What to expect from PAC-Bayes bounds?

Pierre Alquier

RIKEN AIP
Center for
Advanced Intelligence Project

AABI 2022

François Laviolette

1962–2021

PAC-Bayesian Learning of Linear Classifiers

Pascal Germain PASCAL.GERMAIN.1@ULAV
Alexandre Lacasse ALEXANDRE.LACASSE@IFT.ULAV
François Laviolette FRANCOIS.LAVIOLETTE@IFT.ULAV
Mario Marchand MARIO.MARCHAND@IFT.ULAV
Département d'informatique et de génie logiciel, Université Laval, Québec, Canada, G1V-0A6

Abstract

We present a general PAC-Bayes theorem from which all known PAC-Bayes risk bounds are obtained as particular cases. We also propose different learning algorithms for finding linear classifiers that minimize these bounds. These learning algorithms are generally competitive with both AdaBoost and the SVM.

1. Intoduction

For the classification problem, we are given a training set of examples—each generated according to the same (but unknown) distribution D, and the goal is to

their data-dependencies only comes through the ing error of the classifiers. The fact that ther exists VC lower bounds, that are asymptotically tical to the corresponding upper bounds, suggest significantly tighter bounds can only come throu tra data-dependent properties such as the distrib of margins achieved by a classifier on the training

Among the data-dependent bounds have been proposed recently, the PAC- bounds (McAllester, 2003; Seeger, 2002; ford, 2005; Catoni, 2007) seem to be especially These bounds thus appear to be a good starting for the design of a bound-minimizing algorithm this paper, we present a general PAC-Bayes th and show that all known PAC-Bayes bound

PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier

Alexandre Lacasse, François Laviolette and Mario Marchand
Département IFT-GLO
Université Laval
Québec, Canada
Firstname.Secondname@ift.ulaval.ca

Pascal Germain
Département IFT-GLO
Université Laval Québec, Canada
Pascal.Germain.1@ulaval.ca

Nicolas Usunier
Laboratoire d'informatique de Paris 6
Université Pierre et Marie Curie, Paris, France
Nicolas.Usunier@lip6.fr

Abstract

We propose new PAC-Bayes bounds for the risk of the weighted majority vote that depend on the mean and variance of the error of its associated Gibbs classifier. We show that these bounds can be smaller than the risk of the Gibbs classifier and can be arbitrarily close to zero even if the risk of the Gibbs classifier is close to 1/2. Moreover, we show that these bounds can be uniformly estimated on the training data for all possible posteriors Q. Moreover, they can be improved by using a large sample of unlabelled data.

Objective of PAC-Bayes bounds

- empirical risk

$$r(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(f(X_i), Y_i\Big)$$

- generalization risk

$$R(f) = \mathbb{E}_{(X,Y)\sim P}\Big[\ell\Big(f(X), Y\Big)\Big]$$

- randomized prediction / ensemble / ... : $f \sim \rho$,

compare $\mathbb{E}_{f\sim\rho}[R(f)]$ and $\mathbb{E}_{f\sim\rho}[r(f)]$.

In this talk : $\ell(u, v) \in [0, 1]$.

A generic PAC-Bayes bound

Let $\mathcal{S}$ denote the sample $\mathcal{S} = [(X_i, Y_i)]_{i=1}^n$.

Theorem

Let $\mathcal{D} : [0,1]^2 \to \mathbb{R}$ be any convex function. For any $\varepsilon > 0$,

$$\mathbb{P}_{\mathcal{S}}\left[\forall \rho,\ \mathcal{D}\Big(\mathbb{E}_{f\sim\rho}[r(f)], \mathbb{E}_{f\sim\rho}[R(f)]\Big) \leq \frac{KL(\rho\|\pi) + \log \frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\pi} e^{n\mathcal{D}(r(f),R(f))}}{\varepsilon}}{n}\right] \leq \varepsilon.$$

Germain, P., Lacasse, A., Laviolette, F. and Marchand, M. (2009). PAC-Bayesian Learning of Linear Classifiers. *ICML*.

