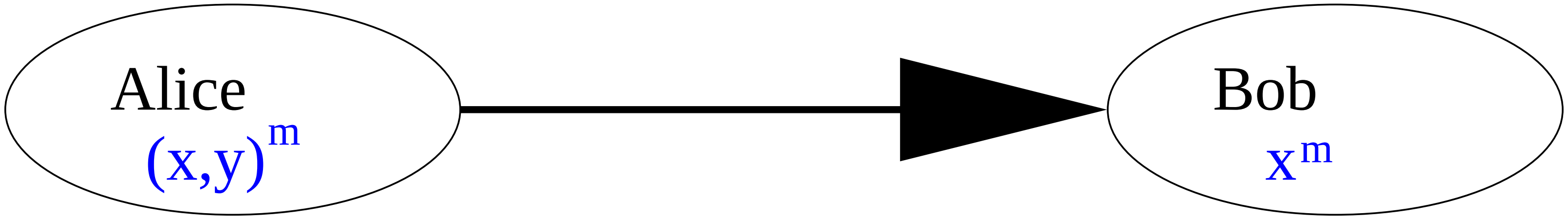
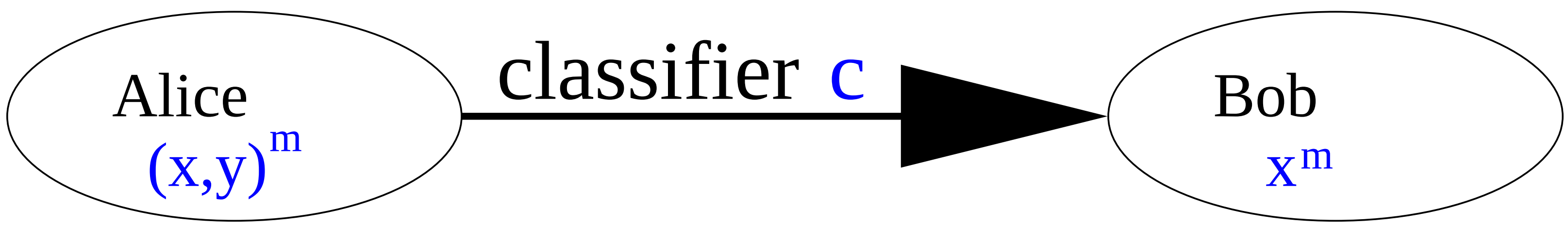
Label Compression implies Verifiable Learning

Avrim Blum and John Langford

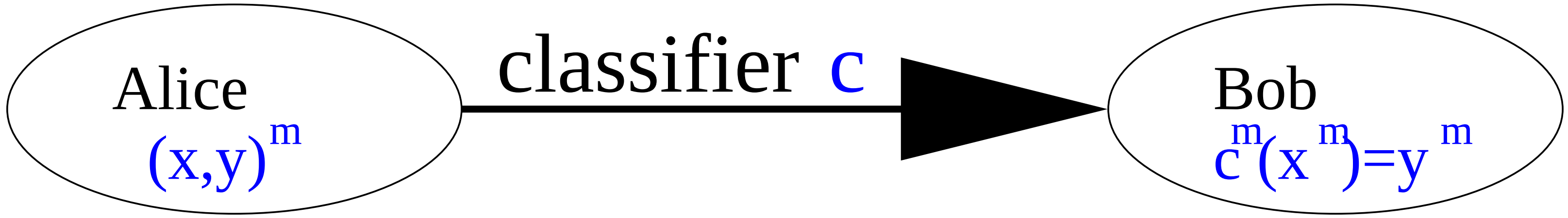
Theorem: (Occam's Razor Bound) Small description length for classifier $\Rightarrow$ verifiable learning



Theorem: (Occam's Razor Bound) Small description length for classifier ⇒ verifiable learning

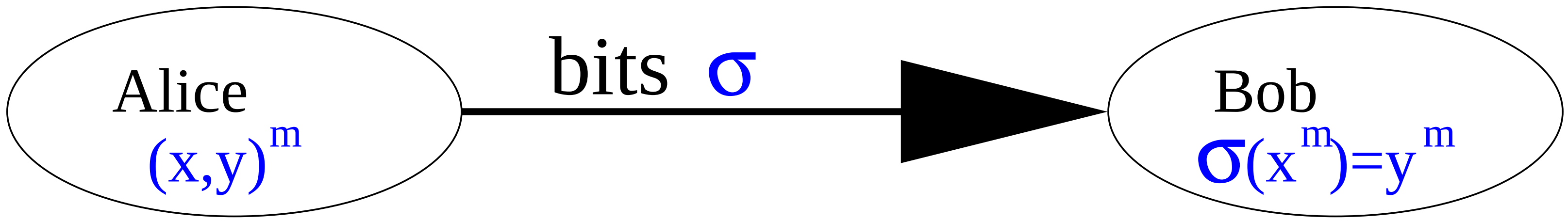


Theorem: (Occam's Razor Bound) Small description length for classifier $\Rightarrow$ verifiable learning



Restated: Small number of bits transmitted $\Rightarrow$verifiable learning

Suppose we transmitted the labels using some other language. Does this still hold?



