A primer for PAC-Bayes

John Langford

1. Occam's Razor bound

2. PAC-Bayes bound

3. Proof

4. Bits-Back

The Model

X = (arbitrary) input space

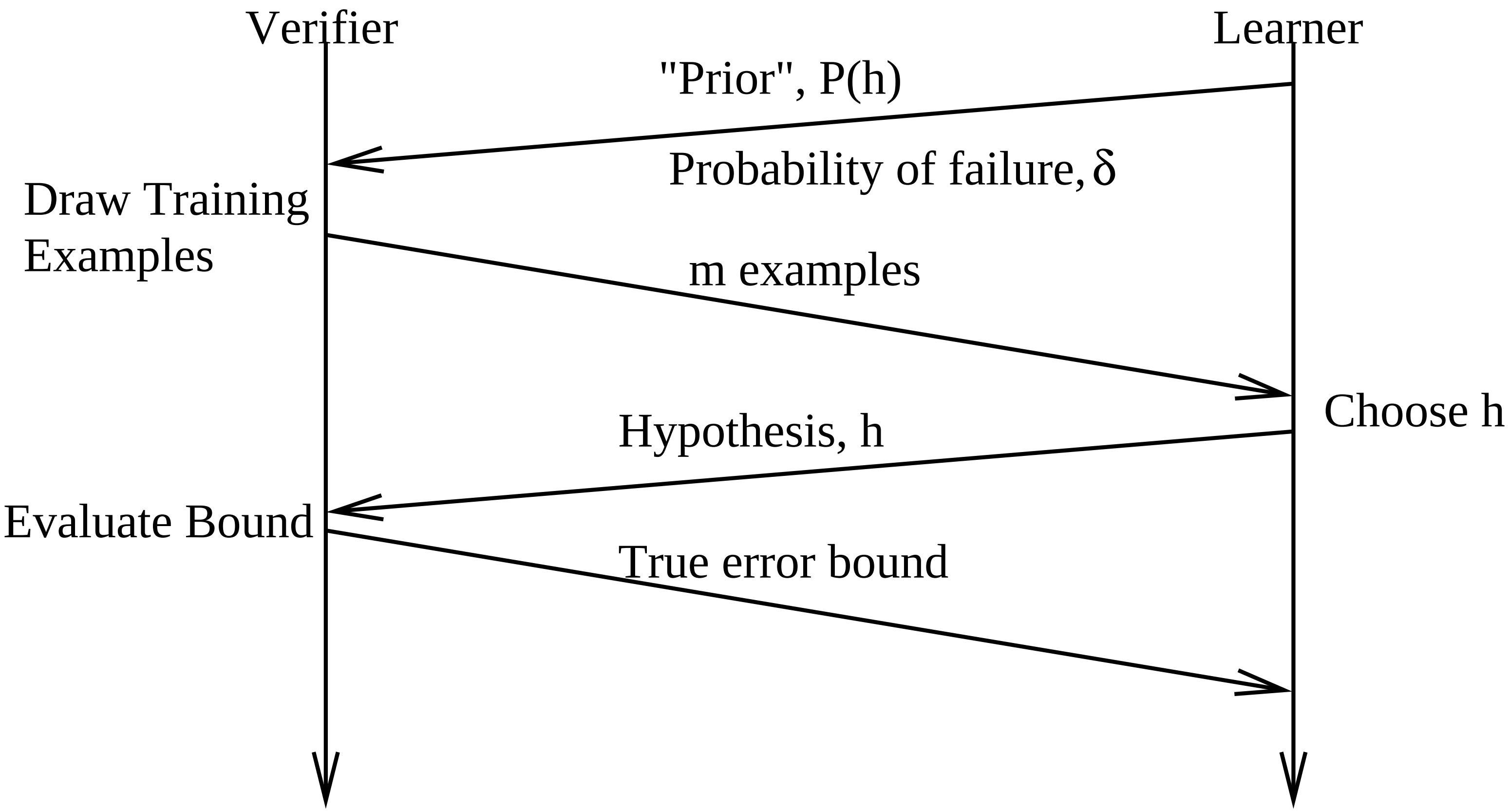
$Y \in \{0, 1\}$ = prediction space

Assume: Examples (X, Y) are drawn independently from distribution D

Terminology

- Samples $= S$
- Learning algorithm $A: \; S \to h$
- True Error rate $e_D(h) = \mathrm{Pr}_D(h(x) \neq y)$
- Empirical Error rate $\hat{e}_S(h) = \mathrm{Pr}_S(h(x) \neq y)$

Training Set Bound



"Occam's Razor" bound (BEHW87)

Theorem:

$$\forall P(h) \quad \Pr_{S \sim D^m}\left(\exists h \in H:\ e_D(h) > \bar{e}(h, \hat{e}_S(h), P(h)\delta)\right) \leq \delta$$

where $\bar{e}(h, \frac{k}{m}, \delta) = \max\{p | \mathrm{Bin}_{p,m}(\frac{k}{m}) \geq \delta\}$

