# 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
- $S = \{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_m, y_m)\}$ - training data set (i.i.d.)

# Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_m, y_m)\}$ - 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
- $S = \{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $\mathcal{A} : S \to \mathcal{H}$ - algorithm

# Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $\mathcal{A} : S \rightarrow \mathcal{H}$ - algorithm
- $h(\mathbf{x})$ - prediction of hypothesis/classifier $h \in \mathcal{H}$ for input sample $\mathbf{x}$

# Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(\mathbf{x}_1, y_1), \dots (\mathbf{x}_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $\mathcal{A} : S \rightarrow \mathcal{H}$ - algorithm
- $h(\mathbf{x})$ - prediction of hypothesis/classifier $h \in \mathcal{H}$ for input sample $\mathbf{x}$
- $l(h, \mathbf{x})$ - instantaneous loss/risk of $h$ on $\mathbf{x}$

# Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $\mathcal{A} : S \to \mathcal{H}$ - algorithm
- $h(\mathbf{x})$ - prediction of hypothesis/classifier $h \in \mathcal{H}$ for input sample $\mathbf{x}$
- $l(h, \mathbf{x})$ - instantaneous loss/risk of $h$ on $\mathbf{x}$
- $l(h, W)$ - expected loss of hypothesis $h$ on entire $\mathcal{X}$, assuming input distribution to be $W$

# Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $\mathcal{A} : S \rightarrow \mathcal{H}$ - algorithm
- $h(\mathbf{x})$ - prediction of hypothesis/classifier $h \in \mathcal{H}$ for input sample $\mathbf{x}$
- $l(h, \mathbf{x})$ - instantaneous loss/risk of $h$ on $\mathbf{x}$
- $l(h, W)$ - expected loss of hypothesis $h$ on entire $\mathcal{X}$, assuming input distribution to be $W$
- $D$ - true but unknown distribution on $\mathcal{X}$

# Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $\mathcal{A} : S \to \mathcal{H}$ - algorithm
- $h(\mathbf{x})$ - prediction of hypothesis/classifier $h \in \mathcal{H}$ for input sample $\mathbf{x}$
- $l(h, \mathbf{x})$ - instantaneous loss/risk of $h$ on $\mathbf{x}$
- $l(h, W)$ - expected loss of hypothesis $h$ on entire $\mathcal{X}$, assuming input distribution to be $W$
- $D$ - true but unknown distribution on $\mathcal{X}$
- $l(h, D) = \mathbb{E}_{\mathbf{x} \sim D}[l(h, \mathbf{x})]$ - expected loss of hypothesis $h$

# Supervised learning - some definitions

- $\mathcal{X}$ - sample space
- $\mathcal{Y}$ - label space
- $m$ - number of training samples
- $S = \{(\mathbf{x}_1, y_1), \dots (\mathbf{x}_m, y_m)\}$ - training data set (i.i.d.)
- $\mathcal{H}$ - hypothesis space
- $\mathcal{A} : S \to \mathcal{H}$ - algorithm
- $h(\mathbf{x})$ - prediction of hypothesis/classifier $h \in \mathcal{H}$ for input sample $\mathbf{x}$
- $l(h, \mathbf{x})$ - instantaneous loss/risk of $h$ on $\mathbf{x}$
- $l(h, W)$ - expected loss of hypothesis $h$ on entire $\mathcal{X}$, assuming input distribution to be $W$
- $D$ - true but unknown distribution on $\mathcal{X}$
- $l(h, D) = \mathbb{E}_{\mathbf{x} \sim D}[l(h, \mathbf{x})]$ - expected loss of hypothesis $h$
- $l(h, S) = \frac{1}{m} \sum_{i=1}^{m} l(h, \mathbf{x}_i)$ - 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

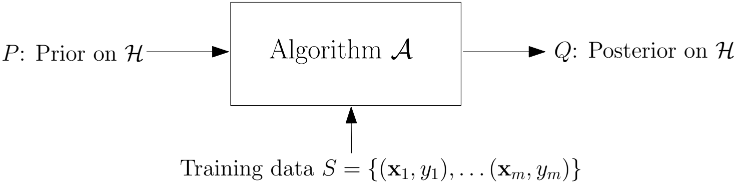
- Start with a prior $P$ on the hypothesis space $\mathcal{H}$.
- After observing $S$, the algorithm $A$ generates a posterior $Q$ on $\mathcal{H}$
- In PAC-Bayes, the classifier is random/stochastic in nature (Gibbs classifier)
  1. For given input $\mathbf{x} \in \mathcal{X}$, draw $h$ from $\mathcal{H}$ acc. to $Q$.
  2. Assign label $y = h(\mathbf{x})$

# 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 $\mathbf{x} \in \mathcal{X}$, draw $h$ from $\mathcal{H}$ acc. to $Q$.
  2. Assign label $y = h(\mathbf{x})$
- 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:




- 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 $\mathcal{A}$.

$$l(Q, S) = \mathbb{E}_Q \left[ l(h, S) \right], \quad \text{where } l(h, S) = \frac{1}{m} \sum_{i=1}^{m} l(h, \mathbf{x}_i)$$

- **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]

$$\left|\mathbb{E}_Q\left[l(h,S)\right] - \mathbb{E}_Q\left[l(h,D)\right]\right|^2 \;\leq\; ??$$

# PAC-Bayesian bounds - different flavors

- Mc Allester bound ['98]

$$\left|\mathbb{E}_Q\left[l(h,S)\right] - \mathbb{E}_Q\left[l(h,D)\right]\right|^2 \;\leq\; ??$$

- Seeger bound ['02]

$$kl\left(\mathbb{E}_Q\left[l(h,S)\right] \;||\; \mathbb{E}_Q\left[l(h,D)\right]\right) \;\leq\; ??$$

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 $S \sim D^m$,

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

# Intuition behind the bound (1/2)

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

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

# Intuition behind the bound (1/2)

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

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

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

# Intuition behind the bound (1/2)

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

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

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

- 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 $S \sim D^m$,

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

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

- 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 $S \sim D^m$,

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

- 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

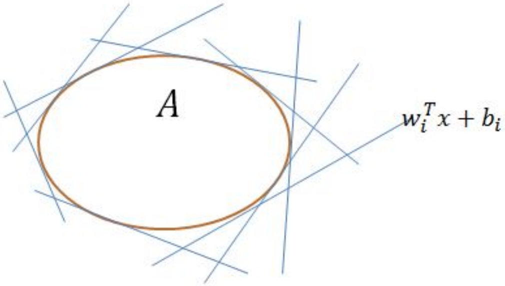
- PAC Bayesian bound:

$$kl\left(l(Q,S)||l(Q,D)\right) \leq \frac{KL(Q||P) + \log\frac{m+1}{\delta}}{m} \qquad \text{w.h.p.}$$

- **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)

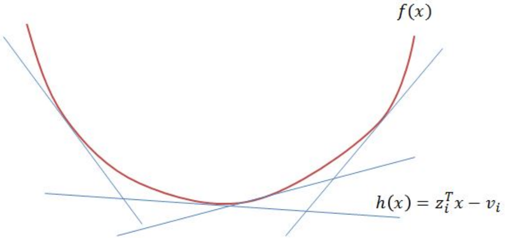
- Dual definition of convex set: [Rockafeller, '70]




- 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** $f$**.**

# 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 $(\mathbf{z}, v)$ such that $h(\mathbf{x}) = \langle \mathbf{x}, \mathbf{z} \rangle - v$ is majorized by $f(\mathbf{x})$, i.e.,

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

  or equivalently,

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

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

# 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 $(\mathbf{z}, v)$ such that $h(\mathbf{x}) = \langle \mathbf{x}, \mathbf{z} \rangle - v$ is majorized by $f(\mathbf{x})$, i.e.,

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

  or equivalently,

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

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

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

# 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 $(\mathbf{z}, v)$ such that $h(\mathbf{x}) = \langle \mathbf{x}, \mathbf{z} \rangle - v$ is majorized by $f(\mathbf{x})$, i.e.,

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

  or equivalently,

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

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

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

- The convex function $f$ and the set $F^*$ convey the same information.

# Conjugate of a convex function (2/2)

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

$$v \geq \sup_{\mathbf{x}} \langle \mathbf{x}, \mathbf{z} \rangle - f(\mathbf{x})$$

# Conjugate of a convex function (2/2)

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

$$v \geq \sup_{\mathbf{x}} \langle \mathbf{x}, \mathbf{z} \rangle - f(\mathbf{x})$$

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

$$f^*(\mathbf{z}) = \sup_{\mathbf{x}} \langle \mathbf{x}, \mathbf{z} \rangle - f(\mathbf{x})$$

# Conjugate of a convex function (2/2)

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

$$v \geq \sup_{\mathbf{x}} \langle \mathbf{x}, \mathbf{z} \rangle - f(\mathbf{x})$$

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

$$f^*(\mathbf{z}) = \sup_{\mathbf{x}} \langle \mathbf{x}, \mathbf{z} \rangle - f(\mathbf{x})$$

- The function $f^*$ is called the **dual** or **convex conjuate** of $f$.

