A PAC-MDL Bound

John Langford (joint work with Avrim Blum)

Probably

Approximately

Correct-

Minimum

Description

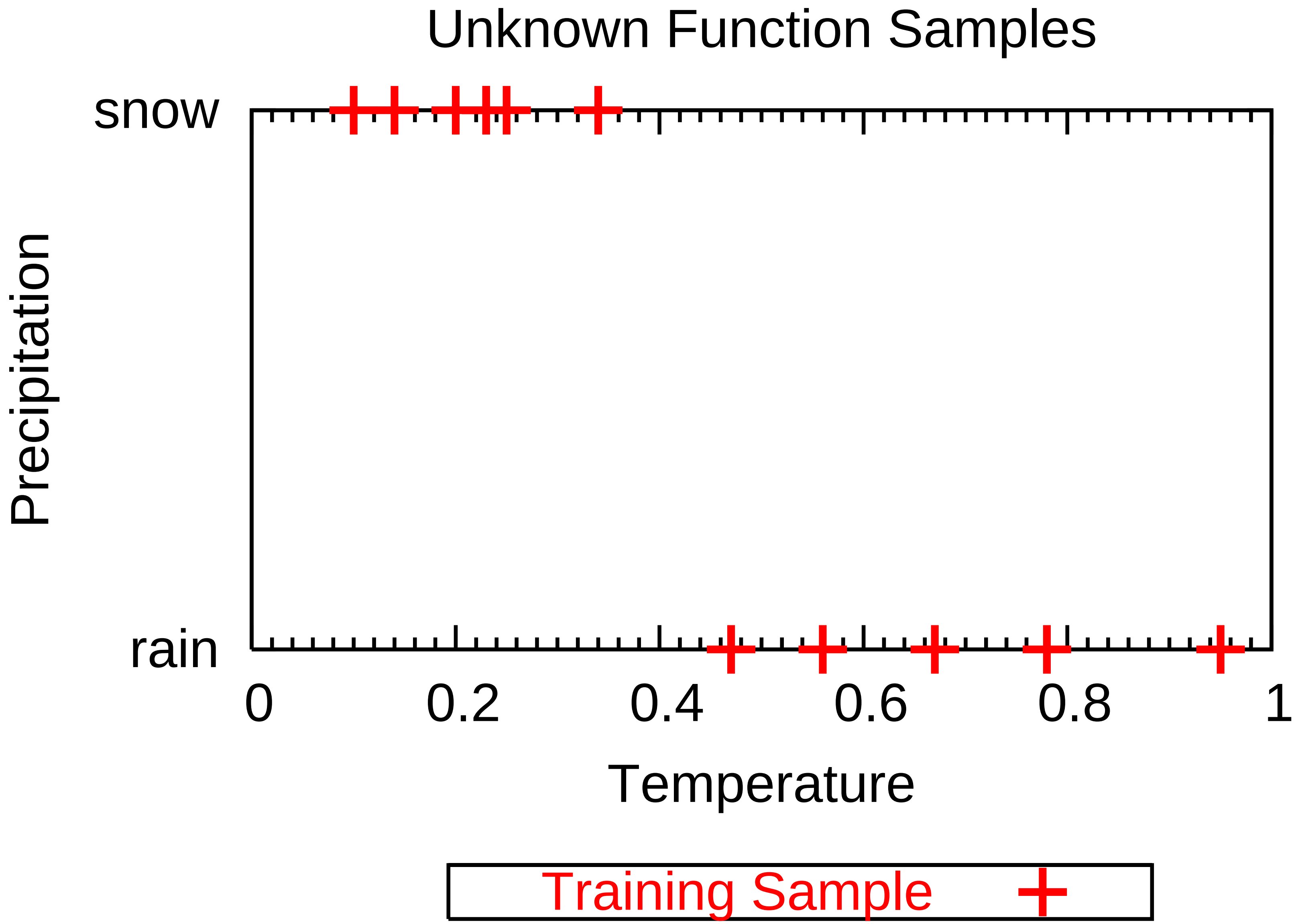
Length

The Machine Learning Problem:

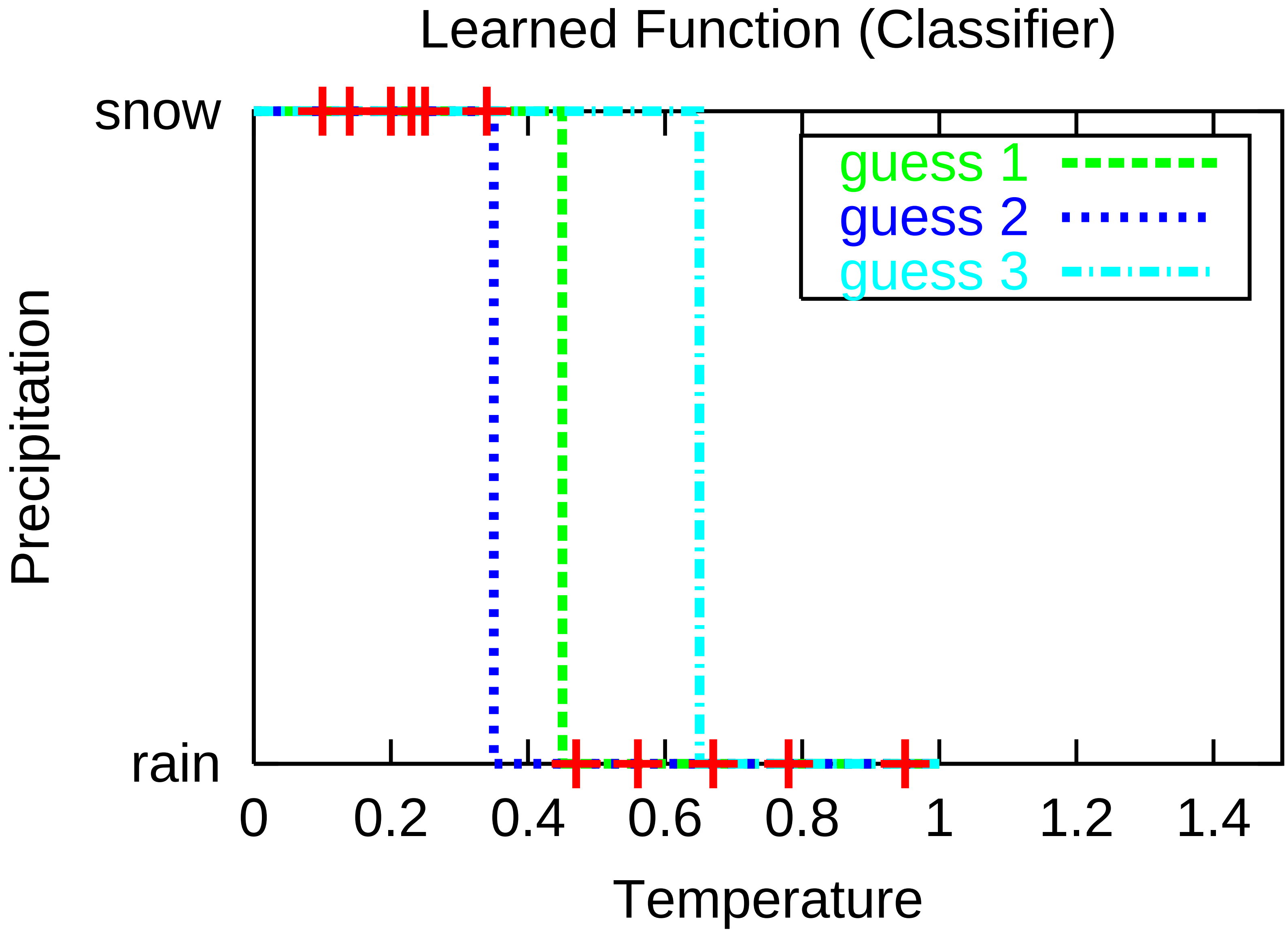
1. Get (input, class label) pairs from unknown function