Proof: by nonuniform bound

Another Representation

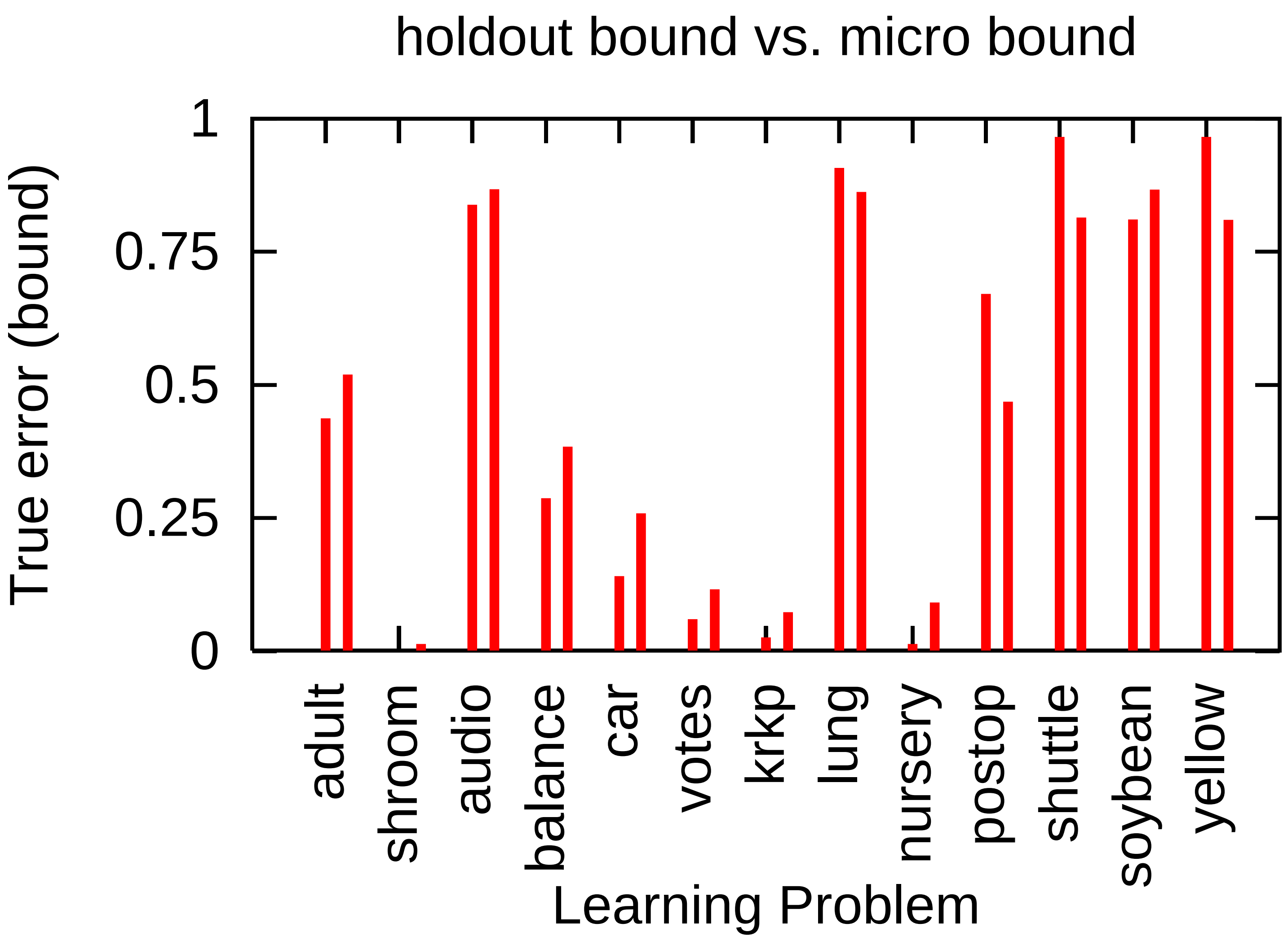
$$\widehat{e}(h_1)\text{———}e(h_1)$$

$$\widehat{e}(h_2)\text{———}e(h_2)$$

$$\widehat{e}(h_3)\text{—————}e(h_3)$$

$$\widehat{e}(h_4)\text{—————}e(h_4)$$

$$\widehat{e}(h_5)\text{——————}e(h_5)$$

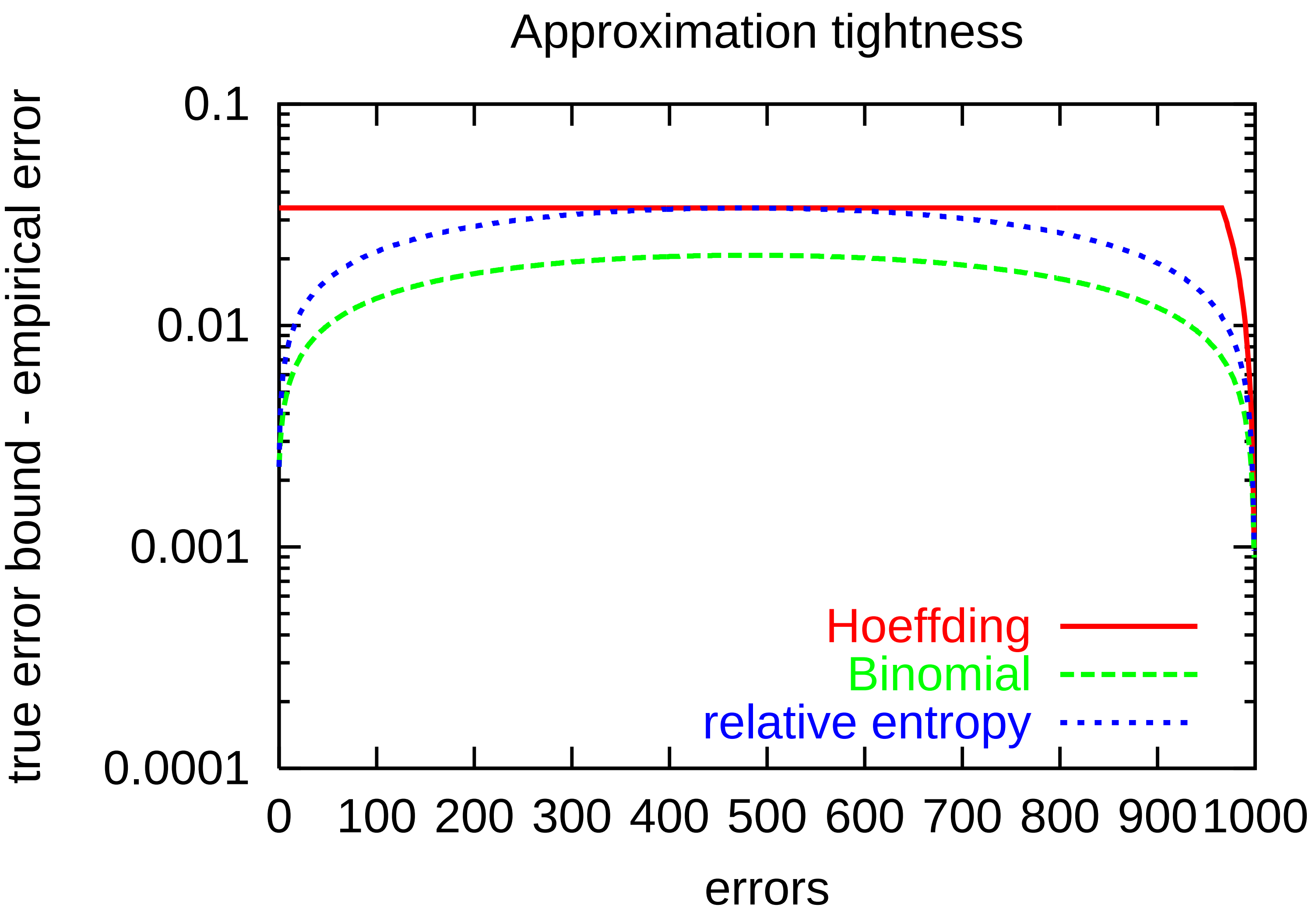

holdout bound vs. micro bound
True error (bound)
1
0.75
0.5
0.25
0
adult
shroom
audio
balance
car
votes
krkp
lung
nursery
postop
shuttle
soybean
yellow
Learning Problem

Corollary:

$$\forall P(h) \quad \Pr_{S\sim D^m}\left(\mathsf{KL}(\hat{e}_S(h)||e_D(h)) \geq \frac{\ln\frac{1}{P(h)} + \ln\frac{1}{\delta}}{m}\right) \leq \delta$$

