Label Compression implies Verifiable Learning

John Langford

(joint work with Avrim Blum)

The skinny

Label compression $\Rightarrow$ Verifiable learning

"all" bounds $\Rightarrow$ Label compression

Outline

1. Review: What are typical learning theory bounds?

2. New Stuff: What is the one bound?

3. Connection: Compare the one bound to older work?

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

$e_D(c) \equiv \mathrm{Pr}_{x,y\sim D}(c(x) \neq y)$ = "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 $D^m(S)$:

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

Review: PAC-Bayes

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

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

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: VC bound

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

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

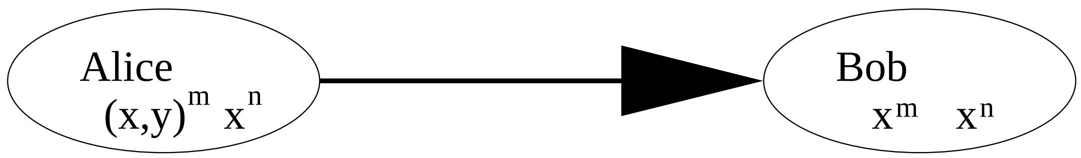
Outline

1. Review: What are typical learning theory bounds?

2. New Stuff: What is the one bound?

3. Connection: Compare the one bound to older work?

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: (The One 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

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

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

New Stuff: Proof II

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

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

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

New Stuff: Proof End

Appeal to algebra and approximation.

Outline

1. Review: What are typical learning theory bounds?

2. New Stuff: What is the one bound?

3. Connection: Compare the one bound to older work?

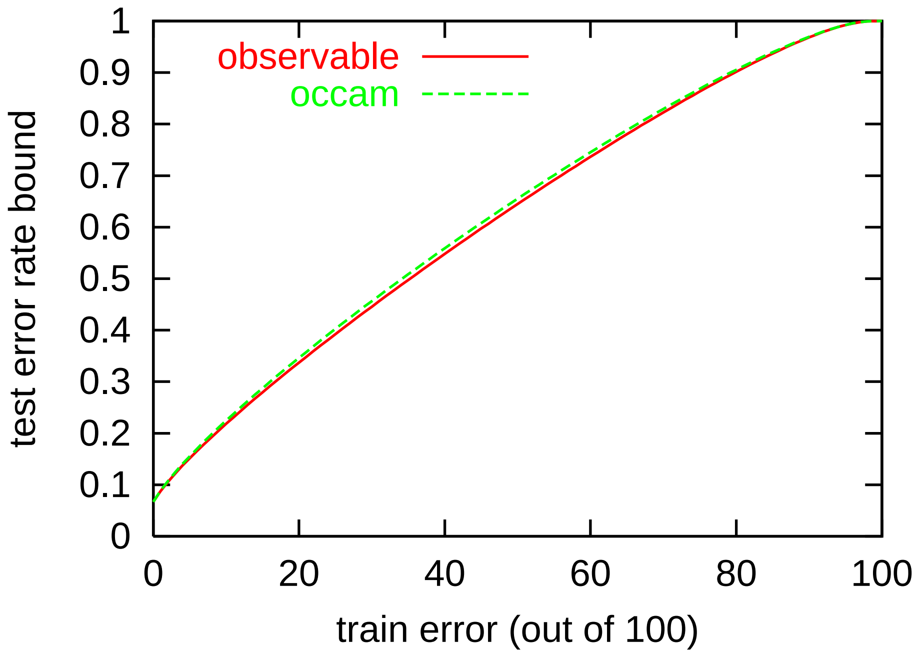
Connection: Occam's Razor Bound

What is the communication language?

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

What is the communication language?

Connection: PAC-Bayes Bound

Ideally:

$$P(y^m|x^m) = \sum_{c:c^m(x^m)=y^m} P(c)$$

$$\mathsf{KL}(Q||P) = E_{c\sim Q} \ln \frac{Q(c)}{P(c)} = \ln \frac{1}{P(y^m|x^m)}$$

Reality:

$$P(y^{m+n}|x^{m+n}) = \sum_{c:c^{m+n}(x^{m+n})=y^{m+n}} P(c)$$

Connection: Compression Bound

What is the communication language?

