Computational and Statistical Learning Theory

TTIC 31120

Prof. Nati Srebro

Lecture 4:

MDL and PAC-Bayes

Uniform vs Non-Uniform Bias

- No Free Lunch: we need some "inductive bias"

- Limiting attention to hypothesis class $\mathcal{H}$: "flat" bias
  - $p(h) = \frac{1}{|\mathcal{H}|}$ for $h \in \mathcal{H}$, and $p(h) = 0$ otherwise

- Non-uniform bias: $p(h)$ encodes bias
  - Can use any $p(h) \geq 0$, s.t. $\sum_h p(h) \leq 1$
  - E.g. choose prefix-disambiguous encoding $d(h)$ and use $p(h) = 2^{-|d(h)|}$
  - Or, choose $c: \mathcal{U} \rightarrow \mathcal{Y}^{\mathcal{X}}$ over prefix-disambiguous programs $\mathcal{U} \subset \{0,1\}^*$ and use $p(h) = 2^{-\min_{c(\sigma)=h} |\sigma|}$
  - Choice of $p(\cdot)$, $d(\cdot)$ or $c(\cdot)$ encodes are expert knowledge/inductive bias

Minimum Description Length Learning

- Choose "prior" $p(h)$ s.t. $\sum_h p(h) \leq 1$ (or description language $d(\cdot)$ or $c(\cdot)$)

- Minimum Description Length learning rule
(based on above prior/description language):

$$MDL_p(S) = \arg\max_{L_S(h)=0} p(h) = \arg\min_{L_S(h)=0} |d(h)|$$

- For any $\mathcal{D}$, w.p. $\geq 1-\delta$,

$$L\left(MDL_p(S)\right) \leq \inf_{h \ s.t. L_{\mathcal{D}}(h)=0} \sqrt{\frac{-\log p(h) + \log 2/\delta}{2m}}$$

Sample complexity: $m = O\left(\frac{-\log p(h)}{\epsilon^2}\right) = O\left(\frac{|d(h)|}{\epsilon^2}\right)$

(more careful analysis: $O\left(\frac{|d(h^*)|}{\epsilon}\right)$)

MDL and Universal Learning

- **Theorem:** For any $\mathcal{H}$ and $p: \mathcal{H} \to [0,1]$, s.t. $\sum_h p(h) \leq 1$, and any source distribution $\mathcal{D}$, if there exists h with $L(h) = 0$ and $p(h) > 0$, then w.p. $\geq 1 - \delta$ over $S \sim \mathcal{D}^m$:

$$L\left(MDL_p(S)\right) \leq \sqrt{\frac{-\log p(h) + \log 2/\delta}{2m}}$$

- Can learn any countable class!
  - Class of all computable functions, with $p(h) = 2^{-\min_{c(\sigma)=h} |\sigma|}$.
  - Class enumerable with $n: \mathcal{H} \to \mathbb{N}$ with $p(h) = 2^{-n(h)}$
- But VCdim(all computable functions)=∞ !

- Why no contradiction to Fundamental Theorem?
  - PAC Learning: Sample complexity $m(\epsilon, \delta)$ is uniform for all $h \in \mathcal{H}$. Depends only on class $\mathcal{H}$, **not** on specific h^*
  - MDL: Sample complexity $m(\epsilon, \delta, \boldsymbol{h})$ depends on h.

Uniform and Non-Uniform Learnability

- **Definition**: A hypothesis class $\mathcal{H}$ is **agnostically PAC-Learnable** if there exists a learning rule A such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$, $\forall \mathcal{D}$, $\forall \boldsymbol{h}$, $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq L_{\mathcal{D}}(h) + \epsilon$$

- **Definition**: A hypothesis class $\mathcal{H}$ is **non-uniformly learnable** if there exists a learning rule A such that $\forall \epsilon, \delta > 0$, $\forall h$, $\exists m(\epsilon, \delta, \boldsymbol{h})$, $\forall \mathcal{D}$, $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta,\boldsymbol{h})}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq L_{\mathcal{D}}(h) + \epsilon$$

- **Theorem:** $\forall \mathcal{D}$, if there exists h with $L(h) = 0$, then $\forall^{\delta}_{S \sim \mathcal{D}^m}$
$$L\left(MDL_p(S)\right) \leq \sqrt{\frac{-\log p(h) + \log 2/\delta}{2m}}$$

Compete also with h s.t. $L(h) > 0$?