Corollary:

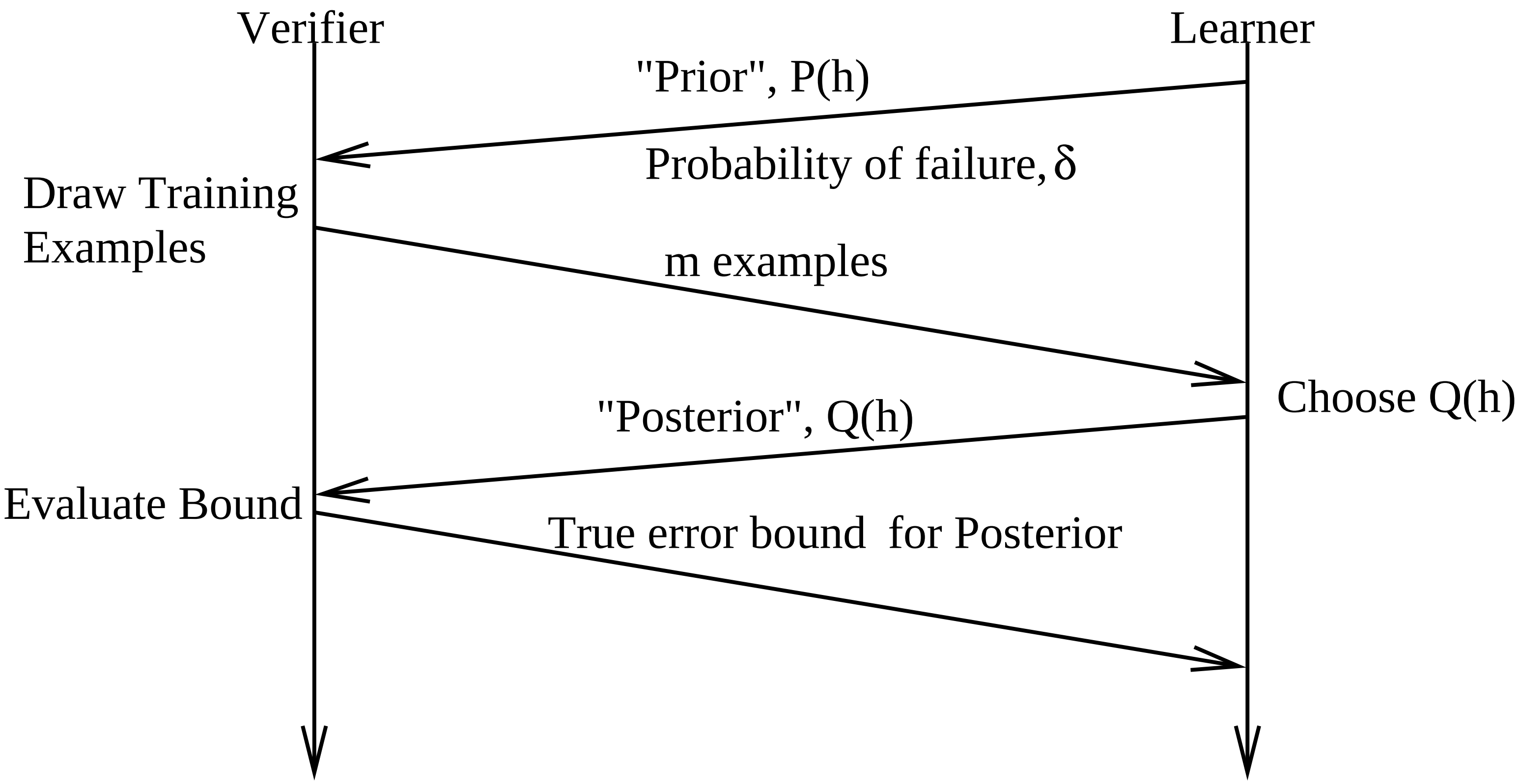
$$\forall P(h) \quad \Pr_{S\sim D^m}\left(e_D(h) \geq \hat{e}_S(h) + \sqrt{\frac{\ln\frac{1}{P(h)} + \ln\frac{1}{\delta}}{2m}}\right) \leq \delta$$



Approximation tightness
true error bound - empirical error
0.1
0.01
0.001
0.0001
0
100
200
300
400
500
600
700
800
900
1000
errors
Hoeffding
Binomial
relative entropy

1. Occam's Razor bound
2. PAC-Bayes bound
3. Proof
4. Bits-Back

PAC−Bayes Bound



PAC-Bayes Terminology

1. $\hat{e}_{Q,S} = E_{h \sim Q} \hat{e}_S(h) =$ expected training error over Q.