2. Guess the unknown function

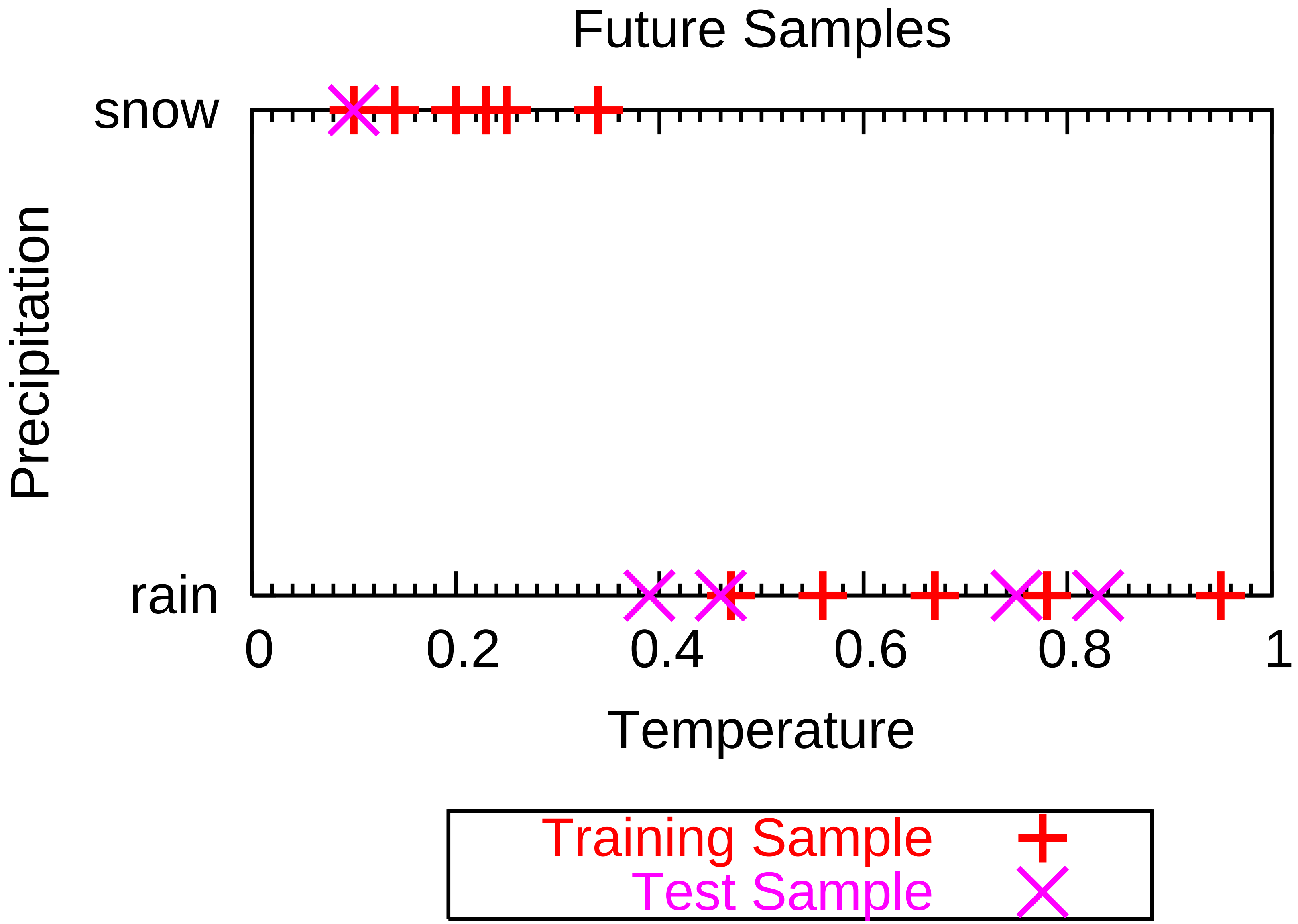
3. ... to minimize errors on future inputs



