A tutorial on the Pac-Bayesian Theory

NIPS workshop - "(Almost) 50 shades of Bayesian Learning: PAC-Bayesian trends and insights"

by François Laviolette

Laboratoire du GRAAL, Université Laval

December 9th 2017

Outline of the Tutorial

- **Definitions and notations**
- some PAC-bayesian bounds
- An historical overview
- Algorithms derived from PAC-Bayesian bound
- Localized PAC-Bayesian bounds
- The transductive setting

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 **observation from distribution** D on $\mathcal{X} \times \mathcal{Y}$.

Learning sample

$$S \;=\; \{\, (x_1, y_1),\; (x_2, y_2),\; \ldots,\; (x_m, y_m) \,\} \;\sim\; D^m$$

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

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

Generalization loss

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

Majority Vote Classifiers

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

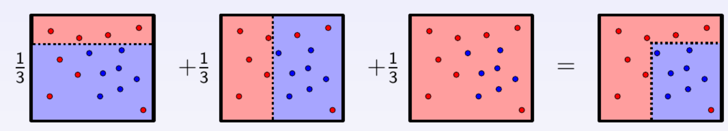
Weighted majority vote

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

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

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

Gibbs Risk / Linear Loss

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

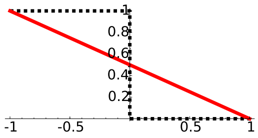
$$\begin{aligned} R_D(G_Q) \;&=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\, \mathop{\mathbf{E}}_{h\sim Q} \mathrm{I}\Big[h(x) \neq y\Big] \\ &=\; \mathop{\mathbf{E}}_{h\sim Q} \mathcal{L}_D^{\ell_{01}}(h)\,, \end{aligned}$$

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

Factor two

It is well-known that

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



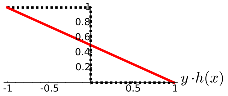
$\mathop{\mathbf{E}}_{h\sim Q} y\cdot h(x)$

From the *Factor 2* to the $\mathcal{C}$-bound

From Markov's inequality ($\Pr\left(X \geq a\right) \leq \frac{\mathbf{E}\, X}{a}$), we obtain:

Factor 2 bound

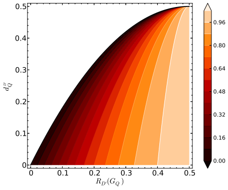
$$
\begin{aligned}
R_D(B_Q) \;&=\; \Pr_{(\mathbf{x},y)\sim D}\Big(1 - y{\cdot}h(x) \geq 1\Big) \\
&\leq\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\Big(1 - y{\cdot}h(x)\Big) \;=\; 2\, R_D(G_Q)\,.
\end{aligned}
$$



From Chebyshev's inequality ($\Pr\left(X - \mathbf{E}\, X \geq a\right) \leq \frac{\mathbf{Var}\, X}{a^2 + \mathbf{Var}\, X}$), we obtain:

The $\mathcal{C}$-bound (Lacasse et al., 2006)

$$
R_D(B_Q) \;\leq\; \mathcal{C}_Q^D \;\stackrel{\text{def}}{=}\; 1 - \frac{\Big(1 - 2\cdot R_D(G_Q)\Big)^2}{1 - 2\cdot d_Q^D}
$$



where d_Q^D is the **expected disagreement**:

$$
d_Q^D \;\stackrel{\text{def}}{=}\; \mathop{\mathbf{E}}_{(x,\cdot)\sim D}\; \mathop{\mathbf{E}}_{h_i\sim Q}\; \mathop{\mathbf{E}}_{h_j\sim Q}\; \mathrm{I}\Big[h_i(x) \neq h_j(x)\Big] \;=\; \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{(x,\cdot)\sim D}\left[\mathop{\mathbf{E}}_{h\sim Q} h(x)\right]^2\right).
$$

Outline of the Tutorial

- Definitions and notations
- **some PAC-bayesian bounds**
- An historical overview
- Algorithms derived from PAC-Bayesian bound
- Localized PAC-Bayesian bounds
- The transductive setting