2. $e_{Q,D} = E_{h \sim Q} e_D(h) =$ expected true error over Q.

3. $\mathrm{KL}(Q||P) = E_{h \sim Q} \ln \frac{Q(h)}{P(h)}$

4. $\mathrm{KL}(\hat{e}_{Q,S}||e_{Q,D}) = \hat{e}_{Q,S} \ln \frac{\hat{e}_{Q,S}}{e_{Q,D}} + (1 - \hat{e}_{Q,S}) \ln \frac{1-\hat{e}_{Q,S}}{1-e_{Q,D}}$

PAC-Bayes bound (McAllester99, me00, Seeger02)

Theorem:

$$\forall P(h) \quad \Pr_{S \sim D^m}\left(\exists Q \sim h : \ \mathrm{KL}(\hat{e}_{Q,S} || e_{Q,D}) > \frac{\mathrm{KL}(Q||P) + \ln \frac{m+1}{\delta}}{m}\right) \leq \delta$$

Proof: later

Stochastic Neural Network

1. Pick a Neural Network

2. For every weight, w

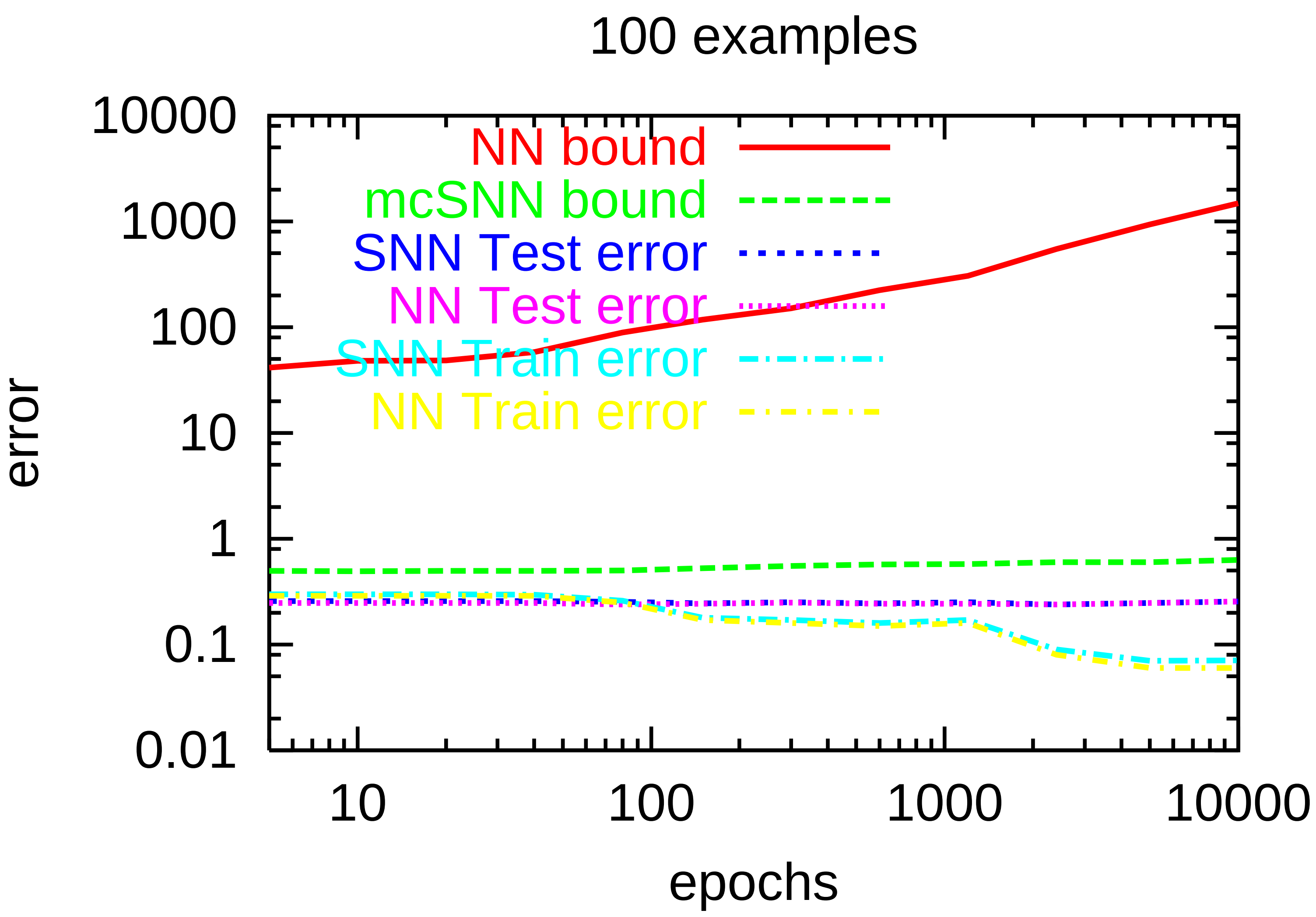
   (a) Find sensitivity, $s(w)$

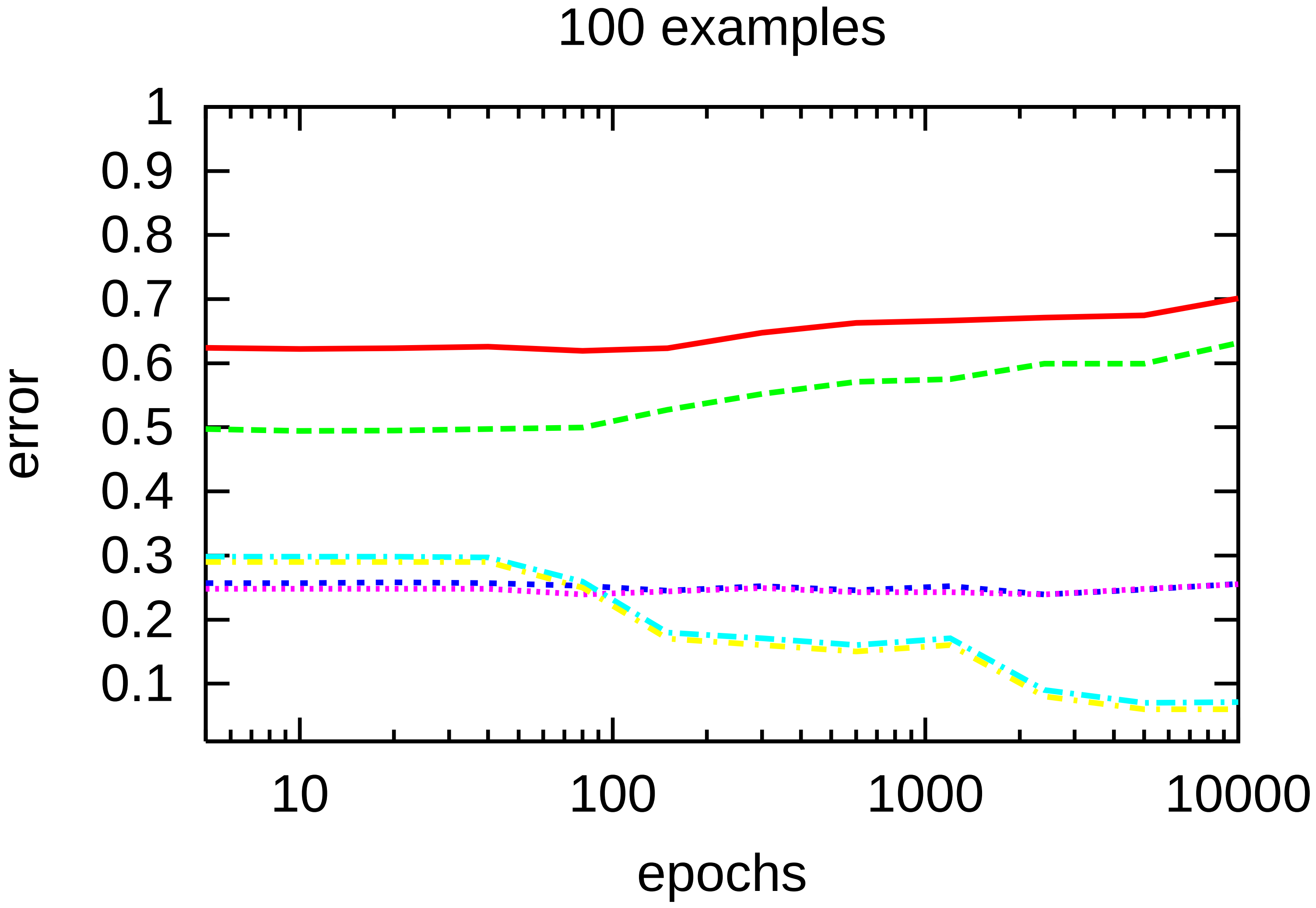
3. Perturb weights by gaussian $N\left(w, \frac{1}{s(w)}\right)$

PAC-Bayes for SNN (Rich & I, NIPS 2001)