Allowing Errors: From MDL to SRM

$$L(h) \le \underbrace{L_S(h)}_{\text{Minimized by ERM}} + \underbrace{\sqrt{\frac{-\log p(h) + \log 2/\delta}{2m}}}_{\text{Minimized by MDL}}$$

- Structural Risk Minimization:

$$SRM_p(S) = \arg\min_h \underbrace{L_S(h)}_{\text{fit the data}} + \underbrace{\sqrt{\frac{-\log p(h)}{2m}}}_{\text{match the prior / simple / short description}}$$

- **Theorem:** For any prior $p(h)$, $\sum_h p(h) \le 1$, and any source distribution $\mathcal{D}$, w.p. $\ge 1 - \delta$ over $S \sim \mathcal{D}^m$:

$$L\left(SRM_p(S)\right) \le \inf_h \left(L(h) + 2\sqrt{\frac{-\log p(h) + \log 2/\delta}{m}} \right)$$

Non-Uniform Learning: Beyond Cardinality

- MDL still essentially based on cardinality ("how many hypothesis are simpler then me") and ignores relationship between predictors.
- Generalizes the cardinality bound: Using $p(h) = \frac{1}{|\mathcal{H}|}$ we get
$$m(\epsilon, \delta, h) = m(\epsilon, \delta) = \frac{\log|\mathcal{H}| + \log 2/\delta}{\epsilon^2}$$
- Can we treat continuous classes (e.g. linear predictors)?
Move from cardinality to "growth function"?
- E.g.:
  - $\mathcal{H} = \left\{ sign\left(f\big(\phi(x)\big)\right) \mid f: \mathbb{R}^d \to \mathbb{R} \text{ is a polynomial} \right\},\ \phi: \mathcal{X} \to \mathbb{R}^d$
  - VCdim($\mathcal{H}$)=∞
  - $\mathcal{H}$ is uncountable, and there is no distribution with $\forall_{h \in \mathcal{H}} p(h) > 0$
  - But what if we bias toward lower order polynomials?

- **Answer 1: prior over hypothesis classes**
  - Write $\mathcal{H} = \cup\, \mathcal{H}_r$ (e.g. $\mathcal{H}_r$ =degree-r polynomials)
  - Use prior $p(H_r)$ over hypothesis classes

Prior Over Hypothesis Classes

- VC bound: $\forall_r \mathbb{P}\left[\forall_{h\in\mathcal{H}_r} L(h) \leq L_S(h) + O\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}_r)+\log{}^1/_{\delta_r}}{m}}\right)\right] \geq 1-\delta_r$
- Setting $\delta_r = p(\mathcal{H}_r)\cdot\delta$ and taking a union bound,

$$\forall^{\delta}_{S\sim\mathcal{D}^m}\forall_{\mathcal{H}_r}\forall_{h\in\mathcal{H}_r} L(h) \leq L_S(h) + O\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}_r) - \log p(\mathcal{H}_r) + \log{}^1/_\delta}{m}}\right)$$

- Structural Risk Minimization over hypothesis classes:

$$SRM_p(S) = \arg\min_{h\in\mathcal{H}_r} L_S(h) + C\sqrt{\frac{-\log p(\mathcal{H}_r) + \mathrm{VCdim}(\mathcal{H}_r)}{m}}$$