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*

$$\mathbb{E}_Q[\Phi] \;\le\; \mathrm{KL}(Q\|P) + \ln \mathbb{E}_P[e^{\Phi}]$$

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

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999a)*

McAllester Bound

For any D, any $\mathcal{H}$, any P of support $\mathcal{H}$, any $\delta \in (0,1]$, we have

$$\Pr_{S\sim D^m}\left(\forall\, Q \text{ on } \mathcal{H}\colon\ 2(R_S(G_Q) - R(G_Q))^2 \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}\right]\right) \geq 1-\delta\,,$$

or

$$\Pr_{S\sim D^m}\left(\forall\, Q \text{ on } \mathcal{H}\colon\ R(G_Q) \leq\ R_S(G_Q) + \sqrt{\frac{\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}\right]}{2m}}\right) \geq 1-\delta\,,$$

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999a)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*

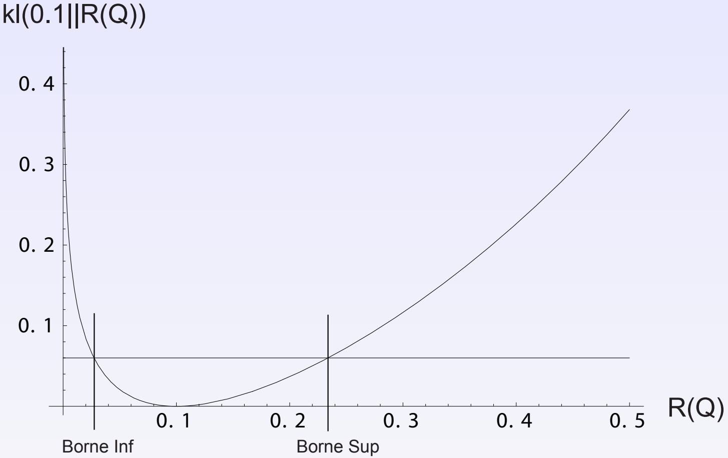
Langford and Seeger Bound

For any D, any $\mathcal{H}$, any P of support $\mathcal{H}$, any $\delta \in (0,1]$, we have

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\, Q \text{ on } \mathcal{H}: \\ \mathrm{kl}(R_S(G_Q)\|R(G_Q)) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}\right]\end{array}\right) \geq 1-\delta\,,$$

where $\quad \mathrm{kl}(q\|p) \overset{\text{def}}{=} q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}.$

Graphical illustration of the Langford/Seeger bound



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. (2014b, 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 $S \sim D^m$,

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

where

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

Proof of the general theorem

General theorem

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

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 m examples

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

$$\Pr_{S\sim D^m}\left(\widehat{\mathcal{L}}_S^\ell(h) = \tfrac{k}{m}\right) \ =\ \binom{m}{k}\Big(\mathcal{L}_D^\ell(h)\Big)^k\Big(1-\mathcal{L}_D^\ell(h)\Big)^{m-k} \ =\ \mathbf{Bin}\Big(k; m, \mathcal{L}_D^\ell(h)\Big)$$

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

$$
\begin{array}{lll}
 & & m\cdot\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S^{\,\ell}(h),\ \mathop{\mathbf{E}}_{h\sim Q}\mathcal{L}_D^{\,\ell}(h)\Big) \\
\textbf{Jensen's Inequality} & \le & \mathop{\mathbf{E}}_{h\sim Q} m\cdot\Delta\Big(\widehat{\mathcal{L}}_S^{\,\ell}(h), \mathcal{L}_D^{\,\ell}(h)\Big) \\
\textbf{Change of measure} & \le & \mathrm{KL}(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} e^{m\Delta\left(\widehat{\mathcal{L}}_S^{\,\ell}(h), \mathcal{L}_D^{\,\ell}(h)\right)} \\
\textbf{Markov's Inequality} & \le_{1-\delta} & \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{S'\sim D^m}\ \mathop{\mathbf{E}}_{h\sim P} e^{m\cdot\Delta(\widehat{\mathcal{L}}_{S'}^{\,\ell}(h), \mathcal{L}_D^{\,\ell}(h))} \\
\textbf{Expectation swap} & = & \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\ \mathop{\mathbf{E}}_{S'\sim D^m} e^{m\cdot\Delta(\widehat{\mathcal{L}}_{S'}^{\,\ell}(h), \mathcal{L}_D^{\,\ell}(h))} \\
\textbf{Binomial law} & = & \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\sum_{k=0}^{m}\mathbf{Bin}\Big(k; m, \mathcal{L}_D^{\,\ell}(h)\Big)e^{m\cdot\Delta(\frac{k}{m}, \mathcal{L}_D^{\,\ell}(h))} \\
\textbf{Supremum over risk} & \le & \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\sup_{r\in[0,1]}\left[\sum_{k=0}^{m}\mathbf{Bin}\Big(k; m, r\Big)e^{m\Delta(\frac{k}{m}, r)}\right] \\
 & = & \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\,\mathcal{I}_\Delta(m)\,. \qquad \square
\end{array}
$$

