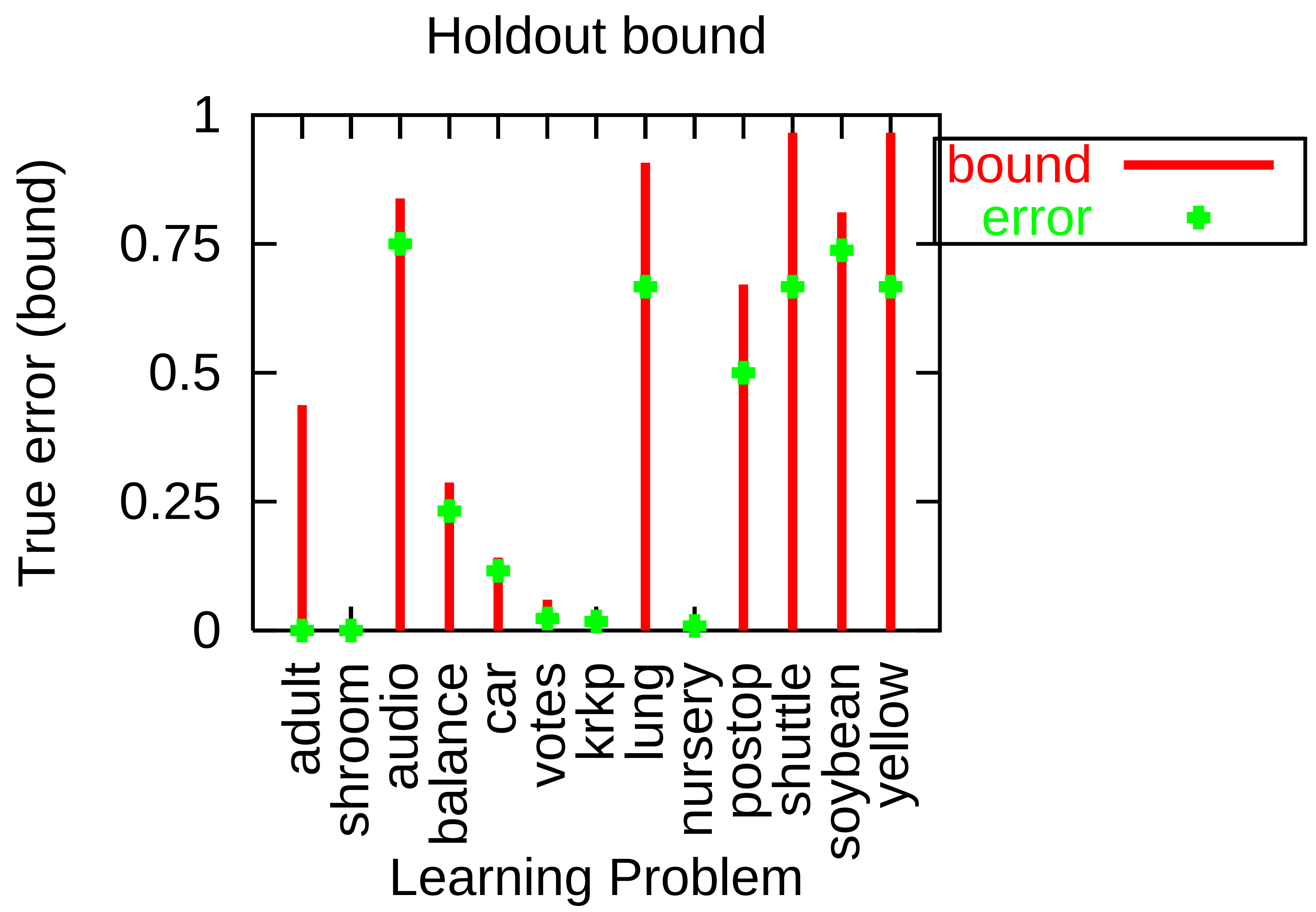
$$\Pr_{S \sim D^m} (e_D(h) \geq \overline{e}(h, \widehat{e}_S(h), \delta)) \leq \delta$$

where $\overline{e}(h, \frac{k}{m}, \delta) = \max\{p | \mathrm{Bin}_{p,m}(k) \geq \delta\}$