- Theorem: w.p. $\geq 1-\delta$,

$$L_\mathcal{D}\left(SRM_p(S)\right) \leq \min_{\mathcal{H}_r, h\in\mathcal{H}_r} L_\mathcal{D}(h) + O\left(\sqrt{\frac{-\log p(\mathcal{H}_r) + \mathrm{VCdim}(\mathcal{H}_r) + \log\frac{1}{\delta}}{m}}\right)$$

Structural Risk Minimization

- Theorem: For a prior $p(\mathcal{H}_r)$ with $\sum_{\mathcal{H}_r} p(\mathcal{H}_r) \leq 1$ and any $\mathcal{D}$, $\forall^{\delta}_{S \sim \mathcal{D}^m}$,

$$L_{\mathcal{D}}\left(SRM_p(S)\right) \leq \min_{\mathcal{H}_r, h \in \mathcal{H}_r} L_{\mathcal{D}}(h) + O\left(\sqrt{\frac{-\log p(\mathcal{H}_r) + \mathrm{VCdim}(\mathcal{H}_r) + \log \frac{1}{\delta}}{m}}\right)$$

- For $\mathcal{H}_i = \{h_i\}$:
  - $VCdim(\mathcal{H}_r) = 0$
  - Reduces to "standard" SRM with a prior over hypothesis
- For $p(\mathcal{H}_r) = 1$
  - Reduces to ERM over a finite-VC class
- More general. Eg for polynomials over $\phi(x) \in \mathbb{R}^d$ with $p(\text{degree } r) = 2^{-r}$,

$$m(\epsilon, \delta, h) = O\left(\frac{\text{degree}(h) + (d+1)^{\text{degree}(h)} + \log \frac{1}{\delta}}{\epsilon^2}\right)$$

- Allows non-uniform learning of a countable union of finite-VC classes

Uniform and Non-Uniform Learnability

- **Definition**: A hypothesis class $\mathcal{H}$ is **agnostically PAC-Learnable** if there exists a learning rule A such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$, $\forall \mathcal{D}$, $\forall \boldsymbol{h}$, $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq L_{\mathcal{D}}(h) + \epsilon$$

- **Definition**: A hypothesis class $\mathcal{H}$ is **non-uniformly learnable** if there exists a learning rule A such that $\forall \epsilon, \delta > 0$, $\forall h$, $\exists m(\epsilon, \delta, \boldsymbol{h})$, $\forall \mathcal{D}$, $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta,\boldsymbol{h})}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq L_{\mathcal{D}}(h) + \epsilon$$

- **Theorem**: A hypothesis class $\mathcal{H}$ is non-uniformly learnable
it is a countable union of finite VC class ($\mathcal{H} = \cup_{i \in \mathbb{N}} \mathcal{H}_i$, $\mathrm{VCdim}(\mathcal{H}_i) < \infty$)

- **Definition**: A hypothesis class $\mathcal{H}$ is **"consistently learnable"** if there exists a learning rule A such that $\forall \epsilon, \delta > 0$, $\forall h \; \forall \boldsymbol{\mathcal{D}}$, $\exists m(\epsilon, \delta, h, \boldsymbol{\mathcal{D}})$, $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta,h,\mathcal{D})}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq L_{\mathcal{D}}(h) + \epsilon$$

Consistency

- $\mathcal{X}$ countable (e.g. $\mathcal{X} = \{0,1\}^*$), $\mathcal{H} = \{\pm 1\}^{\mathcal{X}}$ (all possible functions)
- $\mathcal{H}$ is uncountable, it is *not* a countable union of finite VC classes, and is thus *not* non-uniformly learnable

- Claim: $\mathcal{H}$ is "consistently learnable" using
$$ERM_{\mathcal{H}}(S)(x) = MAJORITY(y_i \ s.t. (x_i, y_i) \in S)$$

- Proof sketch: for any $\mathcal{D}$,
  - Sort $\mathcal{X}$ by decreasing probability. The tail has diminishing probability and thus for any ϵ, there exists some prefix $\mathcal{X}'$ of the sort s.t. the tail $\mathcal{X} \setminus \mathcal{X}'$ has probability mass $\leq \epsilon/2$.
  - We'll give up on the tail. $\mathcal{X}'$ is finite, and so $\{\pm 1\}^{\mathcal{H}}$ is also finite.
- Why only "consistently learnable"?
  - Size of $\mathcal{X}$' required to capture $1 - \epsilon/2$ of mass depends on $\mathcal{D}$.

