Quantitatively Tight Sample Complexity Bounds

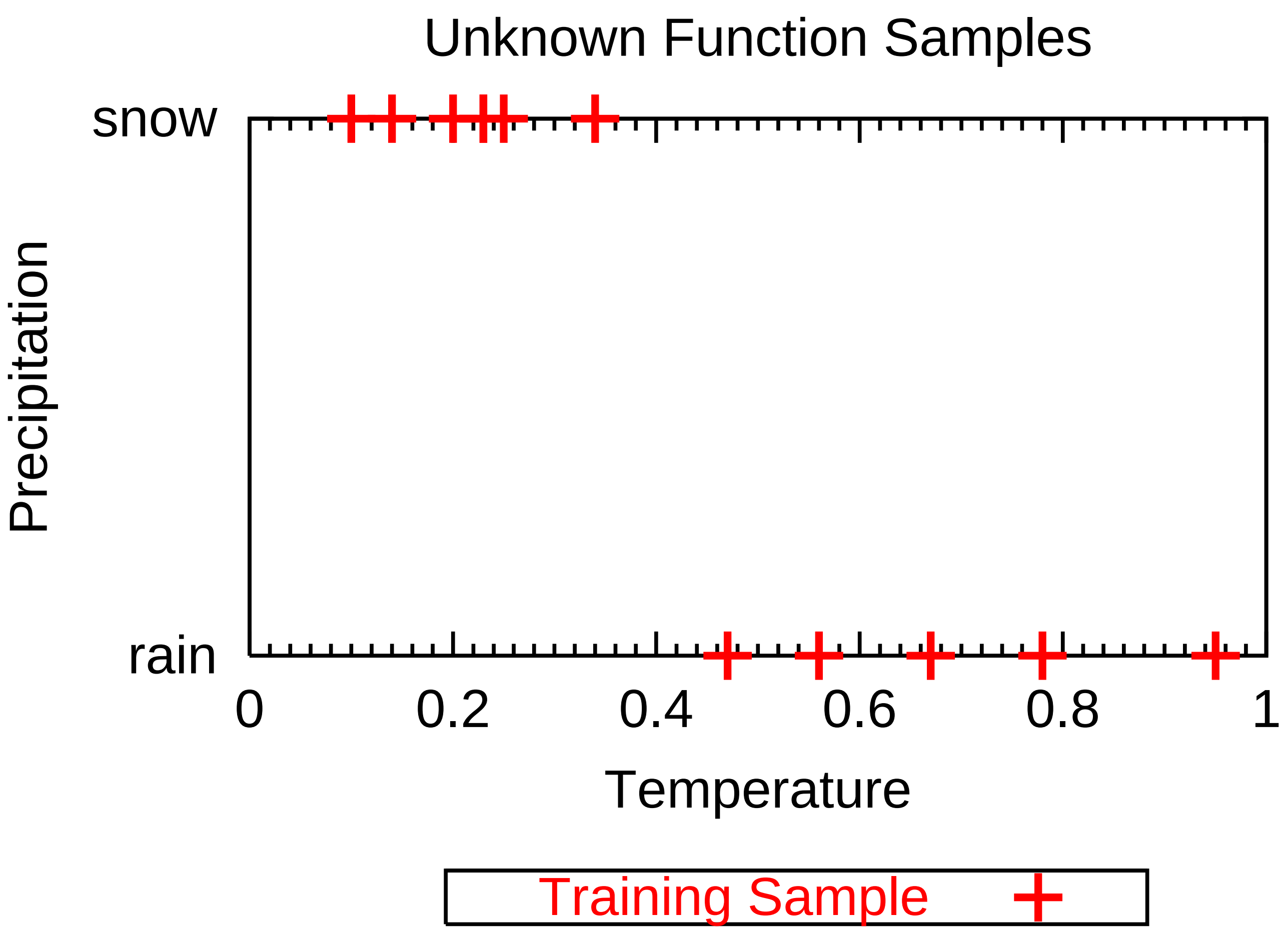
John Langford

The Classification Problem:

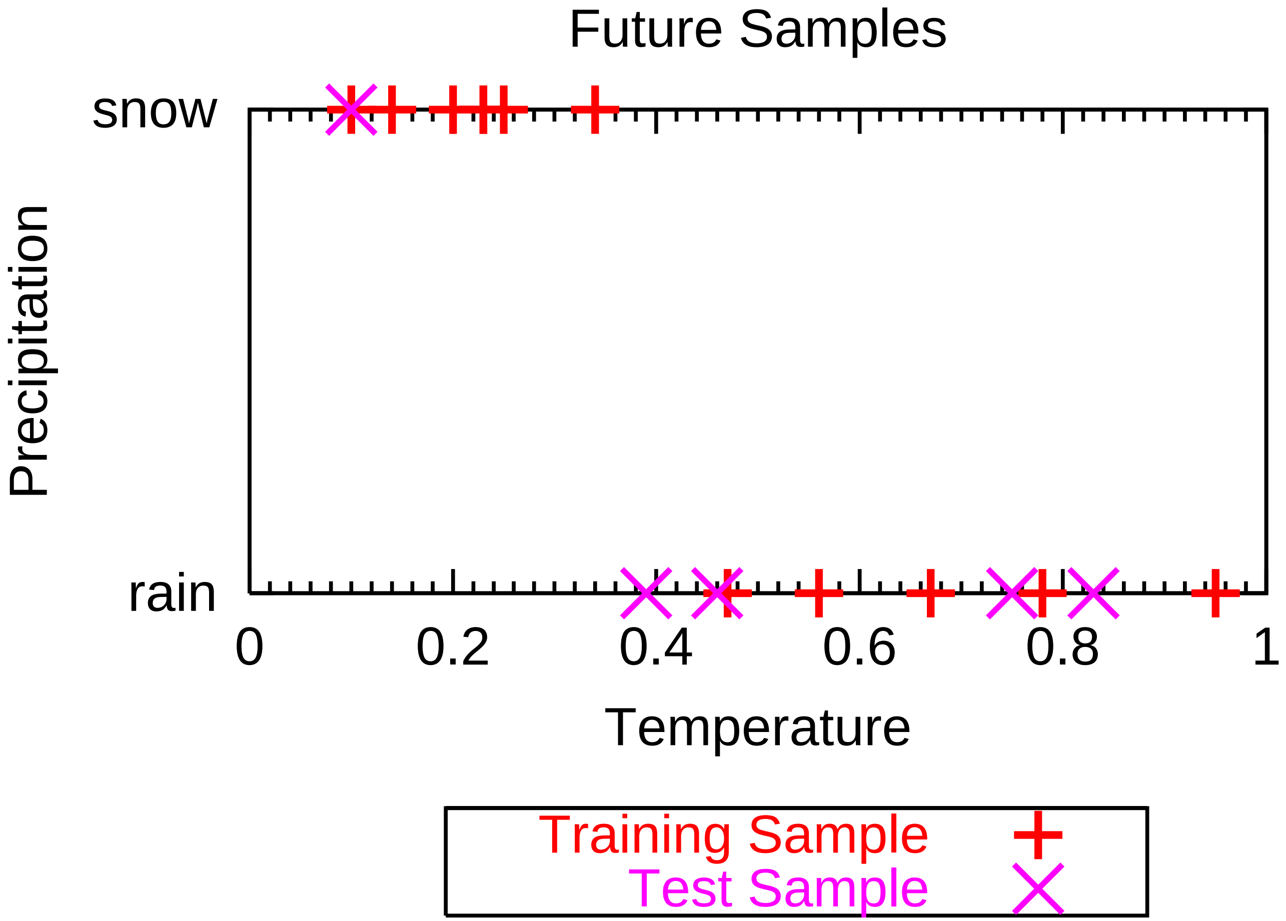
1. Get (input, output) pairs from unknown distribution D

2. Guess a function

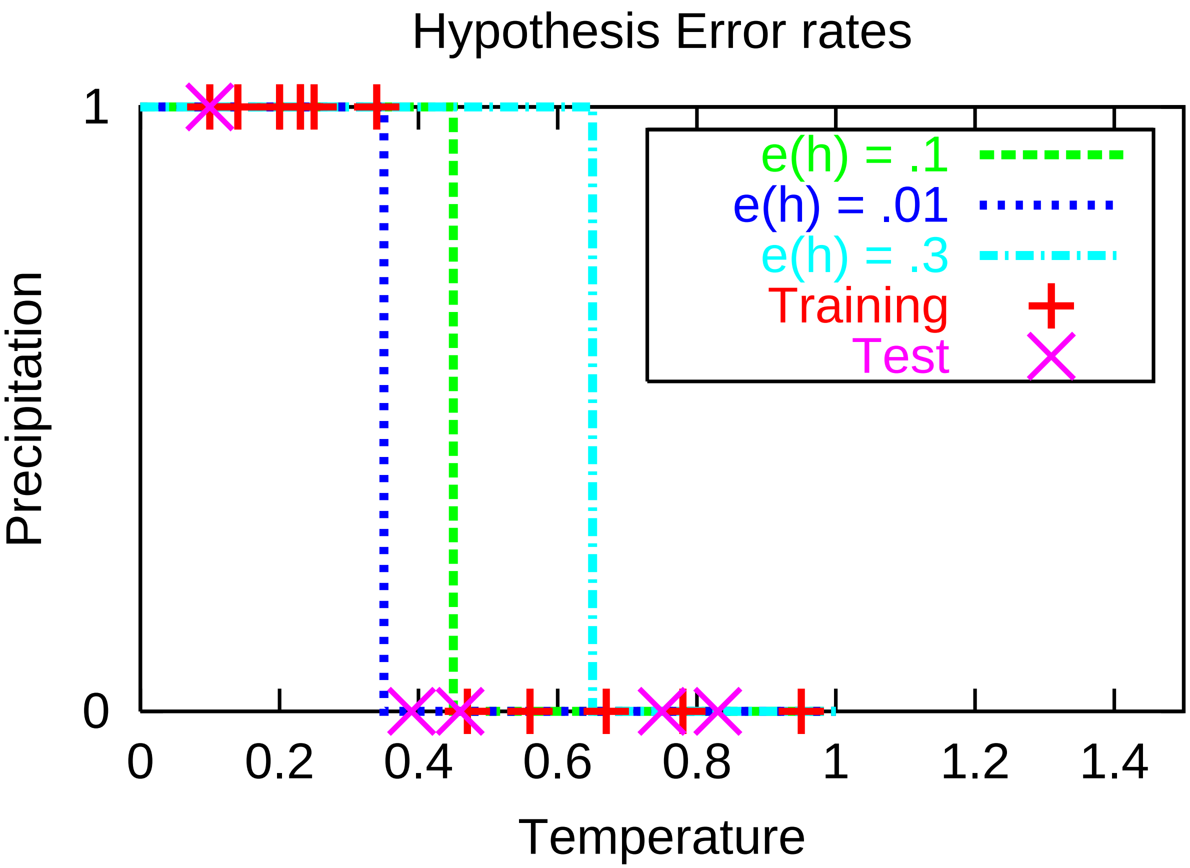
3. ... to minimize output errors on future inputs



Sample $(x, y) \sim D$



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



$h : X \to Y$ $\qquad$ $e(h) = \Pr_{(x,y)\sim D}(h(x) \neq y)$

Problem: We don't know the true error rate, $e(h)$.

How can we know how much learning has happened?

Knowing success: the common answer

The common answer: Use a holdout set of samples and report observed test error.

+ Simple and Easy

- How sure of the answer should you be?

How sure can you be?

Holdout set size = 1000? 100? 10? 1?

How sure can you be?

Holdout set size = 1000? 100? 10? 1?

Common hack: estimate empirical variance

How sure can you be?

Holdout set size = 1000? 100? 10? 1?

Common hack: estimate empirical variance

Hack failure: 0 empirical error $\neq$ absolute confidence

How sure can you be?

Holdout set size = 1000? 100? 10? 1?

Common hack: estimate empirical variance

Hack failure: 0 empirical error $\neq$ absolute confidence

Hack #2: more complicated techniques (cross validation, bagging, etc...)

How sure can you be?