A more explicit bound

For $\mathcal{D}(u,v) = 2(u-v)^2$ we obtain :

$$\left[\mathbb{E}_{f\sim\rho}[R(f)] - \mathbb{E}_{f\sim\rho}[r(f)]\right]^2 \leq \frac{KL(\rho\|\pi) + \log\frac{3}{\varepsilon}}{n}$$

$$\mathbb{E}_{f\sim\rho}[R(f)] \leq \mathbb{E}_{f\sim\rho}[r(f)] + \frac{1}{2}\sqrt{\frac{KL(\rho\|\pi) + \log\frac{3}{\varepsilon}}{n}}$$

McAllester, D. (2003). PAC-Bayesian Learning of Linear Classifiers. *Machine learning*.

A bound in expectation

Theorem

For any (data-dependent) ρ,

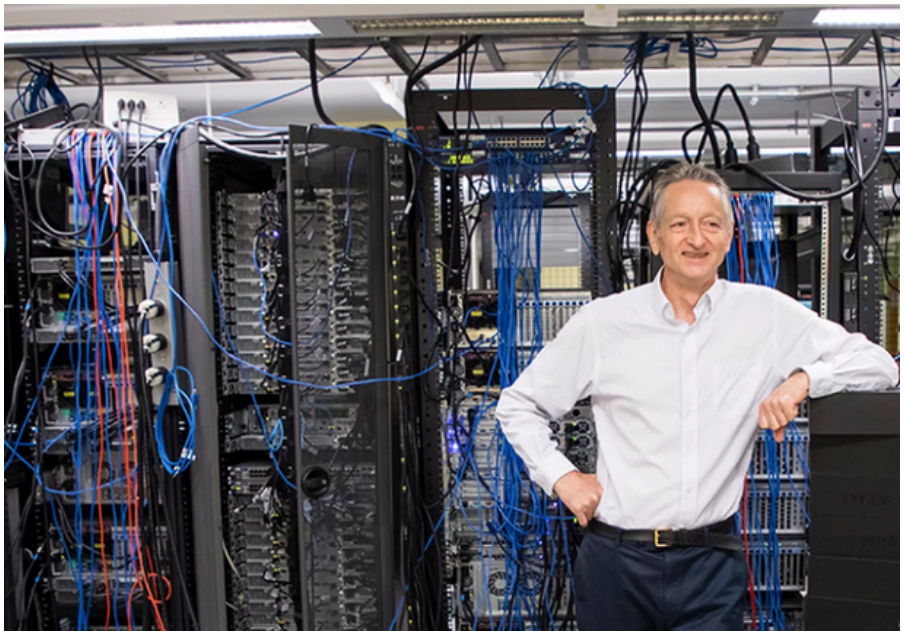
$$\mathbb{E}_S \mathbb{E}_{f\sim\rho}[R(f)] \leq \mathbb{E}_S \mathbb{E}_{f\sim\rho}[r(f)] + \sqrt{\frac{2\mathbb{E}_S KL(\rho\|\pi)}{n}}$$

Important ! This bound does not give a generalization certificate.

An easy problem : find the best neural network

You have one data set $\mathcal{S}$ that you will use as a test set, and two classifiers.

$r(f_1) = 0.15$
$R(f_1) = ?$



$r(f_2) = 0.01$
$R(f_2) = ?$

PAC-Bayes bound for classifier selection

More generally, M classifiers $f_1, \dots, f_M$:

- uniform prior : $\pi = \frac{1}{M} \sum_{i=1}^{M} \delta_{f_i}$
- $\hat{f} = \arg\min_f r(f)$ and $\rho = \delta_{\hat{f}}$

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{f \sim \rho}[R(f)] \leq \mathbb{E}_{\mathcal{S}} \mathbb{E}_{f \sim \rho}[r(f)] + \sqrt{\frac{2 \mathbb{E}_{\mathcal{S}} KL(\rho \| \pi)}{n}}$$

$$\mathbb{E}_{\mathcal{S}} R(\hat{f}) \leq \mathbb{E}_{\mathcal{S}}[\min_f r(f)] + \sqrt{\frac{2 \log(M)}{n}}$$

$$\mathbb{E}_{\mathcal{S}} R(\hat{f}) \leq \min_f R(f) + \sqrt{\frac{2 \log(M)}{n}}$$