Uniform and Non-Uniform Learnability

- **Definition**: A hypothesis class $\mathcal{H}$ is **agnostically PAC-Learnable** if there exists a learning rule A such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$, $\forall \mathcal{D}$, $\forall \boldsymbol{h}$, $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta)}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq L_{\mathcal{D}}(h) + \epsilon$$
- (Agnostically) PAC-Learnable iff $\mathrm{VCdim}(\mathcal{H}) < \infty$

- **Definition**: A hypothesis class $\mathcal{H}$ is **non-uniformly learnable** if there exists a learning rule A such that $\forall \epsilon, \delta > 0$, $\forall h$, $\exists m(\epsilon, \delta, \boldsymbol{h})$, $\forall \mathcal{D}$, $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta,\boldsymbol{h})}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq L_{\mathcal{D}}(h) + \epsilon$$
- Non-uniformly learnable iff $\mathcal{H}$ is a countable union of finite VC classes

- **Definition**: A hypothesis class $\mathcal{H}$ is **"consistently learnable"** if there exists a learning rule A such that $\forall \epsilon, \delta > 0$, $\forall h$ $\forall \boldsymbol{\mathcal{D}}$, $\exists m(\epsilon, \delta, h, \boldsymbol{\mathcal{D}})$, $\forall^{\delta}_{S \sim \mathcal{D}^{m(\epsilon,\delta,h,\mathcal{D})}}$,
$$L_{\mathcal{D}}\big(A(S)\big) \leq L_{\mathcal{D}}(h) + \epsilon$$

SRM In Practice

$$SRM_p(S) = \arg\min_{h\in\mathcal{H}_r} L_S(h) + C\sqrt{\frac{-\log p(\mathcal{H}_r) + \mathrm{VCdim}(\mathcal{H}_r)}{m}}$$

- Bound is loose anyway. Better to view as bi-criteria optimization:

$$\arg\min L_s(h) \quad \textbf{and} \quad \left(-\log p(\mathcal{H}_r) + \mathrm{VCdim}(\mathcal{H}_r)\right)$$

  E.g. serialize as

$$\arg\min L_s(h) + \lambda\left(-\log p(\mathcal{H}_r) + \mathrm{VCdim}(\mathcal{H}_r)\right)$$

- Typically $-\log p(\mathcal{H}_r)$, $\mathrm{VCdim}(\mathcal{H}_r)$ monotone in “complexity” r