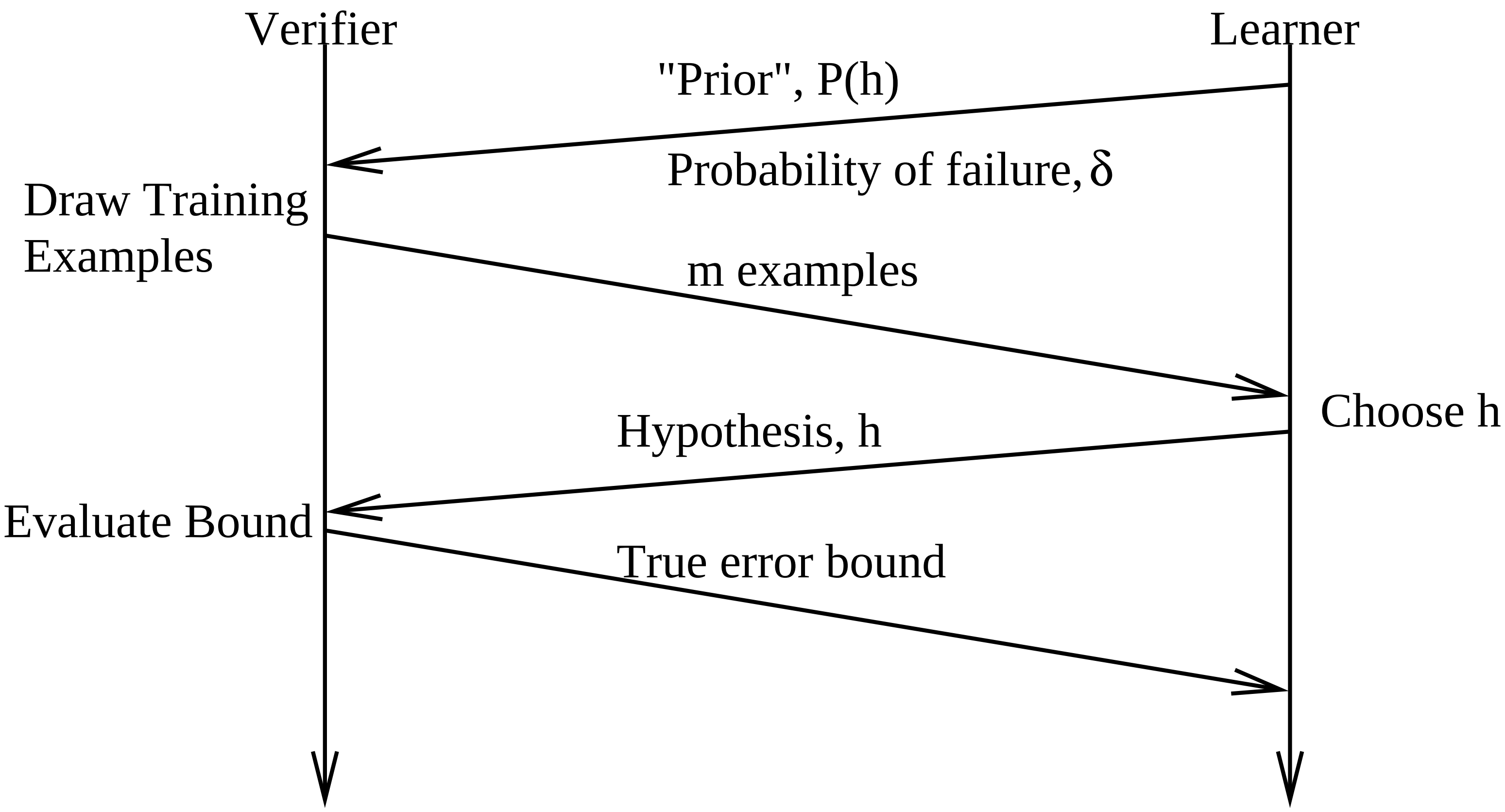
Proof: By definition of cumulative distribution.

