Introduction to PAC Bayesian bounds

Saurabh Khanna,
Signal Processing for Communication, ECE, IISc
Outline

- PAC Bayesian framework
  - Binary classification problem and Gibbs classifier

- PAC Bayesian bounds
  - Statement
  - Insights
  - Theory behind the bound
PAC learning framework [Valiant ’84]

- **PAC** stands for **Probably Approximately Correct**
  - **Approximately**
    - Provide guarantees on the approximation error of empirical estimates
  - **Probably**
    - Guarantees that hold with high probability
Supervised learning - some definitions

- $\mathcal{X}$ - sample space
Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(x_1, y_1), \ldots (x_m, y_m)\}$ - training data set (i.i.d.)
Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(x_1, y_1), \ldots (x_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(x_1, y_1), \ldots (x_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $A : S \rightarrow \mathcal{H}$ - algorithm
Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(x_1, y_1), \ldots (x_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $\mathcal{A} : S \rightarrow \mathcal{H}$ - algorithm
- $h(x)$ - prediction of hypothesis/classifier $h \in \mathcal{H}$ for input sample $x$
Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(x_1, y_1), \ldots (x_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $A : S \rightarrow \mathcal{H}$ - algorithm
- $h(x)$ - prediction of hypothesis/classifier $h \in \mathcal{H}$ for input sample $x$
- $l(h, x)$ - instantaneous loss/risk of $h$ on $x$
Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(x_1, y_1), \ldots (x_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $A : S \rightarrow \mathcal{H}$ - algorithm
- $h(x)$ - prediction of hypothesis/classifier $h \in \mathcal{H}$ for input sample $x$
- $l(h, x)$ - instantaneous loss/risk of $h$ on $x$
- $l(h, \mathcal{W})$ - expected loss of hypothesis $h$ on entire $\mathcal{X}$, assuming input distribution to be $\mathcal{W}$
Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $\mathcal{A} : S \rightarrow \mathcal{H}$ - algorithm
- $h(x)$ - prediction of hypothesis/classifier $h \in \mathcal{H}$ for input sample $x$
- $l(h, x)$ - instantaneous loss/risk of $h$ on $x$
- $l(h, \mathcal{W})$ - expected loss of hypothesis $h$ on entire $\mathcal{X}$, assuming input distribution to be $\mathcal{W}$
- $D$ - true but unknown distribution on $\mathcal{X}$
Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(x_1, y_1), \ldots (x_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $\mathcal{A} : S \rightarrow \mathcal{H}$ - algorithm
- $h(x)$ - prediction of hypothesis/classifier $h \in \mathcal{H}$ for input sample $x$
- $l(h, x)$ - instantaneous loss/risk of $h$ on $x$
- $l(h, W)$ - expected loss of hypothesis $h$ on entire $\mathcal{X}$, assuming input distribution to be $W$
- $D$ - true but unknown distribution on $\mathcal{X}$
- $l(h, D) = \mathbb{E}_{x \sim D} [l(h, x)]$ - expected loss of hypothesis $h$
Supervised learning - some definitions

- \( \mathcal{X} \) - sample space
- \( \mathcal{Y} \) - label space
- \( m \) - number of training samples
- \( S = \{(x_1, y_1), \ldots (x_m, y_m)\} \) - training data set (i.i.d.)
- \( \mathcal{H} \) - hypothesis space
- \( A : S \rightarrow \mathcal{H} \) - algorithm
- \( h(x) \) - prediction of hypothesis/classifier \( h \in \mathcal{H} \) for input sample \( x \)
- \( l(h, x) \) - instantaneous loss/risk of \( h \) on \( x \)
- \( l(h, W) \) - expected loss of hypothesis \( h \) on entire \( \mathcal{X} \), assuming input distribution to be \( W \)
- \( D \) - true but unknown distribution on \( \mathcal{X} \)
- \( l(h, D) = \mathbb{E}_{x \sim D} [l(h, x)] \) - expected loss of hypothesis \( h \)
- \( l(h, S) = \frac{1}{m} \sum_{i=1}^{m} l(h, x_i) \) - empirical loss of hypothesis \( h \)
PAC-Bayesian setting

▶ Start with a prior $P$ on the hypothesis space $\mathcal{H}$.
PAC-Bayesian setting

- Start with a prior $P$ on the hypothesis space $\mathcal{H}$.
- After observing $S$, the algorithm $A$ generates a posterior $Q$ on $\mathcal{H}$.
PAC-Bayesian setting

- Start with a prior $P$ on the hypothesis space $\mathcal{H}$.

