A PAC-MDL Bound

John Langford (joint work with Avrim Blum)

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)

The skinny

"all" bounds $\Rightarrow$ Label compression

Label compression $\Rightarrow$ Verifiable learning

Outline

1. Review: Label Compression in Learning theory.
2. New Stuff: Label Compression implies Learning
3. Connection:PAC-MDL bound vs older bounds

Review: Definitions

X= input space

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

$c : X \rightarrow Y$ = 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"

Review: Occam's Razor

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

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

Review: Occam's Razor

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

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

Label compression: Use $\log \frac{1}{P(c)}$ bits to transmit choice of classifier. Small number of bits $\Rightarrow$ tight bound.

Review: PAC-Bayes

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

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

Review: PAC-Bayes

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

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

Label Compression: The bits-back argument HvC93

Bits-Back

1. Alice picks $c \sim Q$ using *message* M where $|M| \simeq H(Q)$

Bits-Back

1. Alice picks $c \sim Q$ using *message* M where $|M| \simeq H(Q)$

2. Alice transmits c with $\mathsf{H}(Q) + \mathsf{KL}(Q||P)$ bits

Bits-Back

1. Alice picks $c \sim Q$ using *message* M where $|M| \simeq H(Q)$

2. Alice transmits c with $\mathsf{H}(Q) + \mathsf{KL}(Q||P)$ bits

3. Bob decodes labeled examples

Bits-Back

1. Alice picks $c \sim Q$ using *message* M where $|M| \simeq H(Q)$

2. Alice transmits c with $\mathsf{H}(Q) + \mathsf{KL}(Q||P)$ bits

3. Bob decodes labeled examples

4. Bob runs learning algorithm to decode Q

Bits-Back

1. Alice picks $c \sim Q$ using *message* M where $|M| \simeq H(Q)$
2. Alice transmits c with $\mathsf{H}(Q) + \mathsf{KL}(Q||P)$ bits
3. Bob decodes labeled examples
4. Bob runs learning algorithm to decode Q
5. Bob decodes M

Bits-back Summary

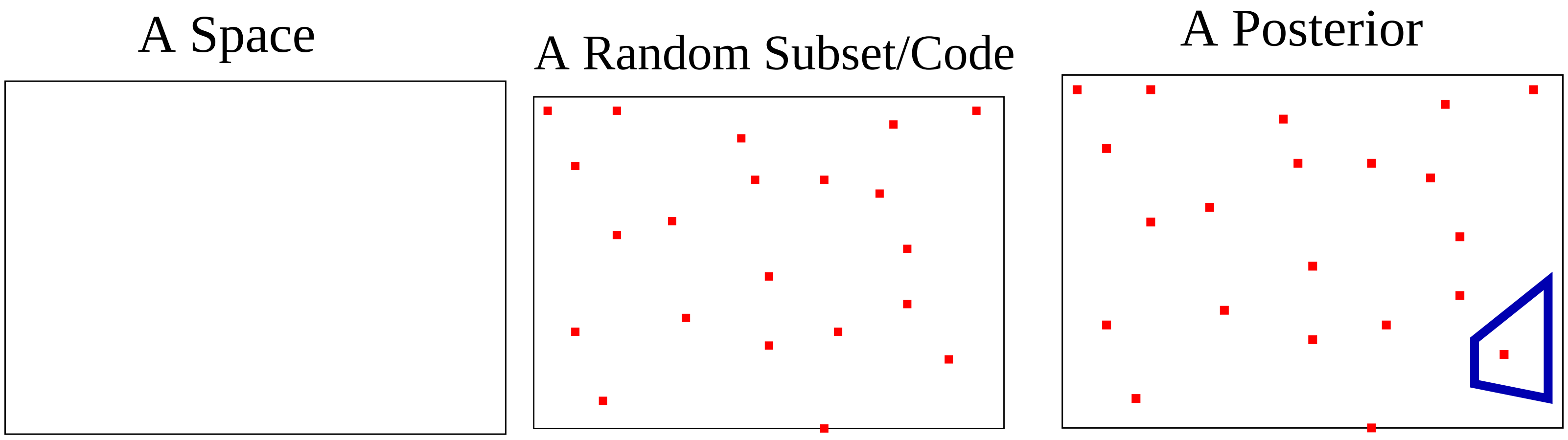
total bits = $\mathsf{H}(Q) + \mathsf{KL}(Q||P)$

total message = M, Q, labels

But this implies a code which uses just $\mathsf{KL}(Q||P)$ bits to send Q and the labels.

Just pick M in advance.

Bits-back, a different viewpoint



Review: Compression

$C(S)$ = "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 $c = A(C(S))$

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

Review: Compression

$C(S)$ = "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 $c = A(C(S))$

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

Label compression: Specify $C(S)$ and labels of $C(S)$. Run learning algorithm to acquire all labels.