Alice and Bob sort the unlabeled data

Alice sends "linear seperator number 2047"

Bob labels his data according to #2047

Does this method overfit?

Note: σ is NOT a classifier (what does "overfit" mean?)

Outline

1. Main Result: Label compression $\Rightarrow$ Verifiable learning

2. Secondary note: For "all" bounds, strong bound $\Leftrightarrow$ Label compression

Basic Definitions

X = input space

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

$c : X \to Y$ = classifier

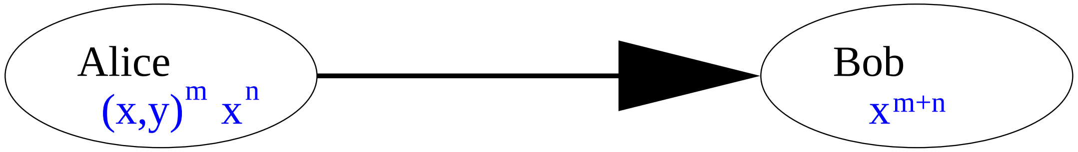
$D(x, y)$ = distribution over X and Y

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

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

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

The compression game



How many bits are used to send the labels to Bob?

New definitions

Learning algorithm is transductive:

S = train set

S' = unlabeled test set

$m = |S|$

$n = |S'|$

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

PAC-MDL Bound

Theorem: (PAC-MDL Bound) For all description languages for σ, with probability $1-\delta$ over draws of $S, S' \sim D^{m+n}$:

$$\forall \sigma : \ \hat{e}_S(\sigma) > 0 \text{ or } \hat{e}_{S'}(\sigma) \leq \frac{|\sigma| + \log_2 \frac{1}{\delta}}{n \log_2 \left(1 + \frac{m}{n}\right)}$$

PAC-MDL Bound: proof intuition

Mix one part VC bound and one part Occam's Razor Bound. Stir.

1. Bob is blind so he cannot distinguish train set from test set.

2. Fix X^{m+n} and randomize over which elements are in train and test set.

3. It's unlikey that we have 0 train errors and t test errors $\Rightarrow$ bound.

4. Bound holds for every X^{m+n} so it holds in expectation over X^{m+n}

Tricky detail: the fundamental distribution is Hypergeometric, not Binomial.

Transduction $\rightarrow$ Induction

Several methods

1. Take limit as $n \rightarrow \infty$

2. Use VC approach for finite n

   (a) Suppose small train error and large test error likely for a classifier.

   (b) Note that this contradicts transductive bound.

Outline

1. Main Result: Label compression $\Rightarrow$ Verifiable learning

2. Secondary note: For "all" bounds, strong bound $\Leftrightarrow$ Label compression

    (a) Occam's Razor Bound

    (b) PAC-Bayes Bound

    (c) VC dimension Bound

    (d) Sample Compression Bound

Occam's Razor Bound

Theorem: (Occam's Razor Bound BEHW87) $\forall P(c)$, for all but δ $D^m(S)$:

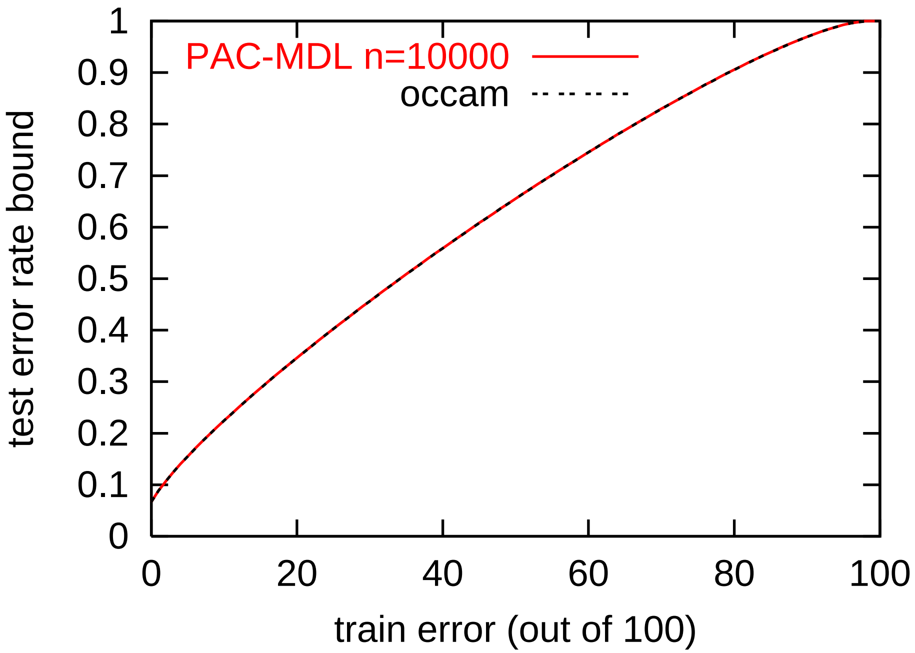
$$\forall c: \; \hat{e}_S(c) > 0 \text{ or } e_D(c) \leq \frac{\ln \frac{1}{P(c)} + \ln \frac{1}{\delta}}{m}$$