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

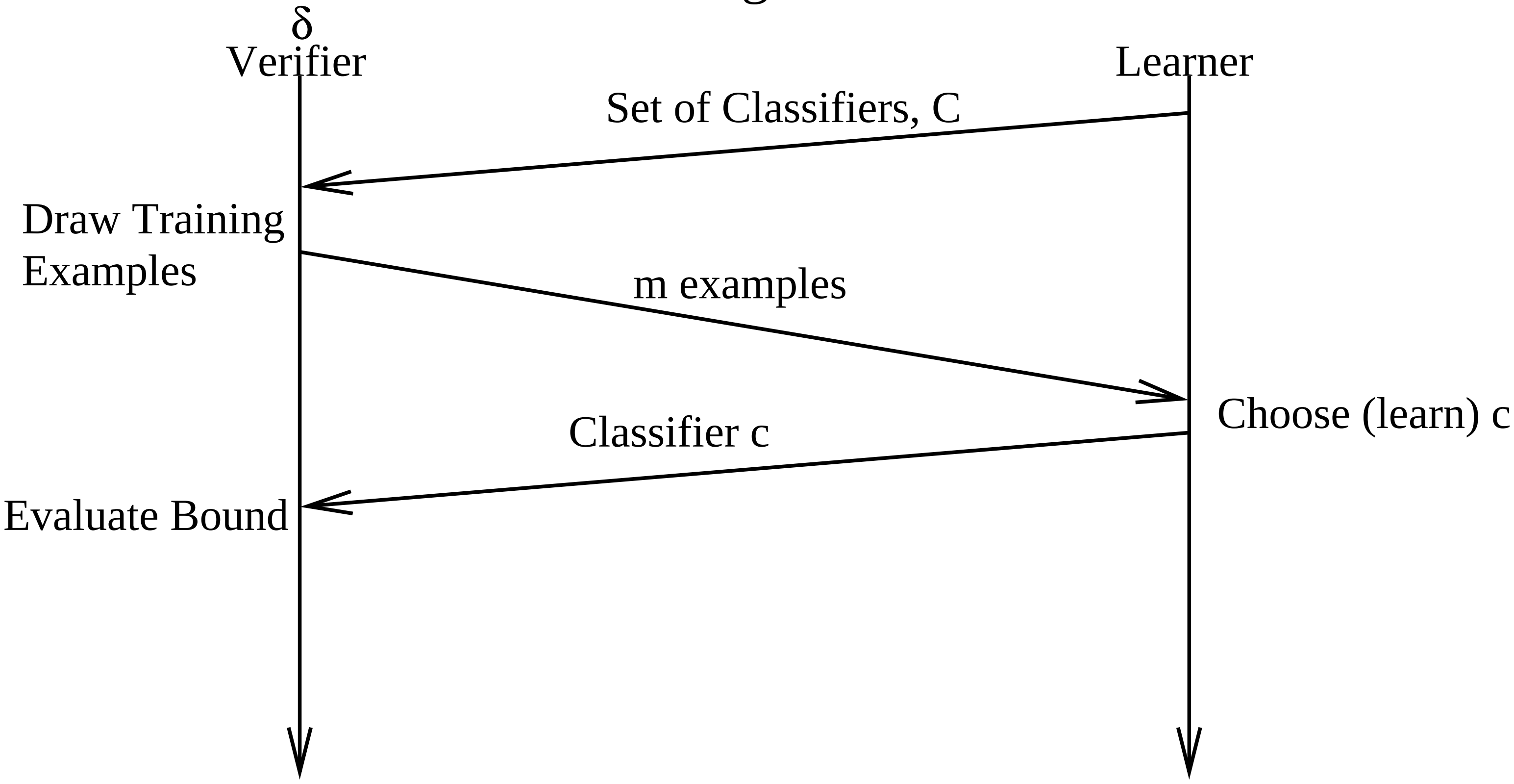
- We can't expect any prediction ability, in general.