General theorem

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

Corollary

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

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

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

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

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

$$\mathrm{kl}(q,p) \;\overset{\text{def}}{=}\; q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p} \;\ge\; 2(q-p)^2\,,$$

$$\Delta_c(q,p) \;\overset{\text{def}}{=}\; -\ln[1-(1-e^{-c})\cdot p] - c\cdot q\,,$$

$$\Delta_\lambda(q,p) \;\overset{\text{def}}{=}\; \frac{\lambda}{m}(p-q)\,.$$

Proof of the Langford/Seeger bound

Follows immediately from General Theorem by choosing $\Delta(q,p) = \mathrm{kl}(q,p)$.

- Indeed, in that case we have

$$
\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\Delta(R_S(h),R(h))} &= \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S\sim D^m} \left(\tfrac{R_S(h)}{R(h)}\right)^{mR_S(h)} \left(\tfrac{1-R_S(h)}{1-R(h)}\right)^{m(1-R_S(h))} \\
&= \mathop{\mathbf{E}}_{h\sim P} \textstyle\sum_{k=0}^{m} \Pr\limits_{S\sim D^m} \left(R_S(h) = \frac{k}{m}\right) \left(\frac{\frac{k}{m}}{R(h)}\right)^{k} \left(\frac{1-\frac{k}{m}}{1-R(h)}\right)^{m-k} \\
&= \textstyle\sum_{k=0}^{m} \binom{m}{k} (k/m)^k (1-k/m)^{m-k}\,, \qquad (1) \\
&\leq 2\sqrt{m}\,. \qquad \square
\end{aligned}
$$

- Note that, in Line (1) of the proof, $\Pr\limits_{S\sim D^m}\left(R_S(h) = \frac{k}{m}\right)$ is replaced by the probability mass function of the binomial.
- This is **only true if** the examples of S are drawn iid. (i.e., $S \sim D^m$)
- So this result is no longuer valid in the non iid case, even if General Theorem is.

Outline of the Tutorial

- Definitions and notations
- some PAC-bayesian bounds
- **An historical overview**
- Algorithms derived from PAC-Bayesian bound
- Localized PAC-Bayesian bounds
- The transductive setting

History

- **Pre-pre-history: Variational Definition of KL-divergence**
- **Pre-history: PAC analysis of Bayesian estimators**
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999a)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
  - **SVMs & linear classifiers** *Langford and Shawe-Taylor (2002); McAllester (2003a); Germain et al. (2009a)*; ...
  - **Theory** *Catoni (2007a); Audibert and Bousquet (2007a); Meir and Zhang (2003); McAllester (2013); Germain et al. (2015, 2016a); London (2017)*; ...
  - **supervised learning algorithms that are bound minimizers** *Ambroladze et al. (2007); Germain et al. (2009b, 2011)*
  - **Regression** *Audibert (2004)*
  - **Transductive learning** *Derbeko et al. (2004); Audibert and Bousquet (2007b); Bégin et al. (2014a)*
  - **Domain adaptation** *Germain et al. (2013, 2016b)*
  - **Non-i.i.d. data** *Ralaivola et al. (2010); Lever et al. (2010); Seldin et al. (2011); Alquier and Guedj (2016)*

    This allows applications to ranking, U-statistic of higher order,

History

- **Pre-pre-history: Variational Definition of KL-divergence**
- **Pre-history: PAC analysis of Bayesian estimators**
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999a)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
  - **SVMs & linear classifiers** *Langford and Shawe-Taylor (2002); McAllester (2003a); Germain et al. (2009a)*; ...
  - **Theory** *Catoni (2007a); Audibert and Bousquet (2007a); Meir and Zhang (2003); McAllester (2013); Germain et al. (2015, 2016a); London (2017)*; ...
  - **supervised learning algorithms that are bound minimizers** *Ambroladze et al. (2007); Germain et al. (2009b, 2011)*
  - **Regression** *Audibert (2004)*
  - **Transductive learning** *Derbeko et al. (2004); Audibert and Bousquet (2007b); Bégin et al. (2014a)*
  - **Domain adaptation** *Germain et al. (2013, 2016b)*
  - **Non-i.i.d. data** *Ralaivola et al. (2010); Lever et al. (2010); Seldin et al. (2011); Alquier and Guedj (2016)*
  - **sample compression setting** *Laviolette and Marchand (2005); Germain et al. (2011)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999a)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
- **Density estimation** *Seldin and Tishby (2010); Higgs and Shawe-Taylor (2010)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999a)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
- **Density estimation** *Seldin and Tishby (2010); Higgs and Shawe-Taylor (2010)*
- **Martingales & reinforcement learning** *Fard and Pineau (2010); Seldin et al. (2011, 2012)*