- After observing $S$, the algorithm $A$ generates a posterior $Q$ on $\mathcal{H}$.

- In PAC-Bayes, the classifier is random/stochastic in nature (Gibbs classifier)
  1. For given input $x \in \mathcal{X}$, draw $h$ from $\mathcal{H}$ acc. to $Q$.
  2. Assign label $y = h(x)$

- Expected loss: $l(Q, D) = \mathbb{E}_{Q}[l(h, D)]$

- Empirical loss: $l(Q, S) = \mathbb{E}_{Q}[l(h, S)]$
PAC-Bayesian setting

- Start with a prior $P$ on the hypothesis space $\mathcal{H}$.

- After observing $S$, the algorithm $A$ generates a posterior $Q$ on $\mathcal{H}$.

- In PAC-Bayes, the classifier is random/stochastic in nature (Gibbs classifier)
  1. For given input $x \in \mathcal{X}$, draw $h$ from $\mathcal{H}$ acc. to $Q$.
  2. Assign label $y = h(x)$

- Expected loss: $l(Q, D) = \mathbb{E}_Q[l(h, D)]$

- Empirical loss: $l(Q, S) = \mathbb{E}_Q[l(h, S)]$
PAC-Bayesian setting

PAC-Bayesian framework:

\[ P: \text{Prior on } \mathcal{H} \rightarrow \text{Algorithm } \mathcal{A} \rightarrow Q: \text{Posterior on } \mathcal{H} \]

Training data \( S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \)

Output of the algorithm is a Gibbs classifier.

Let \( l(Q, S) \) denote the empirical loss/risk of the Gibbs classifier generated by the algorithm \( \mathcal{A} \).

\[
l(Q, S) = \mathbb{E}_Q[l(h, S)], \text{ where } l(h, S) = \frac{1}{m} \sum_{i=1}^{m} l(h, x_i)
\]

Question?
How close is empirical loss \( l(Q, S) \) to the true loss \( l(Q, D) \)?
PAC-Bayesian bounds - different flavors

- Mc Allester bound ['98]

\[ \left| \mathbb{E}_Q [l(h, S)] - \mathbb{E}_Q [l(h, D)] \right|^2 \leq ?? \]
PAC-Bayesian bounds - different flavors

- Mc Allester bound ['98]
  \[ |\mathbb{E}_Q [l(h, S)] - \mathbb{E}_Q [l(h, D)] |^2 \leq \text{??} \]

- Seeger bound ['02]
  \[ kl(\mathbb{E}_Q [l(h, S)] \parallel \mathbb{E}_Q [l(h, D)]) \leq \text{??} \]

where \( kl(q\|p) \) is called the small KL divergence given by
\[ kl(q\|p) = q \log \frac{q}{p} + (1 - q) \log \frac{1-q}{1-p} \]
PAC-Bayesian Bound [Seeger ’02]

- With probability at least \(1 - \delta\) over the choice of \(S \sim D^m\),

\[
\text{kl} \left( l(Q, S) \| l(Q, D) \right) \leq \frac{\text{KL}(Q \| P) + \log \frac{m+1}{\delta}}{m}
\]
Intuition behind the bound (1/2)

- With probability at least \((1 - \delta)\) over the choice of \(S \sim D^m\),

\[
\text{kl}(l(Q, S) \| l(Q, D)) \leq \frac{KL(Q \| P) + \log \frac{m+1}{\delta}}{m}
\]
Intuition behind the bound (1/2)

- With probability at least $(1 - \delta)$ over the choice of $S \sim D^m$,

\[
kl (l(Q, S) \| l(Q, D)) \leq \frac{KL(Q\|P) + \log \frac{m+1}{\delta}}{m}
\]

- $KL(Q\|P) = \langle \mathbb{E}_Q \log \left( \frac{1}{P} \right) \rangle - H(Q)$

  - cross-entropy
  - entropy
Intuition behind the bound (1/2)

- With probability at least \((1 - \delta)\) over the choice of \(S \sim D^m\),
  \[
  kl(l(Q, S) \| l(Q, D)) \leq \frac{KL(Q \| P) + \log \frac{m+1}{\delta}}{m}
  \]

- \(KL(Q \| P) = \langle E_Q \log \left( \frac{1}{P} \right) \rangle - H(Q)\)
  
