Variations on the PAC-Bayesian Bound

followed by some links with the Bayesian theory

Pascal Germain

INRIA Paris (SIERRA Team)

Bayes in Paris
ENSAE

June 2, 2016

Plan

Plan

Definitions

Learning example

An example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is a **description-label** pair.

Data generating distribution

Each example is an **i.i.d. observation from distribution** D on $\mathcal{X} \times \mathcal{Y}$.

Learning sample

$$S \;=\; \{\, (x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n) \,\} \;\sim\; D^n$$

Predictors (or hypothesis)

$$h : \mathcal{X} \to \mathcal{Y}, \quad h \in \mathcal{H}$$

Learning algorithm

$$A(S) \longrightarrow h$$

Loss function

$$\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

Empirical loss

$$\widehat{\mathcal{L}}_S^{\ell}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, x_i, y_i)$$

Generalization loss

$$\mathcal{L}_D^{\ell}(h) = \mathop{\mathbf{E}}_{(x,y)\sim D} \ell(h, x_i, y_i)$$

PAC-Bayesian Theory

Initiated by McAllester (1999), the PAC-Bayesian theory gives **PAC** generalization guarantees to "**Bayesian** like" algorithms.

PAC guarantees (Probably Approximately Correct)

With probability at least "$1-\delta$", the loss of predictor h is less than "ε"

$$\Pr_{S\sim D^n}\left(\mathcal{L}_D^\ell(h) \leq \varepsilon(\widehat{\mathcal{L}}_S^\ell(h), n, \delta, \ldots) \right) \geq 1-\delta$$

Bayesian flavor

Given:

- A **prior** distribution P on $\mathcal{H}$.
- A **posterior** distribution Q on $\mathcal{H}$.

$$\Pr_{S\sim D^n}\left(\mathop{\mathbf{E}}_{h\sim Q} \mathcal{L}_D^\ell(h) \leq \varepsilon(\mathop{\mathbf{E}}_{h\sim Q} \widehat{\mathcal{L}}_S^\ell(h), n, \delta, P, \ldots) \right) \geq 1-\delta$$

A *Classical* PAC-Bayesian Theorem

PAC-Bayesian theorem (adapted from McAllester 1999, 2003)

For any distribution D on $\mathcal{X} \times \mathcal{Y}$, for any set of predictors $\mathcal{H}$, for any loss $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0,1]$, for any distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, we have,

$$\Pr_{S \sim D^n}\left(\forall\, Q \text{ on } \mathcal{H} : \ \mathop{\mathbf{E}}_{h\sim Q} \mathcal{L}_D^\ell(h) \leq \mathop{\mathbf{E}}_{h\sim Q} \widehat{\mathcal{L}}_S^\ell(h) + \sqrt{\frac{1}{2n}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{n}}{\delta}\right]} \right) \geq 1-\delta\,,$$

where $\mathrm{KL}(Q\|P) = \mathop{\mathbf{E}}_{h\sim Q} \ln\frac{Q(h)}{P(h)}$ is the **Kullback-Leibler divergence**.

Training bound

- Gives generalization guarantees **not based on testing sample**.

Valid *for all* posterior Q on $\mathcal{H}$

- Inspiration for conceiving **new learning algorithms**.

Plan

Plan

Majority Vote Classifiers

Consider a binary classification problem, where $\mathcal{Y} = \{-1, +1\}$ and the set $\mathcal{H}$ contains **binary voters** $h : \mathcal{X} \to \{-1, +1\}$

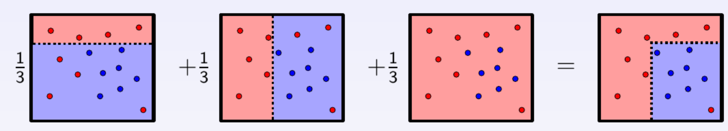
Weighted majority vote

To predict the label of $x \in \mathcal{X}$, the classifier asks for the *prevailing opinion*

$$B_Q(x) \;=\; \operatorname{sgn}\left(\mathop{\mathbf{E}}_{h \sim Q} h(x) \right)$$

Many learning algorithms output majority vote classifiers

AdaBoost, Random Forests, Bagging, ...



A Surrogate Loss

Majority vote risk

$$R_D(B_Q) = \Pr_{(x,y)\sim D}\Big(B_Q(x) \neq y\Big) = \mathop{\mathbf{E}}_{(x,y)\sim D} \mathrm{I}\Big[\mathop{\mathbf{E}}_{h\sim Q} y\cdot h(x) \leq 0\Big]$$

where $\mathrm{I}[a] = 1$ if predicate a is *true*; $\mathrm{I}[a] = 0$ otherwise.

Gibbs Risk

The stochastic Gibbs classifier $G_Q(x)$ draws $h' \in \mathcal{H}$ according to Q and output $h'(x)$.

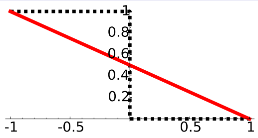
$$\begin{aligned} R_D(G_Q) &= \mathop{\mathbf{E}}_{(x,y)\sim D} \mathop{\mathbf{E}}_{h\sim Q} \mathrm{I}\Big[h(x) \neq y\Big] \\ &= \mathop{\mathbf{E}}_{h\sim Q} \mathcal{L}_D^{\ell_{01}}(h)\,, \end{aligned}$$

where $\ell_{01}(h,x,y) = \mathrm{I}\big[h(x) \neq y\big]$.

Factor two

It is well-known that

$$R_D(B_Q) \leq 2 \times R_D(G_Q)$$



$\mathop{\mathbf{E}}_{h\sim Q} y\cdot h(x)$

See Germain, Lacasse, Laviolette, Marchand, and Roy (2015, JMLR) for an extensive study.