Communication language = classifier description.

Comparison: Occam's Razor Bound

$P(\sigma) = P(c)$

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

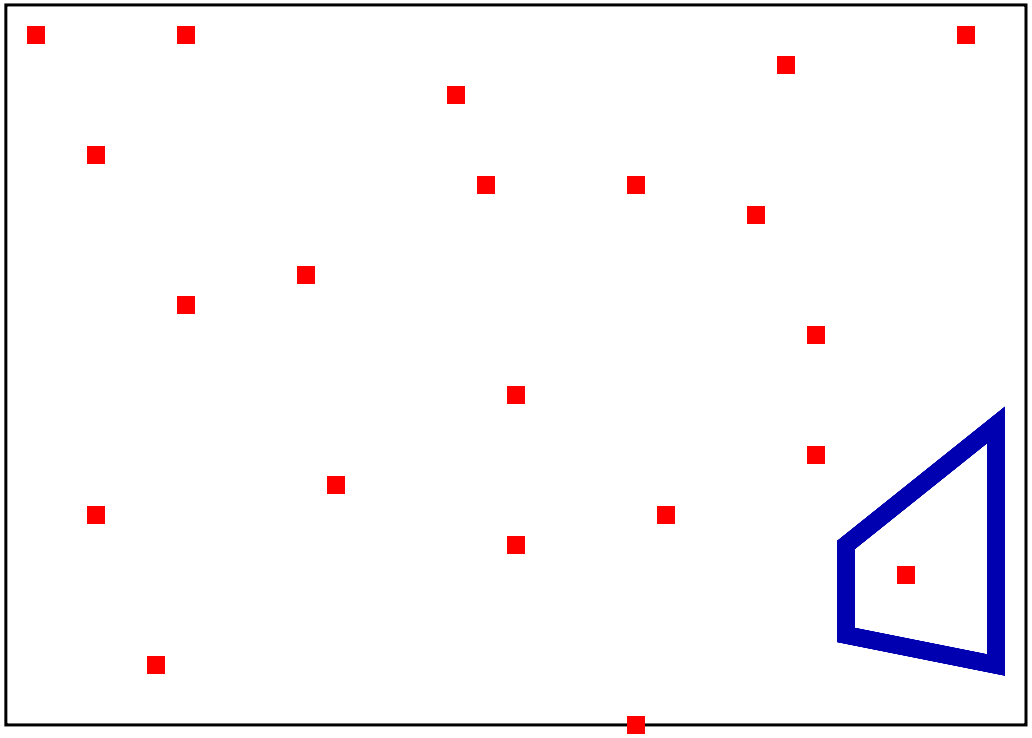


PAC-Bayes Bound

Theorem: (PAC-Bayes Bound M99) $\forall P(c)$, for all but δ $D^m(S)$:

$$\forall Q(c): \;\; E_{Q(c)}\hat{e}_S(c) > 0 \text{ or } E_{Q(c)}e_D(c) \leq \frac{\mathrm{KL}(Q||P) + \ln\frac{1}{\delta}}{m}$$

A Posterior



Communication language = random classifier language drawn from P

Sample Compression Bound

$S' \subseteq S$ = "compression set" = subset of examples that classifier depends upon.

Theorem: (Compression Bound LW89) For all learning algorithms A, For all but δ $D^m(S)$ let $c = A(C(S))$

$$\hat{e}_{S-C(S)}(c) > 0 \text{ or } e_D(c) \leq \frac{\ln \binom{m}{|C(S)|} + \ln \frac{m+1}{\delta}}{m - |C(S)|}$$