  - cross-entropy
  - entropy

- Preferred choice for posterior \(Q\):
  1. has maximum entropy
  2. reduces empirical loss \(l(Q, S)\)
Intuition behind the bound (1/2)

- With probability at least $(1 - \delta)$ over the choice of $S \sim D^m$,

$$\text{KL}(l(Q, S)||l(Q, D)) \leq \frac{\text{KL}(Q||P) + \log \frac{m+1}{\delta}}{m}$$

- $\text{KL}(Q||P) = \langle \mathbb{E}_Q \log \left( \frac{1}{P} \right) \rangle - H(Q)$

- Preferred choice for posterior $Q$:
  1. has maximum entropy
  2. reduces empirical loss $l(Q, S)$

- Preferred choice for prior $P$:
  1. has low complexity
  2. is close to posterior $Q$
Intuition behind the bound (2/2)

- With probability at least \((1 - \delta)\) over the choice of \(S \sim D^m\),

\[
kl \left( I(Q, S) \parallel I(Q, D) \right) \leq \frac{KL(Q \parallel P) + \log \frac{m+1}{\delta}}{m}
\]

- Other key take-away points:
  1. w.h.p. guarantees on expected performance
  2. explicit way to incorporate prior knowledge
  3. non assumption on correctness of prior \(P\)
  4. explicit dependence on the loss function
  5. holds for any posterior \(Q\)
  6. bound is meant for randomized/stochastic classifiers
Theory behind PAC Bayesian bound - major milestones

- **PAC Bayesian bound:**

  \[ kl(I(Q, S)||I(Q, D)) \leq \frac{KL(Q||P) + \log \frac{m+1}{\delta}}{m} \text{ w.h.p.} \]

- **Milestone-1** Fenchel inequality in convex analysis
  [Rockafeller, 70]

- **Milestone-2** Variational factorization of KL divergence
  [Donsker and Varadhan, 75]
  - Also known as Compression Lemma

- **Milestone-3** PAC Bayesian bound [Seeger, 02]
Duality in convex analysis (1/2)

- Dual definition of convex set: [Rockafeller, ’70]

- Any closed convex set $A$ can be defined as an intersection of affine half spaces that contain the set $A$. 
Duality in convex analysis (2/2)

- Dual definition of convex function: [Rockafeller, ’70]

- Any closed convex function can be defined as the pointwise supremum of collection of all affine functions $h$ majorized by $f$. 

\[ f(x) \]

\[ h(x) = a_i^T x - v_i \]
Conjugate of a convex function (1/2)

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function.
Conjugate of a convex function (1/2)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function.

Let $F^*$ be the set of all tuples $(z, v)$ such that $h(x) = \langle x, z \rangle - v$ is majorized by $f(x)$, i.e.,

$$f(x) \geq \langle x, z \rangle - v$$

or equivalently,

$$v \geq \langle x, z \rangle - f(x)$$

for all $x \in \mathbb{R}^d$. 
Conjugate of a convex function (1/2)

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function.

- Let $F^*$ be the set of all tuples $(z, v)$ such that $h(x) = \langle x, z \rangle - v$ is majorized by $f(x)$, i.e.,

$$f(x) \geq \langle x, z \rangle - v$$

or equivalently,

$$v \geq \langle x, z \rangle - f(x)$$

for all $x \in \mathbb{R}^d$.

- Given $z$, if we choose $v \geq \sup_{x \in \mathbb{R}^d} \langle x, z \rangle - f(x)$, then

$$f(x) \geq h(x) \text{ for all } x \in \mathbb{R}^d.$$
Conjugate of a convex function (1/2)

- Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex function.

- Let $F^*$ be the set of all tuples $(z, v)$ such that $h(x) = \langle x, z \rangle - v$ is majorized by $f(x)$, i.e.,

  $$f(x) \geq \langle x, z \rangle - v$$

  or equivalently,

  $$v \geq \langle x, z \rangle - f(x)$$

  for all $x \in \mathbb{R}^d$.

- Given $z$, if we choose $v \geq \sup_{x \in \mathbb{R}^d} \langle x, z \rangle - f(x)$, then

  $$f(x) \geq h(x)$$

  for all $x \in \mathbb{R}^d$.