Ask an undergrad student in statistics

Say $R(f_1) < R(f_2)$,

$$\begin{aligned}\mathbb{E}_{\mathcal{S}} R(\hat{f}) &= \mathbb{E}_{\mathcal{S}}\left[R(f_1)1_{\hat{f}=f_1} + R(f_2)1_{\hat{f}=f_2}\right] \\ &\leq \mathbb{E}_{\mathcal{S}}\left[R(f_1) + 1_{\hat{f}=f_2}\right] \\ &= \min_f R(f) + \mathbb{P}_{\mathcal{S}}[r(f_2) - r(f_1) < 0]\end{aligned}$$

and $r(f_2) - r(f_1) \rightsquigarrow \mathcal{N}\left(\Delta R, \frac{v}{n}\right)$ so

$$\mathbb{P}_{\mathcal{S}}[r(f_2) - r(f_1) < 0] \sim \Phi\left(\Delta R\sqrt{\frac{n}{v}}\right) \sim \frac{\exp\left(-\frac{n[\Delta R]^2}{v}\right)}{\Delta R\sqrt{2\pi\frac{n}{v}}},$$

$\Delta R = R(f_2) - R(f_1)$ and $v = R(f_2)[1 - R(f_2)] + R(f_1)[1 - R(f_1)] - 2\mathbb{P}(f_1(X) = f_2(X) \neq Y)$.

Which is the largest ?

Objective of this talk

- PAC-Bayes bounds (basic version) can be suboptimal in many ways.
- we will see that in the example above, the uniform prior leads to the catastrophe.
- we will discuss "prior improvement" ideas in theory and practice.

Contents

Contents

Optimizing with respect to the prior

In practice, popular choices :

- $\rho = \delta_{\hat{\theta}}$,
- $\rho(f) \propto \exp(-\lambda r(f))p(f)$
- ...

Once ρ is fixed, why not optimize with respect to π ?

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\rho}[R(f)] \leq \mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\rho}[r(f)] + \sqrt{\frac{2\mathbb{E}_{\mathcal{S}}KL(\rho\|\pi)}{n}}$$

$$\mathbb{E}_{\mathcal{S}}KL(\rho\|\pi) = \underbrace{\mathbb{E}_{\mathcal{S}}KL(\rho\|\mathbb{E}_{\mathcal{S}}\rho)}_{=:\mathcal{I}(\rho,\mathcal{S})} + \underbrace{KL(\mathbb{E}_{\mathcal{S}}\rho\|\pi)}_{=0 \text{ if } \pi=\mathbb{E}_{\mathcal{S}}\rho}$$

Catoni, O. (2007). *PAC-Bayesian supervised learning : the thermodynamics of statistical learning.* IMS lecture notes – monograph series.

Mutual information bound

The corresponding bound was re-discovered (independently).

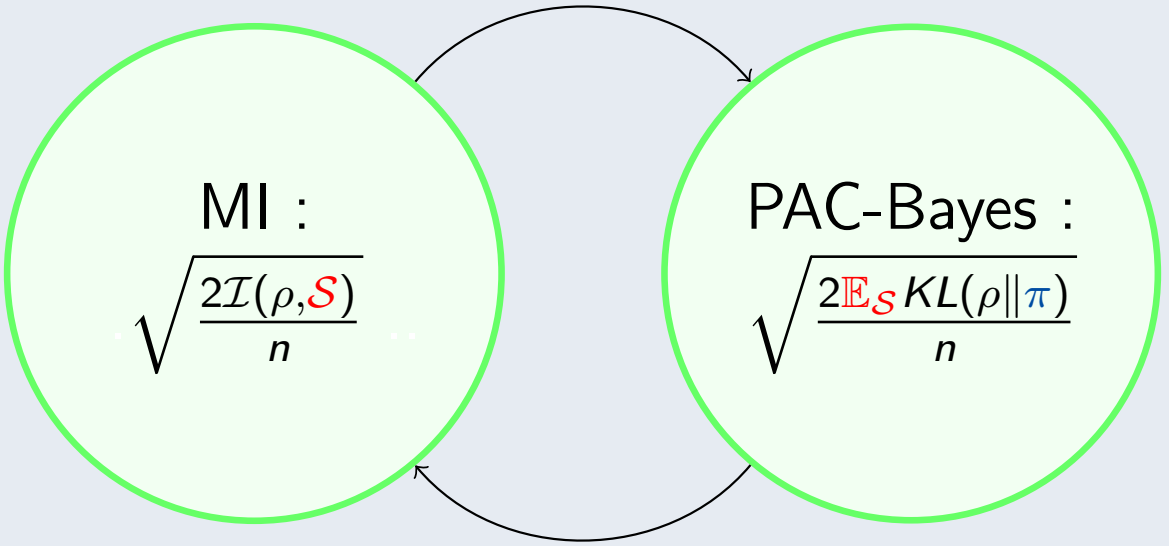
Mutual information bound

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\rho}[R(f)] \leq \mathbb{E}_{\mathcal{S}}\mathbb{E}_{f\sim\rho}[r(f)] + \sqrt{\frac{2\mathcal{I}(\rho,\mathcal{S})}{n}}$$

Russo, D. and Zou, J. (2019). How much does your data exploration overfit ? controlling bias via information usage. *IEEE Transactions on Information Theory*.