History

- **Pre-pre-history: Variational Definition of KL-divergence** *Donsker and Varadhan (1975)*
- **Pre-history: PAC analysis of Bayesian estimators** *Shawe-Taylor and Williamson (1997); Shawe-Taylor et al. (1998)*
- **Birth: First PAC-Bayesian theorems** *McAllester (1998, 1999a)*
- **Introduction of kl form** *Seeger (2002); Langford (2005)*
- **Applications in supervised learning**
- **Density estimation** *Seldin and Tishby (2010); Higgs and Shawe-Taylor (2010)*
- **Martingales & reinforcement learning** *Fard and Pineau (2010); Seldin et al. (2011, 2012)*

- **Sincere apologizes to everybody we could not fit on the slide...**

Outline of the Tutorial

- Definitions and notations
- some PAC-bayesian bounds
- An historical overview
- **Algorithms derived from PAC-Bayesian bound**
- Localized PAC-Bayesian bounds
- The transductive setting

Algorithms derived from PAC-Bayesian Bounds

When given a PAC-Bayes bound, one can easily derive a learning algorithm that will simply consist of finding the posterior Q that minimizes the bound.

Catoni's bound

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\, Q \text{ on } \mathcal{H}: \\ R(G_Q) \le \frac{1}{1-e^{-C}}\Big\{1-\exp\Big[-\big(C\cdot R_S(G_Q) \\ \qquad\qquad +\frac{1}{m}\big[\mathrm{KL}(Q\|P)+\ln\frac{1}{\delta}\big]\big)\Big]\Big\}\end{array}\right) \ge 1-\delta.$$

Interestingly, minimizing the Catoni's bound (when prior and posterior are restricted to Gaussian) give rise to the SVM !
In fact to an SVM where the Hinge loss is replaced by the sigmoid loss.

Algorithms derived from PAC-Bayesian Bounds (cont)

Not only SVM has been rediscover as a PAC-Bayes bound minimizer, we also have:

- **KL-Regularized Adaboost** *Germain et al. (2009b)*
- **Kernel Ridge Regression** *Germain et al. (2011)*
- **the proposed structured output algorithm of Cortes et al. (2007)** *Giguère et al. (2013)*

New algorithms have been found: *Ambroladze et al. (2007); Shawe-Taylor and Hardoon (2009); Germain et al. (2011); Laviolette et al. (2011); Germain et al. (2016b)*, ...

Outline of the Tutorial

- Definitions and notations
- some PAC-bayesian bounds
- An historical overview
- Algorithms derived from PAC-Bayesian bound
- **Localized PAC-Bayesian bounds**
- The transductive setting

What is a localized PAC-Bayesian bound ?

Basically, a PAC-Bayesian bound depends on two quantities:

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln \frac{\xi(m)}{\delta}}{2m}}.$$

- Hence, the bound expresses a tradeoff to be followed for finding *suitable* choices of the posterior distribution Q.
- A tradeoff between "empirical accuracy" and "complexity"; the complexity being quantify by how far a posterior distributions is from our prior knowledge.
- Thus, some "luckiness argument" is involved here.
  This can be good, but one might want to have some guarantees that, even in unlucky situations, the bound does not degrade over some level.
  (In general the KL-divergence can be very large ... even infinite)

Localized PAC-Bayesian bounds : a way to reduce the KL-complexity term

- If something can be done to ensure that the bound remains under control it has to be based on the choice of the prior.

$$L(Q) \lesssim \hat{L}(Q) + \sqrt{\frac{\mathrm{KL}(Q\|{\color{red}P}) + \ln\frac{\xi(m)}{\delta}}{2m}}.$$

- However, recall that the prior is not allowed to depend in any way on the training set.