Plan

A General PAC-Bayesian Theorem

Δ-function: «distance» between $\widehat{R}_S(G_Q)$ et $R_D(G_Q)$

Convex function $\Delta : [0,1] \times [0,1] \to \mathbb{R}$.

General theorem (Bégin et al. 2014, 2016; Germain 2015)

For any distribution D on $\mathcal{X} \times \mathcal{Y}$, for any set $\mathcal{H}$ of voters, for any distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, and for any Δ-function, we have, with probability at least $1-\delta$ over the choice of $S \sim D^n$,

$$\forall\, Q \text{ on } \mathcal{H} : \quad \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \;\leq\; \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{I}_\Delta(n)}{\delta}\right],$$

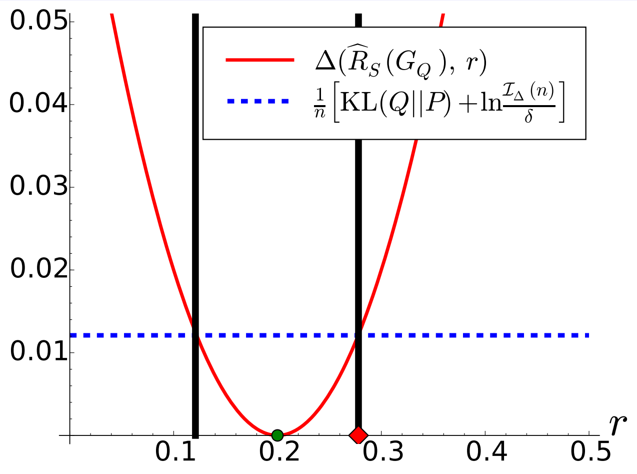
where

$$\mathcal{I}_\Delta(n) \;=\; \sup_{r\in[0,1]} \left[\, \sum_{k=0}^{n} \underbrace{\binom{n}{k} r^k (1-r)^{n-k}}_{\mathbf{Bin}\left(k;n,r\right)} e^{n\Delta(\frac{k}{n},\, r)} \right].$$

General theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\left(\widehat{R}_S(G_Q), R_D(G_Q)\right) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \geq 1-\delta\,.$$

Interpretation.



General theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \geq 1{-}\delta\,.$$

Proof ideas.

Change of Measure Inequality

For any P and Q on $\mathcal{H}$, and for any measurable function $\phi : \mathcal{H} \to \mathbb{R}$, we have

$$\mathop{\mathbf{E}}_{h\sim Q}\phi(h) \ \leq\ \mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{h\sim P} e^{\phi(h)}\right).$$

Markov's inequality

$$\Pr\left(X \geq a\right) \leq \frac{\mathbf{E}\,X}{a} \quad \Longleftrightarrow \quad \Pr\left(X \leq \frac{\mathbf{E}\,X}{\delta}\right) \geq 1{-}\delta\,.$$

Probability of observing k misclassifications among n examples

Given a voter h, consider a **binomial variable** of n trials with **success** $\mathcal{L}_D^{\ell_{01}}(h)$:

$$\begin{aligned}\Pr_{S\sim D^n}\left(\widehat{\mathcal{L}}_S^{\ell_{01}}(h) = \tfrac{k}{n}\right) &= \binom{n}{k}\left(\mathcal{L}_D^{\ell_{01}}(h)\right)^k\left(1-\mathcal{L}_D^{\ell_{01}}(h)\right)^{n-k} \\ &= \mathbf{Bin}\left(k; n, \mathcal{L}_D^{\ell_{01}}(h)\right)\end{aligned}$$

General theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H} :\ \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \le \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \ge 1-\delta\,.$$

Proof.

$$n\cdot\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S^{\ell}(h),\ \mathop{\mathbf{E}}_{h\sim Q}\mathcal{L}_D^{\ell}(h)\Big)$$

Jensen's Inequality $\le \mathop{\mathbf{E}}_{h\sim Q} n\cdot\Delta\Big(\widehat{\mathcal{L}}_S^{\ell}(h), \mathcal{L}_D^{\ell}(h)\Big)$

Change of measure $\le \mathrm{KL}(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} e^{n\Delta\left(\widehat{\mathcal{L}}_S^{\ell}(h), \mathcal{L}_D^{\ell}(h)\right)}$

Markov's Inequality $\le_{1-\delta} \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{S'\sim D^n}\mathop{\mathbf{E}}_{h\sim P} e^{n\cdot\Delta(\widehat{\mathcal{L}}_{S'}^{\ell}(h), \mathcal{L}_D^{\ell}(h))}$

Expectation swap $= \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\mathop{\mathbf{E}}_{S'\sim D^n} e^{n\cdot\Delta(\widehat{\mathcal{L}}_{S'}^{\ell}(h), \mathcal{L}_D^{\ell}(h))}$

Binomial law $= \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\sum_{k=0}^{n}\mathbf{Bin}\Big(k; n, \mathcal{L}_D^{\ell}(h)\Big)\, e^{n\cdot\Delta(\frac{k}{n}, \mathcal{L}_D^{\ell}(h))}$

Supremum over risk $\le \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\sup_{r\in[0,1]}\left[\sum_{k=0}^{n}\mathbf{Bin}\Big(k; n, r\Big)\, e^{n\Delta(\frac{k}{n}, r)}\right]$

$= \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathcal{I}_\Delta(n)\,.$ □

General theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\left(\widehat{R}_S(G_Q), R_D(G_Q)\right) \le \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \ \ge\ 1{-}\delta\,.$$