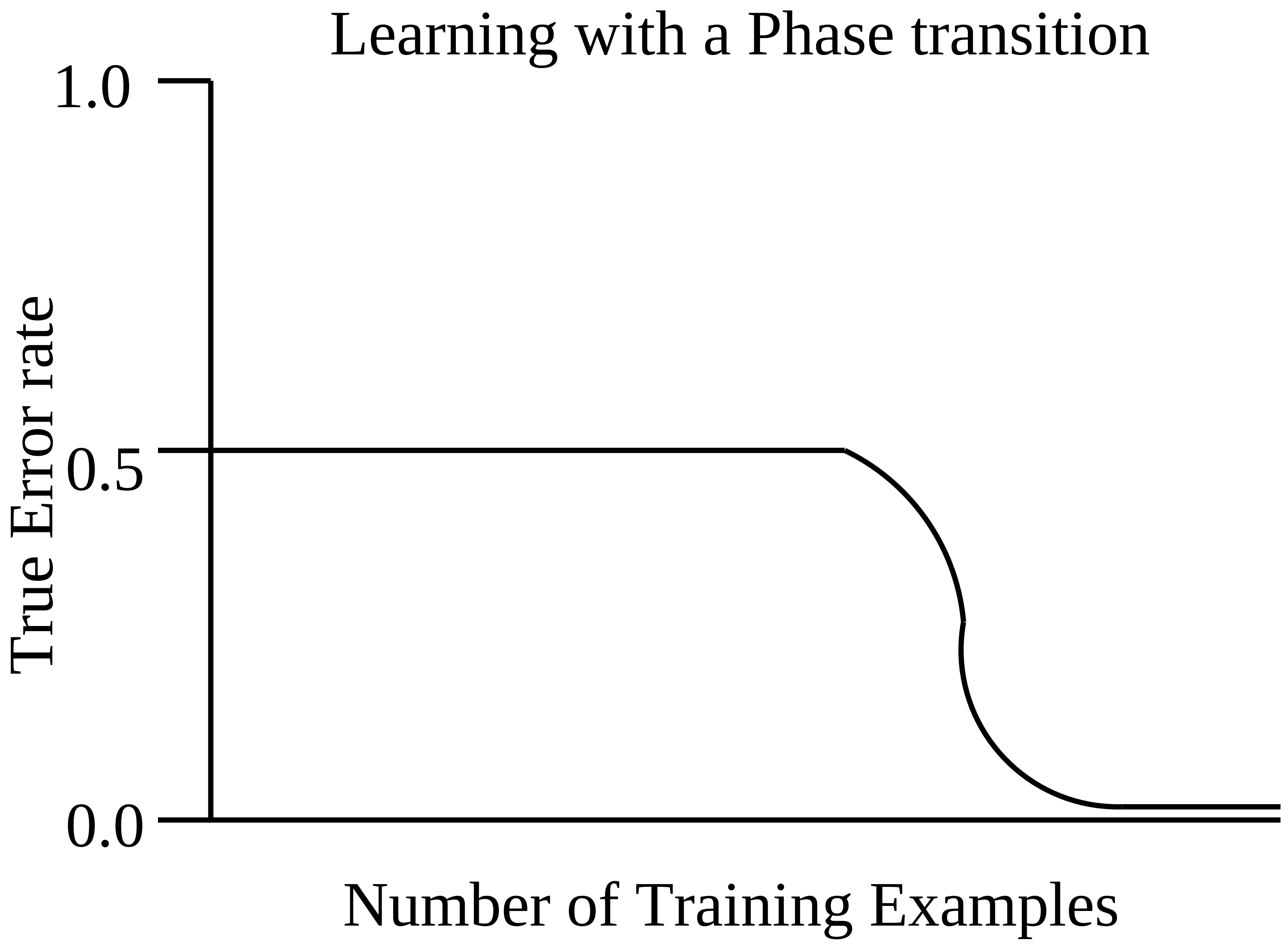
Holdout set size = 1000? 100? 10? 1?

Common hack: estimate empirical variance

Hack failure: 0 empirical error $\neq$ absolute confidence

Hack #2: more complicated techniques (cross validation, bagging, etc...)

Hack failure: still no guarantees (in general) and none can exist


Learning with a Phase transition
1.0
0.5
0.0
True Error rate
Number of Training Examples

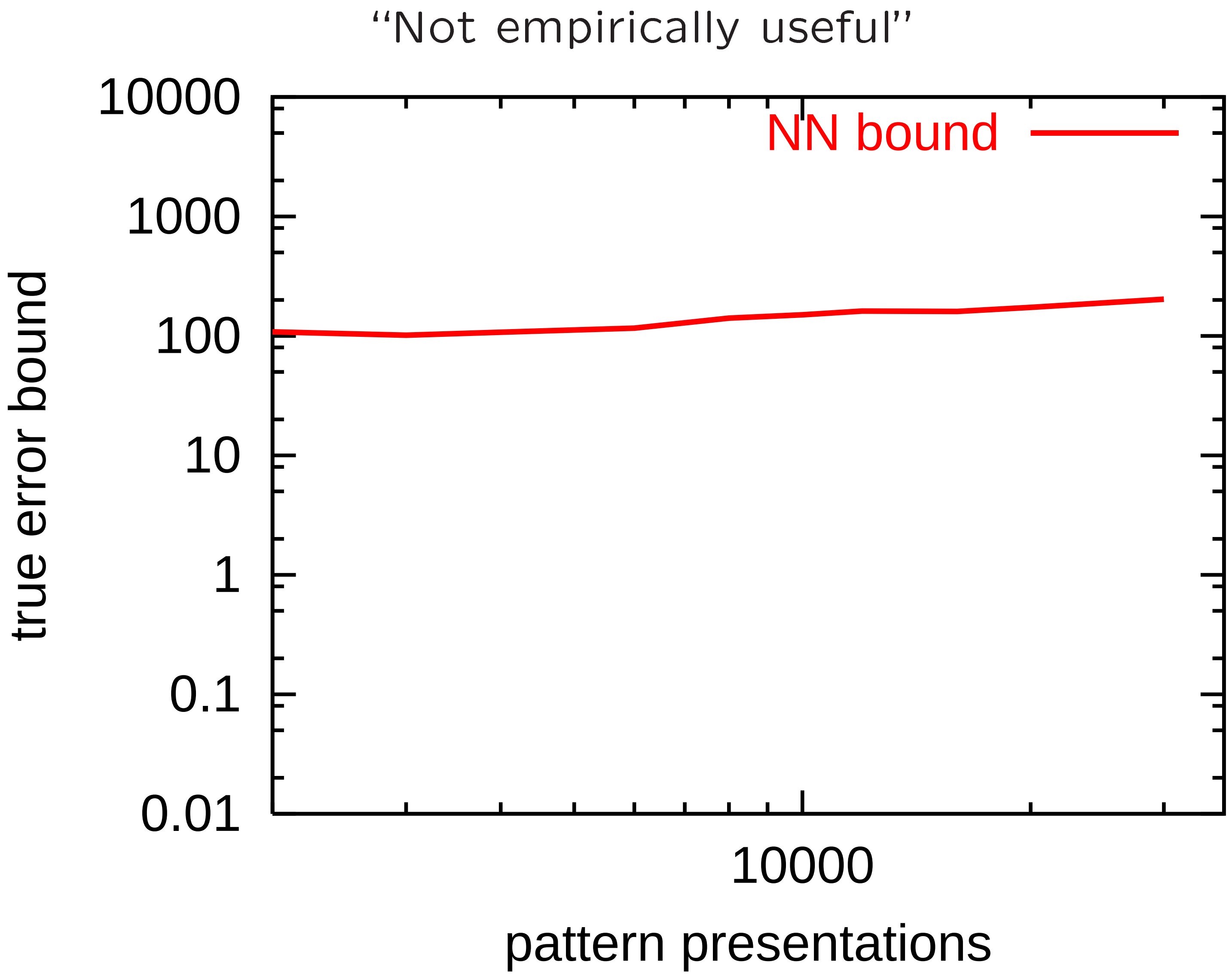
Knowing Success: the theoretical answer

1. Assume sample independence
2. Fix a hypothesis set
3. Upper bound the number of examples required so that the empirical error of *every* hypothesis is near the true error.

+ Guarantee exists subject to explicit assumptions

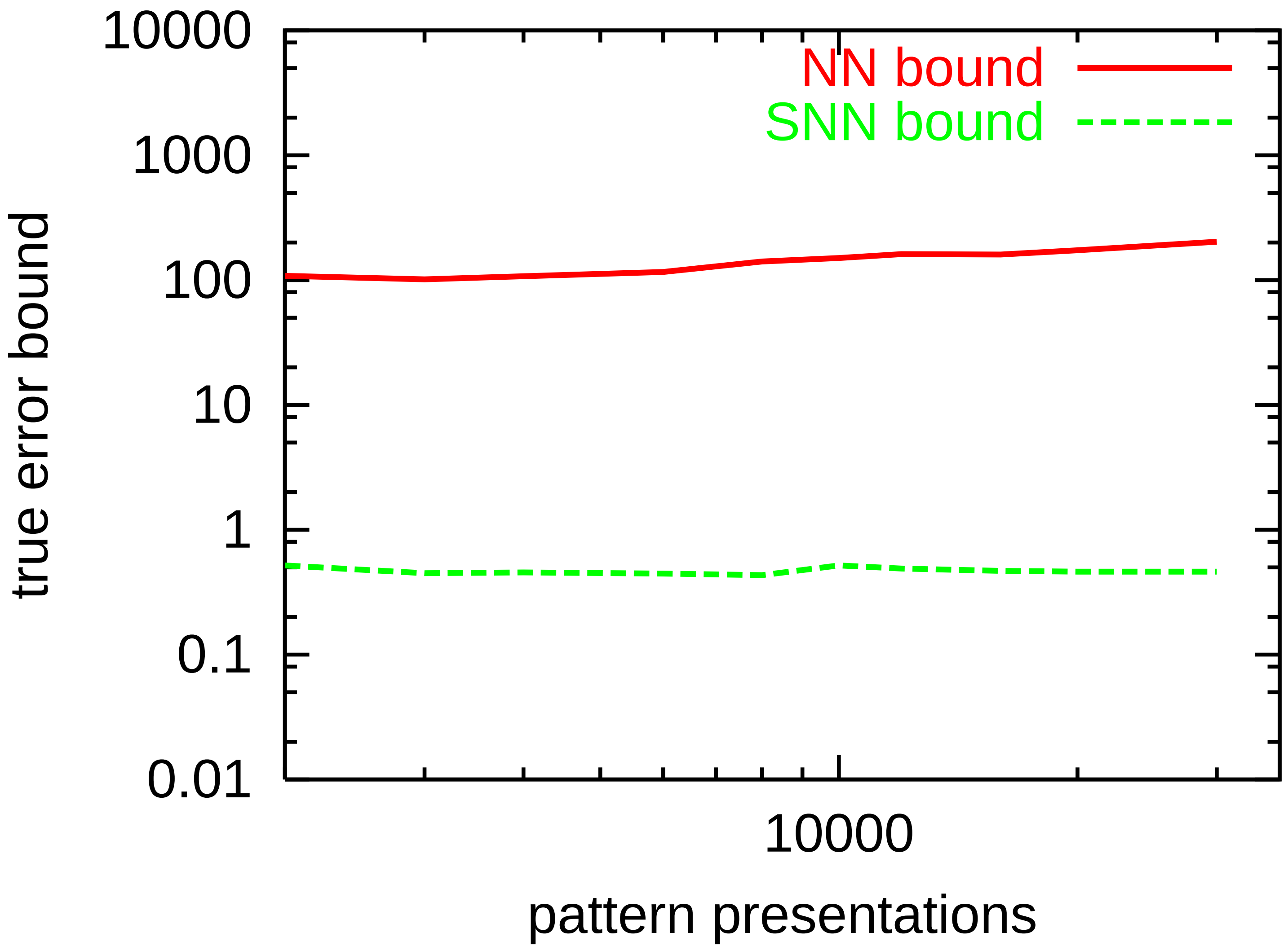
+ (Near) optimal asymptotic rates

- (Typically) Not empirically useful



"Not empirically useful"
NN bound
true error bound
10000
1000
100
10
1
0.1
0.01
10000
pattern presentations

A teaser



NN bound
SNN bound
true error bound
10000
1000
100
10
1
0.1
0.01
10000
pattern presentations

Thesis

The guarantees of theoretical approaches can be made both practical and useful.

