General Generalization Bounds, Part II

John Langford

Basic Definitions: Quick Review

X = input space

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

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

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

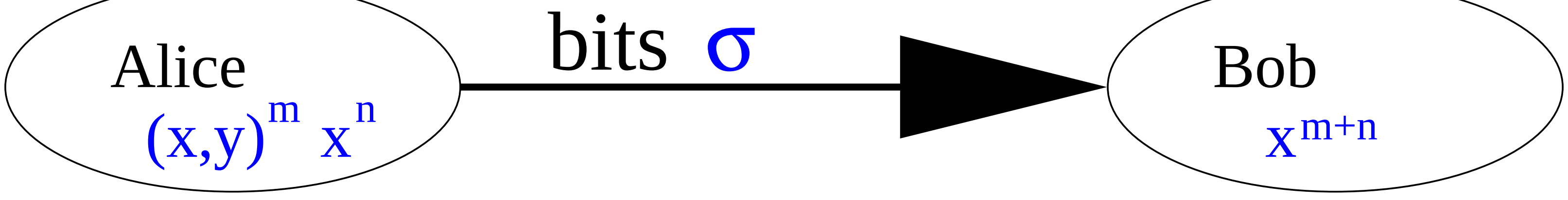
$S' = (x, y)^n \sim D^n$ = test set

$\sigma : X^{|S|+|S'|} \rightarrow Y^{|S|+|S'|}$ = transductive classifier

$\hat{\sigma}_S \equiv \sum_{x,y \in S} I(\sigma(S, S')(x) \neq y)$ = train error

$\hat{\sigma}_{S'} \equiv \sum_{x,y \in S'} I(\sigma(S, S')(x) \neq y)$ = test error

The compression game



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

PAC-MDL Bound: Simplified

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

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

Outline

1. Relationship between Transduction and Induction

   (a) infinite sample limit

   (b) finite sample case

2. For “all” bounds, strong bound $\Leftrightarrow$ Label compression

Induction setting

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

Distribution of error = Binomial

$$\begin{aligned} \mathrm{Bin}\,(m,k,c_D) &= \mathrm{Pr}_{S\sim D^m}\left(\hat{c}_S \leq \frac{k}{m}\middle|\, c_D\right) \\ &= \sum_{i=0}^{k} \binom{m}{i} c_D^i (1-c_D)^{m-i} \end{aligned}$$

Worst consistent true error rate = upper bound.

$$\overline{\mathrm{Bin}}\,(m,k,\delta) = \max\left\{p : \mathrm{Bin}\,(m,k,p) \geq \delta\right\}$$

Transduction $\rightarrow$ Induction: Infinite samples

$$\lim_{|S'| \to \infty} \frac{\widehat{\sigma}_{S'}}{|S'|} = c_D$$

Transduction $\rightarrow$ Induction: Finite Samples

Theorem: For all measures $P(\sigma)$, with probability $1 - \delta$ over draws of $S \sim D^m$:

$$\forall c \;\; c_D \leq \overline{\mathrm{Bin}}\left(m, \frac{\max_{\sigma \text{ consistent with } c} |\sigma| + \log_2 \frac{1}{\delta}}{\log_2\left(1 + \frac{m}{n}\right)}, \frac{1}{2}\right)$$

Proof

(PAC-MDL bound)

$$\Pr_{S,S'\sim D^{m+n}}\left(\forall\sigma:\ \hat{\sigma}_S = 0 \Rightarrow \hat{\sigma}_{S'} \leq \frac{|\sigma| + \log_2\frac{1}{\delta}}{\log_2\left(1+\frac{m}{n}\right)}\right) \geq 1 - \frac{\delta}{2}$$

Let $b(\sigma) = \frac{|\sigma|+\log_2\frac{1}{\delta}}{\log_2(1+\frac{m}{n})}$

$$\Pr_{S,S'\sim D^{m+n}}(\forall\sigma:\ \hat{\sigma}_S = 0 \Rightarrow \hat{\sigma}_{S'} \leq b(\sigma)) \geq 1 - \frac{\delta}{2}$$

since $\Pr_{A,B}(\Phi(A,B)) > 1-\epsilon$

$$\Rightarrow \forall\alpha > 1:\ \Pr_A(\Pr_B(\Phi(A,B)) > 1-\alpha\epsilon) > 1 - \frac{1}{\alpha}$$

$$\Rightarrow \Pr_{S\sim D^m}\left(\Pr_{S'\sim D^n}(\forall\sigma:\ \hat{\sigma}_S = 0 \Rightarrow \hat{\sigma}_{S'} \leq b(\sigma)) > \frac{1}{2}\right) \geq 1 - \delta$$