Corollary

[...] with probability at least $1{-}\delta$ *over the choice of* $S \sim D^n$*, for all* Q *on* $\mathcal{H}$ *:*

(a) $\mathrm{kl}\left(\widehat{R}_S(G_Q), R_D(G_Q)\right) \le \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{n}}{\delta}\right]$, *(Langford and Seeger 2001)*

(b) $R_D(G_Q) \le \widehat{R}_S(G_Q) + \sqrt{\frac{1}{2n}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{n}}{\delta}\right]}$, *(McAllester 1999, 2003)*

(c) $R_D(G_Q) \le \frac{1}{1-e^{-c}}\left(c\cdot\widehat{R}_S(G_Q) + \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\right]\right)$, *(Catoni 2007)*

(d) $R_D(G_Q) \le \widehat{R}_S(G_Q) + \frac{1}{\lambda}\left[\mathrm{KL}(Q\|P) + \ln\frac{1}{\delta} + f(\lambda, n)\right]$. *(Alquier et al. 2015)*

$$\begin{aligned}\mathrm{kl}(q,p) &= q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p} \ \ge\ 2(q-p)^2\,,\\ \Delta_c(q,p) &= -\ln[1-(1-e^{-c})\cdot p] - c\cdot q\,,\\ \Delta_\lambda(q,p) &= \frac{\lambda}{n}(p-q)\,.\end{aligned}$$

Plan

Transductive Learning

Assumption

Examples are drawn *without replacement* from a finite set Z of size N.

$$\begin{array}{lcllll} S & = & \{\,(x_1, y_1), & (x_2, y_2), & \ldots, & (x_n, y_n)\,\} \quad \subset Z \\ U & = & \{\,(x_{n+1}, \cdot), & (x_{n+2}, \cdot), & \ldots, & (x_N, \cdot)\,\} \quad = Z \setminus S \end{array}$$

Inductive learning: n draws with replacement according to $D \Rightarrow$ Binomial law.

Transductive learning: n draws without replacement in $Z \Rightarrow$ Hypergeometric law.

Theorem (Bégin et al. 2014)

For any set Z of N examples, [...] with probability at least $1{-}\delta$ over the choice of n examples among Z,

$$\forall\, Q \text{ on } \mathcal{H}: \quad \Delta\Big(\widehat{R}_S(G_Q), \widehat{R}_Z(G_Q)\Big) \le \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{T}_\Delta(n, N)}{\delta}\right],$$

where

$$\mathcal{T}_\Delta(n, N) = \max_{K=0\ldots N}\left[\sum_{k=\max[0,K+n-N]}^{\min[n,K]} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} e^{n\Delta\left(\frac{k}{n}, \frac{K}{N}\right)}\right].$$

Theorem

$$\Pr_{S\sim[Z]^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\Big(\widehat{R}_S(G_Q), \widehat{R}_Z(G_Q)\Big) \le \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{T}_\Delta(n,N)}{\delta}\right]\right) \ge 1-\delta\,.$$

Proof.

$$n\cdot\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}^{\ell}_S(h), \mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}^{\ell}_Z(h)\Big)$$

Jensen's inequality $\le \mathop{\mathbf{E}}_{h\sim Q} n\cdot\Delta\Big(\widehat{\mathcal{L}}^{\ell}_S(h), \widehat{\mathcal{L}}^{\ell}_Z(h)\Big)$

Change of measure $\le \mathrm{KL}(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} e^{n\Delta\left(\widehat{\mathcal{L}}^{\ell}_S(h), \widehat{\mathcal{L}}^{\ell}_Z(h)\right)}$

Markov's inequality $\le_{1-\delta} \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\mathop{\mathbf{E}}_{S'\sim[Z]^n}\mathop{\mathbf{E}}_{h\sim P} e^{n\cdot\Delta(\widehat{\mathcal{L}}^{\ell}_{S'}(h), \widehat{\mathcal{L}}^{\ell}_Z(h))}$

Expectations swap $= \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\mathop{\mathbf{E}}_{S'\sim[Z]^n} e^{n\cdot\Delta(\widehat{\mathcal{L}}^{\ell}_{S'}(h), \widehat{\mathcal{L}}^{\ell}_Z(h))}$

Hypergeometric law $= \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\sum_k \frac{\binom{N\cdot\widehat{\mathcal{L}}^{\ell}_Z(h)}{k}\binom{N-N\cdot\widehat{\mathcal{L}}^{\ell}_Z(h)}{n-k}}{\binom{N}{n}} e^{n\cdot\Delta(\frac{k}{n}, \widehat{\mathcal{L}}^{\ell}_Z(h))}$

Supremum over risk $\le \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\max_{K=0\ldots N}\left[\sum_k \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} e^{n\Delta(\frac{k}{n}, \frac{K}{N})}\right]$

$= \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\mathcal{T}_\Delta(n,N)\,.$ □

A New Transductive Bound for the Gibbs Risk

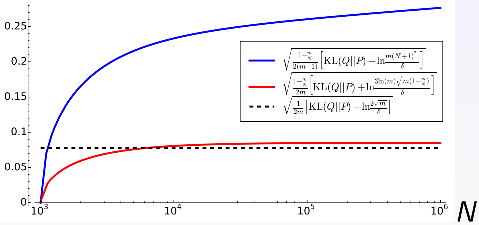
Corollary (Bégin et al. 2014)

[...] with probability at least $1-\delta$ over the choice of n examples among Z,

$$\forall\, Q \text{ on } \mathcal{H} : \widehat{R}_Z(G_Q) \le \widehat{R}_S(G_Q) + \sqrt{\frac{1-\frac{n}{N}}{2n}\left[\mathrm{KL}(Q\|P) + \ln \frac{3\ln(n)\sqrt{n(1-\frac{n}{N})}}{\delta}\right]}\,.$$