Connection: Compression Bound

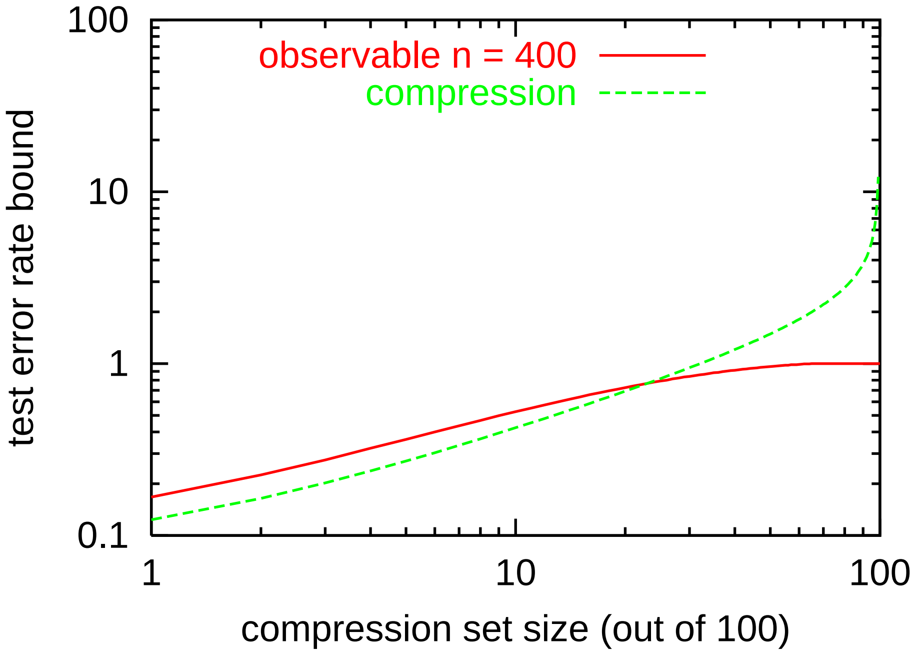
Ideally: $P(y^m|x^m) =$ communication complexity of compression set.

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

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

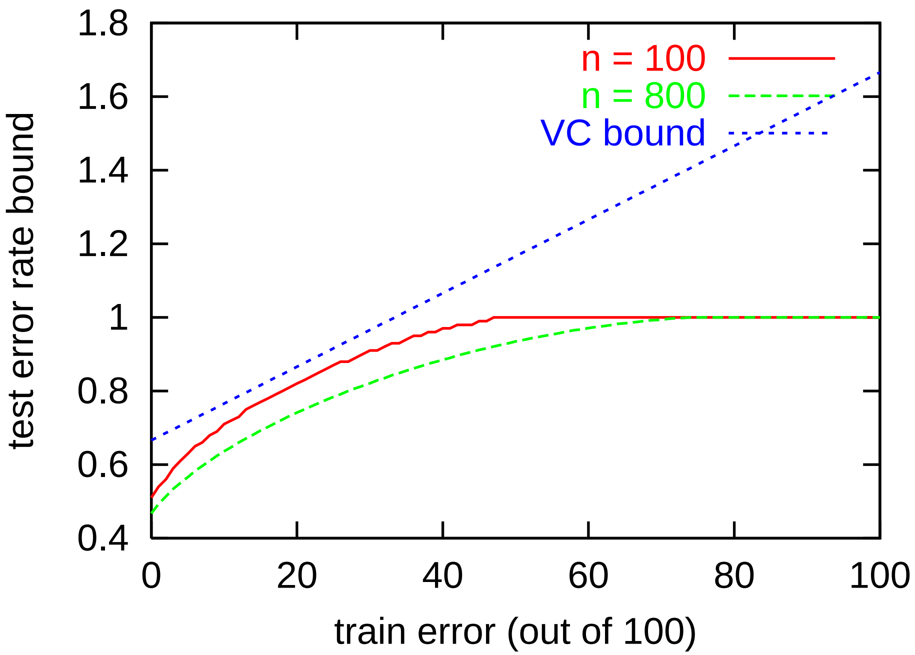


Connection: VC bound

What is the communication language?

Connection: VC bound

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



Conclusion

1. Applies to transduction and induction

2. (partial) Bound unification

3. Simple proof

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 PAC-Bayes and Compression bound communication complexity?

2. What algorithms does this motivate?

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