- We can expect it, if examples come independently, sometimes.

A "Bound" is a high probability confidence interval on the future error rate of a classifier.

A "Bound" provides a yardstick for evaluation and can guide learning algorithm design

Training Set Bound



Outline

1. Old bounds tight $\Leftrightarrow$ few bits are required to send class labels in particular language

2. PAC-MDL bound: For all class label communication languages, few bits required to send class labels $\Rightarrow$ tight bound

3. PAC-MDL bound (approximately) implies older bounds.

Review: Definitions

X= input space

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

$c : X \to Y$ = classifier

$D(x, y)$ = distribution over x and y

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

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

$\hat{e}_S(c) \equiv \Pr_{x,y\sim U(S)}(c(x) \neq y)$ = "train error"

Review: VC bound

$VC(C)$ = the maximal number of points which can be classified in all ways by different classifiers in C.

Theorem: (VC bound) For all classifier sets, C, with probability 1-δ over training sets, for all classifiers $c \in C$:

$$\text{train error} = 0 \;\; \Rightarrow \;\; \text{true error} \leq 2\frac{VC(C)\left(1 + \ln \frac{2m}{VC(C)}\right) + \ln \frac{4}{\delta}}{m}$$

m = number of training examples

δ = probability of bound failure

Review: VC bound

$VC(C)$ = the maximal number of points which can be classified in all ways by different classifiers in C.

Theorem: (VC bound) For all classifier sets, C, with probability 1-δ over training sets, for all classifiers $c \in C$:

$$\text{train error} = 0 \;\Rightarrow\; \text{true error} \leq 2\frac{VC(C)\left(1 + \ln\frac{2m}{VC(C)}\right) + \ln\frac{4}{\delta}}{m}$$

Where is the compression?

VC dimension $d \Rightarrow$ at most $\left(\frac{me}{d}\right)^d$ ways to label m examples.

Pick canonical order and use $d \log \frac{me}{d}$ bits to send the index.

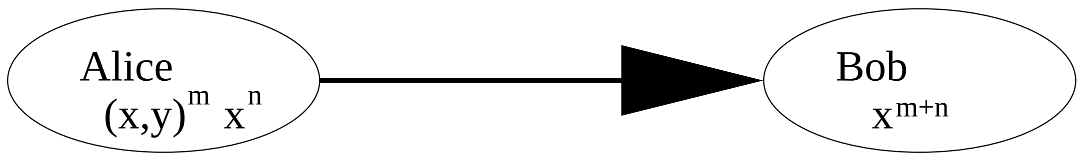
Outline

1. Old bounds tight $\Leftrightarrow$ few bits are required to send class labels in particular language

2. PAC-MDL bound: For all class label communication languages, few bits required to send class labels $\Rightarrow$ tight bound

3. PAC-MDL bound (approximately) implies older bounds.

The compression game



How many bits does Alice need to send the class labels to Bob?

- Bob knows the unlabeled data
- Bob doesn't know which examples are in training set or test set