Theorem (Derbeko et al. 2004)

$$\forall\, Q \text{ on } \mathcal{H} : \widehat{R}_Z(G_Q) \le \widehat{R}_S(G_Q) + \sqrt{\frac{1-\frac{n}{N}}{2(n-1)}\left[\mathrm{KL}(Q\|P) + \ln \frac{n(N+1)^7}{\delta}\right]}\,.$$



Plan

A New Change of Measure

Kullback-Leibler Change of Measure Inequality

For any P and Q on $\mathcal{H}$, and for any $\phi : \mathcal{H} \to \mathbb{R}$, we have

$$\mathop{\mathbf{E}}_{h\sim Q} \phi(h) \;\leq\; \mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{h\sim P} e^{\phi(h)}\right).$$

Rényi Change of Measure Inequality (Atar and Merhav 2015)

For any P and Q on $\mathcal{H}$, any $\phi : \mathcal{H} \to \mathbb{R}$, and for any $\alpha > 1$, we have

$$\frac{\alpha}{\alpha-1} \ln \mathop{\mathbf{E}}_{h\sim Q} \phi(h) \;\leq\; D_\alpha(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{h\sim P} \phi(h)^{\frac{\alpha}{\alpha-1}}\right),$$

with $\quad D_\alpha(Q\|P) \;=\; \dfrac{1}{\alpha-1} \ln\left[\mathop{\mathbf{E}}_{h\sim P}\left(\dfrac{Q(h)}{P(h)}\right)^{\alpha}\right] \;\geq\; \mathrm{KL}(Q\|P),$

and $\quad \lim_{\alpha\to 1} D_\alpha(Q\|P) \;=\; \mathrm{KL}(Q\|P).$

Rényi-Based General Theorem

Theorem (Bégin et al. 2016)

[...] for any $\alpha > 1$, with probability at least $1-\delta$ over the choice of $S \sim D^n$,

$$\forall Q \text{ on } \mathcal{H}: \quad \ln \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \leq \frac{1}{\alpha'}\left[D_\alpha(Q\|P) + \ln \frac{\mathcal{I}^{\mathrm{R}}_{\Delta}(n,\alpha')}{\delta}\right],$$

with

$$\mathcal{I}^{\mathrm{R}}_{\Delta}(n,\alpha') \;=\; \sup_{r\in[0,1]} \left[\sum_{k=0}^{n} \mathbf{Bin}\big(k;n,r\big)\,\Delta\big(\tfrac{k}{n},\, r\big)^{\alpha'}\right],$$

and $\alpha' := \frac{\alpha}{\alpha-1} > 1\,.$

Rényi-Based General Theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \ln \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \le \frac{1}{\alpha'}\left[D_\alpha(Q\|P) + \ln \frac{\mathcal{I}^{\mathrm{R}}_{\Delta}(n,\alpha')}{\delta}\right]\right) \ \ge\ 1-\delta\,.$$

Proof. $\alpha' := \frac{\alpha}{\alpha-1}$

$$\alpha' \cdot \ln \Delta\Big(\mathop{\mathbf{E}}_{h\sim Q} \widehat{\mathcal{L}}^{\ell}_S(h), \mathop{\mathbf{E}}_{h\sim Q} \mathcal{L}^{\ell}_D(h)\Big)$$

Jensen's Inequality $\le\ \alpha' \cdot \ln \mathop{\mathbf{E}}_{h\sim Q} \Delta\Big(\widehat{\mathcal{L}}^{\ell}_S(h), \mathcal{L}^{\ell}_D(h)\Big)$

Change of measure $\le\ D_\alpha(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} \Delta\big(\widehat{\mathcal{L}}^{\ell}_S(h), \mathcal{L}^{\ell}_D(h)\big)^{\alpha'}$

Markov's Inequality $\le_{1-\delta}\ D_\alpha(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{S'\sim D^n} \mathop{\mathbf{E}}_{h\sim P} \Delta\big(\widehat{\mathcal{L}}^{\ell}_{S'}(h), \mathcal{L}^{\ell}_D(h)\big)^{\alpha'}$

Expectation swap $=\ D_\alpha(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S'\sim D^n} \Delta\big(\widehat{\mathcal{L}}^{\ell}_{S'}(h), \mathcal{L}^{\ell}_D(h)\big)^{\alpha'}$

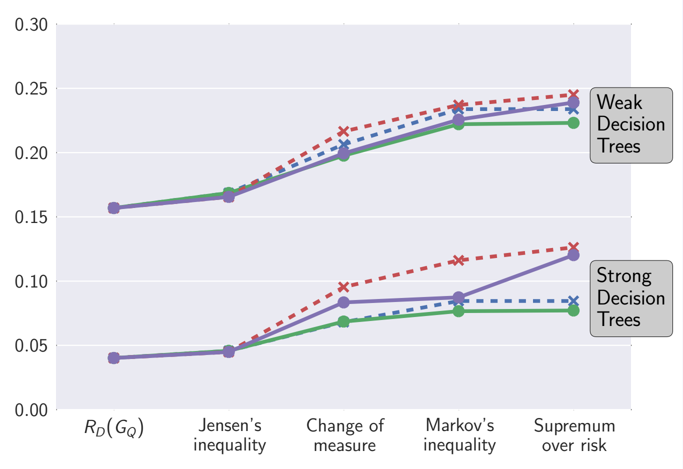
Binomial law $=\ D_\alpha(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \sum_{k=0}^{n} \mathbf{Bin}\big(k; n, \mathcal{L}^{\ell}_D(h)\big)\Delta\big(\tfrac{k}{n}, \mathcal{L}^{\ell}_D(h)\big)^{\alpha'}$