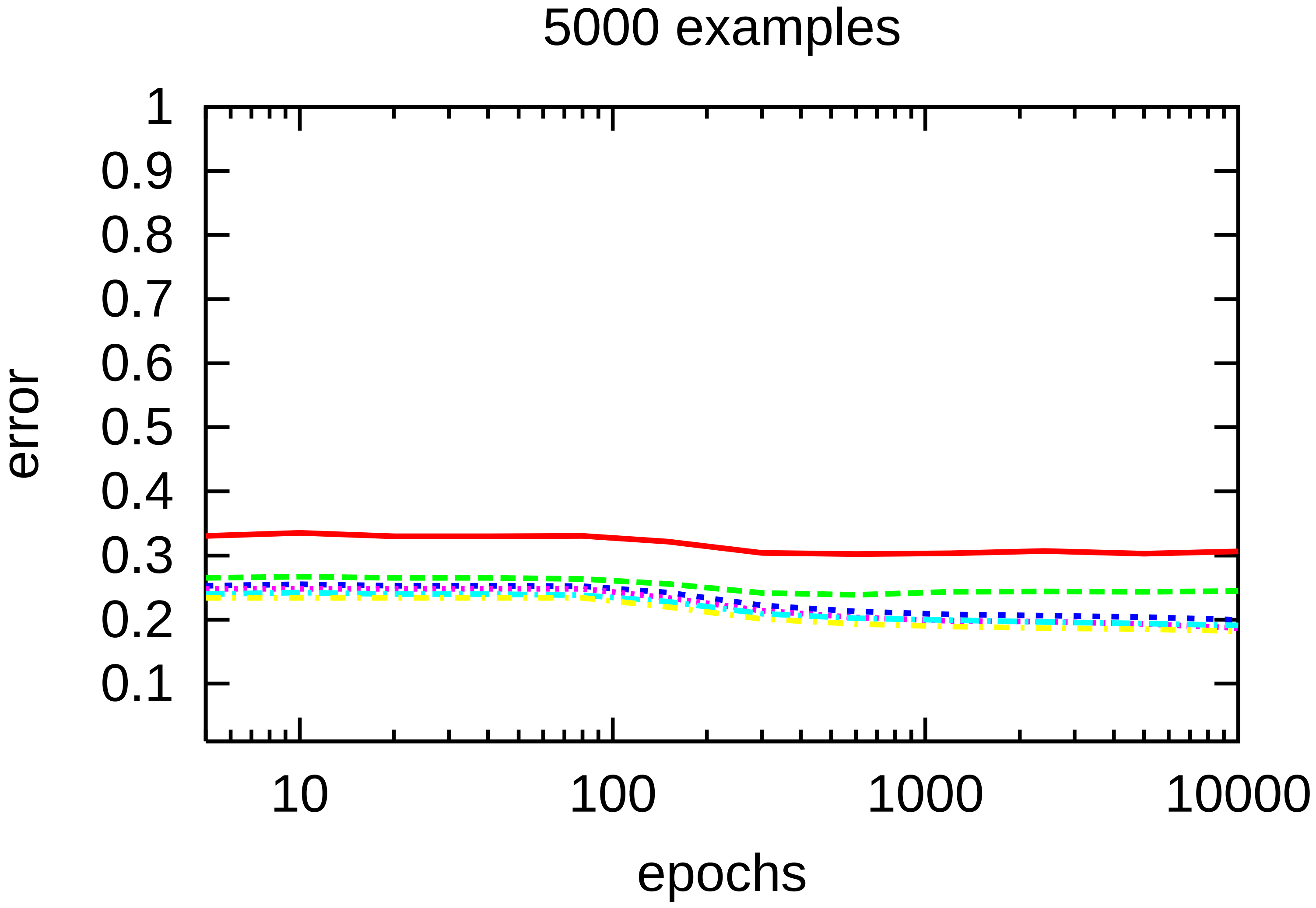
let n = number of weights

1. Pick Prior = $N(0, \alpha)^n$
   - actually, a mixture.
2. Pick Posterior = $N\left(w, \frac{1}{s(w)}\right)$
3. Monte carlo estimate and bound $\hat{e}_{Q,S}$


100 examples
NN bound
mcSNN bound
SNN Test error
NN Test error
SNN Train error
NN Train error
error
10000
1000
100
10
1
0.1
0.01
10
100
1000
10000
epochs



100 examples
error
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
10
100
1000
10000
epochs


5000 examples
error
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
10
100
1000
10000
epochs

1. Occam's Razor bound

2. PAC-Bayes bound

3. Proof

4. Bits-Back

Lemma: $\forall P \ \ \mathrm{Pr}_{S \sim D^m} \left(E_{h \sim P} e^{m \mathrm{KL}(\hat{e}(h) || e(h))} \leq \frac{m+1}{\delta} \right) \geq 1 - \delta$

Lemma: $\forall P \;\; \Pr_{S\sim D^m}\left(E_{h\sim P} e^{m\mathsf{KL}(\hat{e}(h)||e(h))} \leq \frac{m+1}{\delta}\right) \geq 1-\delta$

Lemma: $\forall Q, P, f \;\; E_{h\sim Q} f(h) \leq \mathsf{KL}(Q||P) + \ln E_{h\sim P} e^{f(h)}$

Lemma: $\forall P \; \Pr_{S\sim D^m}\left(E_{h\sim P} e^{m\mathsf{KL}(\hat{e}(h)||e(h))} \leq \frac{m+1}{\delta}\right) \geq 1-\delta$

Lemma: $\forall Q, P, f \; E_{h\sim Q} f(h) \leq \mathsf{KL}(Q||P) + \ln E_{h\sim P} e^{f(h)}$

Proof: Let $f(h) = m\mathsf{KL}(\hat{e}(h)||e(h))$

$\Rightarrow m E_{h\sim Q} \mathsf{KL}(\hat{e}(h)||e(h)) \leq \mathsf{KL}(Q||P) + \ln E_{h\sim P} e^{m\mathsf{KL}(\hat{e}(h)||e(h))}$

Lemma: $\forall P \;\; \Pr_{S\sim D^m}\left(E_{h\sim P} e^{m\mathsf{KL}(\hat{e}(h)||e(h))} \leq \frac{m+1}{\delta}\right) \geq 1-\delta$

Lemma: $\forall Q, P, f \;\; E_{h\sim Q} f(h) \leq \mathsf{KL}(Q||P) + \ln E_{h\sim P} e^{f(h)}$

Proof: Let $f(h) = m\mathsf{KL}(\hat{e}(h)||e(h))$

$\Rightarrow m E_{h\sim Q}\mathsf{KL}(\hat{e}(h)||e(h)) \leq \mathsf{KL}(Q||P) + \ln E_{h\sim P} e^{m\mathsf{KL}(\hat{e}(h)||e(h))}$

$\Rightarrow \Pr_{S\sim D^m}\left(m E_{h\sim Q}\mathsf{KL}(\hat{e}(h)||e(h)) \leq \mathsf{KL}(Q||P) + \ln \frac{m+1}{\delta}\right) \geq 1-\delta$

Lemma: $\forall P \ \Pr_{S\sim D^m}\left(E_{h\sim P} e^{m\mathrm{KL}(\hat{e}(h)||e(h))} \leq \frac{m+1}{\delta}\right) \geq 1-\delta$

Lemma: $\forall Q, P, f \ E_{h\sim Q} f(h) \leq \mathrm{KL}(Q||P) + \ln E_{h\sim P} e^{f(h)}$

Proof: Let $f(h) = m\mathrm{KL}(\hat{e}(h)||e(h))$

$\Rightarrow mE_{h\sim Q}\mathrm{KL}(\hat{e}(h)||e(h)) \leq \mathrm{KL}(Q||P) + \ln E_{h\sim P} e^{m\mathrm{KL}(\hat{e}(h)||e(h))}$