Transduction definitions

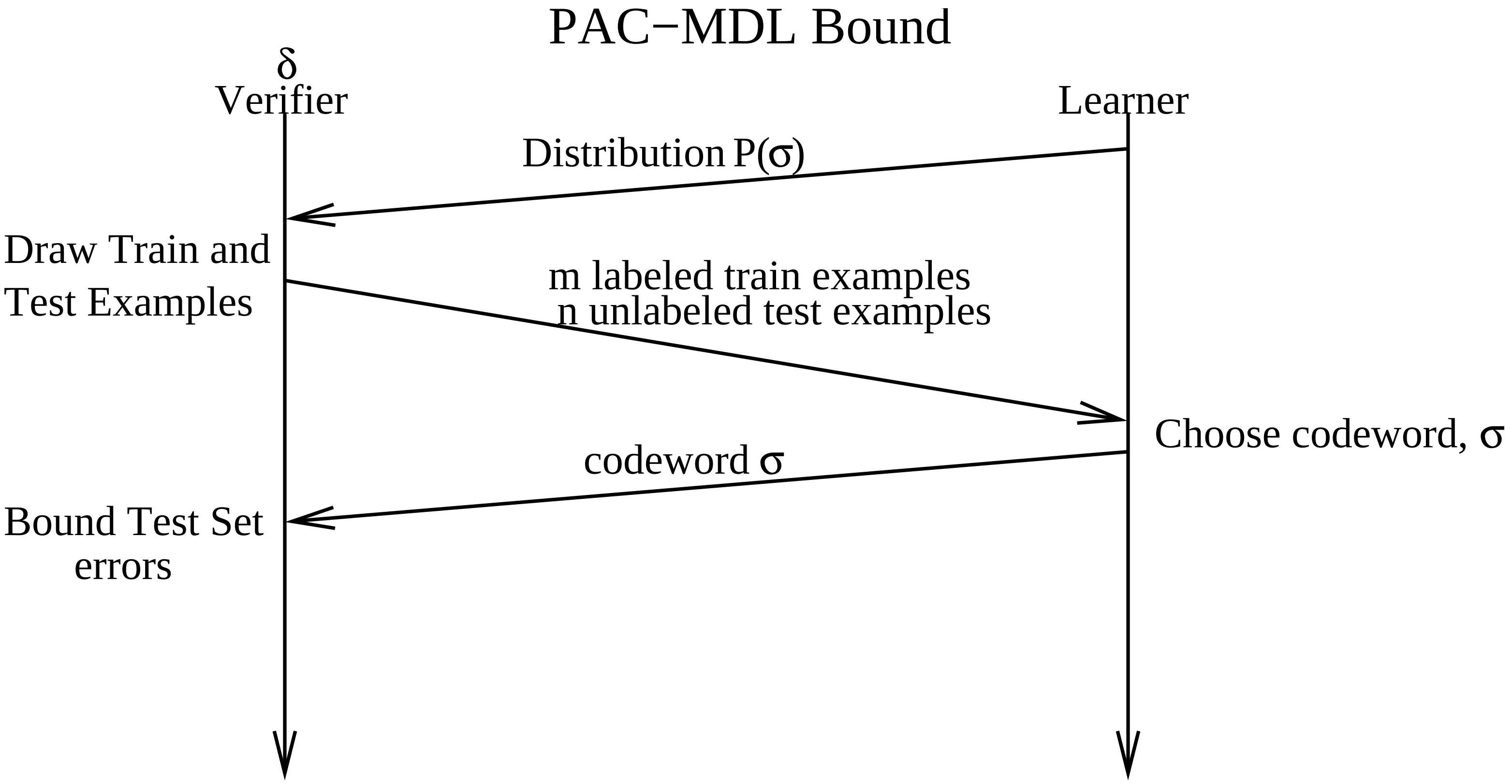
S = train set

S' = unlabeled test set

$m = |S|$

$n = |S'|$

$\sigma : X^{m+n} \to Y^{m+n}$ = transductive classifier


PAC−MDL Bound
δ
Verifier
Learner
Distribution P(σ)
Draw Train and
Test Examples
m labeled train examples
n unlabeled test examples
Choose codeword, σ
codeword σ
Bound Test Set
errors

PAC-MDL bound

Theorem: (PAC-MDL Bound)
For all distributions over transductive classifiers, $P(\sigma)$,
with probability $1-\delta$ over the train and test samples,
for all transductive classifiers σ:

$$\text{train error} = 0 \;\Rightarrow\; \text{test error rate} \leq \frac{\ln \frac{1}{P(\sigma)} + \ln \frac{1}{\delta}}{n \ln\left(1 + \frac{m}{n}\right)}$$

Here, $P(\sigma)$ is the (information theoretic) communication language. About $\log \frac{1}{P(\sigma)}$ bits are required to send σ.