Supremum over risk $\le\ D_\alpha(Q\|P) + \ln \frac{1}{\delta} \sup_{r\in[0,1]} \left[\sum_{k=0}^{n} \mathbf{Bin}\big(k; n, r\big)\Delta\big(\tfrac{k}{n}, r\big)^{\alpha'}\right]$

, $=\ D_\alpha(Q\|P) + \ln \frac{1}{\delta} \mathcal{I}^{\mathrm{R}}_{\Delta}(n,\alpha')\,.$ □

Empirical Study

Majority votes of 500 decision trees on *Mushroom* dataset



× $\mathrm{KL}(Q\|P)$ and $\Delta := 2(q-p)^2$ • $D_\alpha(Q\|P)$ and $\Delta := 2(q-p)^2$

× $\mathrm{KL}(Q\|P)$ and $\Delta := \mathrm{kl}(q,p)$ • $D_\alpha(Q\|P)$ and $\Delta := \mathrm{kl}(q,p)$

Plan

PAC-Bayesian Bounds for Regression

Lemma (Maurer 2004)

For any $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0,1]$, and convex $\Delta : [0,1] \times [0,1] \to \mathbb{R}$,

$$\mathop{\mathbf{E}}_{S' \sim D} e^{n \cdot \Delta(\widehat{\mathcal{L}}^{\ell}_{S'}(h), \mathcal{L}^{\ell}_{D}(h))} \leq \sum_{k=0}^{n} \mathbf{Bin}\big(k; n, \mathcal{L}^{\ell}_{D}(h)\big) e^{n \cdot \Delta(\frac{k}{n}, \mathcal{L}^{\ell}_{D}(h))}$$

General theorem for regression (with bounded losses)

For any distribution D on $\mathcal{X} \times \mathcal{Y}$, for any set $\mathcal{H}$ of predictors, for any $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0,1]$ for any distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, and for any Δ-function, we have, with probability at least $1-\delta$ over the choice of $S \sim D^n$,

$$\forall\, Q \text{ on } \mathcal{H} : \quad \Delta\Big(\mathop{\mathbf{E}}_{h \sim Q} \widehat{\mathcal{L}}^{\ell}_{S}(h), \mathop{\mathbf{E}}_{h \sim Q} \mathcal{L}^{\ell}_{D}(h)\Big) \leq \frac{1}{n}\left[\mathrm{KL}(Q \| P) + \ln \frac{\mathcal{I}_{\Delta}(n)}{\delta}\right].$$

General theorem for regression (with bounded losses)

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\left(\mathbf{E}_{h\sim Q}\widehat{\mathcal{L}}_S^{\ell}(h),\ \mathbf{E}_{h\sim Q}\mathcal{L}_D^{\ell}(h)\right) \le \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \ \ge\ 1-\delta\,.$$

Proof.

$$n\cdot\Delta\left(\mathbf{E}_{h\sim Q}\widehat{\mathcal{L}}_S^{\ell}(h),\ \mathbf{E}_{h\sim Q}\mathcal{L}_D^{\ell}(h)\right)$$

Jensen's Inequality $\quad\le\quad \mathbf{E}_{h\sim Q}\, n\cdot\Delta\left(\widehat{\mathcal{L}}_S^{\ell}(h), \mathcal{L}_D^{\ell}(h)\right)$

Change of measure $\quad\le\quad \mathrm{KL}(Q\|P) + \ln \mathbf{E}_{h\sim P}\, e^{n\Delta\left(\widehat{\mathcal{L}}_S^{\ell}(h),\mathcal{L}_D^{\ell}(h)\right)}$

Markov's Inequality $\quad\le_{1-\delta}\quad \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathbf{E}_{S'\sim D^n}\mathbf{E}_{h\sim P}\, e^{n\cdot\Delta(\widehat{\mathcal{L}}_{S'}^{\ell}(h),\mathcal{L}_D^{\ell}(h))}$

Expectation swap $\quad=\quad \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathbf{E}_{h\sim P}\mathbf{E}_{S'\sim D^n}\, e^{n\cdot\Delta(\widehat{\mathcal{L}}_{S'}^{\ell}(h),\mathcal{L}_D^{\ell}(h))}$

Maurer's Lemma $\quad\le\quad \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathbf{E}_{h\sim P}\sum_{k=0}^{n}\mathbf{Bin}\left(k; n, \mathcal{L}_D^{\ell}(h)\right)e^{n\cdot\Delta(\frac{k}{n},\mathcal{L}_D^{\ell}(h))}$

Supremum over risk $\quad\le\quad \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\sup_{r\in[0,1]}\left[\sum_{k=0}^{n}\mathbf{Bin}\left(k; n, r\right)e^{n\Delta(\frac{k}{n}, r)}\right]$

$\quad=\quad \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathcal{I}_\Delta(n)\,.$ □

PAC-Bayesian Bounds for Regression

General theorem for regression (with bounded losses)

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\Big(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S^{\ell}(h),\ \mathop{\mathbf{E}}_{h\sim Q}\mathcal{L}_D^{\ell}(h)\Big) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \geq 1-\delta\,.$$

Corollary

[...] with probability at least $1-\delta$ *over the choice of* $S\sim D^n$*, for all* Q *on* $\mathcal{H}$ *:*

(a) $$\mathrm{kl}\Big(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S^{\ell}(h),\ \mathop{\mathbf{E}}_{h\sim Q}\mathcal{L}_D^{\ell}(h)\Big) \leq \tfrac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\tfrac{2\sqrt{n}}{\delta}\right],$$ *(Langford and Seeger 2001)*