Theoretical approach

1. Focus on *only* answering "How much have I learned?"

2. Fundamental assumption: Sample independence

3. Remove approximations wherever possible

The Master Plan

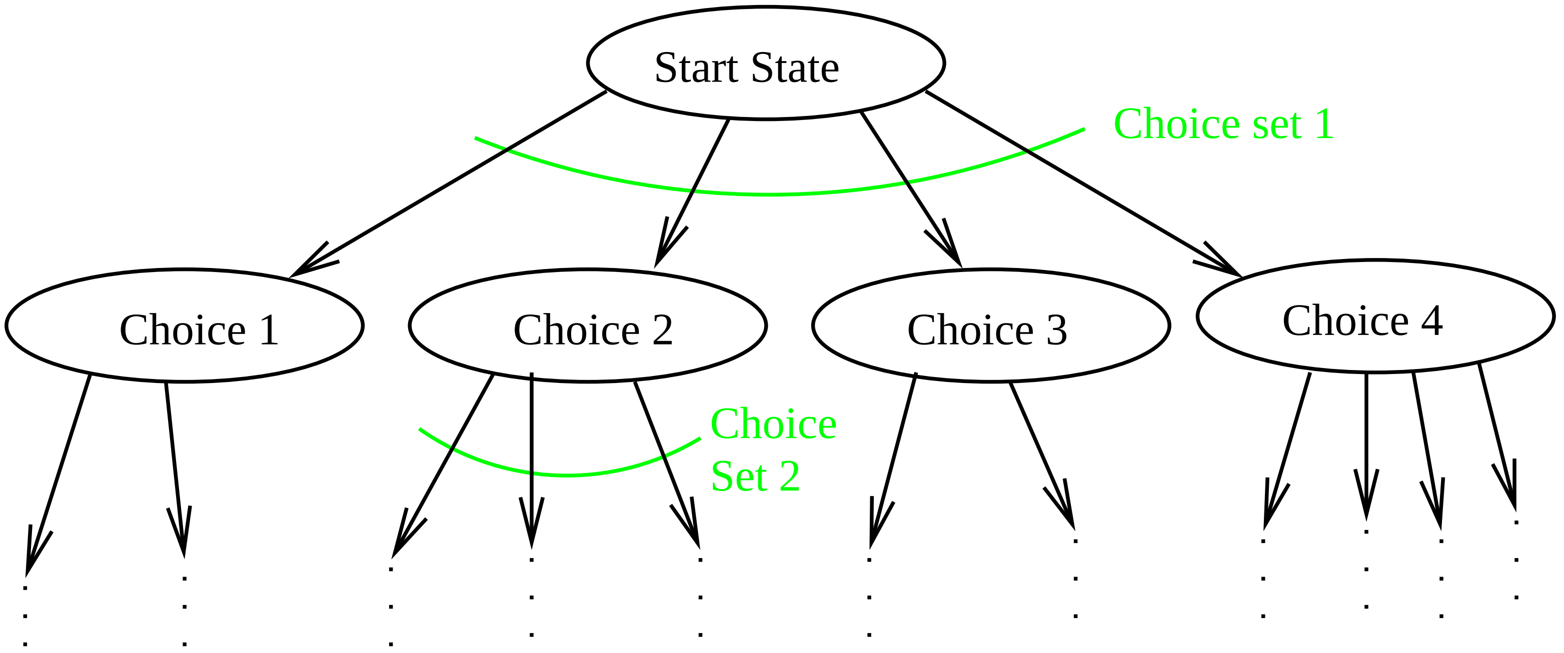
1. Microchoice Bound Detail

2. Bound Overview

3. Empirical Results

Microchoice details

Learning algorithm = sequence of data-dependent choices ending in a hypothesis



Observation

There “ought” to exist a bound parameterized by the choice sets.

Microchoice Bound

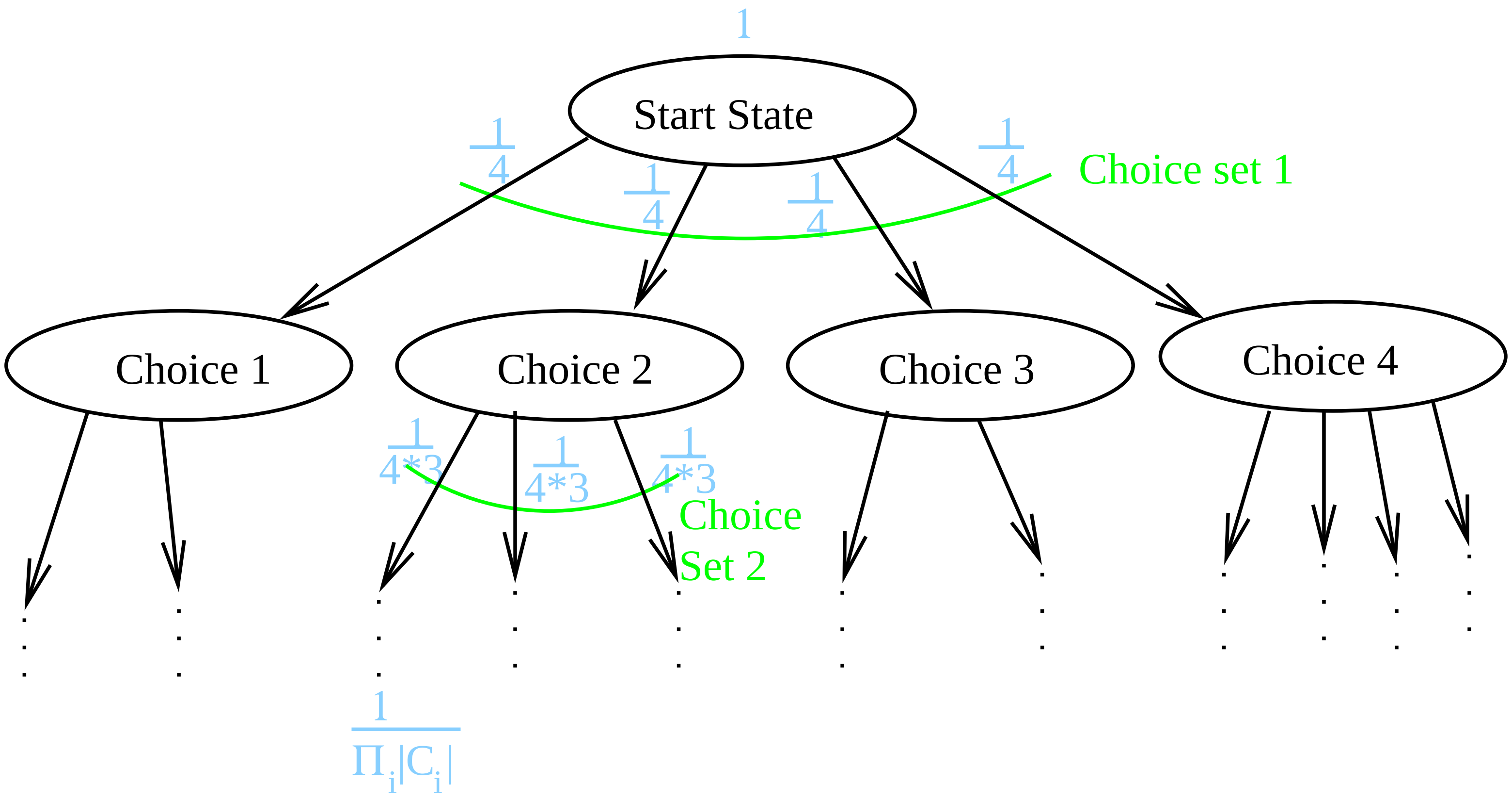
Let $C_i(\text{path})$ = ith choice set on path.

Let $P(h) = \sum_{\text{paths to h}} \frac{1}{\prod_{j=1}^{d(h)} |C_j(\text{path})|}$.

Cool fact: $\sum_h P(h) = 1$ and we know (BEHW87, realizable case):

$$\bar{e}(h) \leq \frac{\ln \frac{1}{P(h)} + \ln \frac{1}{\delta}}{m}$$

The Microchoice Algorithm



$$P(h) \geq \frac{1}{\prod_i |C_i(\text{path to } h)|}$$

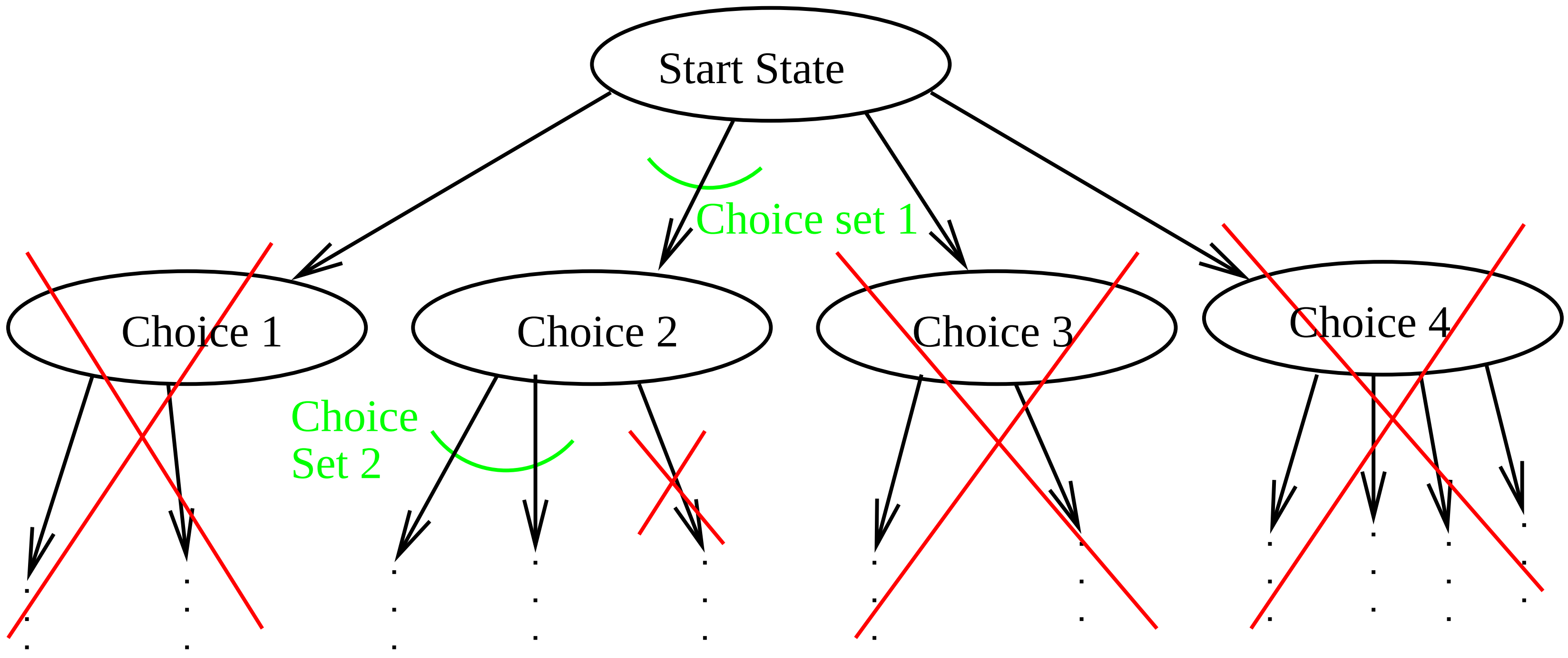
Adaptive Microchoice

What about many extremely low probability choices?

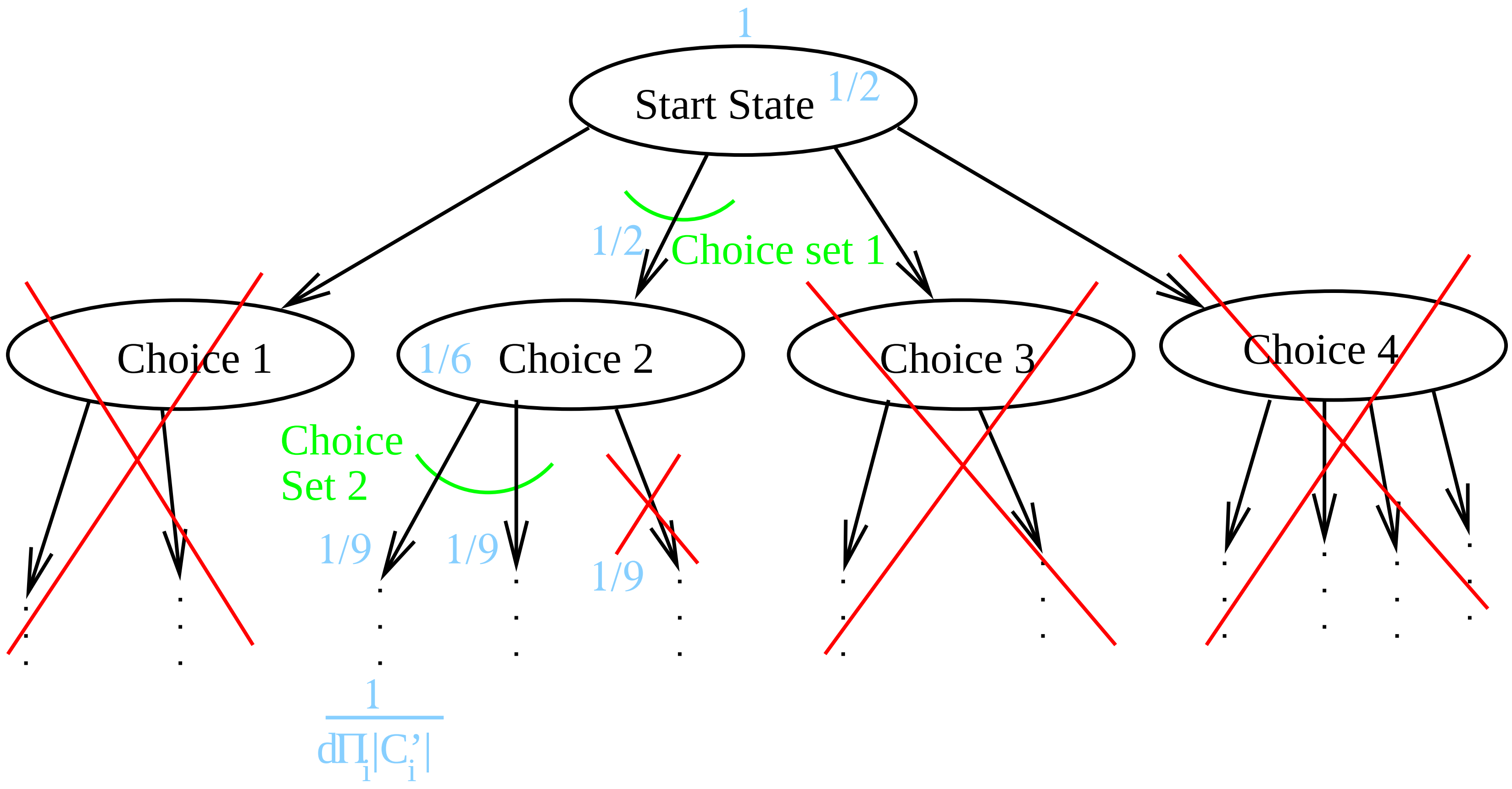
1. Throw away extreme low probability choices.