Outline

1. Old bounds tight $\Leftrightarrow$ few bits are required to send class labels in particular language

2. PAC-MDL bound: For all class label communication languages, few bits required to send class labels $\Rightarrow$ tight bound

3. PAC-MDL bound (approximately) implies older bounds.

Older bounds for which class label compression $\Leftrightarrow$ bound is tight

1. Occam's Razor Bound (Blumer, Ehrenfucht, Haussler, Warmuth 87)

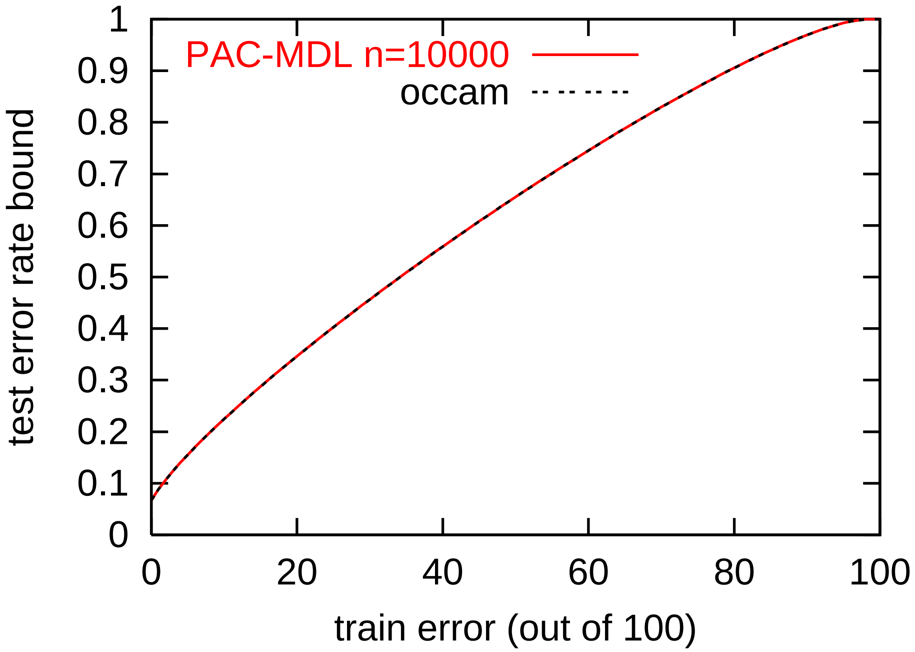
2. PAC-Bayes Bound (McAllester 99)

3. Compression Bound (Littlestone and Warmuth 89)

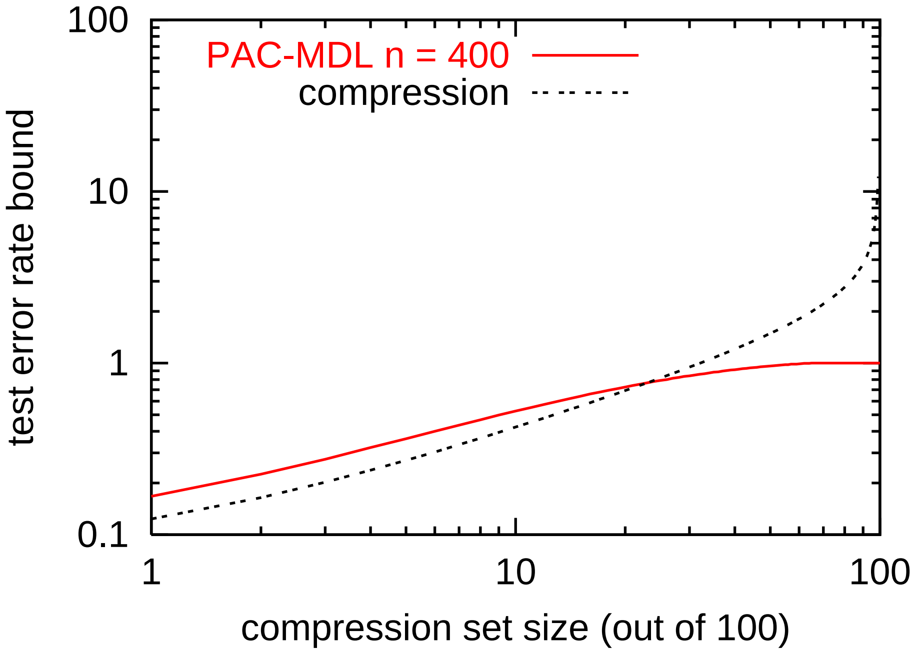
4. VC bound (Vapnik and Chervonenkis 74)