- The convex function $f$ and the set $F^*$ convey the same information.
For convex function $f$, the set $F^*$ is the collection of tuples $(z, v)$ such that

$$v \geq \sup_x \langle x, z \rangle - f(x)$$
Conjugate of a convex function (2/2)

- For convex function $f$, the set $F^*$ is the collection of tuples $(z, v)$ such that

$$v \geq \sup_x \langle x, z \rangle - f(x)$$

- The set $F^*$ is also the epigraph of the convex function $f^*$

$$f^*(z) = \sup_x \langle x, z \rangle - f(x)$$
Conjugate of a convex function (2/2)

▶ For convex function $f$, the set $F^*$ is the collection of tuples $(z, v)$ such that

$$v \geq \sup_x \langle x, z \rangle - f(x)$$

▶ The set $F^*$ is also the epigraph of the convex function $f^*$

$$f^*(z) = \sup_x \langle x, z \rangle - f(x)$$

▶ The function $f^*$ is called the **dual** or **convex conjugate** of $f$. 
Properties of conjugate functions

- $f^*$ is also a convex function
- $(f^*)^* = f$
- $f(x) + f^*(y) \geq \langle x, y \rangle, \quad \forall x, y$
Properties of conjugate functions

- $f^*$ is also a convex function
- $(f^*)^* = f$
- $f(x) + f^*(y) \geq \langle x, y \rangle$, $\forall x, y$

In fact, the conjugate pair $f$ and $f^*$ are the best pair to satisfy the below inequality:

$$f(x) + g(y) \geq \langle x, y \rangle$$

Proof: We work out.
Fenchel’s inequality

- The convex conjugate pair $f$ and $f^*$ always satisfy:

$$f(x) + f^*(y) \geq \langle x, y \rangle \quad \forall x, y$$
Compression Lemma [McAllester, ’03]

Let $\mathcal{H}$ be a parameter space.

For any measurable function $\phi(h)$ on $\mathcal{H}$ and any distributions $P$ and $Q$ on $\mathcal{H}$, we have:

$$
\mathbb{E}_Q[\phi(h)] - \log \mathbb{E}_P[\exp \phi(h)] \leq KL(Q\|P)
$$

Further,

$$
\sup_{\phi} \left( \mathbb{E}_Q[\phi(h)] - \log \mathbb{E}_P[\exp \phi(h)] \right) = KL(Q\|P)
$$

Also known by following names:

1. Change of measure inequality
2. Donsker-Varadhan formula
Compression Lemma - Proof

\[ \mathbb{E}_Q [\phi(h)] \]
\[ = \mathbb{E}_Q \left[ \log \left( \frac{Q(h)}{P(h)} \exp(\phi(h)) \frac{P(h)}{Q(h)} \right) \right] \]
\[ = \mathbb{E}_Q \left[ \log \left( \frac{Q(h)}{P(h)} \right) \right] + \mathbb{E}_Q \left[ \log \left( \exp(\phi(h)) \frac{P(h)}{Q(h)} \right) \right] \]
\[ = KL(Q \| P) + \mathbb{E}_Q \left[ \log \left( \exp(\phi(h)) \frac{P(h)}{Q(h)} \right) \right] \]
\[ \leq \text{Jensen ineq.} \]
\[ KL(Q \| P) + \log \left( \mathbb{E}_Q \left[ \exp(\phi(h)) \frac{dP(h)}{dQ(h)} \right] \right) \]
\[ = KL(Q \| P) + \log \left( \mathbb{E}_P \left[ \exp(\phi(h)) \right] \right) \]
For any measurable function $\phi : \mathcal{H} \to \mathbb{R}$, define

$$f(\phi) = \log \mathbb{E}_P [\exp (\phi(h))]$$
Connection b/w Compression Lemma and Fenchel’s Inequality

- For any measurable function $\phi : \mathcal{H} \rightarrow \mathbb{R}$, define

$$f(\phi) = \log \mathbb{E}_P \left[ \exp (\phi(h)) \right]$$

- $f$ is convex with respect to $\phi$

- Choose $\phi^*$ to be the probability density corresponding to a distribution $Q$ on $\mathcal{H}$ so that $\langle \phi, \phi^* \rangle = \mathbb{E}_{h \sim Q}[\phi(h)]$

- The conjugate of $f$ is:

$$f^*(\phi^*) = \sup \phi (\langle \phi, \phi^* \rangle - f(\phi)) = \sup \phi \left( \mathbb{E}_Q[\phi(h)] - \log \mathbb{E}_P \left[ \exp (\phi(h)) \right] \right) = \text{KL}(Q || P)$$
Connection b/w Compression Lemma and Fenchel’s Inequality