$\Rightarrow \Pr_{S\sim D^m}\left(mE_{h\sim Q}\mathrm{KL}(\hat{e}(h)||e(h)) \leq \mathrm{KL}(Q||P) + \ln \frac{m+1}{\delta}\right) \geq 1-\delta$

$\Rightarrow \Pr_{S\sim D^m}\left(\mathrm{KL}(E_{h\sim Q}\hat{e}(h)||E_{h\sim Q}e(h)) \leq \frac{\mathrm{KL}(Q||P) + \ln \frac{m+1}{\delta}}{m}\right) \geq 1-\delta$
(jensen)

Lemma: $\forall P \ \Pr_{S\sim D^m}\left(E_{h\sim P} e^{m\mathsf{KL}(\hat{e}(h)||e(h))} \leq \frac{m+1}{\delta}\right) \geq 1-\delta$

Lemma: $\forall P$ $\Pr_{S\sim D^m}\left(E_{h\sim P}e^{m\mathsf{KL}(\hat{e}(h)||e(h))} \leq \frac{m+1}{\delta}\right) \geq 1-\delta$

Proof: $\Pr_{\mathrm{error}}(S) = p^{\#heads}(1-p)^{\#tails}$

Lemma: $\forall P$ $\Pr_{S\sim D^m}\left(E_{h\sim P}e^{m\mathsf{KL}(\hat{e}(h)||e(h))} \leq \frac{m+1}{\delta}\right) \geq 1-\delta$

Proof: $\Pr_{\text{error}}(S) = p^{\#heads}(1-p)^{\#tails}$

$= p^{\hat{p}m}(1-p)^{(1-\hat{p})m}$

Lemma: $\forall P\ \ \mathrm{Pr}_{S\sim D^m}\left(E_{h\sim P}e^{m\mathsf{KL}(\hat{e}(h)||e(h))} \le \frac{m+1}{\delta}\right) \ge 1-\delta$

Proof: $\mathrm{Pr}_{\mathrm{error}}(S) = p^{\#heads}(1-p)^{\#tails}$

$= p^{\hat{p}m}(1-p)^{(1-\hat{p})m}$

$= e^{\log p^{\hat{p}m} + \log(1-p)^{(1-\hat{p})m}}$

Lemma: $\forall P \;\; \Pr_{S\sim D^m}\left(E_{h\sim P} e^{m \mathsf{KL}(\hat{e}(h)||e(h))} \leq \frac{m+1}{\delta}\right) \geq 1-\delta$

Proof: $\Pr_{\text{error}}(S) = p^{\#heads}(1-p)^{\#tails}$

$= p^{\hat{p}m}(1-p)^{(1-\hat{p})m}$

$= e^{\log p^{\hat{p}m} + \log (1-p)^{(1-\hat{p})m}}$

$= e^{-m(\hat{p}\log\frac{1}{p} - (1-\hat{p})\log\frac{1}{1-p})}$

Lemma: $\forall P$ $\Pr_{S\sim D^m}\left(E_{h\sim P} e^{m\mathrm{KL}(\hat{e}(h)||e(h))} \leq \frac{m+1}{\delta}\right) \geq 1-\delta$

Proof: $\Pr_{\mathrm{error}}(S) = p^{\#heads}(1-p)^{\#tails}$

$= p^{\hat{p}m}(1-p)^{(1-\hat{p})m}$

$= e^{\log p^{\hat{p}m} + \log(1-p)^{(1-\hat{p})m}}$

$= e^{-m(\hat{p}\log\frac{1}{p} - (1-\hat{p})\log\frac{1}{1-p})}$

$= e^{m(-(\hat{p}\log\frac{\hat{p}}{p} - (1-\hat{p})\log\frac{1-\hat{p}}{1-p}) - (\hat{p}\log\frac{1}{\hat{p}} - (1-\hat{p})\log\frac{1}{1-\hat{p}}))}$

Lemma: $\forall P \;\; \Pr_{S\sim D^m}\left(E_{h\sim P} e^{m\mathsf{KL}(\hat{e}(h)||e(h))} \leq \frac{m+1}{\delta}\right) \geq 1-\delta$

Proof: $\Pr_{\text{error}}(S) = p^{\#heads}(1-p)^{\#tails}$

$= p^{\hat{p}m}(1-p)^{(1-\hat{p})m}$

$= e^{\log p^{\hat{p}m} + \log(1-p)^{(1-\hat{p})m}}$

$= e^{-m(\hat{p}\log\frac{1}{p} - (1-\hat{p})\log\frac{1}{1-p})}$

$= e^{m(-(\hat{p}\log\frac{\hat{p}}{p} - (1-\hat{p})\log\frac{1-\hat{p}}{1-p}) - (\hat{p}\log\frac{1}{\hat{p}} - (1-\hat{p})\log\frac{1}{1-\hat{p}}))}$

$= e^{m(-\mathsf{KL}(\hat{p}||p) - \mathsf{H}(\hat{p}))}$

$$E_S e^{m\mathsf{KL}(\hat{p}||p)} \leq \sum_{\hat{p}} e^{m\mathsf{KL}(\hat{p}||p)} \mathsf{Pr}(S)|\{S : \hat{p}\}|$$

$$E_S e^{m\mathsf{KL}(\widehat{p}||p)} \leq \sum_{\widehat{p}} e^{m\mathsf{KL}(\widehat{p}||p)}\,\mathsf{Pr}(S)|\{S : \widehat{p}\}|$$

$$= \sum_{\widehat{p}} e^{m\mathsf{KL}(\widehat{p}||p)} e^{m(-\mathsf{KL}(\widehat{p}||p)-\mathsf{H}(\widehat{p}))} e^{m\mathsf{H}(\widehat{p})}$$