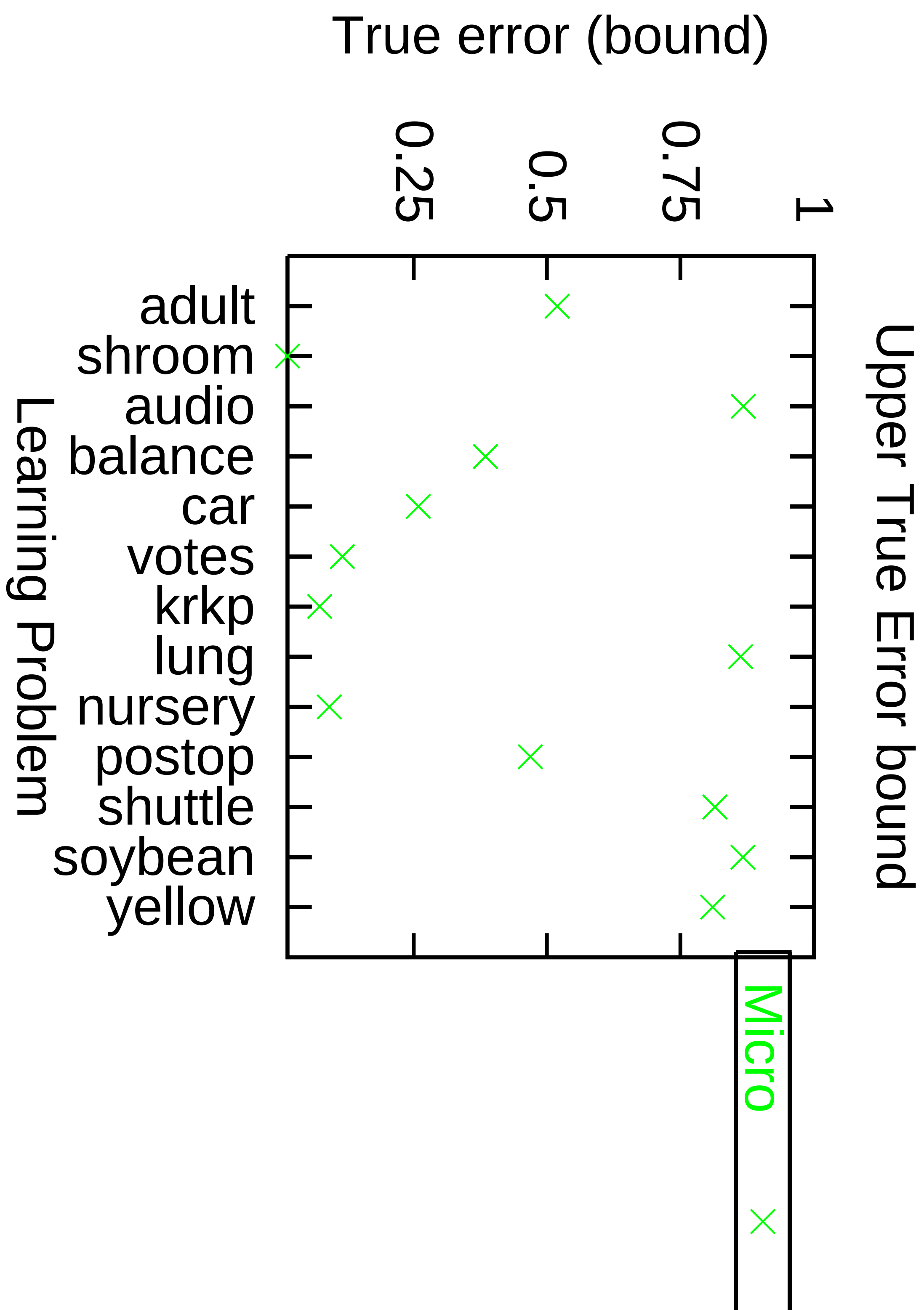
2. Bound probability of falling off the pruned choice tree.

Adaptive Microchoice Tree



Adaptive Microchoice Algorithm




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

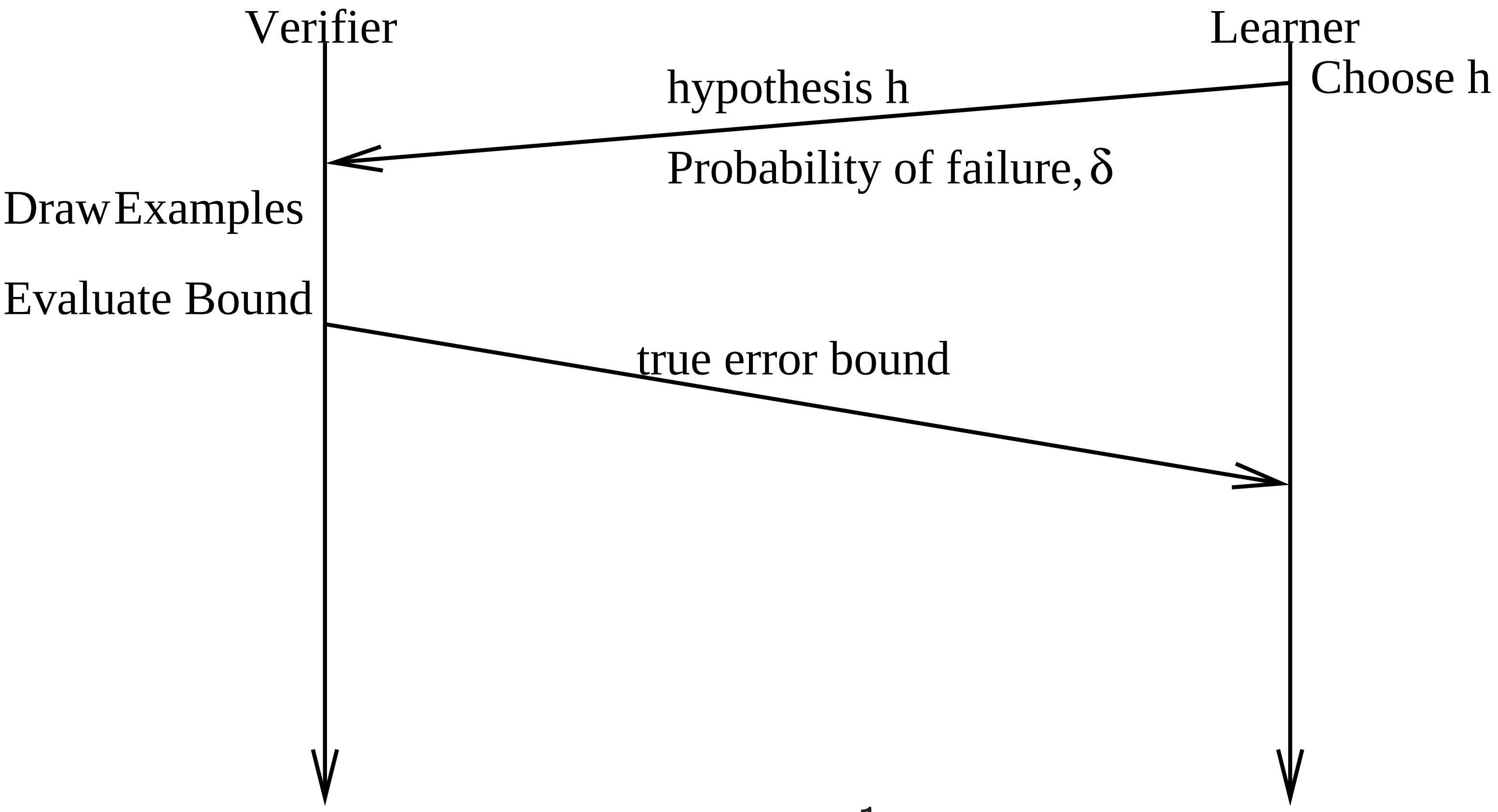
The Master Plan

1. Microchoice Bound Detail
2. Bound Overview
3. Empirical Results

Bound Overview

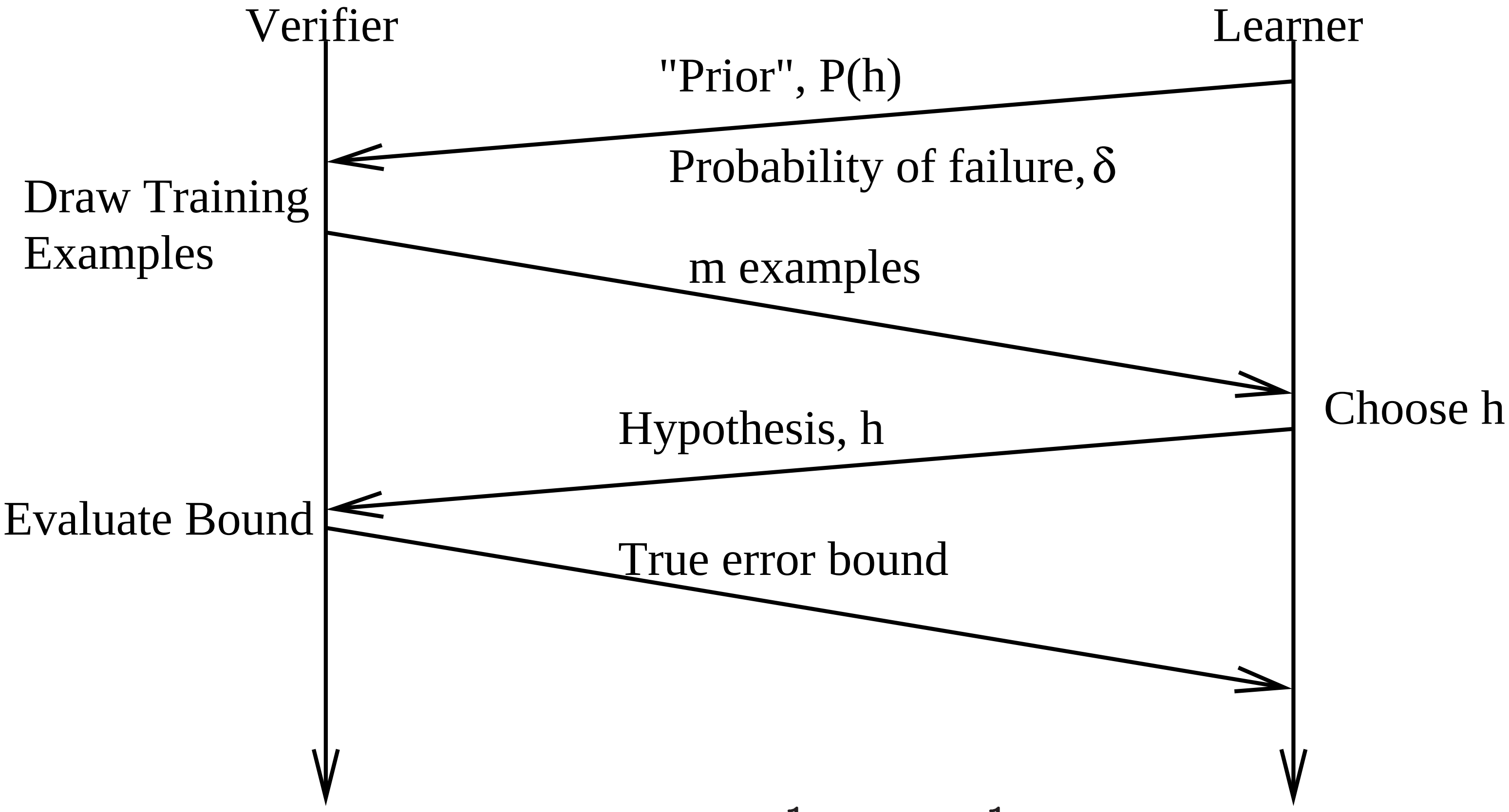
bound	type	Work done	Coauthors	Builds upon
Test set	Test	Ref./None		Chernoff?
Train set	D. Train	Ref./None		BEHW
Microchoice	D. Train	Built	Avrim	BEHW/F
PAC-Bayes	C. Train	Refined		McAllester
Shell	C. Train	Built	McAllester	S&J
Averaging	C. Train	Built	S&M	SFBL
Progressive	Test	Built	A&K	Chernoff
Train and Test	Meta	Built		