Review: VC bound

$VC(H)$ = the maximal number of points which can be labeled in any way by different classifiers in H.

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

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

Review: VC bound

$VC(H)$ = the maximal number of points which can be labeled in any way by different classifiers in H.

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

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

Label compression: Embedded in the argument.

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.

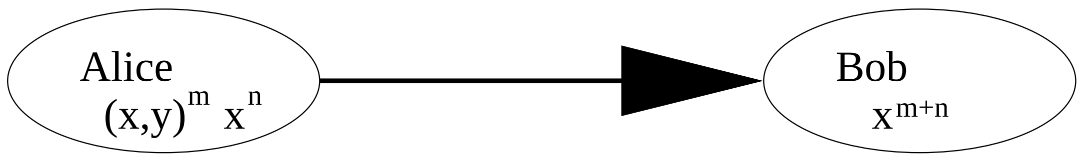
Outline

1. Review: Label Compression in Learning theory

2. New Stuff: Label Compression implies Learning

3. Connection: Labeling bound vs older bounds

New Stuff: The compression game



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

New Stuff: 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

New Stuff: The One Bound

Theorem: (PAC-MDL Bound) $\forall P(\sigma)$, for all but δ $D^{m+n}(S, S')$:

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

New Stuff: The proof

Fix X^{m+n}. Let $\pi(m,n)$ = uniform distributions over partitions

$$\forall x^{m+n}, y^m \forall \sigma \quad \Pr_{S,S' \sim \pi(m,n)} (\text{0 train and } t \text{ test errors}) = \frac{\binom{n}{t}}{\binom{m+n}{t}}$$

$$\Rightarrow \forall x^{m+n}, y^m \forall \sigma \quad \Pr_{S,S' \sim \pi(m,n)} \left(\text{0 train and } \frac{\binom{n}{t}}{\binom{m+n}{t}} \leq \delta \right) \leq \delta$$

$$\Rightarrow \forall x^{m+n}, y^{m} \quad \Pr_{S,S'\sim\pi(m,n)}\left(\exists\sigma \;\; \text{0 train and } \frac{\binom{n}{t}}{\binom{m+n}{t}} \leq \delta P(\sigma)\right) \leq \delta$$

$$\Rightarrow \Pr_{S,S'\sim D^{m+n}}\left(\exists\sigma \;\; \text{0 train and } \frac{\binom{n}{t}}{\binom{m+n}{t}} \leq \delta P(\sigma)\right) \leq \delta$$

$$\Rightarrow \Pr_{S,S'\sim D^{m+n}}\left(\forall\sigma \;\; \text{1 or more train or } \frac{\binom{n}{t}}{\binom{m+n}{t}} \geq \delta P(\sigma)\right) \geq 1-\delta$$

New Stuff: Proof End

Appeal to algebra and approximation.

Outline

1. Review: Label Compression in Learning theory.

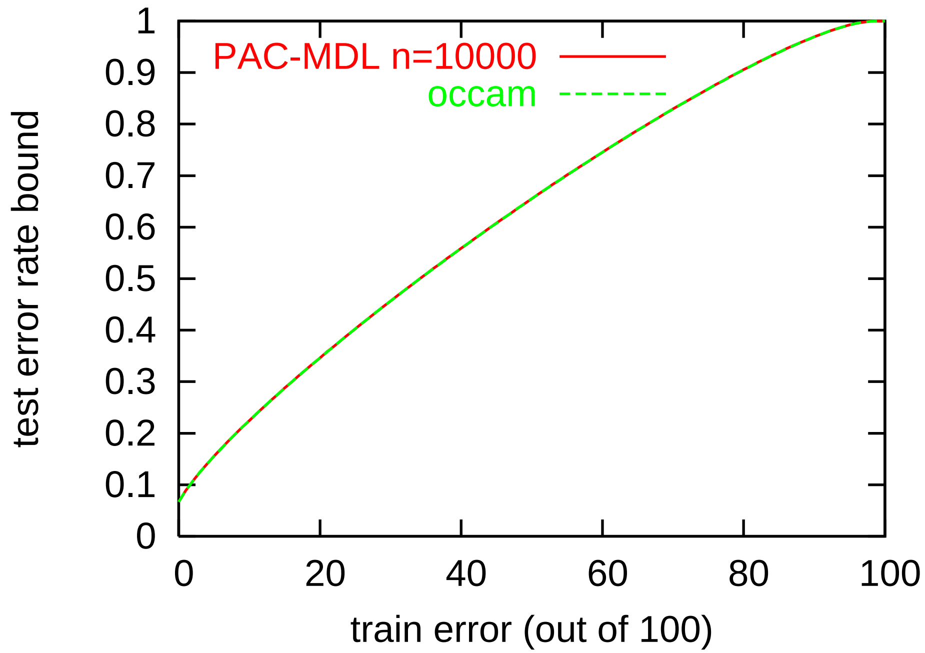
2. New Stuff: Label Compression implies Learning

3. Connection: Labeling bound vs older bounds

Connection: Occam's Razor Bound

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

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



Connection: PAC-Bayes Bound

Pick a random subset of the classifiers. Send the classifier in the random subset.