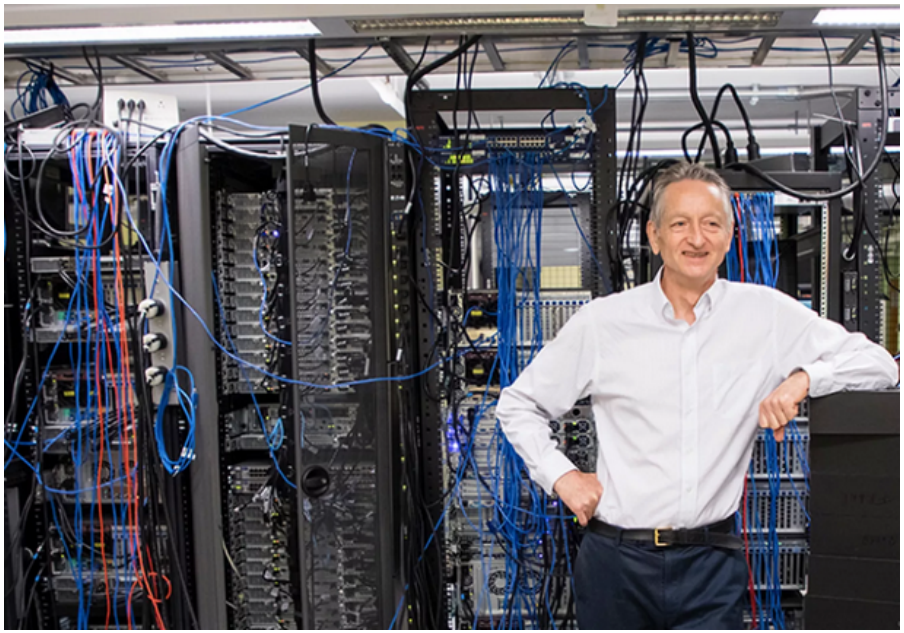
PAC-Bayes and MI bounds

$$\mathcal{I}(\rho, \mathcal{S}) = \mathbb{E}_{\mathcal{S}} KL(\rho \| \mathbb{E}_{\mathcal{S}} \rho) \leq \mathbb{E}_{\mathcal{S}} KL(\rho \| \pi)$$



Classifier selection

$r(f_1) = 0.15$



$r(f_2) = 0.01$

Application in the selection problem

Prior $\pi_\alpha(f) = \alpha\delta_{f_1} + (1-\alpha)\delta_{f_2}$.
Say $R(f_1) < R(f_2)$. For any α,

$$\mathbb{E}_S R(\hat{f}) \leq \min_f R(f) + \sqrt{\frac{2\mathbb{E}_S KL(\rho\|\pi_\alpha)}{n}}$$

$$= \min_f R(f) + \sqrt{\frac{2\mathbb{E}_S\left[1_{\hat{f}=f_1}\log\frac{1}{\alpha} + 1_{\hat{f}=f_2}\log\frac{1}{1-\alpha}\right]}{n}}$$

$$\leq \min_f R(f) + \sqrt{\frac{2\left[\log\frac{1}{\alpha} + \Phi\left(\frac{n\Delta R}{2v}\right)\log\frac{1}{1-\alpha}\right]}{n}}$$

Take $\alpha = \exp\left[-\Phi\left(\frac{n\Delta R}{2v}\right)\right]$...

Application in the selection problem

Theorem

In the case of M functions $f_1, \dots, f_M$, put

$$\Delta = \min_{i : R(f_i) \neq \min_f R(f)} R(f_i) - \min R(f).$$

Then

$$\mathbb{E}_S R(\hat{f}) \leq \min_f R(f) + \frac{16}{n\Delta} \log\left(1 + M\mathrm{e}^{-\frac{n\Delta^2}{32}}\right)$$

For $\Delta \simeq 1\sqrt{n}$ we recover the $\sqrt{\log(M)/n}$ rate...