Localized PAC-Bayesian bounds :

(1) Let us simply learn the prior !

- one may leave a part of the training set in order to learn the prior, and only use the remaining part of it to calculate the PAC-Bayesian bound.

    - A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In *Advances in Neural Information Processing Systems* 18, (2006) Pages 9-16.
    - P. Germain, A. Lacasse, F. Laviolette and M. Marchand. PAC-Bayesian learning of linear classifiers, in *Proceedings of the 26nd International Conference on Machine Learning* (ICML'09, Montréal, Canada.). ACM Press (2009), 382, Pages 453-460.

Localized PAC-Bayesian bounds: (2) distribution-dependent!

- Even if the prior can not be data dependent, it can depend on the distribution D that generates the data.
  - How can this be possible ? D is supposed to be unknown !
  - Thus, P will have to remain unknown !
  - But may be we can manage to nevertheless estimate $\mathrm{KL}(Q\|P)$. This is all we need here.

This has been proposed in

- A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In *Advances in Neural Information Processing Systems* 18, (2006) Pages 9-16.
  The chosen prior was: $\mathbf{w}_p = \mathbb{E}_{(\mathbf{x},y)\sim D}(y\,\boldsymbol{\phi}(\mathbf{x}))$.
- O. Catoni. A PAC-Bayesian approach to adaptive classification. Preprint n.840, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2003.
- G. Lever, F. Laviolette, J. Shawe-Taylor. Distribution-Dependent PAC-Bayes Priors. Proceedings of the 21st International Conference on Algorithmic Learning Theory (ALT 2010), 119-133.

Localized PAC-Bayesian bounds :

(2) Distribution-Dependent PAC-Bayes Priors (cont)

- in particular, Lever et al propose a distribution dependent prior of the form:

$$P(h) \;=\; \frac{1}{Z}\exp(-\gamma R(h))\,,$$

  for some a priori chosen hyper-parameter gamma.
- Such distribution dependent priors are designed to put more weight on accurate hypothesis and exponentially decrease the weight as the accuracies are decreasing. (A "wise" choice).
- Then, we can bound the KL-term under the restriction that the posterior is of the form

$$Q(h) \;=\; \frac{1}{Z'}\exp(-\gamma R_S(h))\,.$$

  Again a suitable form for a posterior (and which this time is a known quantity).

Localized PAC-Bayesian bounds :

(2) Distribution-Dependent PAC-Bayes Priors (cont)

The KL-term is bounded as follows:

$$\mathrm{KL}(Q\|P) \;\le\; \frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{2\xi(m)}{\delta}} + \frac{\gamma^2}{4m}\,.$$

The trick: we apply a second PAC-bayesian bound and applied it to the KL-term.

This gives rise to a very *tight* localized PAC-Bayesian bound:

Lever et al. (2010)

For any D, any $\mathcal{H}$, any P of support $\mathcal{H}$, any $\delta \in (0,1]$, we have

$$\Pr_{S\sim D^m}\Bigg(\forall\, Q \text{ on } \mathcal{H}\colon\; \mathrm{kl}(R_S(G_Q), R(G_Q)) \le \frac{1}{m}\left[\frac{\gamma}{\sqrt{m}}\sqrt{\ln\frac{2\xi(m)}{\delta/2}} + \frac{\gamma^2}{4m} + \ln\frac{\xi(m)}{\delta/2}\right]\Bigg) \ge 1-\delta\,.$$

Localized PAC-Bayesian bounds :

(3) Let us do magic and let us simply make the KL-term disappear

Consider any auto-complemented set $\mathcal{H}$ of hypothesis. We say that Q is **aligned** on P iff for all $h \in \mathcal{H}$, we have

$$Q(h) + Q(-h) = P(h) + P(-h)\,.$$

Note: we can construct any (almost any if $\mathcal{H}$ is uncountable) majority vote with aligned posteriors.

In other words, for any posterior Q, there is a posterior Q', aligned on P such that

$$B_Q(\mathbf{x}) \;=\; B_{Q'}(\mathbf{x})\,.$$

So, same classification capacity if one restrict itself to aligned posterior. But then, the KL-term vanishes from the PAC-Bayesian bound !!!

MAGIC !!!

Theorem

If $\Delta(p,q) = \mathrm{kl}(p,q)$ *or* $2(q-p)^2$, *then*

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

Proof.