Test Set Bound



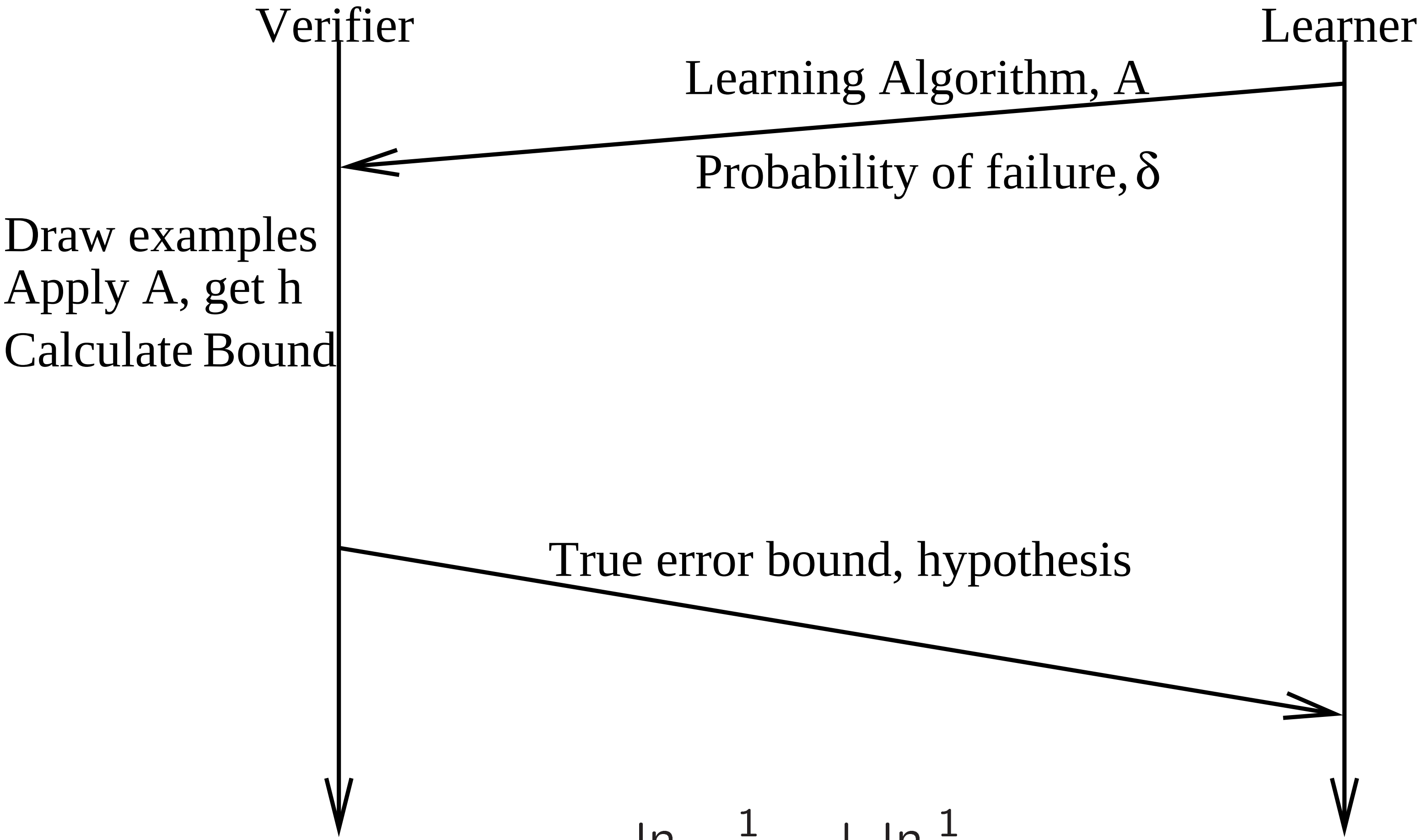
$$\bar{e}(h) \leq \frac{\ln \frac{1}{\delta}}{m}$$

Training Set Bound



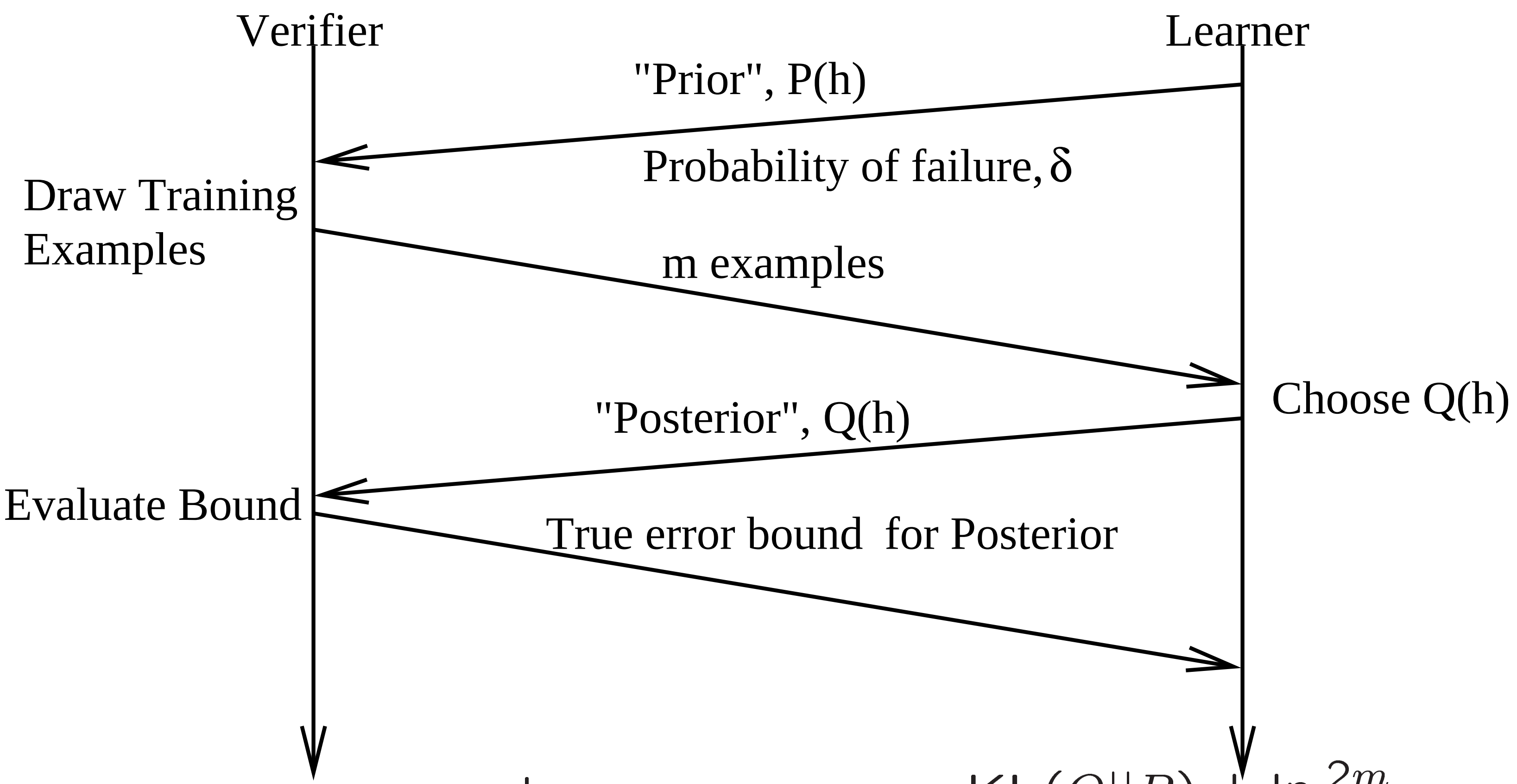
$$\bar{e}(h) \leq \frac{\ln \frac{1}{P(h)} + \ln \frac{1}{\delta}}{m}$$

Microchoice Bound



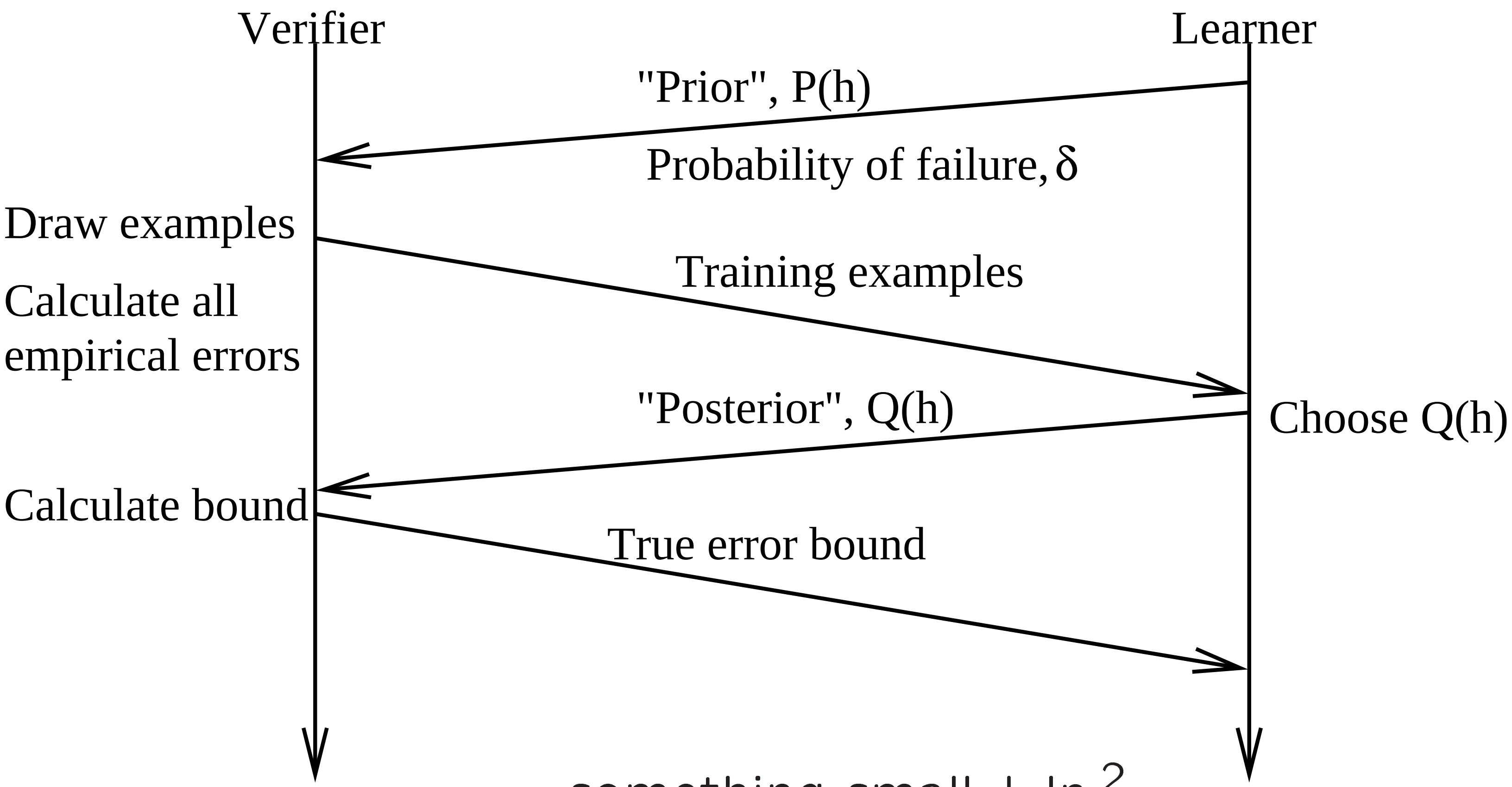
$$\bar{e}(h) \leq \frac{\ln \frac{1}{P_A(h)} + \ln \frac{1}{\delta}}{m}$$