(b) $$\mathop{\mathbf{E}}_{h\sim Q}\mathcal{L}_D^{\ell}(h) \leq \mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S^{\ell}(h) + \sqrt{\tfrac{1}{2n}\left[\mathrm{KL}(Q\|P) + \ln\tfrac{2\sqrt{n}}{\delta}\right]},$$ *(McAllester 1999, 2003)*

(c) $$\mathop{\mathbf{E}}_{h\sim Q}\mathcal{L}_D^{\ell}(h) \leq \tfrac{1}{1-e^{-c}}\left(c\cdot\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S^{\ell}(h) + \tfrac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\tfrac{1}{\delta}\right]\right),$$ *(Catoni 2007)*

(d) $$\mathop{\mathbf{E}}_{h\sim Q}\mathcal{L}_D^{\ell}(h) \leq \mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S^{\ell}(h) + \tfrac{1}{\lambda}\left[\mathrm{KL}(Q\|P) + \ln\tfrac{1}{\delta} + f(\lambda,n)\right].$$ *(Alquier et al. 2015)*

Plan

Plan

Optimal Gibbs Posterior

Corollary

[...] with probability at least $1-\delta$ *over the choice of* $S \sim D^n$, *for all* Q *on* $\mathcal{H}$:

$$\text{(c)} \quad \mathop{\mathbf{E}}_{h\sim Q} \mathcal{L}_D^{\ell}(h) \le \frac{1}{1-e^{-c}} \left(c \cdot \mathop{\mathbf{E}}_{h\sim Q} \widehat{\mathcal{L}}_S^{\ell}(h) + \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \right] \right), \quad \textit{(Catoni 2007)}$$

$$\text{(d)} \quad \mathop{\mathbf{E}}_{h\sim Q} \mathcal{L}_D^{\ell}(h) \le \mathop{\mathbf{E}}_{h\sim Q} \widehat{\mathcal{L}}_S^{\ell}(h) + \frac{1}{\lambda}\left[\mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} + f(\lambda, n) \right]. \quad \textit{(Alquier et al. 2015)}$$

From an algorithm design perspective, Corollary **(c)** suggests optimizing the following trade-off:

$$c\, n\, \widehat{R}_S(G_Q) + \mathrm{KL}(Q\|P)\,,$$

which also minimizes **(d)**, with $\lambda := c\,n$.

The *optimal Gibbs posterior* is given by

$$Q_c^*(h) \;=\; \frac{1}{Z_S} P(h)\, e^{-c\,n\,\widehat{\mathcal{L}}_S^{\ell}(h)}\,.$$

(See Catoni 2007, Alquier et al. 2015,...)

Tying the Concepts

Let us denote Θ as the set of all possible model parameters.

Bayesian Rule

$$p(\theta|X,Y) \;=\; \frac{p(\theta)\,p(Y|X,\theta)}{p(Y|X)} \;\propto\; p(\theta)\,p(Y|X,\theta)\,,$$

where $X=\{x_1,\ldots,x_n\}$, $Y=\{y_1,\ldots,y_n\}$, and

- $p(\theta)$ is the prior for each $\theta\in\Theta$ (similar to P over $\mathcal{H}$)
- $p(\theta|X,Y)$ is the posterior for each $\theta\in\Theta$ (similar to Q over $\mathcal{H}$)
- $p(Y|X,\theta)$ is the *likelihood* of the parameters θ given the sample S.

Negative log-likelihood loss function

$$\ell_{\rm nll}(\theta,x,y) \;=\; \ln\frac{1}{p(y|x,\theta)}\,.$$

Then,

$$\widehat{\mathcal{L}}_S^{\,\ell_{\rm nll}}(\theta) \;=\; \frac{1}{n}\sum_{i=1}^n \ell_{\rm nll}(\theta,x_i,y_i) \;=\; -\frac{1}{n}\sum_{i=1}^n \ln p(y_i|x_i,\theta) \;=\; -\frac{1}{n}\ln p(Y|X,\theta)\,.$$

Rediscovering the Marginal Likelihood

With the negative log-likelihood loss, the Bayesian and PAC-Bayesian posteriors align:

$$p(\theta|X,Y) \;=\; \frac{p(\theta)\,p(Y|X,\theta)}{p(Y|X)} = \frac{P(\theta)\,e^{-n\,\widehat{\mathcal{L}}_S^{\ell_{\rm nll}}(\theta)}}{Z_S} = Q^*(\theta)\,.$$

The normalization constant Z_S corresponds to the *marginal likelihood*

$$Z_S \;=\; p(Y|X) \;=\; \int_\Theta P(\theta)\,e^{-n\,\widehat{\mathcal{L}}_S^{\ell_{\rm nll}}(\theta)}d\theta\,.$$

Putting back the posterior inside the PAC-Bayesian bounds, we obtain:

$$n \mathop{\mathbf{E}}_{\theta\sim Q^*} \widehat{\mathcal{L}}_S^{\ell_{\rm nll}}(\theta) + \mathrm{KL}(Q^*\|P)$$

$$= \quad n\int_\Theta \frac{P(\theta)\,e^{-n\,\widehat{\mathcal{L}}_S^{\ell_{\rm nll}}(\theta)}}{Z_S}\widehat{\mathcal{L}}_S^{\ell_{\rm nll}}(\theta)\,d\theta + \int_\Theta \frac{P(\theta)\,e^{-n\,\widehat{\mathcal{L}}_S^{\ell_{\rm nll}}(\theta)}}{Z_S}\ln\left[\frac{P(\theta)\,e^{-n\,\widehat{\mathcal{L}}_S^{\ell_{\rm nll}}(\theta)}}{P(\theta)\,Z_S}\right]d\theta$$

$$= \quad \int_\Theta \frac{P(\theta)\,e^{-n\,\widehat{\mathcal{L}}_S^{\ell_{\rm nll}}(\theta)}}{Z_S}\left[\ln\frac{1}{Z_S}\right]d\theta \;=\; \frac{Z_S}{Z_S}\ln\frac{1}{Z_S} \;=\; -\ln Z_S\,.$$