Sample Compression Bound

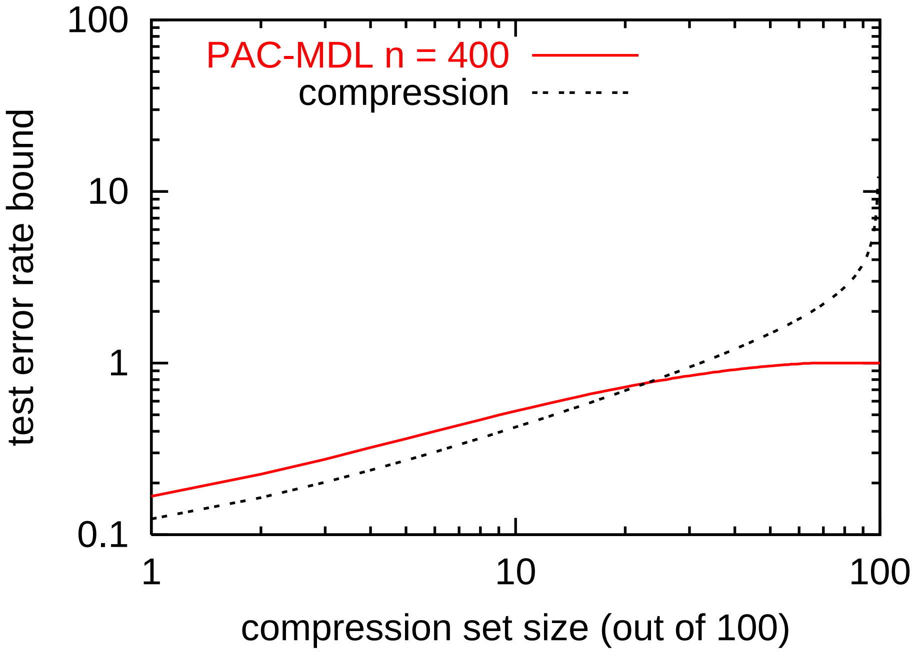
Ideally: $P(y^m|x^m) =$ communication complexity of compression set.

$$\ln P(y^m|x^m) = \ln \binom{m}{|C(S)|} + \ln(m+1) + |C(S)| \ln 2$$

Reality: $P(y^m|x^m) =$ communication complexity in train+test set.

$$\ln P(y^{m+n}|x^{m+n}) = \ln \binom{m+n}{|C(S)|} + \ln(m+n+1) + |C(S)| \ln 2$$

Sample Compression Bound II



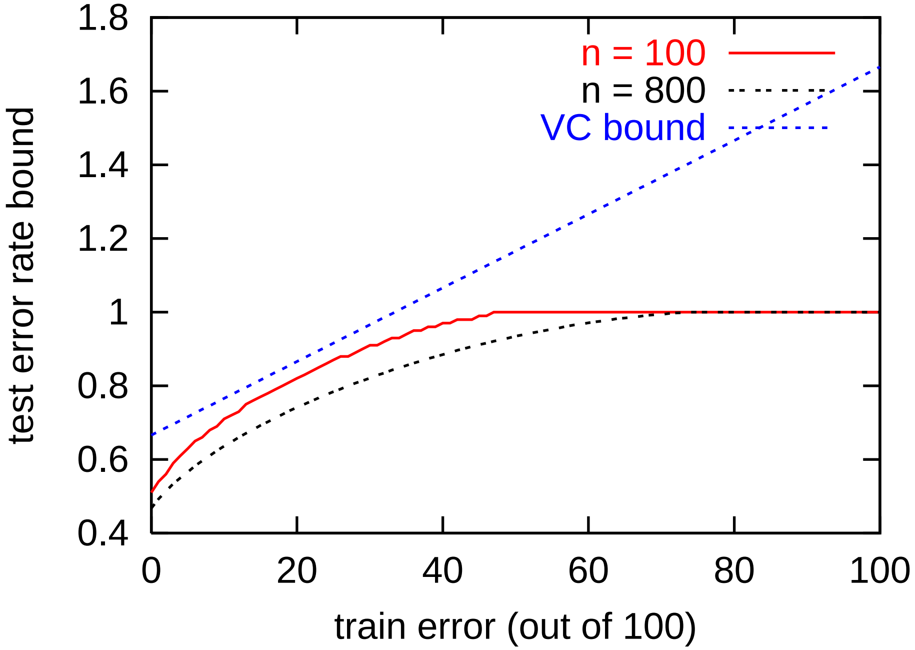
Review: VC bound

Theorem: (VC bound VC71) For all sets of classifiers, H, for all but δ $D^m(S)$, let $d = VC(H)$:

$$\forall c \in H: \quad \hat{e}_S(c) > 0 \text{ or } e_D(c) \leq 2\frac{d\left(1 + \ln \frac{2m}{d}\right) + \ln \frac{4}{\delta}}{m}$$

Connection: VC bound

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



Conclusion

1. Applies to transduction and induction

2. (partial) Bound unification: Learning is fundamentally related to compression.

3. Simple proof

4. Holds for all distributions, but the "worst case" is conditioned on the unlabeled data ($\Rightarrow$ remove bound slack)