PAC−Bayes Bound



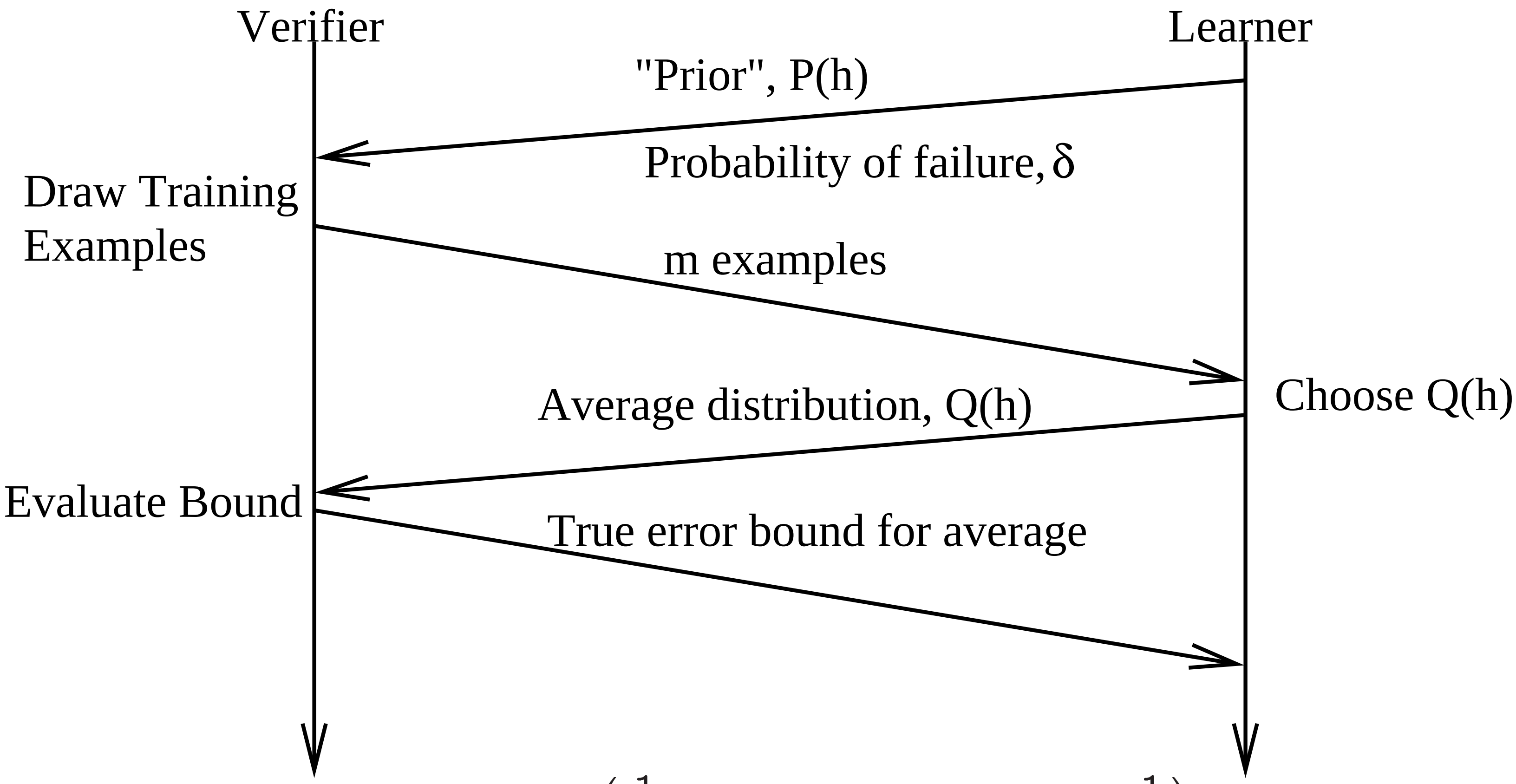
$$\mathsf{KL}(Q||P) = E_Q \frac{\ln q_i}{\ln p_i} \qquad \bar{e}(h_Q) \le \frac{\mathsf{KL}(Q||P) + \ln \frac{2m}{\delta}}{m-1}$$

Shell Bound



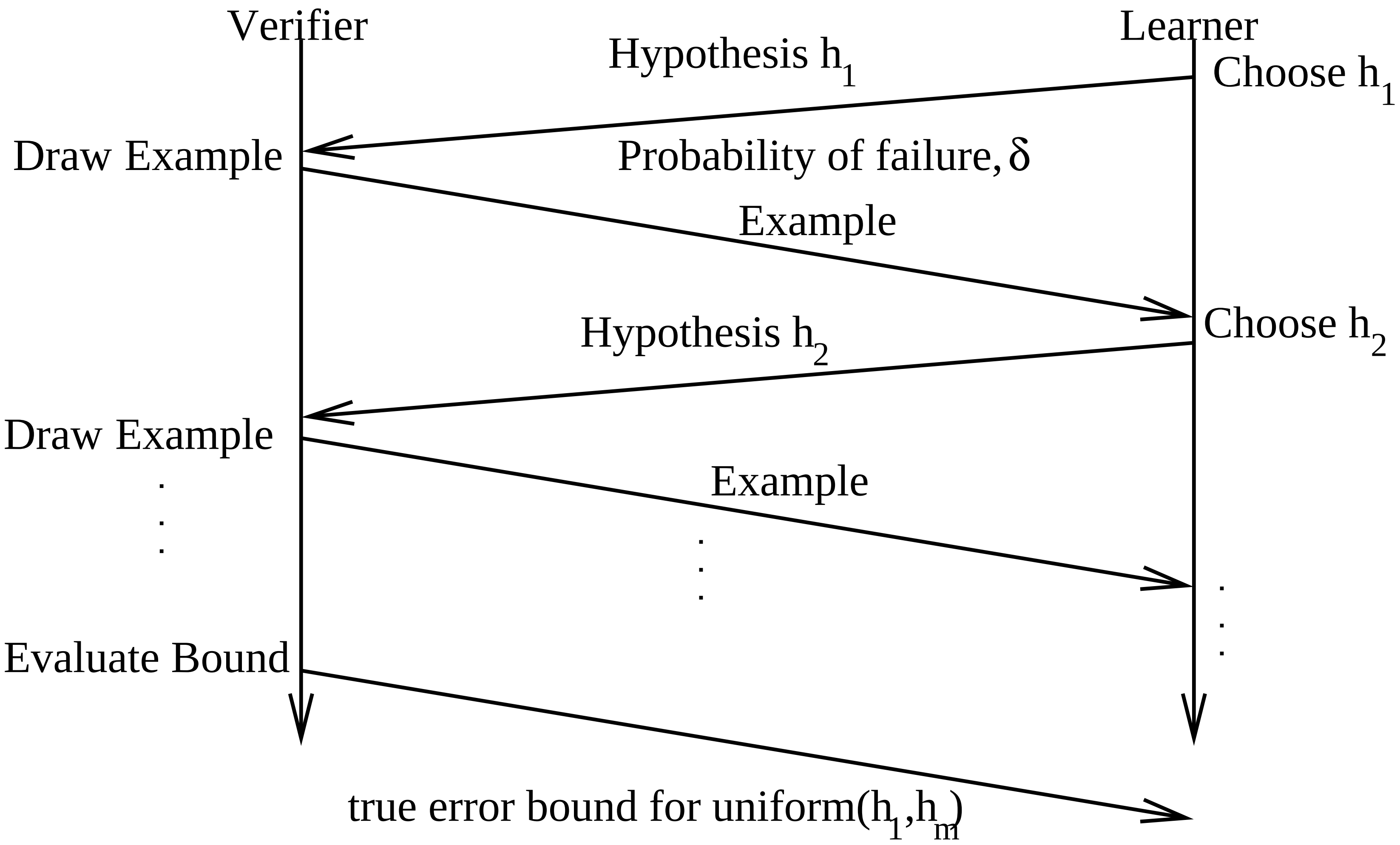
$$\bar{e}(h) \leq \frac{\text{something small} + \ln \frac{2}{\delta}}{m}$$

Averaging Bound



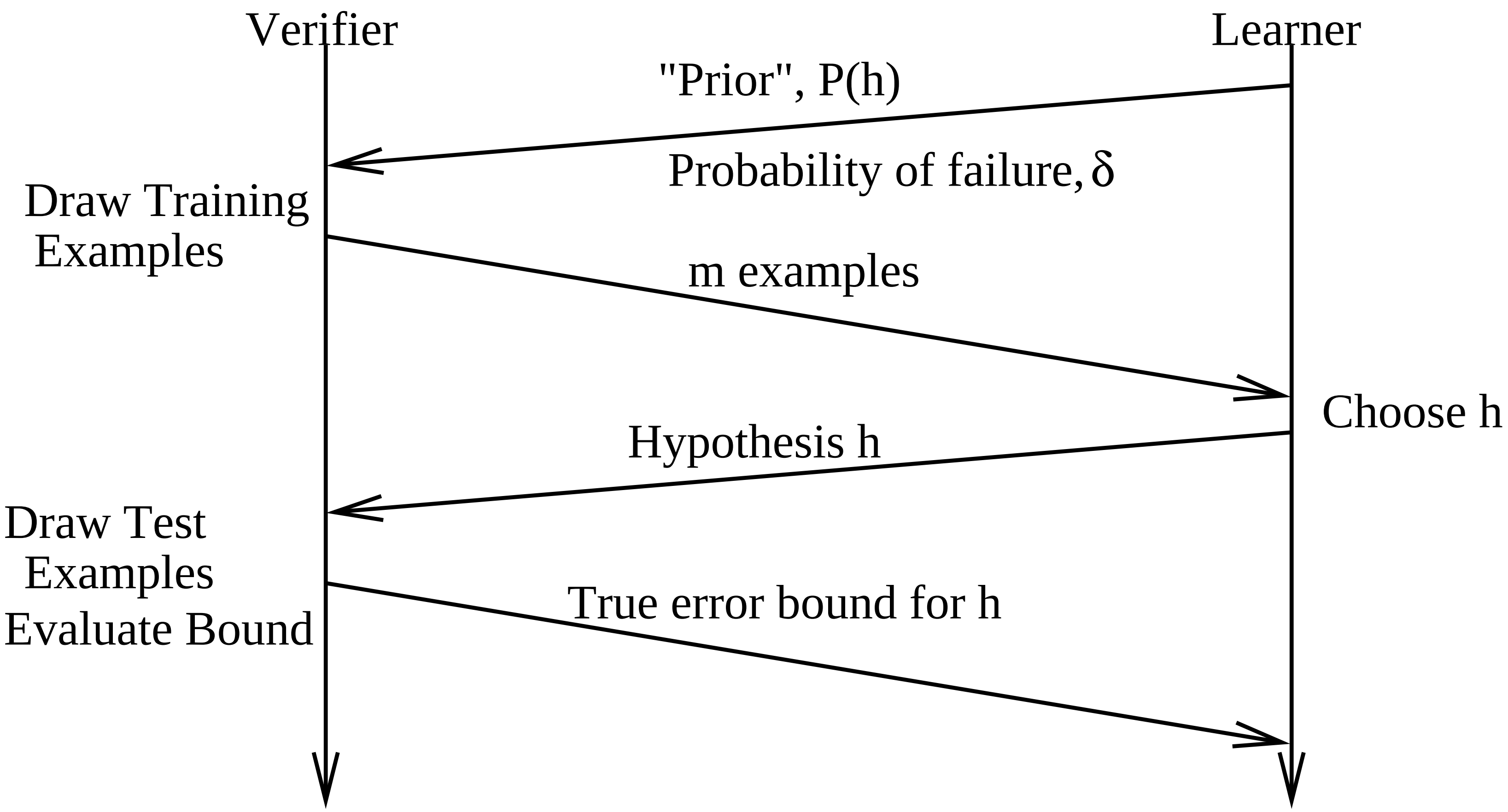
$$\bar{e}(h_Q) \leq O\left(\frac{\frac{1}{\theta^2}\mathsf{KL}(Q||P)\ln m + \ln\frac{1}{\delta}}{m}\right)$$