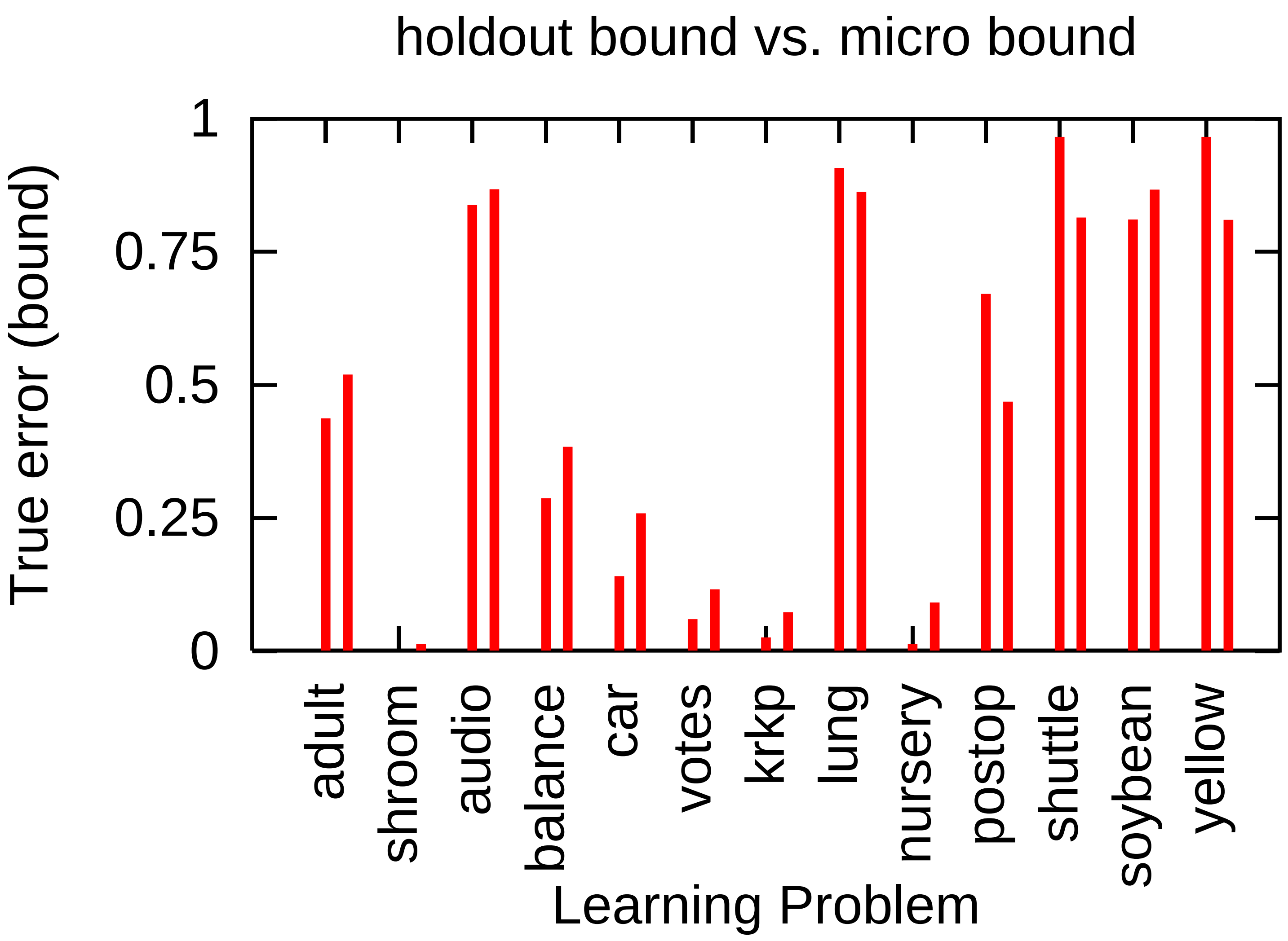


Examples per Learning Problem
Learning Problem
Example Count
yellow
soybean
shuttle
postop
nursery
lung
krkp
votes
car
balance
audio
shroom
adult
1
4
16
64
256
1024
4096


Holdout bound
True error (bound)
1
0.75
0.5
0.25
0
bound
error
adult
shroom
audio
balance
car
votes
krkp
lung
nursery
postop
shuttle
soybean
yellow
Learning Problem

Training Set Bound




holdout bound vs. micro bound
True error (bound)
0
0.25
0.5
0.75
1
adult
shroom
audio
balance
car
votes
krkp
lung
nursery
postop
shuttle
soybean
yellow
Learning Problem