(quantifier relaxation)

$$\Rightarrow \Pr_{S\sim D^m}\left(\forall\sigma:\ \Pr_{S'\sim D^n}(\hat{\sigma}_S = 0 \Rightarrow \hat{\sigma}_{S'} \le b(\sigma)) > \frac{1}{2}\right) \ge 1-\delta$$

(renaming)

$$\Rightarrow \Pr_{S\sim D^m}\left(\forall c: \forall\sigma \text{ consistent with } c\ \Pr_{S'\sim D^n}(\hat{\sigma}_S = 0 \Rightarrow \hat{\sigma}_{S'} \le b(\sigma)) > \frac{1}{2}\right) \ge 1-\delta$$

(Taking worst case)

$$\Rightarrow \Pr_{S\sim D^m}\left(\forall c:\ c_D \le \overline{\mathrm{Bin}}\left(|S|, \max_{\sigma \text{ consistent with } c} b(\sigma), \frac{1}{2}\right)\right) \ge 1-\delta$$

Outline

1. Relationship between Transduction and Induction

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

   (a) Occam's Razor Bound

   (b) PAC-Bayes Bound

   (c) Sample Compression Bound

   (d) VC dimension Bound

Occam's Razor Bound

Theorem:(Occam's Razor Bound BEHW87) $\forall$ description languages on c, with probability $1-\delta$ over $S \sim D^m$:

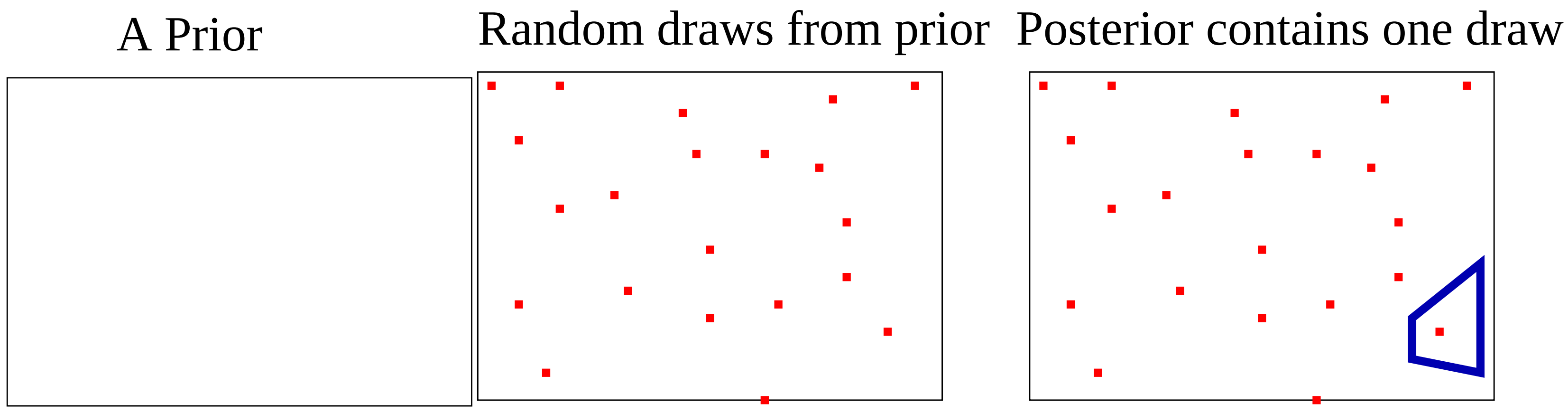
$$\forall c: \ \hat{c}_S = 0 \Rightarrow c_D \leq \frac{|c| + \ln \frac{1}{\delta}}{m}$$

Communication language = classifier description.