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

Jensen's inequality	$\le$	$\mathop{\mathbf{E}}_{h\sim Q} m\cdot\Delta\Big(\widehat{\mathcal{L}}_S^{\,\ell}(h), \mathcal{L}_D^{\,\ell}(h)\Big)$
Jensen's inequality	$=$	$\mathop{\mathbf{E}}_{h\sim Q}\ln\left(e^{m\Delta\left(\widehat{\mathcal{L}}_S^{\,\ell}(h),\mathcal{L}_D^{\,\ell}(h)\right)}\right) \le \ln\mathop{\mathbf{E}}_{h\sim Q} e^{m\cdot\Delta\left(\widehat{\mathcal{L}}_S^{\,\ell}(h),\mathcal{L}_D^{\,\ell}(h)\right)}$
Change of measure	$\le$	$\ln\mathop{\mathbf{E}}_{h\sim P} e^{m\cdot\Delta\left(\widehat{\mathcal{L}}_S^{\,\ell}(h),\mathcal{L}_D^{\,\ell}(h)\right)}$
Markov's Inequality	$\le_{1-\delta}$	$\ln\frac{1}{\delta}\mathop{\mathbf{E}}_{S'\sim D^m}\mathop{\mathbf{E}}_{h\sim P} e^{m\cdot\Delta(\widehat{\mathcal{L}}_{S'}^{\,\ell}(h),\mathcal{L}_D^{\,\ell}(h))}$
Expectation swap	$=$	$\ln\frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\mathop{\mathbf{E}}_{S'\sim D^m} e^{m\cdot\Delta(\widehat{\mathcal{L}}_{S'}^{\,\ell}(h),\mathcal{L}_D^{\,\ell}(h))}$
Binomial law	$=$	$\ln\frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\sum_{k=0}^{m}\mathbf{Bin}\big(k;m,\mathcal{L}_D^{\,\ell}(h)\big)e^{m\cdot\Delta(\frac{k}{m},\mathcal{L}_D^{\,\ell}(h))}$
Supremum over risk	$\le$	$\ln\frac{1}{\delta}\sup_{r\in[0,1]}\left[\sum_{k=0}^{m}\mathbf{Bin}\big(k;m,r\big)e^{m\Delta(\frac{k}{m},r)}\right] = \ln\frac{1}{\delta}\mathcal{I}_\Delta(m)$

Absence of KL for Aligned Posteriors

Let $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$ with $\mathcal{H}_1 \cap \mathcal{H}_2 = \emptyset$ such that for each $h \in \mathcal{H}_1 : \ -h \in \mathcal{H}_2$.

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{h\sim P} e^{m\cdot 2(R_S(h)-R(h))^2} \\
&\quad = \int_{h\in\mathcal{H}_1} dP(h) e^{m\cdot 2(R_S(h)-R(h))^2} + \int_{h\in\mathcal{H}_2} dP(h) e^{m\cdot 2(R_S(h)-R(h))^2} \\
&\quad = \int_{h\in\mathcal{H}_1} dP(h) e^{m\cdot 2(R_S(h)-R(h))^2} + \int_{h\in\mathcal{H}_1} dP(-h) e^{m\cdot 2((1-R_S(h))-(1-R(h)))^2} \\
&\quad = \int_{h\in\mathcal{H}_1} dP(h) e^{m\cdot 2(R_S(h)-R(h))^2} + \int_{h\in\mathcal{H}_1} dP(-h) e^{m\cdot 2(R_S(h)-R(h))^2} \\
&\quad = \int_{h\in\mathcal{H}_1} \left(dP(h) + dP(-h)\right) e^{m\cdot 2(R_S(h)-R(h))^2} \\
&\quad = \int_{h\in\mathcal{H}_1} \left(dQ(h) + dQ(-h)\right) e^{m\cdot 2(R_S(h)-R(h))^2} \\
&\quad \vdots \\
&\quad = \mathop{\mathbf{E}}_{h\sim Q} e^{m\cdot 2(R_S(h)-R(h))^2}\,.
\end{aligned}
$$