From the Marginal Likelihood to PAC-Bayesian Bounds

Corollary (Germain, Bach, et al. 2016)

Given a data distribution D, a parameter set Θ, a prior distribution P over Θ, a $\delta \in (0,1]$, if $\ell_{\rm nll}$ lies in $[a,b]$, we have, with probability at least $1-\delta$ over the choice of $S \sim D^n$,

(c) $\displaystyle \mathop{\mathbf{E}}_{\theta\sim Q^*} \mathcal{L}_D^{\ell_{\rm nll}}(\theta) \;\leq\; a + \frac{b-a}{1-e^{a-b}} \left[1 - e^a \sqrt[n]{Z_S\,\delta}\,\right],$

(d) $\displaystyle \mathop{\mathbf{E}}_{\theta\sim Q^*} \mathcal{L}_D^{\ell_{\rm nll}}(\theta) \;\leq\; \frac{1}{2}(b-a)^2 - \frac{1}{n}\ln\left(Z_S\,\delta\right).$

Take home message!

The marginal likelihood minimizes (some) PAC-Bayesian Bounds (under the negative log-likelihood loss function)

Plan

Model Comparaison

Consider

- a discrete set of L models $\{\mathcal{M}_i\}_{i=1}^L$ with parameters $\{\Theta_i\}_{i=1}^L$,
- a prior $p(\mathcal{M}_i)$ over these models,
- for each model $\mathcal{M}_i$, a prior $p(\theta|\mathcal{M}_i) = P_i(\theta)$ over Θ_i

Bayesian Rule

$$p(\theta|X,Y,\mathcal{M}_i) \;=\; \frac{p(\theta|\mathcal{M}_i)\,p(Y|X,\theta,\mathcal{M}_i)}{p(Y|X,\mathcal{M}_i)}\,,$$

where the *model evidence* is

$$p(Y|X,\mathcal{M}_i) \;=\; \int_{\Theta_i} p(\theta|\mathcal{M}_i)\,p(Y|X,\theta,\mathcal{M}_i)\,d\theta = Z_{S,i}\,.$$

Frequentist Bounds for Bayesian Model Selection

Corollary (Germain, Bach, et al. 2016)

[...] with probability at least $1-\delta$ *over the choice of* $S \sim D^n$,

$\forall\, i \in \{1,\ldots,L\}$:

(c) $\underset{\theta\sim Q_i^*}{\mathbf{E}} \mathcal{L}_D^{\ell_{\rm nll}}(\theta) \;\leq\; a + \frac{b-a}{1-e^{a-b}}\left[1 - e^a \sqrt[n]{Z_{S,i}\,\frac{\delta}{L}}\,\right],$

(d) $\underset{\theta\sim Q^*}{\mathbf{E}} \mathcal{L}_D^{\ell_{\rm nll}}(\theta) \;\leq\; \frac{1}{2}(b-a)^2 - \frac{1}{n}\ln\left(Z_{S,i}\,\frac{\delta}{L}\right).$

Alternative explanation for the *Bayesian Occam's Razor* phenomena...

Plan

Bayesian Linear Regression

Consider a mapping function $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^d$. Given $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$, model parameters $\theta := \mathbf{w} \in \mathbb{R}^d$ and a fixed noise σ, we consider the likelihood

$$p(y|x, \mathbf{w}) = \mathcal{N}(y|\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y - \mathbf{w}\cdot\boldsymbol{\phi}(x))^2}$$

Thus, the negative log-likelihood loss function is

$$\ell_{\mathrm{nll}}(\mathbf{w}, x, y) \;=\; \ln \frac{1}{p(y|x, \mathbf{w})} \;\; = \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}(y - \mathbf{w} \cdot \boldsymbol{\phi}(x))^2$$

We also consider an isotropic Gaussian prior of mean $\mathbf{0}$ and variance σ_P^2

$$p(\mathbf{w}|\sigma_P) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \sigma_P^2) = \frac{1}{\sqrt{(2\pi)^d \sigma_P^2}} e^{-\frac{1}{2\sigma_P^2}\|\mathbf{w}\|^2}.$$

Bayesian Linear Regression

The **Gibbs optimal posterior** is given by

$$Q^*(\mathbf{w}) = p(\mathbf{w}|\sigma,\sigma_P) \;=\; \frac{p(\mathbf{w}|\sigma,\sigma_P)\,p(X,Y|\mathbf{w},\sigma,\sigma_P)}{p(Y|X,\sigma,\sigma_P)} \;=\; \mathcal{N}(\mathbf{w}\,|\,\widehat{\mathbf{w}}, A^{-1}),$$

where $A := \frac{1}{\sigma^2}\mathbf{\Phi}^T\mathbf{\Phi} + \frac{1}{\sigma_P^2}\mathbf{I}$ and $\widehat{\mathbf{w}} := \frac{1}{\sigma^2}A^{-1}\mathbf{\Phi}^T\mathbf{y}$.