$$\arg\min L_s(h) \quad \textbf{and} \quad r(h)$$

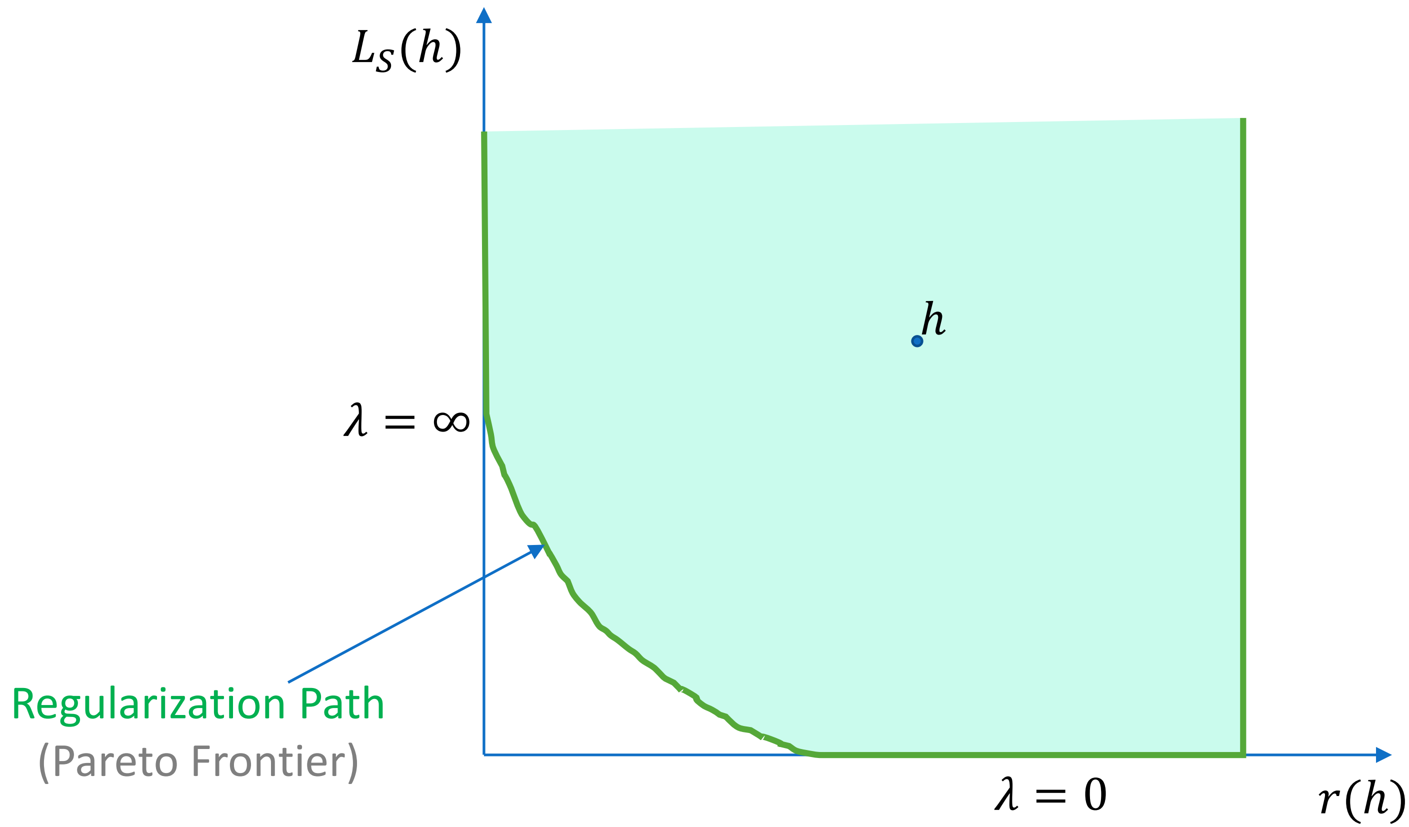
  where

$$r(h) = \min r \;\; s.t.\; h \in \mathcal{H}_r$$

SRM as a Bi-Criteria Problem

$$\arg\min L_s(h) \quad \textbf{and} \quad r(h)$$

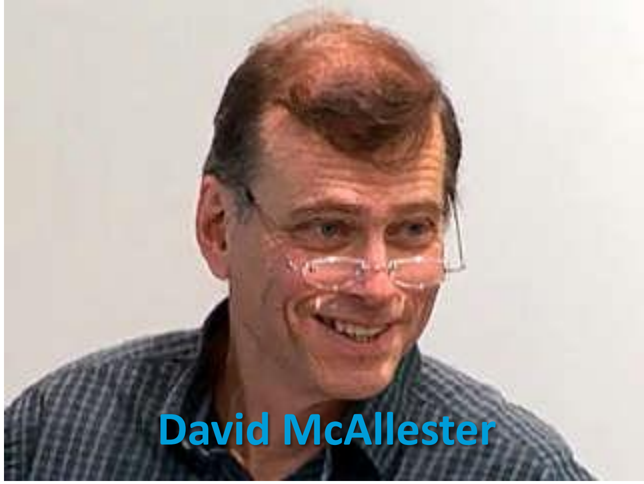
Regularization Path = $\{\arg\min_h L_S(h) + \lambda \cdot r(h) \;\;|\;\; 0 \leq \lambda \leq \infty\}$