Outline of the Tutorial

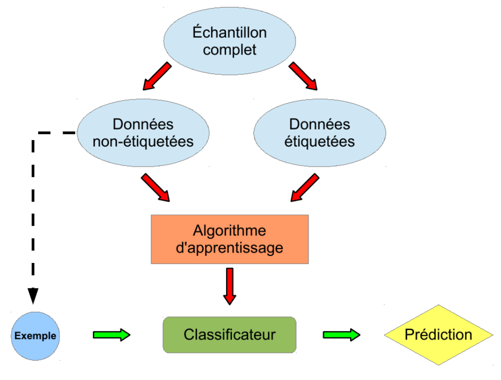
- Definitions and notations
- some PAC-bayesian bounds
- An historical overview
- Algorithms derived from PAC-Bayesian bound
- Localized PAC-Bayesian bounds
- **The transductive setting**

Transductive Learning

Assumption

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

$$\begin{array}{lllllll} S & = & \{\,(x_1,y_1), & (x_2,y_2), & \ldots, & (x_m,y_m)\,\} & \subset Z \\ U & = & \{\,(x_{m+1},\cdot), & (x_{m+2},\cdot), & \ldots, & (x_N,\cdot)\,\} & = Z \setminus S \end{array}$$



Transductive learning

Assumption

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

$$\begin{array}{rclllll} S & = & \{\, (x_1, y_1), & (x_2, y_2), & \ldots, & (x_m, y_m) \,\} & \subset Z \\ U & = & \{\, (x_{m+1}, \cdot), & (x_{m+2}, \cdot), & \ldots, & (x_N, \cdot) \,\} & = Z \setminus S \end{array}$$

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

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

Theorem (*Bégin et al. (2014b)*)

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

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

where

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

Theorem

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

Proof.

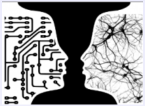
$$
\begin{aligned}
& & & m\cdot\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}^{\ell}_{S}(h), \mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}^{\ell}_{Z}(h)\Big)\\
&\text{Jensen's inequality} & \le\; & \mathop{\mathbf{E}}_{h\sim Q} m\cdot\Delta\Big(\widehat{\mathcal{L}}^{\ell}_{S}(h), \widehat{\mathcal{L}}^{\ell}_{Z}(h)\Big)\\
&\text{Change of measure} & \le\; & \mathrm{KL}(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} e^{m\Delta\left(\widehat{\mathcal{L}}^{\ell}_{S}(h), \widehat{\mathcal{L}}^{\ell}_{Z}(h)\right)}\\
&\text{Markov's inequality} & \le_{1-\delta}\; & \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\mathop{\mathbf{E}}_{S'\sim[Z]^m}\mathop{\mathbf{E}}_{h\sim P} e^{m\cdot\Delta(\widehat{\mathcal{L}}^{\ell}_{S'}(h), \widehat{\mathcal{L}}^{\ell}_{Z}(h))}\\
&\text{Expectations swap} & =\; & \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\mathop{\mathbf{E}}_{S'\sim[Z]^m} e^{m\cdot\Delta(\widehat{\mathcal{L}}^{\ell}_{S'}(h), \widehat{\mathcal{L}}^{\ell}_{Z}(h))}\\
&\text{Hypergeometric law} & =\; & \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\sum_k \frac{\binom{N\cdot\widehat{\mathcal{L}}^{\ell}_{Z}(h)}{k}\binom{N-N\cdot\widehat{\mathcal{L}}^{\ell}_{Z}(h)}{n-k}}{\binom{N}{m}} e^{m\cdot\Delta(\frac{k}{m}, \widehat{\mathcal{L}}^{\ell}_{Z}(h))}\\
&\text{Supremum over risk} & \le\; & \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\max_{K=0\ldots N}\left[\sum_k \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{m}} e^{m\Delta(\frac{k}{m}, \frac{K}{N})}\right]\\
& & =\; & \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\mathcal{T}_\Delta(m,N)\,.
\end{aligned}
$$

□

Aknowledgements

A big thank's to Mario Marchand that initiated me to PAC-Bayes theory and that have been my main PAC-Bayes collaborator since then.

Thank's also to all Graal's members (actual, past and future ones).
Namely for some of the present slides.



Thank's also to John Shawe-Taylor, Liva Ralaivola, David McAllester, Guy Lever, Yevgeny Seldin, Iliya Tolstikhin and John Langford for more than insightful discussions about the subject that often lead to nice collaborations.

Pierre Alquier and Benjamin Guedj. Simpler pac-bayesian bounds for hostile data. http://arXiv:1610.07193v1, 2016.

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

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

J.-Y. Audibert. *Théorie Statistique de l?Apprentissage : une approche PAC-Bayésienne*. thèse de doctorat de l?Université Paris VI, 2004.