- For any measurable function $\phi : \mathcal{H} \to \mathbb{R}$, define
  \[
  f(\phi) = \log \mathbb{E}_P [\exp (\phi(h))]
  \]

- $f$ is convex with respect to $\phi$

- Choose $\phi^*$ to be the probability density corresponding to a distribution $Q$ on $\mathcal{H}$ so that
  \[
  \langle \phi, \phi^* \rangle = \mathbb{E}_{h \sim Q} [\phi(h)]
  \]
Connection b/w Compression Lemma and Fenchel’s Inequality

▶ For any measurable function $\phi : \mathcal{H} \to \mathbb{R}$, define

$$f(\phi) = \log \mathbb{E}_P[\exp(\phi(h))]$$

▶ $f$ is convex with respect to $\phi$

▶ Choose $\phi^*$ to be the probability density corresponding to a distribution $Q$ on $\mathcal{H}$ so that

$$\langle \phi, \phi^* \rangle = \mathbb{E}_{h \sim Q}[\phi(h)]$$

▶ The conjugate of $f$ is:

$$f^*(\phi^*) = \sup_{\phi} (\langle \phi, \phi^* \rangle - f(\phi))$$

$$= \sup_{\phi} (\mathbb{E}_Q[\phi(h)] - \log \mathbb{E}_P[\exp(\phi(h))])$$

$$= KL(Q||P)$$
PAC-Bayesian Bound

- With probability at least \((1 - \delta)\) over the choice of \(S \sim D^m\),

\[
\text{kl} \left( I(Q, S) \mid\mid I(Q, D) \right) \leq \frac{KL(Q \mid\mid P) + \log \frac{m+1}{\delta}}{m}
\]

- Can be derived as a special case of Compression Lemma.
PAC-Bayesian Bound - derivation (1/4)

- From compression lemma, for any measurable function $\phi(h)$, we have

$$\mathbb{E}_Q[\phi(h)] \leq KL(Q||P) + \log(\mathbb{E}_P[\exp(\phi(h))])$$

- Let $\phi(h) \triangleq m.kl(l(h, S)||l(h, D))$, where $S$ is the sample distribution and $D$ is the true distribution. Then,

$$\mathbb{E}_Q[kl(l(h, S)||l(h, D))] \leq \frac{KL(Q||P) + \log(\mathbb{E}_P[\exp(m.kl(l(h, S)||l(h, D)))])}{m}$$

- We first fix the LHS.
Since relative entropy is jointly convex in both its arguments, by using Jensen’s inequality

\[ \text{KL}(I(Q, S) | | I(Q, D)) \leq \mathbb{E}_Q [\text{KL}(I(h, S) | | I(h, D))] \]

We next fix the RHS.
PAC-Bayesian Bound - derivation (3/4)

- We need to show that
\[ \mathbb{E}_P \left[ \exp \left( m \cdot k \cdot l(h, S) \| l(h, D) \right) \right] \leq \frac{m+1}{\delta} \text{ w.h.p.} \]
We need to show that
\[ \mathbb{E}_P \left[ \exp \left( m.kl (l(h, S)\|l(h, D))) \right) \right] \leq \frac{m+1}{\delta} \text{ w.h.p.} \]

From Markov's inequality:
\[
\mathbb{E}_P \left[ \exp \left( m.kl (l(h, S)\|l(h, D))) \right) \right] \\
\leq \frac{\mathbb{E}_{S\sim D^m} \mathbb{E}_P \left[ \exp \left( m.kl (l(h, S)\|l(h, D))) \right) \right]}{\delta} \\
\text{with probability at least } 1 - \delta.
\]

Next we will show that
\[ \mathbb{E}_{S\sim D^m} \mathbb{E}_P \left[ \exp \left( m.kl (l(h, S)\|l(h, D))) \right) \right] \leq m + 1. \]
Next we will show that
\[
\mathbb{E}_{S \sim D^m} \mathbb{E}_P \left[ \exp \left( m \cdot k l (l(h, S) \mid \mid l(h, D)) \right) \right] \leq m + 1
\]
PAC-Bayesian Bound - derivation (4/4)

Next we will show that
\[ \mathbb{E}_{S \sim D^m} \mathbb{E}_P [\exp (m.kl (l(h, S)||l(h, D)))] \leq m + 1 \]