# Properties of conjugate functions

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

# Properties of conjugate functions

- $f^*$ is also a convex function
- $(f^*)^* = f$
- $f(\mathbf{x}) + f^*(\mathbf{y}) \geq \langle \mathbf{x}, \mathbf{y} \rangle, \quad \forall \mathbf{x}, \mathbf{y}$
- In fact, the conjugate pair $f$ and $f^*$ are the best pair to satisfy the below inequality:

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

**Proof:** We work out.

# Fenchel's inequality

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

$$f(\mathbf{x}) + f^*(\mathbf{y}) \geq \langle \mathbf{x}, \mathbf{y} \rangle \quad \forall \mathbf{x}, \mathbf{y}$$

# Compression Lemma [McAllester, '03]

- Let $\mathcal{H}$ be a parameter space.
- For any measureable 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

$$
\begin{aligned}
&\mathbb{E}_Q\left[\phi(h)\right] \\
&\quad = \mathbb{E}_Q\left[\log\left(\frac{Q(h)}{P(h)}\exp(\phi(h))\frac{P(h)}{Q(h)}\right)\right] \\
&\quad = \mathbb{E}_Q\left[\log\left(\frac{Q(h)}{P(h)}\right)\right] + \mathbb{E}_Q\left[\log\left(\exp(\phi(h))\frac{P(h)}{Q(h)}\right)\right] \\
&\quad = KL(Q||P) + \mathbb{E}_Q\left[\log\left(\exp(\phi(h))\frac{P(h)}{Q(h)}\right)\right] \\
&\quad \underset{\text{Jensen ineq.}}{\leq} KL(Q||P) + \log\left(\mathbb{E}_Q\left[\exp(\phi(h))\frac{dP(h)}{dQ(h)}\right]\right) \\
&\quad = KL(Q||P) + \log\left(\mathbb{E}_P\left[\exp(\phi(h))\right]\right)
\end{aligned}
$$

# Connection b/w Compression Lemma and Fenchel's Inequality

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

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

# Connection b/w Compression Lemma and Fenchel's Inequality

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

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

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

# Connection b/w Compression Lemma and Fenchel's Inequality

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

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

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

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

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

# Connection b/w Compression Lemma and Fenchel's Inequality

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

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

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

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

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

- The conjugate of $f$ is:

$$\begin{aligned} f^*(\phi^*) &= \sup_{\phi} \left(\langle \phi, \phi^* \rangle - f(\phi)\right) \\ &= \sup_{\phi} \left(\mathbb{E}_Q\left[\phi(h)\right] - \log \mathbb{E}_P\left[\exp\left(\phi(h)\right)\right]\right) \\ &= KL(Q||P) \end{aligned}$$

# PAC-Bayesian Bound

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

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

- 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

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

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

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

- We first fix the LHS.

# PAC-Bayesian Bound - derivation (2/4)

- Since relative entropy is jointly convex in both its arguments, by using Jensen's inequality

$$kl\left(l(Q,S)||l(Q,D)\right) \leq \mathbb{E}_Q\left[kl\left(l(h,S)||l(h,D)\right)\right]$$

- We next fix the RHS.

# PAC-Bayesian Bound - derivation (3/4)

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

# PAC-Bayesian Bound - derivation (3/4)

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

- From Markov's inequality:

$$\mathbb{E}_P \left[\exp\left(m.kl\left(l(h,S)||l(h,D)\right)\right)\right] \leq \frac{\mathbb{E}_{S\sim D^m}\mathbb{E}_P \left[\exp\left(m.kl\left(l(h,S)||l(h,D)\right)\right)\right]}{\delta}$$

  with probability at least $1-\delta$.

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

# PAC-Bayesian Bound - derivation (4/4)

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

# PAC-Bayesian Bound - derivation (4/4)

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

# PAC-Bayesian Bound - derivation (4/4)

- Next we will show that
  $\mathbb{E}_{S\sim D^m}\mathbb{E}_P\left[\exp\left(m.kl\left(l(h,S)||l(h,D)\right)\right)\right] \leq m+1$
- Or equivalently, [Fubini's theorem]
  $\mathbb{E}_P\mathbb{E}_{S\sim D^m}\left[\exp\left(m.kl\left(l(h,S)||l(h,D)\right)\right)\right] \leq m+1$
- Since $m.l(h,S)$ is binomial distributed with probability $\pi = l(h,D)$, we have:

$$
\begin{aligned}
&\mathbb{E}_{S\sim D^m}\left[\exp\left(m.kl\left(l(h,S)||l(h,D)\right)\right)\right] \\
&= \sum_{s\sim Binomial(\pi,m)} p(s)\ \exp\left(m.kl(l(h,s)||\pi)\right) \\
&= \sum_{n=0}^{m} \binom{m}{n} \pi^n (1-\pi)^{m-n} \exp\left(m.kl\left(\frac{n}{m}||\pi\right)\right) \\
&= \sum_{n=0}^{m} \binom{m}{n} \exp\left(-mH\left(\frac{n}{m}\right)\right) \leq \sum_{n=1}^{m} 1 = m+1
\end{aligned}
$$

# References


- On Bayesian Bounds, Arindam Banerjee, ICML, 2006
- PAC Bayesian Analysis: Background and Applications, Yevgeny Seldin, John Shawe-Taylor, Francois Laviolette