$$E_S e^{m\mathsf{KL}(\hat{p}||p)} \leq \sum_{\hat{p}} e^{m\mathsf{KL}(\hat{p}||p)} \mathsf{Pr}(S)|\{S : \hat{p}\}|$$

$$= \sum_{\hat{p}} e^{m\mathsf{KL}(\hat{p}||p)} e^{m(-\mathsf{KL}(\hat{p}||p) - \mathsf{H}(\hat{p}))} e^{m\mathsf{H}(\hat{p})}$$

$$= \sum_{\hat{p}} 1$$

$$E_S e^{m\mathsf{KL}(\widehat{p}||p)} \leq \sum_{\widehat{p}} e^{m\mathsf{KL}(\widehat{p}||p)}\, \mathsf{Pr}(S)|\{S : \widehat{p}\}|$$

$$= \sum_{\widehat{p}} e^{m\mathsf{KL}(\widehat{p}||p)} e^{m(-\mathsf{KL}(\widehat{p}||p) - \mathsf{H}(\widehat{p}))} e^{m\mathsf{H}(\widehat{p})}$$

$$= \sum_{\widehat{p}} 1$$

$$= m + 1$$

$$E_S e^{m\mathsf{KL}(\hat{p}||p)} \leq \sum_{\hat{p}} e^{m\mathsf{KL}(\hat{p}||p)} \mathsf{Pr}(S)|\{S : \hat{p}\}|$$

$$= \sum_{\hat{p}} e^{m\mathsf{KL}(\hat{p}||p)} e^{m(-\mathsf{KL}(\hat{p}||p)-\mathsf{H}(\hat{p}))} e^{m\mathsf{H}(\hat{p})}$$

$$= \sum_{\hat{p}} 1$$

$$= m + 1$$

$$\Rightarrow E_S E_{h\sim P} e^{m\mathsf{KL}(\hat{e}(h)||e(h))} \leq m + 1$$

$$E_S e^{m\mathsf{KL}(\hat{p}||p)} \leq \sum_{\hat{p}} e^{m\mathsf{KL}(\hat{p}||p)} \mathsf{Pr}(S)|\{S : \hat{p}\}|$$

$$= \sum_{\hat{p}} e^{m\mathsf{KL}(\hat{p}||p)} e^{m(-\mathsf{KL}(\hat{p}||p) - \mathsf{H}(\hat{p}))} e^{m\mathsf{H}(\hat{p})}$$

$$= \sum_{\hat{p}} 1$$

$$= m + 1$$

$$\Rightarrow E_S E_{h\sim P} e^{m\mathsf{KL}(\hat{e}(h)||e(h))} \leq m + 1$$

$$\Rightarrow \forall P \ \ \mathsf{Pr}_S(E_{h\sim P} e^{m\mathsf{KL}(\hat{e}(h)||e(h))} \leq \frac{m+1}{\delta}) \geq 1 - \delta \ \ \text{(markov)}$$

Lemma: $\forall Q, P, f \;\; E_{h \sim Q} f(h) \leq \mathsf{KL}(Q||P) + \ln E_{h \sim P} e^{f(h)}$

Lemma: $\forall Q, P, f \; E_{h\sim Q} f(h) \leq \mathrm{KL}(Q||P) + \ln E_{h\sim P} e^{f(h)}$

Proof: Let $P_G(h) = \frac{e^{f(h)}}{E_P e^{f(h)}} P(h)$

Lemma: $\forall Q, P, f \;\; E_{h\sim Q} f(h) \leq \mathsf{KL}(Q||P) + \ln E_{h\sim P} e^{f(h)}$

Proof: Let $P_G(h) = \frac{e^{f(h)}}{E_P e^{f(h)}} P(h)$

$0 \leq \mathsf{KL}(Q||P_G) = E_{h\sim Q} \ln \frac{E_{h\sim P} e^{f(h)}}{e^{f(h)}} \frac{Q(h)}{P(h)}$

Lemma: $\forall Q, P, f \;\; E_{h\sim Q} f(h) \leq \mathrm{KL}(Q||P) + \ln E_{h\sim P} e^{f(h)}$

Proof: Let $P_G(h) = \frac{e^{f(h)}}{E_P e^{f(h)}} P(h)$

$0 \leq \mathrm{KL}(Q||P_G) = E_{h\sim Q} \ln \frac{E_{h\sim P} e^{f(h)}}{e^{f(h)}} \frac{Q(h)}{P(h)}$

$= \mathrm{KL}(Q||P) - E_{h\sim Q} f(h) + \ln E_{h\sim P} e^{f(h)}$

1. Occam's Razor bound

2. PAC-Bayes bound

3. Proof

4. Bits-Back

PAC-Bayes versus MDL

1. Shorter description length?

2. But $Q \sim h$ may be very unshort

Bits-Back (Hinton and van Camp 93)

1. Alice talks to Bob a lot.

2. Alice has labeled examples.

3. Bob has same examples, but no labels.

How many *extra* bits are required for Alice to send the labels?

Suppose Alice and Bob share a learning algorithm with posterior, Q.

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

Suppose Alice and Bob share a learning algorithm with posterior, Q.

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

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

Suppose Alice and Bob share a learning algorithm with posterior, Q.

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

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

3. Bob decodes labeled examples

Suppose Alice and Bob share a learning algorithm with posterior, Q.

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

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

3. Bob decodes labeled examples

4. Bob runs learning algorithm to decode Q

Suppose Alice and Bob share a learning algorithm with posterior, Q.

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

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

3. Bob decodes labeled examples

4. Bob runs learning algorithm to decode Q

5. Bob decodes M

Bits-back Summary

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

total message $= M,\ Q,$ labels

PAC-Bayes open questions

Tightness:

1. Can we lose the $m+1$ term?

2. Can we reduce it to the CDF of the Binomial?

Application: Where else does it apply?

1. Matthias Seeger says: Gaussian Processes