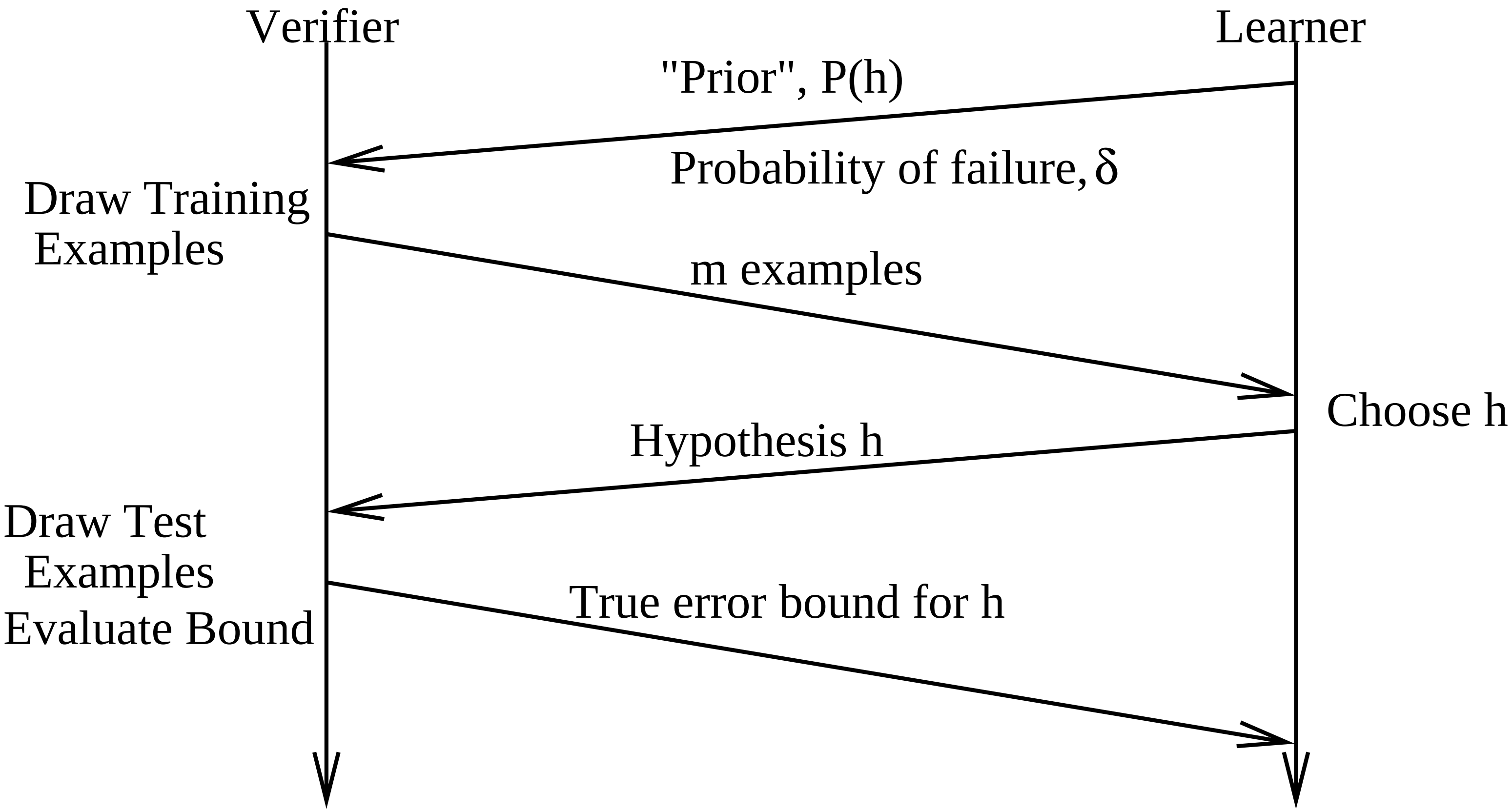
1. Test & Train set bounds

2. Combinations

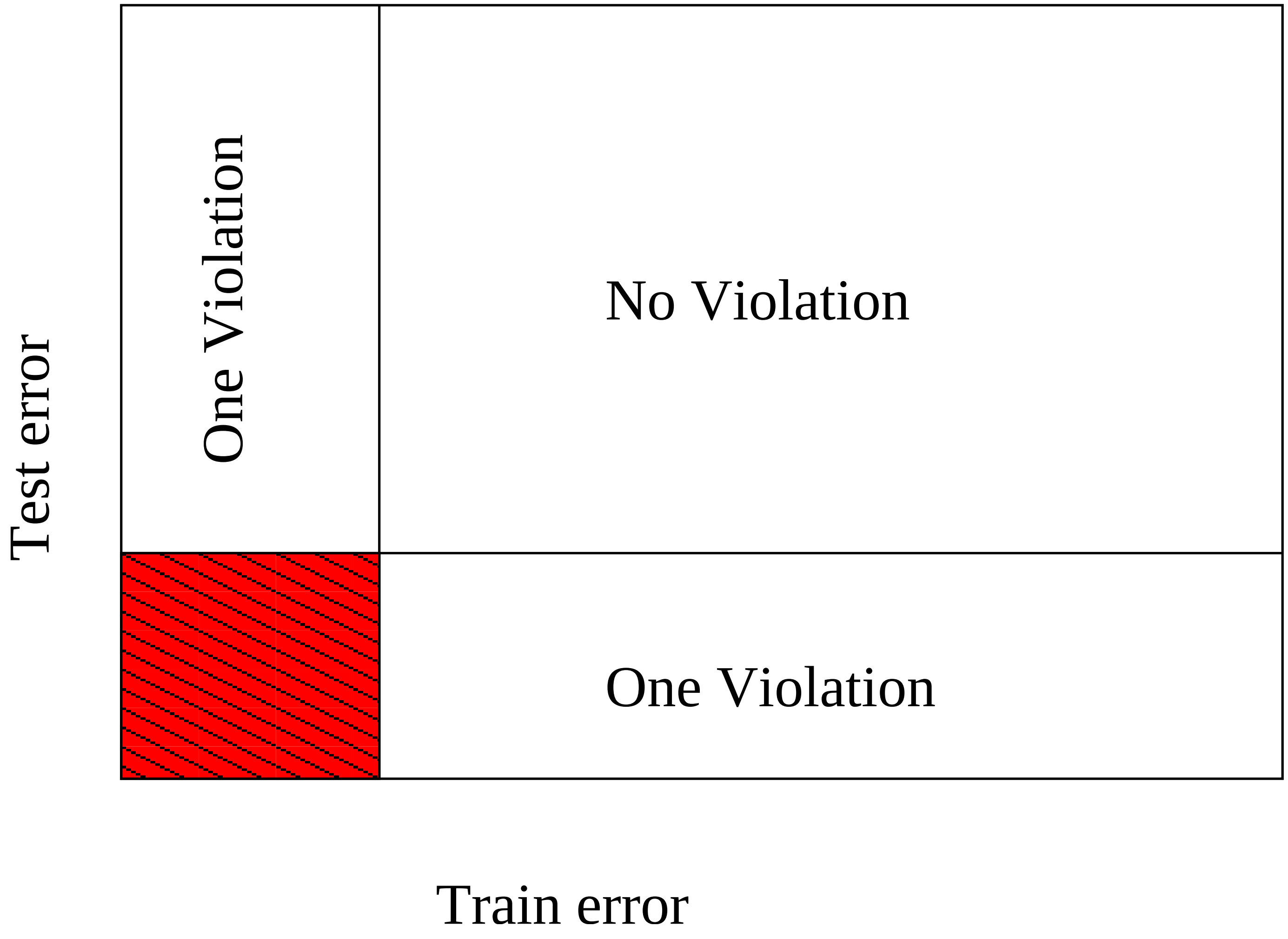
3. A combination theorem

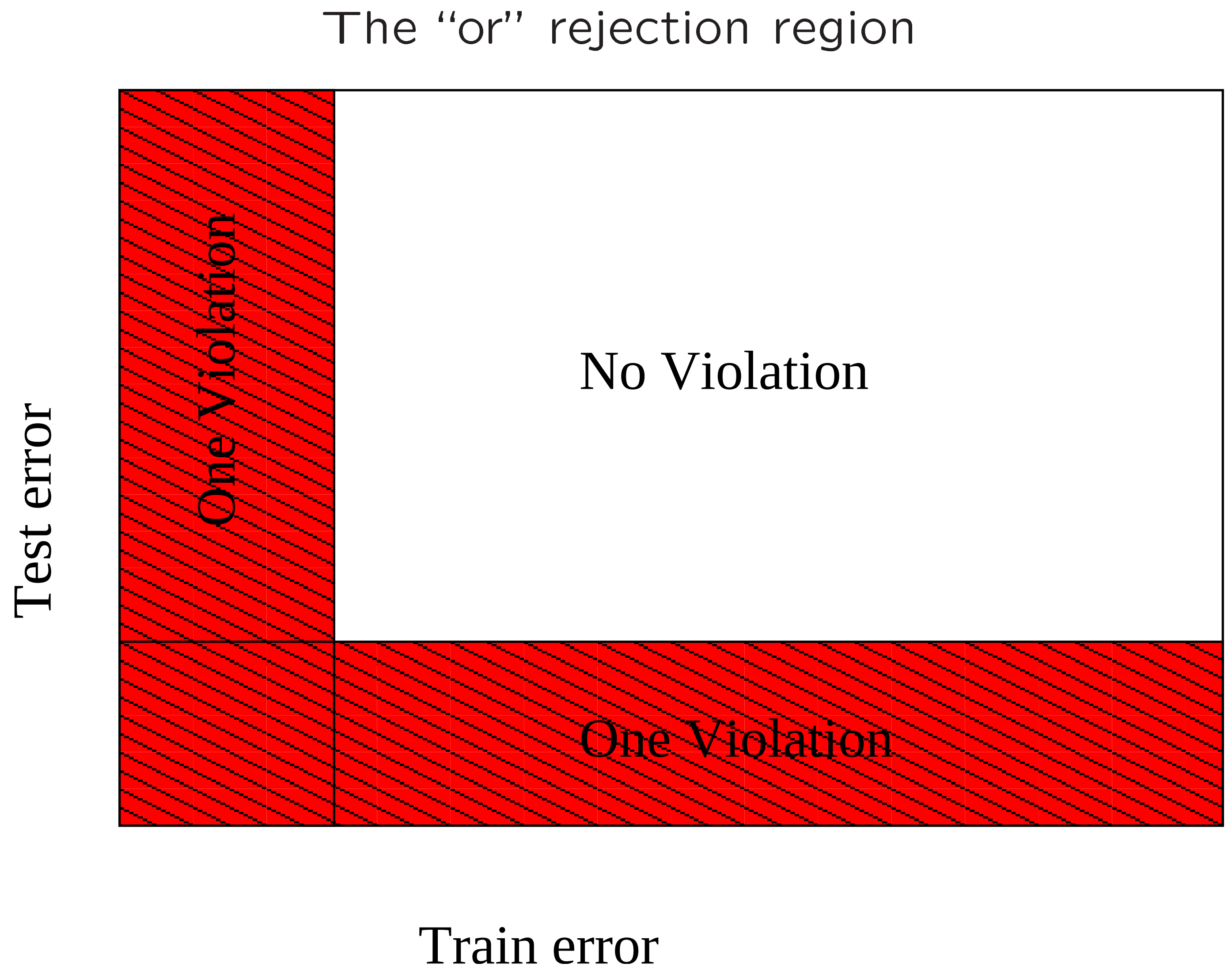
4. Technique and Method

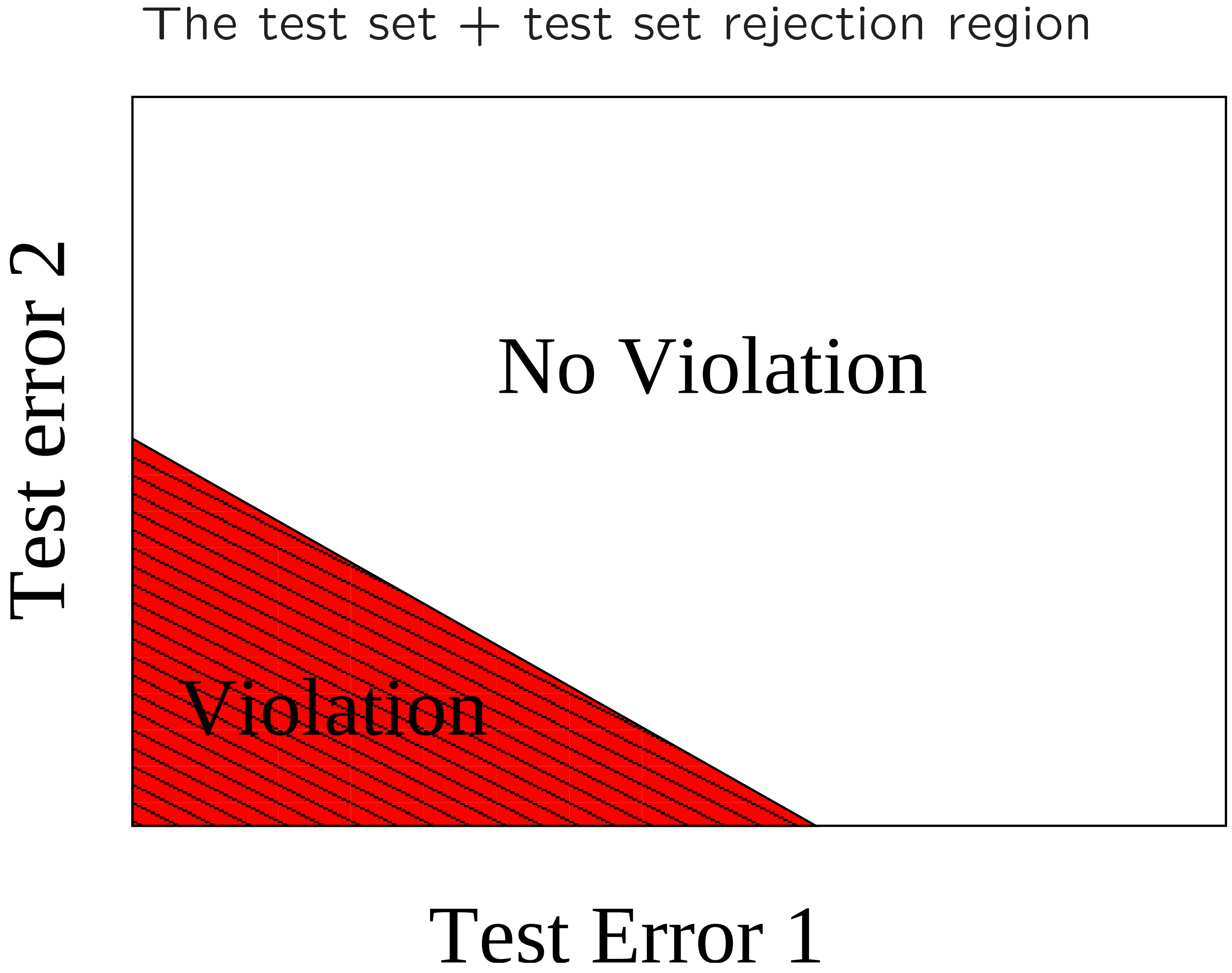
Train and Test Bound

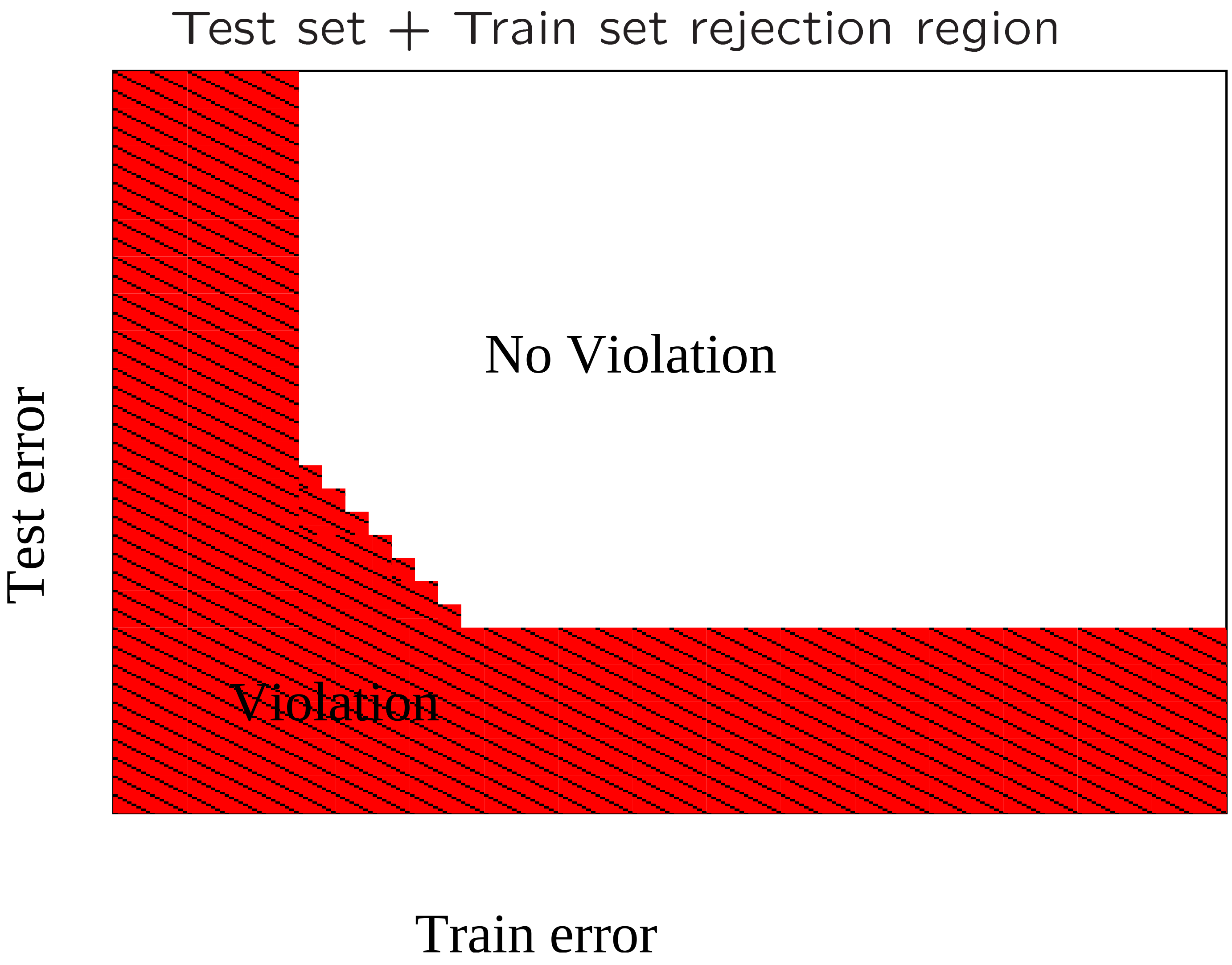


The "and" rejection region




The "or" rejection region
One Violation
No Violation
