PAC-Bayesian Learning: A tutorial

Pascal Germain
www.pascalgermain.info

Université Laval, département d'informatique et de génie logiciel
Canada CIFAR AI Chair

Workshop PAC-Bayes meets Interactive Learning @ ICML 2023

Acknowledgment

This tutorial material has been developed in collaboration with Benjamin Guedj.
https://bguedj.github.io/

The Alan Turing Institute

Inria

The Inria-London Programme

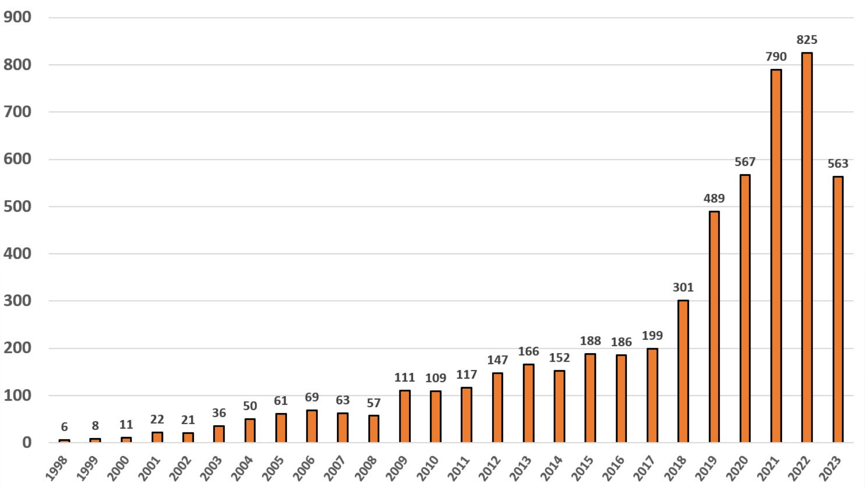
fbc FRANCO-BRITISH COUNCIL

UCL

Acknowledgment

This tutorial is greatly inspired by my mentor, François Laviolette.

PAC-Bayes Publications



Number of search results per year for “PAC-Bayes(ian)” keywords on Google Scholar.

Plan

Plan

What is PAC-Bayes?

- A statistical learning theory
- A frequentist approach with a Bayesian twist (notions of prior and posterior).
- A generic framework to (re)think generalisation of machine learning algorithms.

PAC-Bayes Theorems

High-confidence bounds on the generalization loss of a predictor/model obtained from its performance on the training sample.

- PAC-Bayes bounds are safety checks; numerical certificates!

PAC-Bayes Algorithms

Optimizing the PAC-Bayes bounds lead to self-certified learning algorithms.

- Numerous existing learning algorithms can be cast as PAC-Bayes ones, ...
- ... and new algorithms can be conceived this way!

Why PAC-Bayes?

- PAC-Bayes is modular:
  - Choose your own predictor/model, loss, data assumptions, etc.
- PAC-Bayes is inclusive:
  - Reconciliates Frequentists and Bayesians
  - Bridges machine learning and information theory
  - Welcomes both modeling cultures: data modeling and algorithmic modeling (Breiman 2001)
  - Offers a playground for those developing equations and those running experiments.
  - Adapts to many existing learning approaches, from boosting to deep neural networks
- Plus:
  - The proofs are (relatively) simple
  - The bounds can be tight (numerically non-vacuous)
  - Deriving self-certified learning algorithms is a noble and fun journey!

Plan

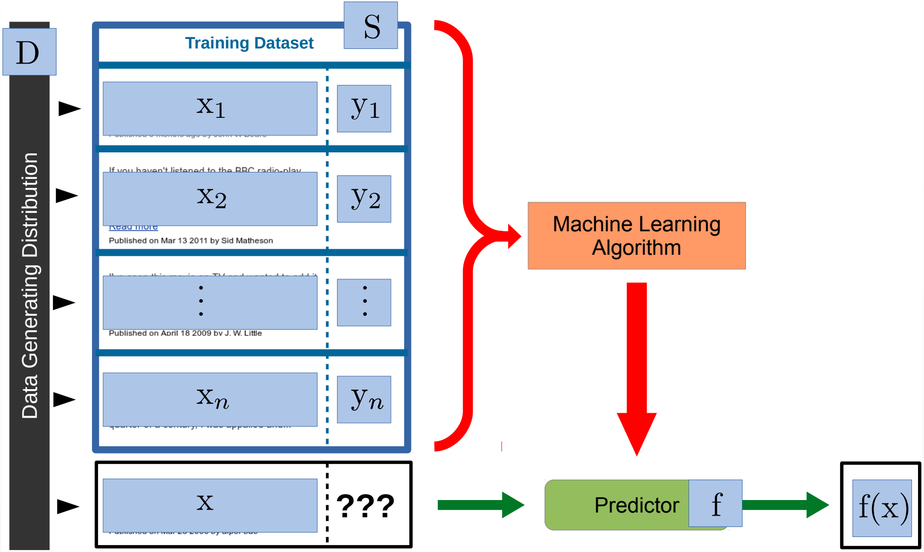
Historical landmarks

- **Pre-history: PAC analysis of Bayesian estimators** (Shawe-Taylor and Williamson 1997)
- **Birth: First PAC-Bayesian theorems** (McAllester 1998, 1999)
  - **Empirical bounds**
    - PAC-Bayes *kl* bound (Langford and Seeger 2001)
    - Neural Networks (Langford and Caruana 2001)
    - SVM & Margins (Langford and Shawe-Taylor 2002)
  - **Oracle bounds**
    - PAC-Bayes *tempered* bound, *localized* prior, link with mutual information, . . . (Catoni 2003, 2004, 2007)
  - **Self-certified learning algorithms**
    - "PAC-Bayesian learning of linear classifiers" (Germain, Lacasse, Laviolette, and Marchand 2009)
    - "Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks..." (Dziugaite and D. M. Roy 2017)
    - "Tighter Risk Certificates for Neural Networks" (Pérez-Ortiz et al. 2021)

Plan

Plan

Machine Learning: The Prediction Problem (non-interactive setting)



Definitions

A **learning example** $z := (x, y) \in \mathcal{Z}$ is a **description-label** pair.