Bits-back $\Rightarrow$ communication complexity of about $\mathsf{KL}(Q||P)$

Connection: Compression Bound

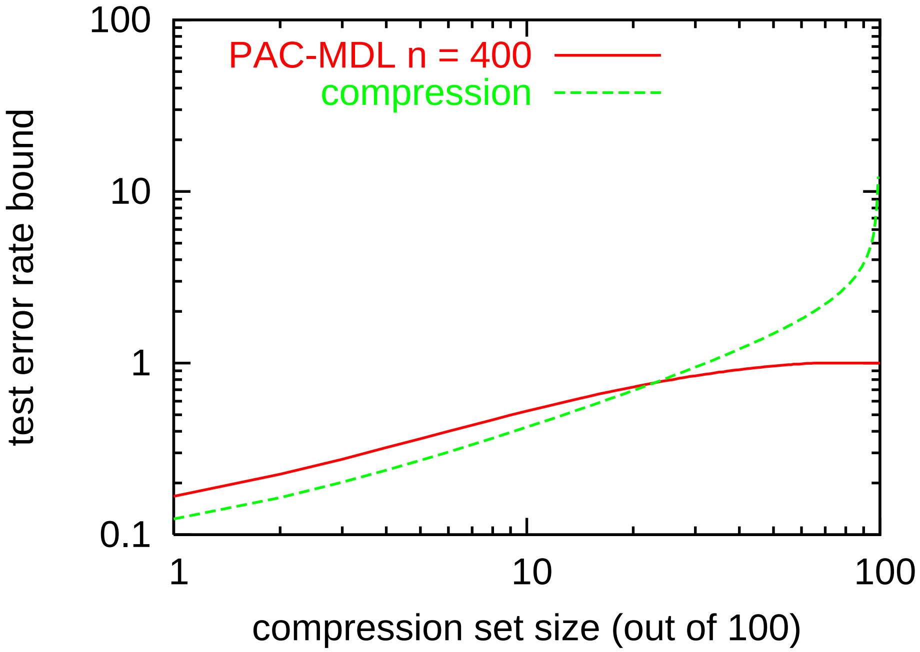
Ideally: $P(y^m|x^m) =$ communication complexity of compression set.

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

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

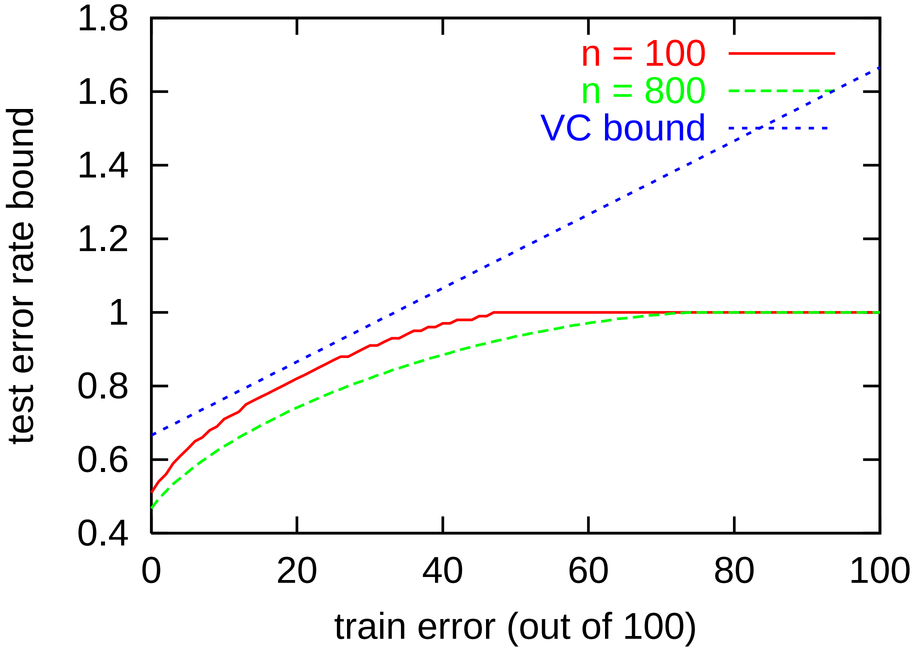
$$\ln P(y^{m+n}|x^{m+n}) = \ln \binom{m+n}{|C(S)|} + \ln(m+n+1) + |C(S)| \ln 2$$

Connection: Compression Bound II



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. (approximate-to-perfect) Unification of most standard bounds

3. Proof complexity: Occam's Razor $<$ Compression $<$ PAC-MDL $<$ PAC-Bayes $<$ VC

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

TODO

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

2. What algorithms does this motivate?

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