One Violation
Test error
Train error


The test set + test set rejection region
No Violation
Violation
Test error 2
Test Error 1



Test set + Train set rejection region
No Violation
Violation
Test error
Train error

1. Test & Train set bounds

2. Combinations

3. A combination theorem

4. Technique and Method

Bound = worst case rejection region

Let $\phi_P(\delta)$ = a set of events dependent on P and δ.

$$\Pr_{x \sim P}(x \in \phi_P(\delta)) \leq \delta$$

Randomized bound = Randomized rejection region

$$E_{w\sim Q} \Pr_{x\sim P}(x \in \phi_{P,w}(\delta)) \leq \delta$$

$$\Rightarrow \Pr_{w\sim Q,\ x\sim P}(x \in \phi_{P,w}(\delta)) \leq \delta$$

Exact test and train bound

Theorem: Let $\phi_D(\delta)$ be any bound satisfying:

$$\forall D\ \forall \delta \geq 0\ \ \Pr_{S \sim D^m}(S \in \phi_D(\delta)) \leq \delta$$

Let $f(S_{\text{test}})$ be any function satisfying:

$$E_{S_{\text{test}} \sim D^{m_{\text{test}}}} \Pr_{S \sim D^m}(S \in \phi_D(f(S_{\text{test}}) * \delta)) \leq \delta$$

then:

$$\Pr_{S, S_{\text{test}} \sim D^{m+m_{\text{test}}}}(S \in \phi_D(f(S_{\text{test}}) * \delta)) \leq \delta$$

Problem: Sometimes we use approximate bounds instead of $\mathsf{Bin}_{p,m}(k)$.