Or equivalently, [Fubini’s theorem]
\[ \mathbb{E}_P \mathbb{E}_{S \sim D^m} [\exp (m.kl (l(h, S)||l(h, D)))] \leq m + 1 \]
Next we will show that
\[ \mathbb{E}_{S \sim D^m} \mathbb{E}_P [\exp (m \cdot kl (l(h, S) \| l(h, D)))] \leq m + 1 \]

Or equivalently, [Fubini’s theorem]
\[ \mathbb{E}_P \mathbb{E}_{S \sim D^m} [\exp (m \cdot kl (l(h, S) \| l(h, D)))] \leq m + 1 \]

Since \( m \cdot l(h, S) \) is binomial distributed with probability \( \pi = l(h, D) \), we have:
\[
\mathbb{E}_{S \sim D^m} [\exp (m \cdot kl (l(h, S) \| l(h, D)))]
= \sum_{s \sim \text{Binomial}(\pi, m)} p(s) \exp (m \cdot kl(l(h, s) \| \pi))
= \sum_{n=0}^{m} \binom{m}{n} \pi^n (1 - \pi)^{m-n} \exp \left( m \cdot kl \left( \frac{n}{m} \| \pi \right) \right)
= \sum_{n=0}^{m} \binom{m}{n} \exp \left( -m H \left( \frac{n}{m} \right) \right) \leq \sum_{n=1}^{m} 1 = m + 1
\]
References

▶ On Bayesian Bounds, Arindam Banerjee, ICML, 2006

▶ PAC Bayesian Analysis: Background and Applications, Yevgeny Seldin, John Shawe-Taylor, Francois Laviolette

▶ A note on PAC Bayesian Theorem, Andreas Maurer, 2006
Literature Survey (1/6)

- Bayesian Model Comparison and Backprop nets, David MacKay, ’91. Link between evidence and generalization hypothesized by David McKay.


- Some PAC-Bayesian Theorems, David McAllester, ’98. Introduces PAC Bayesian bounds which give PAC type guarantees for ”Bayesian” algorithms - algorithms that optimize risk expressions involving a prior probability on the concept (model) and a likelihood on the training data.
Literature Survey (2/6)

- PAC Bayesian Model Averaging, David McAllester, ’99. PAC Bayes generalization bounds for weighted mixture of concepts instead of stochastic concept.

- PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification, Mathias Seeger, ’02. PAC baysian bounds using small kl divergence with simplified proof. Also derives bounds specific to non-parametric Gaussian Process models used for classification.

- PAC-Bayesian stochastic model selection, David McAllester, ’03. Investigates the nature of the posterior distribution providing the best performance guarantee for stochastic model selection. It is shown that the optimal posterior is a Gibbs distribution. However, it is also shown that simpler posterior distributions are nearly optimal.
Literature Survey (3/6)


- On Bayesian Bounds, Arindam Banerjee, ’06. Simple derivation of PAC Bayesian bounds using Donsker-Varadhan inequality which in turn is rooted in Fenchel’s inequality from convex analysis.

- A Note on the PAC Bayesian Theorem, Andreas Maurer, ’06. Further tightening of Seeger bound specific to 0-1 loss.
▶ Catoni’s monograph: PAC Bayesian supervised classification, 2007. Introduces a special construction of prior and posterior which helps in bounding the KL divergence term in PAC bound.


▶ PAC-Bayesian Generalization Bound for Density Estimation with Application to Co-clustering, Yevgeny Seldin and Naftali Tishby, ’09. PAC Bayesian bounds for generalization error for discrete density estimation.
PAC Bayesian Learning of Linear Classifiers, Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario Marchand, ICML, 2009. A more general form of PAC Bayesian bound with second term upper bounded by incomplete binomial term which is $O(\sqrt{m})$.


PAC-Bayes Bounds with Data Dependent Priors, Emilio Parrado-Hernandez, Amiran Ambroladze and John-Shawe Taylor, 12 To learn the prior from a separate data set, or to consider an expectation prior that does not need this separate data set.

Distribution-Dependent PAC-Bayes Priors, Guy Lever, Francois Laviolette, and John Shawe-Taylor, 13. Insights on how to choose localized priors for tighter generalization error bounds.
Papers by Tong Zhang

- Information-Theoretic Upper and Lower Bounds for Statistical Estimation, 2006

- From $\epsilon$-entropy to KL-entropy: ANALYSIS OF MINIMUM INFORMATION COMPLEXITY DENSITY ESTIMATION, 2007