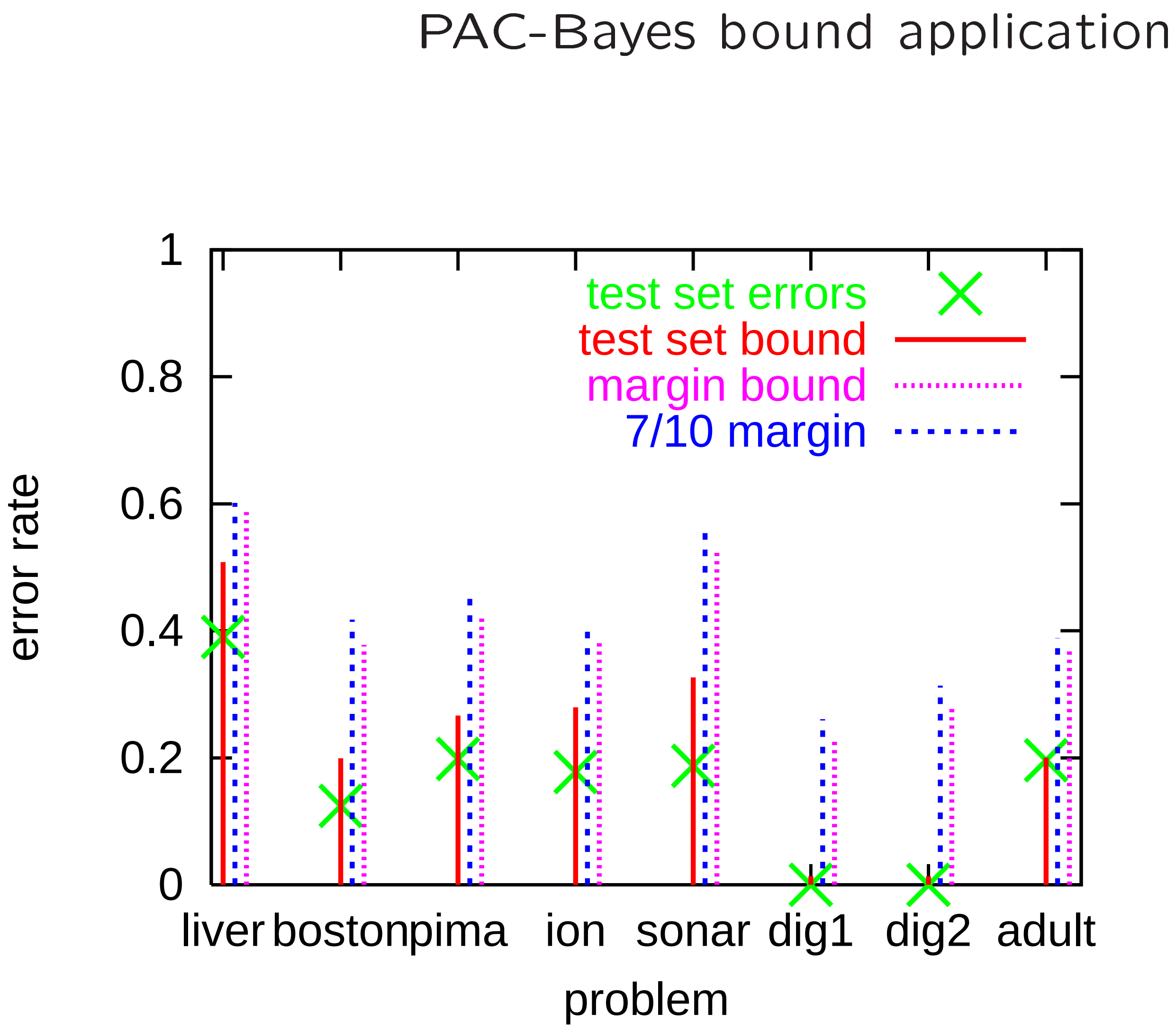
PAC-Bayes Bound

Theorem: (PAC-Bayes Bound M99) $\forall P(c)$, with probability $1-\delta$ over $S \sim D^m$:

$$\forall Q(c): \quad \hat{c}_S = 0 \Rightarrow E_{c\sim Q} c_D \leq \frac{\mathrm{KL}(Q||P) + \ln \frac{m+1}{\delta}}{m}$$



Communication language = random classifier language drawn from P



PAC-Bayes bound application
test set errors
test set bound
margin bound
7/10 margin
error rate
1
0.8
0.6
0.4
0.2
0
liver
boston
pima
ion
sonar
dig1
dig2
adult
problem

Sample Compression Bound

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

Theorem: (Compression Bound LW86) For all learning algorithms A, with probability $1-\delta$ over $S \sim D^m$ let $c = A(S')$

$$\hat{e}_{S-S'}(c) = 0 \Rightarrow e_D(c) \leq \frac{\ln \binom{m}{|S'|} + \ln \frac{m+1}{\delta}}{m - |S'|}$$