Jean-Yves Audibert and Olivier Bousquet. Combining PAC-Bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 2007a.

Jean-Yves Audibert and Olivier Bousquet. Combining pac-bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8: 863–889, 2007b.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian Theory for Transductive Learning. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International*

Conference on Artificial Intelligence and Statistics, volume 33 of *Proceedings of Machine Learning Research*, pages 105–113, Reykjavik, Iceland, 22–25 Apr 2014a. PMLR. URL `http://proceedings.mlr.press/v33/begin14.html`.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian theory for transductive learning. In *AISTATS*, 2014b.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian bounds based on the Rényi divergence. In *AISTATS*, 2016.

Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *IMS Lecture Notes Monograph Series*, 56, 2007a.

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

Philip Derbeko, Ran El-Yaniv, and Ron Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of Artificial Intelligence Research*, 22, 2004.

Monroe D. Donsker and S.R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28, 1975.

Mahdi Milani Fard and Joelle Pineau. PAC-Bayesian model selection for reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Pascal Germain. *GÃl'nÃl'ralisations de la thÃl'orie PAC-bayÃl'sienne pour l'apprentissage inductif, l'apprentissage transductif et l'adaptation de domaine.* PhD thesis, Université Laval, 2015. URL http://www.theses.ulaval.ca/2015/31774/.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009a.

Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Sara Shanian. From pac-bayes bounds to kl regularization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 603–610. 2009b. URL http://books.nips.cc/papers/files/nips22/NIPS2009_0456.pdf.

Pascal Germain, Alexandre Lacoste, François Laviolette, Mario Marchand, and Sara Shanian. A pac-bayes sample-compression approach to kernel methods. In *ICML*, pages 297–304, 2011.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *ICML*, 2013.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *J. Mach. Learn. Res.*, 16 (1):787–860, January 2015. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2789272.2831140`.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. Pac-bayesian theory meets bayesian inference. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1884–1892. Curran Associates, Inc., 2016a. URL `http://papers.nips.cc/paper/6569-pac-bayesian-theory-meets-bayesian-inference.pdf`.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A new pac-bayesian perspective on domain adaptation. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 859–868, 2016b. URL `http://jmlr.org/proceedings/papers/v48/germain16.html`.

Sébastien Giguère, François Laviolette, Mario Marchand, and Khadidja Sylla. Risk bounds and learning algorithms for the regression approach to structured output prediction. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 107–114, 2013. URL `http://jmlr.org/proceedings/papers/v28/giguere13.html`.

Matthew Higgs and John Shawe-Taylor. A PAC-Bayes bound for tailored density estimation. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2010.

John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

John Langford and Matthias Seeger. Bounds for averaging classifiers. Technical report, Carnegie Mellon, Departement of Computer Science, 2001.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems (NIPS)*, 2002.

François Laviolette and Mario Marchand. PAC-Bayes risk bounds for sample-compressed Gibbs classifiers. *Proc. of the 22nth International Conference on Machine Learning (ICML)*, pages 481–488, 2005.

François Laviolette, Mario Marchand, and Jean-Francis Roy. From pac-bayes bounds to quadratic programs for majority votes. In *ICML*, pages 649–656, 2011.

Guy Lever, François Laviolette, and John Shawe-Taylor. Distribution-dependent PAC-Bayes priors. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2010.

Ben London. A pac-bayesian analysis of randomized learning with application to stochastic gradient descent. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2935–2944. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/6886-a-pac-bayesian-analysis-of-randomized-learning-with-ap` `pdf`.

David McAllester. Some PAC-Bayesian theorems. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 1998.

David McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37, 1999a.

David McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37 (3), 1999b.

David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1), 2003a.

David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1), 2003b.

David McAllester. A pac-bayesian tutorial with a dropout bound. http://arXiv:1307.2118v1, 2013.

Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic PAC-Bayes bounds for non-IID data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 2010.

Matthias Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 2002.

Yevgeny Seldin and Naftali Tishby. PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11, 2010.

Yevgeny Seldin, Peter Auer, François Laviolette, John Shawe-Taylor, and Ronald Ortner. PAC-Bayesian analysis of contextual bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 2012. Accepted. Preprint available at http://arxiv.org/abs/1110.6886.

John Shawe-Taylor and David Hardoon. Pac-bayes analysis of maximum entropy classification. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 1997.

John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5), 1998.