Comparison: Occam's Razor Bound

Pick $m = 100$, $n = \infty$, $\delta P(\sigma) = 0.001$ and get:

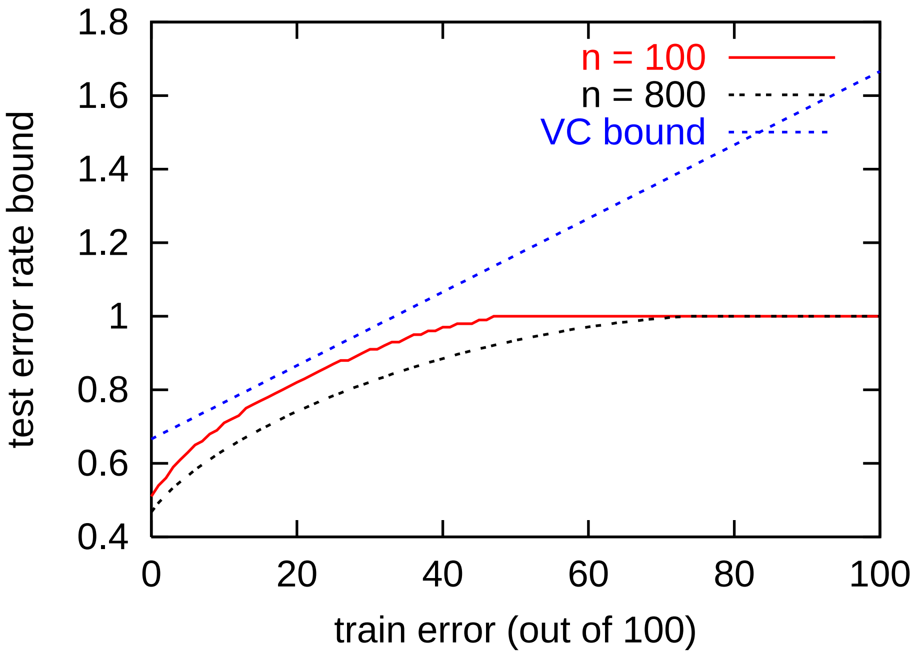


Comparison: Compression Bound



Comparison: VC bound

$P(y^{m+n}|x^{m+n}) =$uniform over all allowed possibilities $\geq \left(\frac{d}{em}\right)^d$



One bound to rule them all, one bound to find them,

One bound to bring them all, and in the darkness bind them.
-J.R.R. Tolkein (roughly)

1. Old bounds tight $\Leftrightarrow$ few bits are required to send class labels in particular language

2. PAC-MDL bound: For all class label communication languages, few bits required to send class labels $\Rightarrow$ tight bound

3. PAC-MDL bound (approximately) implies older bounds.

TODO

1. Can we find an argument which matches exactly the Compression bound communication complexity?

2. What algorithms does this motivate?

3. Are any other obvious choices missing from the set of languages?