Progressive Validation Bound



$$\bar{e}(U(h_1,..,h_m)) \leq \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}$$

Train and Test Bound



$$\bar{e}(h) \simeq \min(\bar{e}_{\mathsf{train}}(h), \bar{e}_{\mathsf{test}}(h))$$

The Master Plan

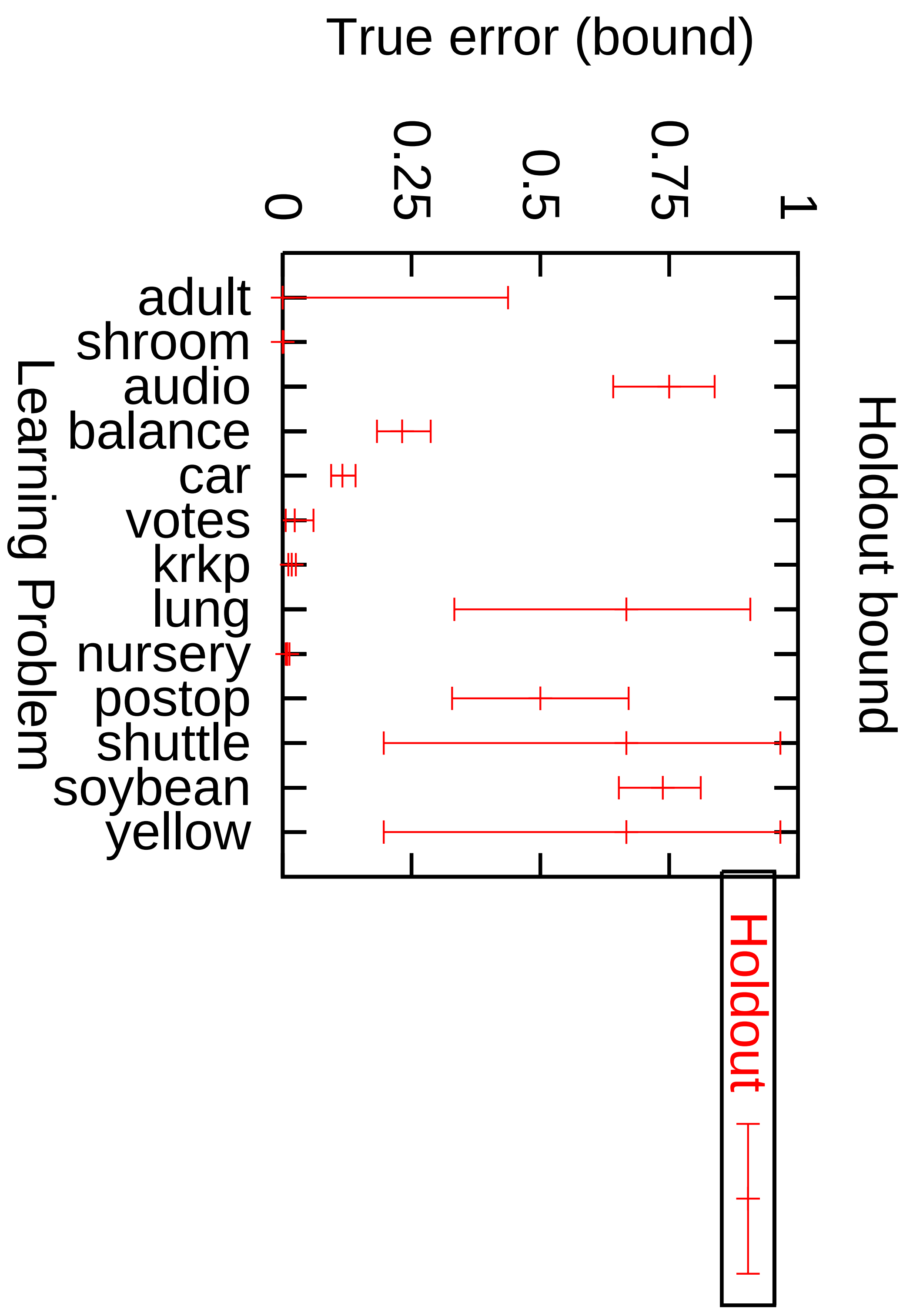
1. Microchoice Bound Detail

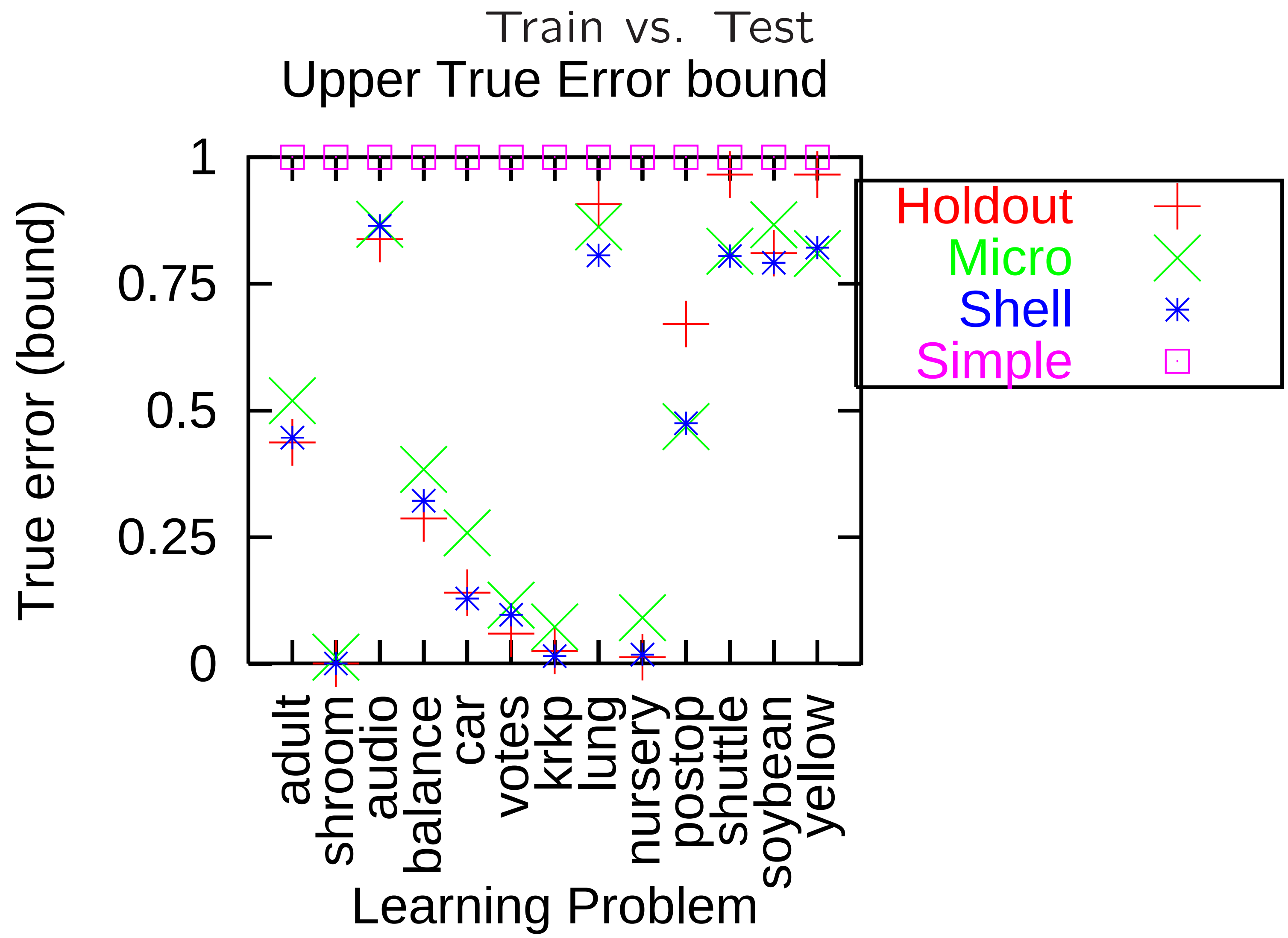
2. Bound Overview

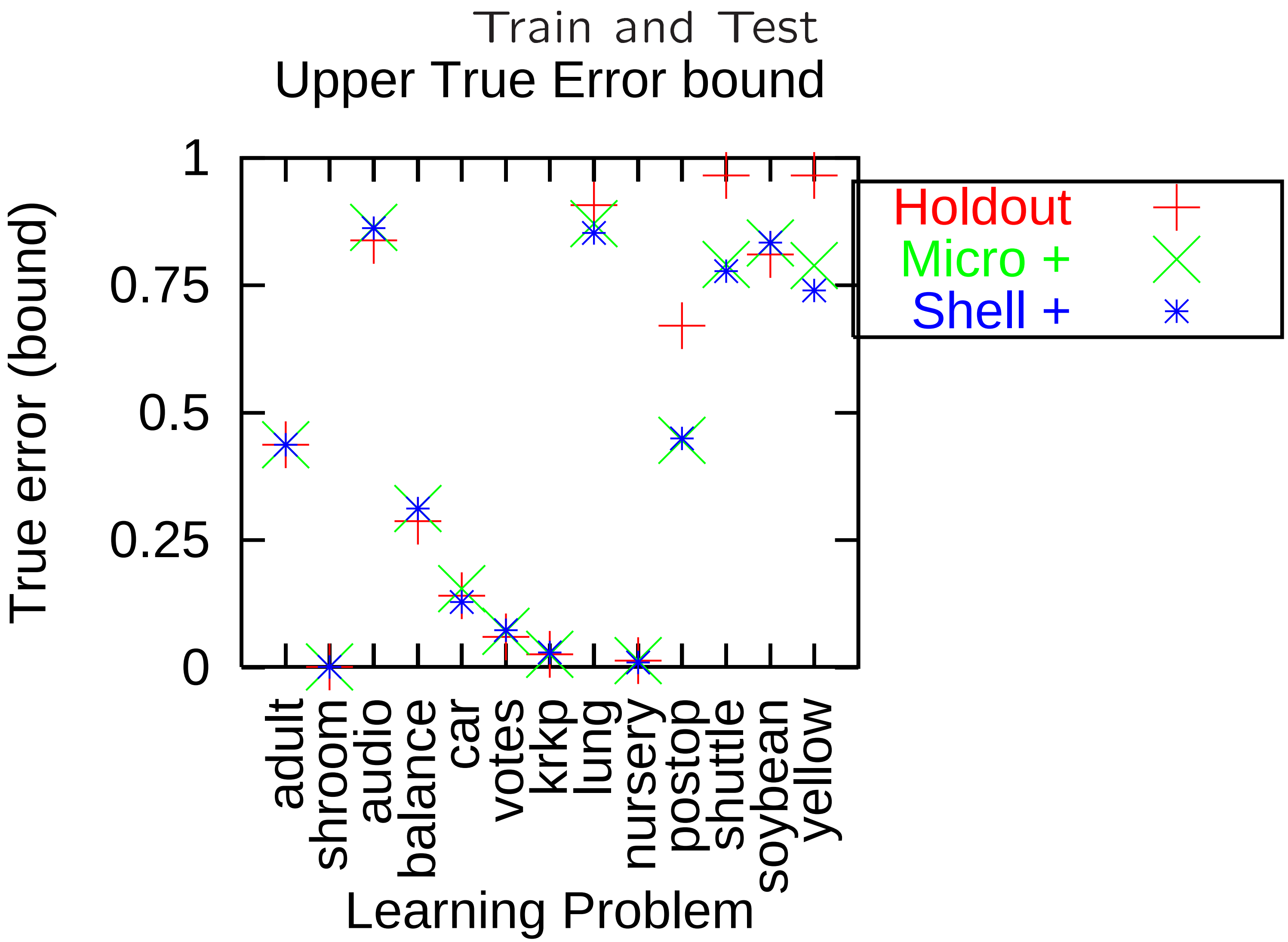
3. Empirical Results

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


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
Holdout bound
Holdout


Train vs. Test
Upper True Error bound
True error (bound)
1
0.75
0.5
0.25
0
adult
shroom
audio
balance
car
votes
krkp
lung
nursery
postop
shuttle
soybean
yellow
Learning Problem
Holdout
Micro
Shell
Simple


Train and Test
Upper True Error 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
Holdout
Micro +
Shell +

Results: Neural Networks (with Rich Caruana)