Sample Compression Bound

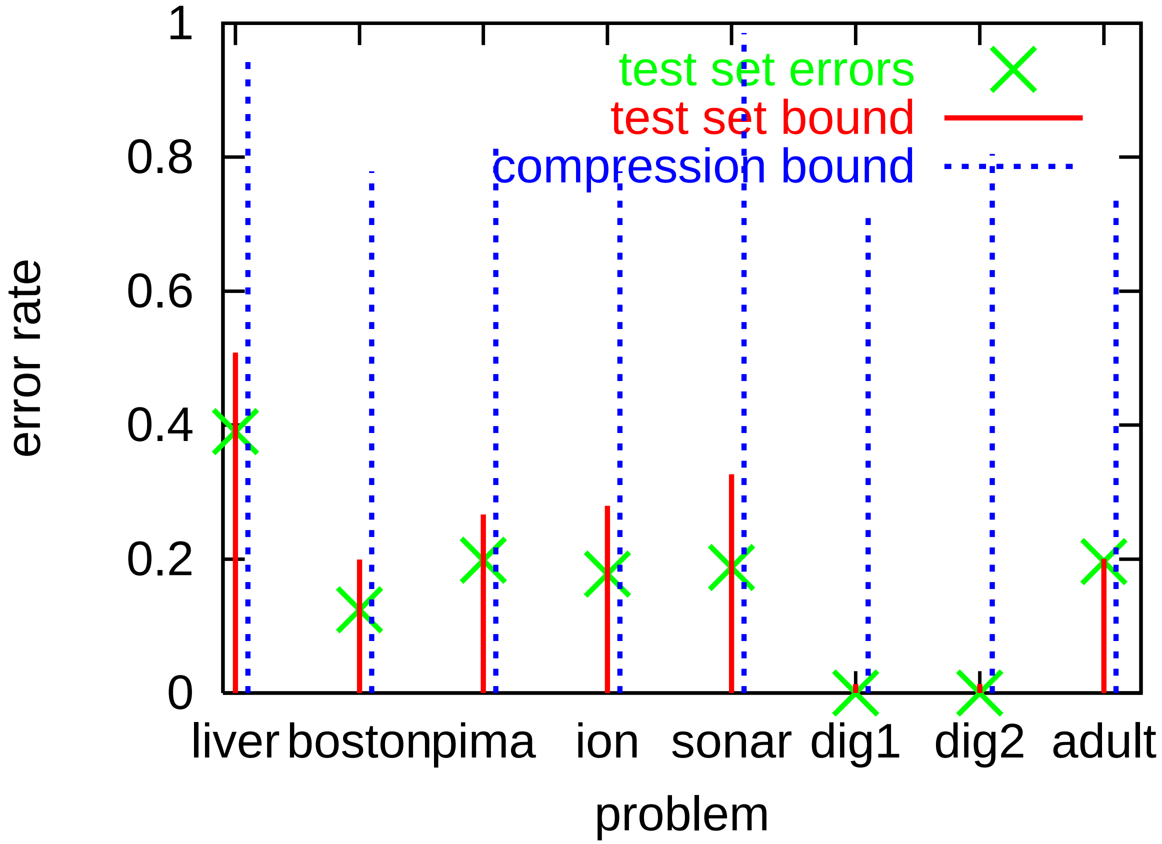
Ideally: communication complexity of compression set.

$$|\sigma| = \log_2 \binom{m}{|S'|} + \log_2(m+1) + |S'|$$

Reality: communication complexity in train+test set.

$$|\sigma| = \log_2 \binom{m+n}{|S'|} + \log_2(m+n+1) + |S'|$$

Sample Compression Bound Application



VC bound

$VC(H)$ = VC dimension = the size of the largest subset of X which can be labeled arbitrarily by elements c of H.

Theorem: (VC bound VC71) For all sets of classifiers H, let $d = VC(H)$ with probability $1 - \delta$:

$$\forall c \in H : \quad \hat{c}_S = 0 \Rightarrow c_D \leq 2 \frac{d\left(1 + \ln \frac{2m}{d}\right) + \ln \frac{4}{\delta}}{m}$$

Core Lemma of VC bound

Sauer's Lemma: If there are m examples, then there are at most $\left(\frac{VC(H)}{em}\right)^{VC(H)}$ ways to label the examples using elements of c in H.

Communication language = uniform over all transductive classifiers consistent with $c \in H$.

VC bound comparison