Select λ using a validation set—exact bound not needed

Non-Uniform Learning: Beyond Cardinality

- MDL still essentially based on cardinality ("how many hypothesis are simpler then me") and ignores relationship between predictors.
- Can we treat continuous classes (e.g. linear predictors)?
  Move from cardinality?
  Take into account that many predictors are similar?
- **Answer 1: prior $p(\mathcal{H})$ over hypothesis class**
- **Answer 2: PAC-Bayes Theory**
  - Prior distribution P (not necessarily discrete) over $\mathcal{H}$



PAC-Bayes

- Until now (MDL, SRM) we used a discrete "prior" (discrete "distribution" $p(h)$ over hypothesis, or discrete "distribution" $p(\mathcal{H}_r)$ over hypothesis classes)
- Instead: encode inductive bias as distribution P over hypothesis

- Use randomized (averaged) predictor h_Q, where for each prediction chooses $h \sim Q$ and predicts $h(x)$
  - $h_Q(x) = y \text{ w.p. } \mathbb{P}_{h\sim Q}(h(x) = y)$
  - $L_{\mathcal{D}}(h_Q) = \mathbb{E}_{h\sim Q}[L_{\mathcal{D}}(h)]$

- **Theorem**: for any distribution P over hypothesis and any $\mathcal{D}$, $\forall^{\delta}_{S\sim\mathcal{D}^m}$

$$\left|L_{\mathcal{D}}(h_Q) - L_S(h_Q)\right| \leq \sqrt{\frac{KL(Q||P) + \log {}^{2m}/_{\delta}}{2(m-1)}}$$

KL-Divergence

$$KL(Q||P) = \mathbb{E}_{h\sim Q}\left[\log\frac{dQ}{dP}\right]$$

$$= \sum_h q(h)\log\frac{q(h)}{p(h)}$$ for discrete dist with pmf p, q

$$= \int f_Q(h)\log\frac{f_Q(h)}{f_P(h)}dh$$ for continuous distributions

- Measures how much Q deviates from Q
- $KL(Q||P) \geq 0$, and$KL(Q||P) = 0$ if and only if $Q = P$
- If $Q(A) > 0$ while $P(A) = 0$, $KL(Q||P) = \infty$ (other direction is allowed)
- $KL(H_1||H_0)$=information per sample for rejecting H_0 when H_1 is true
- $KL\big(Q||\mathrm{Unif}(n)\big) = \log n - H(Q)$
- $I(X,Y) = KL\big(P(X,Y)||P(X)P(Y)\big)$

PAC-Bayes

- For any distribution $\mathcal{P}$ over hypothesis and any $\mathcal{D}$, $\forall^{\delta}_{S\sim\mathcal{D}^m}$

$$\left|L_{\mathcal{D}}(h_Q) - L_S(h_Q)\right| \leq \sqrt{\frac{KL(Q||P) + \log {}^{2m}/_{\delta}}{2(m-1)}}$$

- Can only use hypothesis in the support of P (otherwise $KL(Q||P) = \infty$)
- For a finite $\mathcal{H}$ with $P = \mathrm{Unif}(\mathcal{H})$
  - Consider Q $=$ point mass on h
  - $KL(Q||P) = \log|\mathcal{H}|$
  - Generalizes cardinality bound (up to $\log m$)
- More generally, for a discrete P and Q $=$ point mass on h
  - $KL(Q||P) = \sum q(h) \log \frac{q(h)}{p(h)} = \frac{1}{p(h)}$
  - Generalizes MDL/SRM (up to $\log m$)
- For continuous P (eg over linear predictors or polynomials)
  - For Q=point-mass (or any discrete), $KL(Q||P) = \infty$
  - Take h_Q as average over similar hypothesis (eg with same behavior on S)