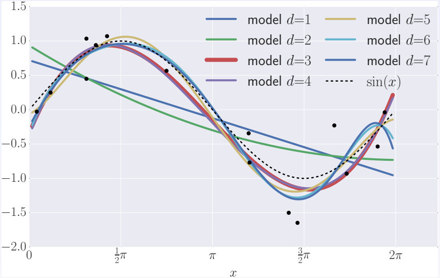
The negative log **marginal likelihood** is

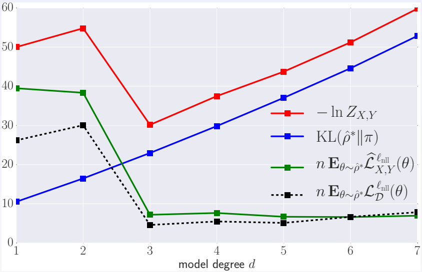
$$\begin{aligned}
&-\ln\Big(Z_S(\sigma,\sigma_P)\Big)\\
&\quad= \tfrac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{\Phi}\widehat{\mathbf{w}}\|^2 + \tfrac{n}{2}\ln(2\pi\sigma^2) + \tfrac{1}{2\sigma_P^2}\|\widehat{\mathbf{w}}\|^2 + \tfrac{1}{2}\log|A| + d\ln\sigma_P\\
&\quad= \underbrace{n\,\widehat{\mathcal{L}}_S^{\ell_{\rm nll}}(\widehat{\mathbf{w}}) + \tfrac{1}{2\sigma^2}\operatorname{tr}(\mathbf{\Phi}^T\mathbf{\Phi}A^{-1})}_{n\,\mathbf{E}_{\mathbf{w}\sim Q^*}\widehat{\mathcal{L}}_S^{\ell_{\rm nll}}(\mathbf{w})} + \underbrace{\tfrac{1}{2\sigma_P^2}\operatorname{tr}(A^{-1}) - \tfrac{d}{2} + \tfrac{1}{2\sigma_P^2}\|\widehat{\mathbf{w}}\|^2 + \tfrac{1}{2}\log|A| + d\ln\sigma_P}_{\mathrm{KL}\left(\mathcal{N}(\widehat{\mathbf{w}},A^{-1})\,\|\,\mathcal{N}(\mathbf{0},\sigma_P^2\mathbf{I})\right)}.
\end{aligned}$$

Fitting $y = \sin(x) + \epsilon$ with polynomial models
(Inspired by Bishop 2006)

Illustrate the decomposition of the marginal likelihood into the empirical loss and KL-divergence.

$$-\ln Z_S \; = \; n \mathop{\mathbf{E}}_{\theta \sim Q^*} \widehat{\mathcal{L}}_S^{\ell_{\mathrm{nll}}}(\theta) + \mathrm{KL}(Q^* \| P)$$





Plan

Conclusion and future works

I talked about..

- A General theorem from which we recover existing results;
- My modular proof, easy to adapt to various frameworks;
- A direct link between PAC-Bayesian (frequentist) bounds and Bayesian model selection.

I did not talk about...

- Our learning algorithms inspired by PAC-Bayesian Bounds;

  see Germain, Lacasse, Laviolette, and Marchand 2009 (ICML) and Germain, Habrard, et al. 2016 (ICML)

- Our PAC-Bayesian theorems for unbounded losses.

  see Germain, Bach, et al. 2016 (arXiv)

I plan to...

- Study other Bayesian techniques from a PAC-Bayes perspective (empirical Bayes, variational Bayes, etc.)

References I

Alquier, Pierre, James Ridgway, and Nicolas Chopin (2015). "On the properties of variational approximations of Gibbs posteriors". In: *ArXiv e-prints*. url: http://arxiv.org/abs/1506.04091.

Atar, Rami and Neri Merhav (2015). "Information-theoretic applications of the logarithmic probability comparison bound". In: *IEEE International Symposium on Information Theory (ISIT)*.

Bégin, Luc, Pascal Germain, François Laviolette, and Jean-Francis Roy (2014). "PAC-Bayesian Theory for Transductive Learning". In: *AISTATS*.

— (2016). "PAC-Bayesian Bounds based on the Rényi Divergence". In: *AISTATS*.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Catoni, Olivier (2007). *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Vol. 56. Inst. of Mathematical Statistic.

Derbeko, Philip, Ran El-Yaniv, and Ron Meir (2004). "Explicit Learning Curves for Transduction and Application to Clustering and Compression Algorithms". In: *J. Artif. Intell. Res. (JAIR)* 22.

Germain, Pascal (2015). "Généralisations de la théorie PAC-bayésienne pour l'apprentissage inductif, l'apprentissage transductif et l'adaptation de domaine." PhD thesis. Université Laval. url: http://www.theses.ulaval.ca/2015/31774/.

References II

Germain, Pascal, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien (2016). "PAC-Bayesian Theory Meets Bayesian Inference". In: *ArXiv e-prints*. url: http://arxiv.org/abs/1605.08636.

Germain, Pascal, Amaury Habrard, François Laviolette, and Emilie Morvant (2016). "A New PAC-Bayesian Perspective on Domain Adaptation". In: *ICML*. url: http://arxiv.org/abs/1506.04573.

Germain, Pascal, Alexandre Lacasse, Francois Laviolette, and Mario Marchand (2009). "PAC-Bayesian learning of linear classifiers". In: *ICML*.

Germain, Pascal, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Jean-Francis Roy (2015). "Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm". In: *JMLR* 16.

Langford, John and Matthias Seeger (2001). *Bounds for averaging classifiers*. Tech. rep. Carnegie Mellon, Departement of Computer Science.

Maurer, Andreas (2004). "A Note on the PAC-Bayesian Theorem". In: *CoRR* cs.LG/0411099.

McAllester, David (1999). "Some PAC-Bayesian Theorems". In: *Machine Learning* 37.3.

— (2003). "PAC-Bayesian Stochastic Model selection". In: *Machine Learning* 51.1.