Solution:

1. Restrain $f(S_{\text{test}})$ to be monotonic decreasing in $\hat{e}_{\text{test}}$

2. if $\forall x\ F_1(x) \leq F_2(x)$ and $f(x) \geq 0$ then

$$\int f(x) dF_1(x) \leq \int f(x) dF_2(x)$$

1. Test & Train set bounds

2. Combinations

3. A combination theorem

4. Technique and Method

The Combined Bound

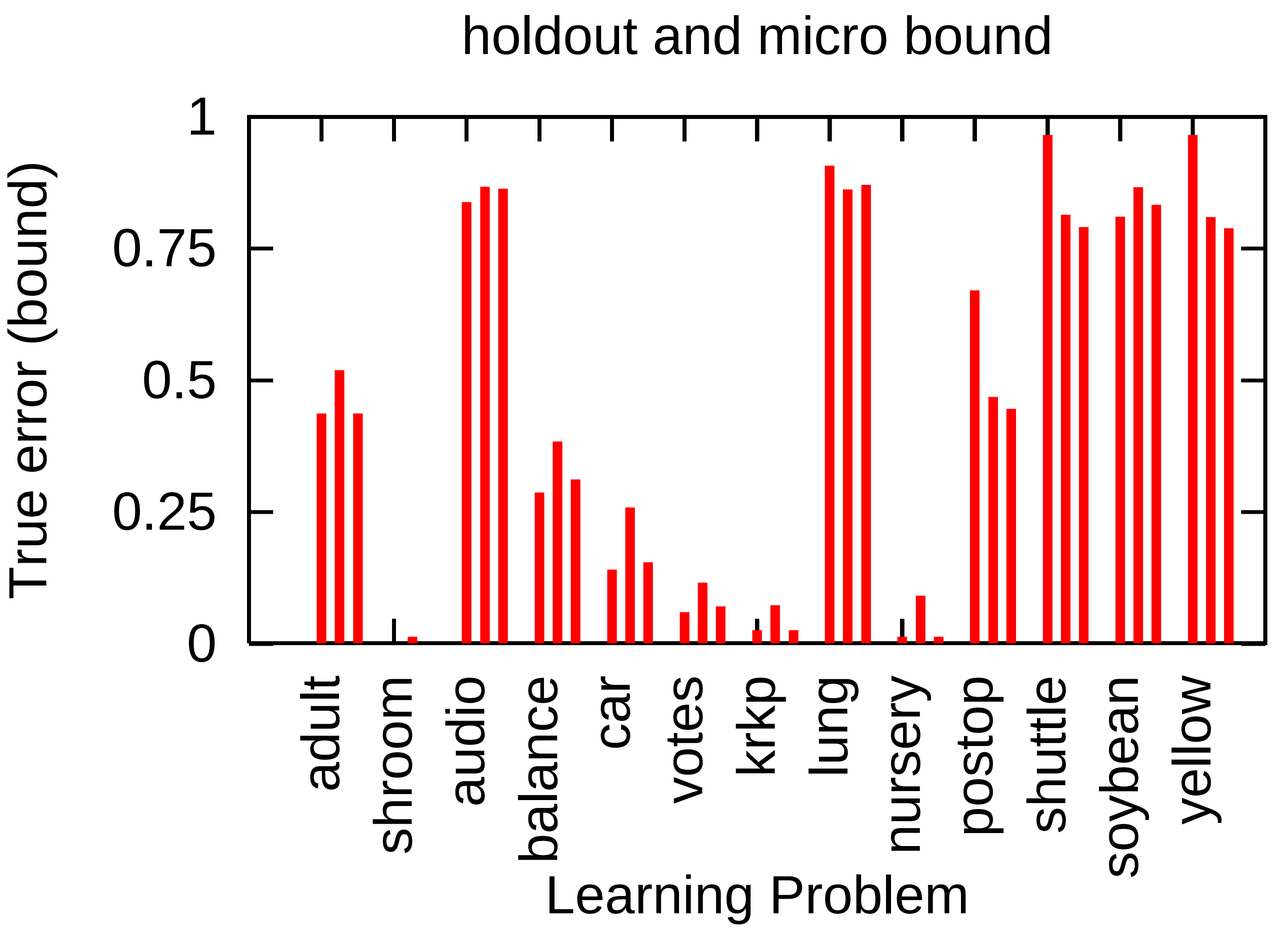
1. Fix bound.

2. Bound implies CDF, $F_{\text{train}}^{-1}(\delta)$, $F_{\text{test}}^{-1}(\delta)$.

3. Let rejection region $= F_{\text{train}}(\hat{e}_{\text{train}}(h))F_{\text{test}}(\hat{e}_{\text{test}}(h)) \leq t(\delta)$ where $t(\delta)$ satisfies:

$$\Pr_{x_{\text{test}},x_{\text{train}}}(F_{\text{test}}(x_{\text{test}})F_{\text{train}}(x_{\text{train}}) \leq t(\delta)) \leq \delta$$

4. Binary search on true error rate $\Rightarrow$ worst consistent true error rate = bound



left = test, middle = train, right = train and test

Conclusion: This approach works, and (empirically) never harms much while sometimes helping a lot.

Open:

1. Can we speed up the combined bound calculation?

2. Is this the right combination?