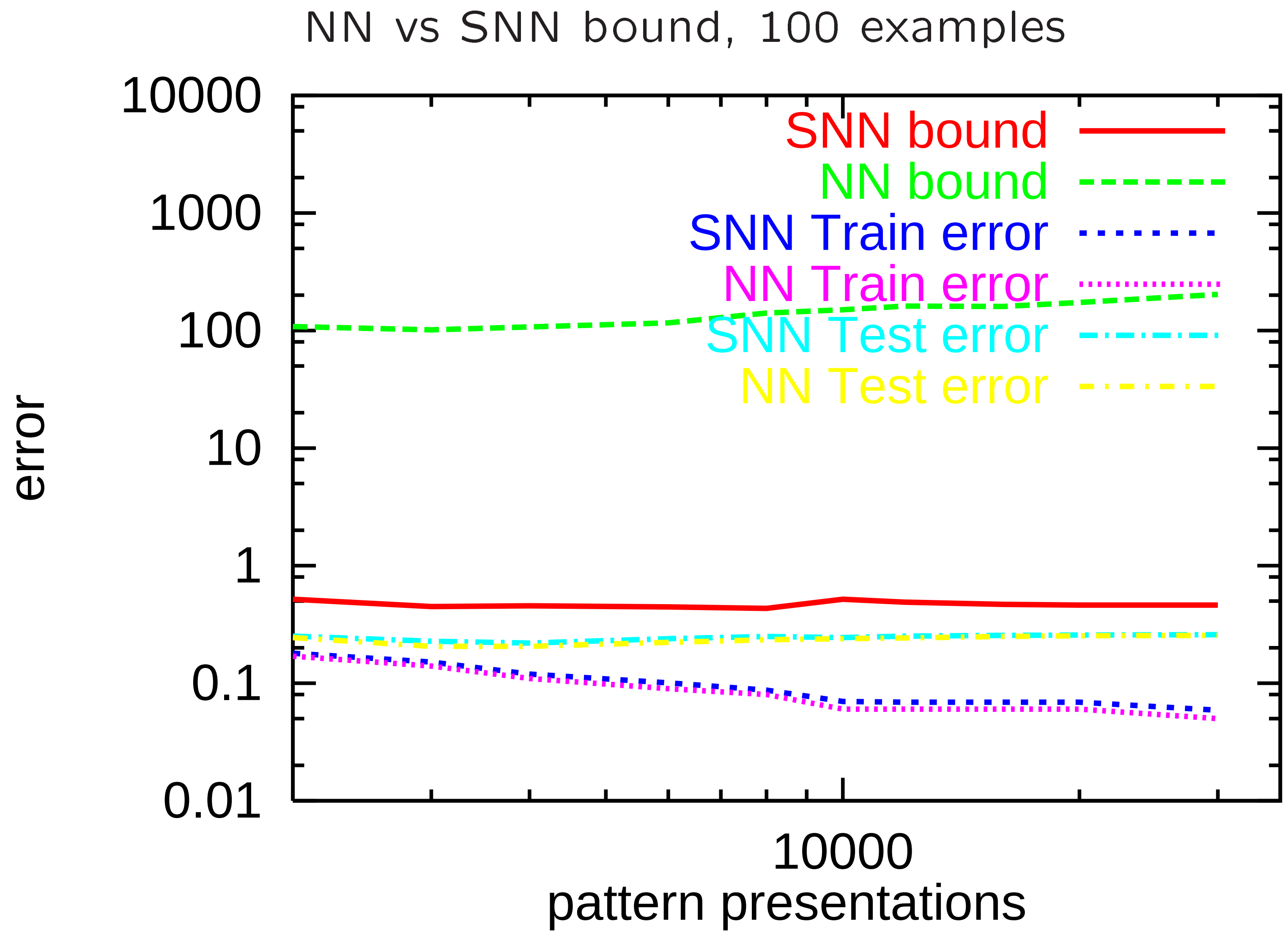
1. Noisy artificial problem

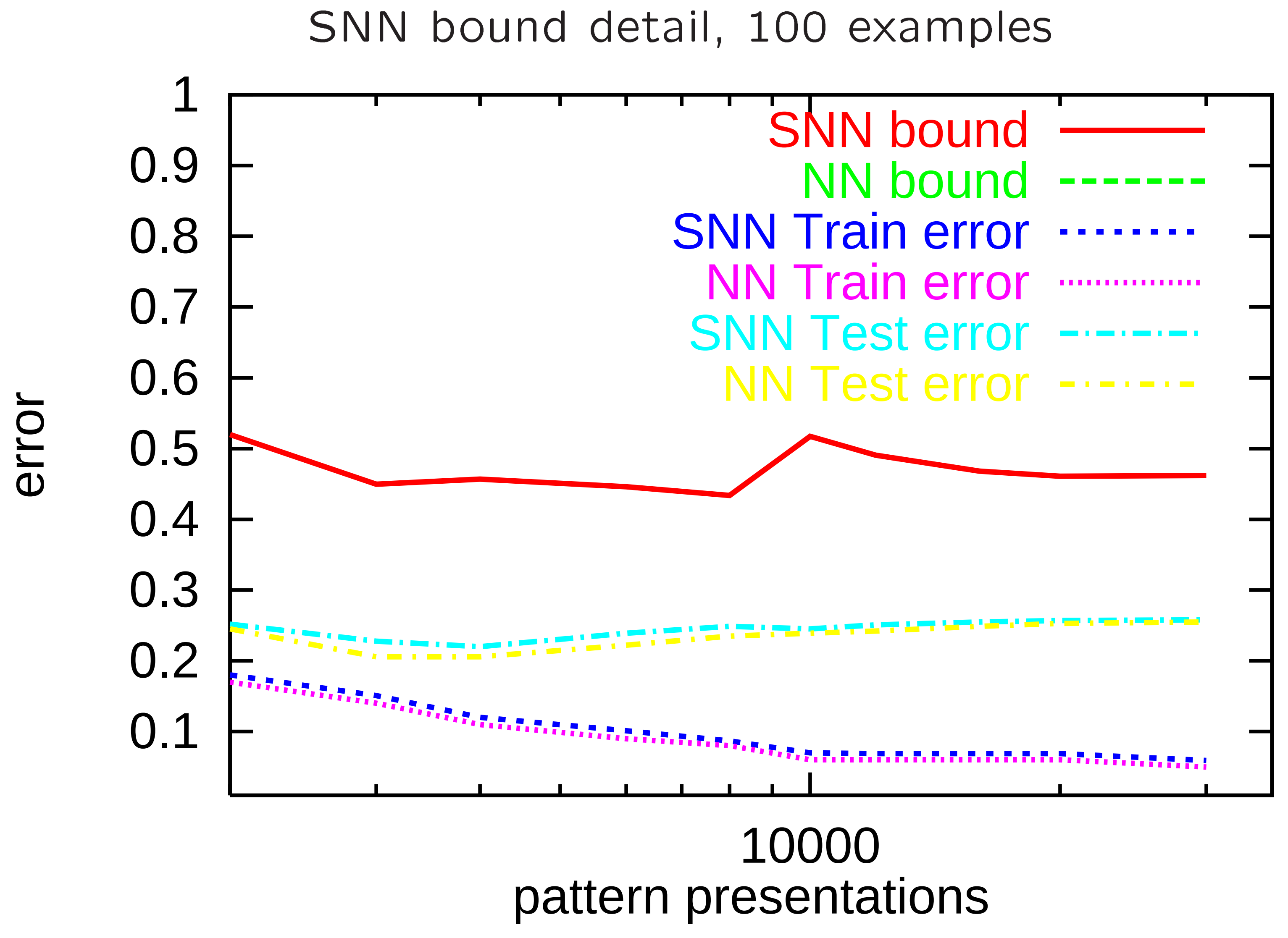
2. Train 2 hidden unit NN via gradient descent

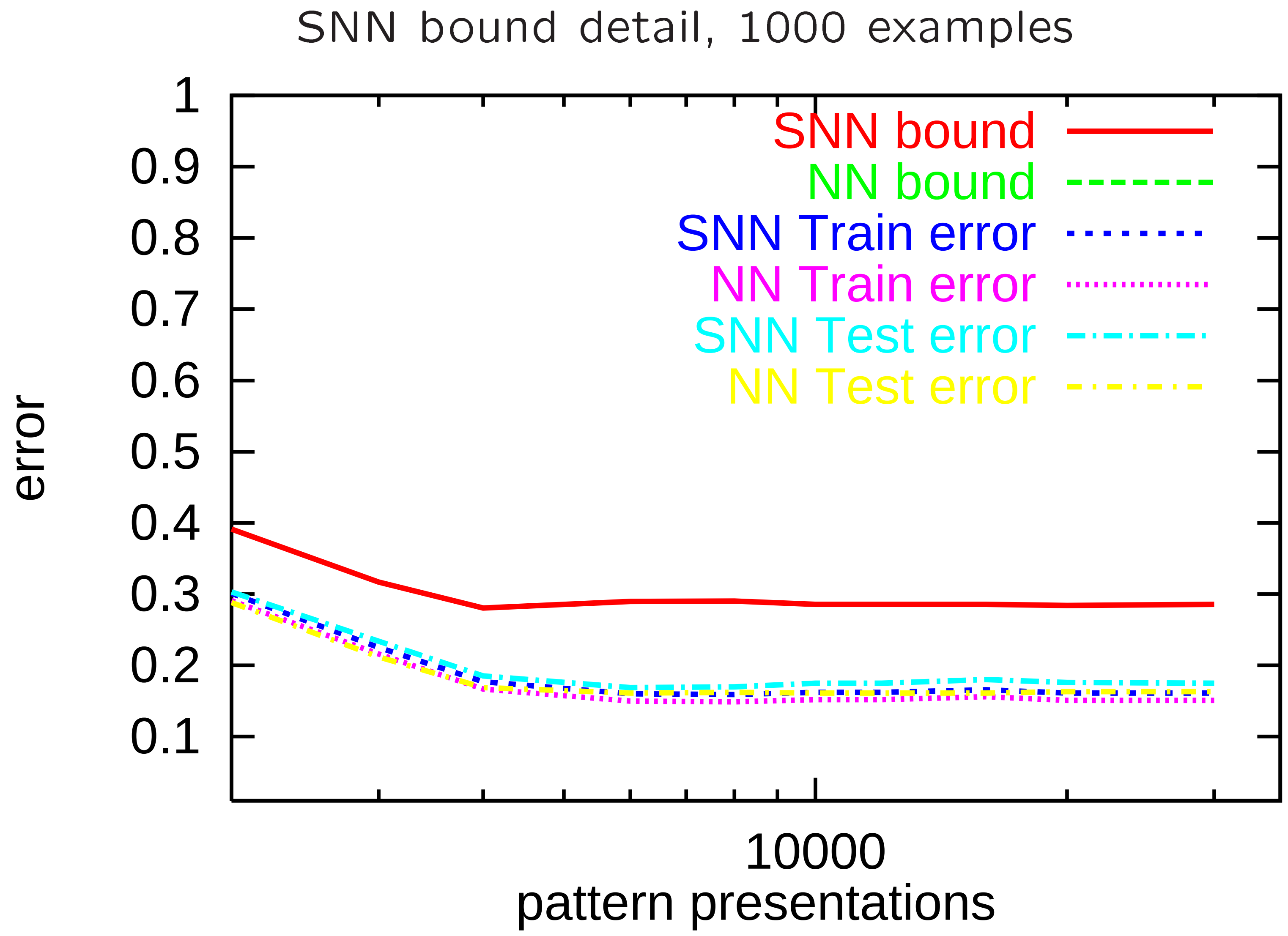
3. Do sensitivity analysis on weights

4. Choose "posterior" $Q =$ multidimensional gaussian with mean $=$ weights and variance $= \frac{1}{\text{sensitivity}}$

5. probability of failure $= \delta = 0.1$


NN vs SNN bound, 100 examples
error
10000
1000
100
10
1
0.1
0.01
10000
pattern presentations
SNN bound
NN bound
SNN Train error
NN Train error
SNN Test error
NN Test error


SNN bound detail, 100 examples
error
pattern presentations
10000
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
SNN bound
NN bound
SNN Train error
NN Train error
SNN Test error
NN Test error



SNN bound detail, 1000 examples
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
error
10000
pattern presentations
SNN bound
NN bound
SNN Train error
NN Train error
SNN Test error
NN Test error

Demo: holdout bound calculation

Conclusion

We can do machine learning and provide strong guarantees. With proper attention to details, training set based bounds can be useful, especially in conjunction with test set based bounds.

Code for bound calculation at:

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

Open Problems

1. (v. hard) Find a *quantitative* solution to the correlation problem. (PAC-Bayes? Alternate approach in chapter 8)

2. (hard, easy) Improved PAC-Bayes is not entirely tight. Can we make it tighter? What is a PAC-Bayes lower bound?

3. (v. hard) Find a true error bound for cross-validation with *arbitrary* learning algorithms.

4. (hard) Tighten the Progressive Validation bound so that it doesn't use the Hoeffding/Azuma approximation.