Alquier, P. (2021). *User-friendly introduction to PAC-Bayes bounds*. Preprint arXiv.

Optimization of the prior : more cases

When $\rho(f) \propto \exp(-\lambda r(f))p(f)$, Catoni suggests to use the (almost optimal) "localized prior"

$$\pi_{-\beta R}(f) \propto \exp(-\beta R(f))p(f).$$

situation	uniform prior	localized prior
$\dim(\Theta) = d$	$\sqrt{\frac{d}{n}\log\frac{n}{d}}$	$\sqrt{\frac{d}{n}}$
(MA) $+ \dim(\Theta) = d$	$\frac{d}{n}\log\frac{n}{d}$	$\frac{d}{n}$

(MA) = margin assumption, includes noiseless classification

Alquier, P. (2021). *User-friendly introduction to PAC-Bayes bounds*. Preprint arXiv.

Contents

Towards tighter generalization certificates

- define an architecture : $\{f_w, w \in \mathcal{W}\}$.
- minimize (in ρ) a PAC-Bayes bound on $\mathbb{E}_{w\sim\rho}[R(f_w)]$.
- outcome :
  - the posterior ρ,
  - a numerical certificate for $\mathbb{E}_{w\sim\rho}[R(f_w)]$.

Dziugaite, G. K. and Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *UAI*.

Pérez-Ortiz, M., Rivasplata, O., Shawe-Taylor, J. and Szepesvári, C. (2020). *Tighter risk certificates for neural networks*. Preprint arXiv.

- (standard) PAC-Bayes bounds lead to vacuous certificates,
- propose many improvements to obtain tighter bounds.

Some ideas

Prior on the weights : $w \sim \mathcal{N}(w_0, \sigma I)$.

- w_0 chosen randomly = the initialization of the SGD,
- $\sigma = c \exp(-j/b)$, $j \in \mathbb{N}$ (with union bound).
- other option : sample splitting. Learn a prior on the first half, learn ρ via minimizing a PAC-Bayes bound on the second half only.
- etc.

Empirical bound with a localized prior

Theorem

Put

$$\pi_{-\beta r}(f) \propto \exp(-\beta r(f))p(f).$$

Fix $\lambda \in [0, n]$ and $\xi \in [0, 1)$. For any $\varepsilon > 0$,

$$\mathbb{P}_{\mathcal{S}}\left(\forall \rho, \ \mathbb{E}_{f\sim\rho}[R(f)] \right.$$
$$\left. \leq \frac{(1-\xi)\mathbb{E}_{f\sim\rho}[r(f)] + KL\left(\rho \| \pi_{-\xi\lambda r}\right) + (1+\xi)\log\frac{2}{\varepsilon}}{(1-\xi)\lambda + (1+\xi)\frac{\lambda^2}{n}} \right) \geq 1-\varepsilon$$

Catoni, O. (2003). *A PAC-Bayesian approach to adaptive classification*. Preprint LPMA.

Pros and cons of Catoni's localized bounds

$$\pi_{-\beta r}(f) \propto \exp(-\beta r(f))p(f).$$

- pros : we recover the rates in $\sqrt{d/n}$ and d/n above...
- cons :

$$KL\left(\rho \| \pi_{-\xi\lambda r}\right) = KL(\rho\|p) + \beta\mathbb{E}_{f\sim\rho}[r(f)] + \log \mathbb{E}_{f\sim p}\left[\exp(-\beta r(f))\right]$$

Should be tried in practice.

Future objective

On the "theoretical bounds" side, most issues with PAC-Bayes and MI bounds are solved in

Grünwald, P., Steinke, T. and Zakynthinou, L. (2021). *MAC-Bayes and Conditional Mutual Information : Fast rate bounds that handle general VC classes*. Preprint arXiv.

How much can we improve the numerical certificate via data-dependent priors ? Besides Catoni's bound we mentioned earlier, more recent works :

Parrado-Hernández, E., Ambroladze, A., Shawe-Taylor, J. and Sun., S. (2012). PAC-Bayes bounds with data dependent priors. *JMLR*.

Dziugaite, G. K. and Roy, D. M. (2018). Data-dependent PAC-Bayes priors via differential privacy. *NeurIPS*.