PAC-Bayes

$$L_{\mathcal{D}}(h_Q) \leq L_S(h_Q) + \sqrt{\frac{KL(Q||P) + \log {}^{2m}/_{\delta}}{2(m-1)}}$$

- What learning rule does the PAC-Bayes bound suggest?

$$Q_\lambda = \arg\min_Q L_S(h_Q) + \lambda \cdot KL(Q||P)$$

- **Theorem**:

$$q_\lambda(h) \propto p(h)e^{-\beta L_S(h)}$$

  for some "inverse temperature" β

- As $\lambda \to \infty$ we ignore the data, corresponding to infinite temperature, $\beta \to 0$
- As $\lambda \to 0$ we insist on minimizing $L_S(h_Q)$, corresponding to zero temperature, $\beta \to \infty$, and the prediction becomes ERM (or rather, a distribution over the ERM hypothesis in the support of P)

PAC-Bayes vs Bayes

Bayesian approach:

- Assume $h \sim \mathcal{P}$,
- $y_1, \ldots, y_m$ iid conditioned on h, with $y_i | x_i, h = \begin{cases} h(x_i), & w.p.\, 1-\nu \\ -h(x_i), & w.p.\, \nu \end{cases}$

Use posterior:

$$
\begin{aligned}
p(h|S) &\propto p(h)p(S|h) \\
&= p(h) \textstyle\prod_i p(x_i)p(y_i|x_i) \\
&\propto p(h) \textstyle\prod_i \left(\frac{\nu}{1-\nu}\right)^{[[h(x_i) \neq y_i]]} \\
&= p(h)e^{-\beta L_s(h)}
\end{aligned}
$$

where $\beta = m \log \frac{1-\nu}{\nu}$

PAC-Bayes vs Bayes

PAC-Bayes

- P encodes inductive bias, not assumption about reality

- SRM-type bound minimized by Gibbs distribution
$$q_\lambda(h) \propto p(h) e^{-\beta L_S(h)}$$

- Post-hoc guarantee always valid ($\forall_S^\delta$), with no assumption about reality
$$L_{\mathcal{D}}(h_Q) \leq L_S(h_Q) + \sqrt{\frac{KL(Q||P) + \log {}^{2m}/_{\delta}}{2(m-1)}}$$

- Bound valid for any Q

- If inductive bias very different from reality, bound will be high

Bayesian Approach

- $\mathcal{P}$ is prior over reality

- Posterior given by Gibbs distribution
$$q_\lambda(h) \propto p(h) e^{-\beta L_S(h)}$$

- Risk analysis assuming prior

PAC-Bayes: Tighter Version

- For any distribution P over hypothesis and any source distribution $\mathcal{D}$, $\forall^{\delta}_{S\sim\mathcal{D}^m}$

$$KL\left(L_S(h_Q)||L_{\mathcal{D}}(h_Q)\right) \leq \frac{KL(Q||P) + \log {}^{2m}/_{\delta}}{m-1}$$

where $KL(\alpha||\beta) = \alpha \log\frac{\alpha}{\beta} + (1-\alpha)\log\frac{1-\alpha}{1-\beta}$ for $\alpha, \beta \in [0,1]$

➔

$$L_{\mathcal{D}}(h_Q) \leq L_S(h_Q) + \sqrt{\frac{2L_s(h_Q)KL(QP) + \log\frac{2m}{\delta}}{m-1}} + \frac{2\left(KL(QP) + \log\frac{2m}{\delta}\right)}{m-1}$$

- This generalizes the realizable case ($L_S(h_Q) = 0$, and so only the $\frac{1}{m}$ term appears) and the agnostic case (where the $\sqrt{{}^1/_m}$ term is dominant)
- Numerically much tighter
- Can also be used as a tail bound instead of Hoeffding or Bernstein also with cardinality or VC-based guarantees. Arises naturally in PAC-Bayes.