Data generating distribution

Each example is an **observation from distribution** D on $\mathcal{Z}$.

Learning sample

$$S \;:=\; \{\, z_1,\; z_2,\; \ldots,\; z_n \,\} \;\sim\; D^n$$

Predictors (or hypothesis)

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

Learning algorithm

$$A(S) \longrightarrow h$$

Loss function

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

Empirical loss

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

Generalization loss

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

The Generalization Challenge

Goal: Minimize the generalization loss on D

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

The learning algorithm see *only* the **empirical loss** on S:

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

Plan

PAC (without Bayes) Learning

PAC guarantees (Probably Approximately Correct)

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

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

- Single hypothesis h (building block):

$$\mathcal{L}_D(h) \leq \widehat{\mathcal{L}}_S(h) + \sqrt{\tfrac{1}{2n}\log\left(\tfrac{1}{\delta}\right)}.$$

- Finite function class $\mathcal{H}$ (worst-case approach):

$$\forall h \in \mathcal{H}, \quad \mathcal{L}_D(h) \leq \widehat{\mathcal{L}}_S(h) + \sqrt{\tfrac{1}{2n}\log\left(\tfrac{|\mathcal{H}|}{\delta}\right)}$$

- Structural risk minimisation; hypotheses h_i associated with prior weight p_i:

$$\forall h_i \in \mathcal{H}, \quad \mathcal{L}_D(h_i) \leq \widehat{\mathcal{L}}_S(h_i) + \sqrt{\tfrac{1}{2n}\log\left(\tfrac{1}{p_i\delta}\right)}$$

- Uncountably infinite function class: VC dimension, Rademacher complexity...

Plan

PAC-Bayesian Learning

Classical PAC approaches are suited to analyze the performance of individual functions,
$\longrightarrow$ Extension: PAC-Bayes allows to consider *distributions* over hypotheses.

It tastes Bayesian...

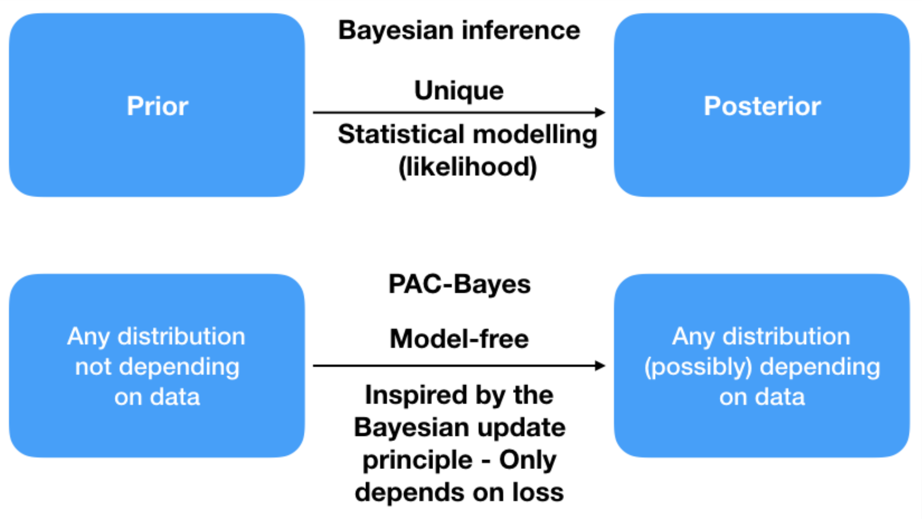
Given a **prior** distribution P on $\mathcal{H}$ and a **posterior** distribution Q on $\mathcal{H}$..

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

... but it's not!

- **Prior**
  - PAC-Bayes: bounds hold for any prior distribution
  - Bayes: prior choice impacts inference
- **Posterior**
  - PAC-Bayes: bounds hold for any posterior distribution
  - Bayes: posterior uniquely defined by prior and likelihood
- **Data**
  - PAC-Bayes: observations come from an unknown data distribution (*iid* assumption)
  - Bayes: observations are generated by a model from a specified family

PAC-Bayes bounds vs. Bayesian inference



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,*

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

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** as we can optimise for Q.

One can predict with...

- The "Maximum-A-Posteriori (MAP)" predictor:

$$MAP_Q(x) = h^* \text{ with } h^* = \operatorname*{argmax}_h(Q(h)).$$

- The (so-called) "Bayes" majority vote predictor (classification only):

$$B_Q(x) = \max_{y \in \mathcal{Y}} \left[\int_{\mathcal{H}} Q(h) I[h(x) = y] dh \right] \text{ with } h \sim Q.$$

- The (so-called) "Gibbs" stochastic predictor:

$$G_Q(x) = h(x) \text{ with } h \sim Q.$$

- The "Aggregated" predictor :

$$H_Q(x) = \int_{\mathcal{H}} Q(h) dh.$$

Plan

Plan

A General PAC-Bayesian Theorem

Δ-function: "distance" between $\mathbf{E}_{h\sim Q}\widehat{\mathcal{L}}_S(h)$ and $\mathbf{E}_{h\sim Q}\mathcal{L}_D(h)$

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

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

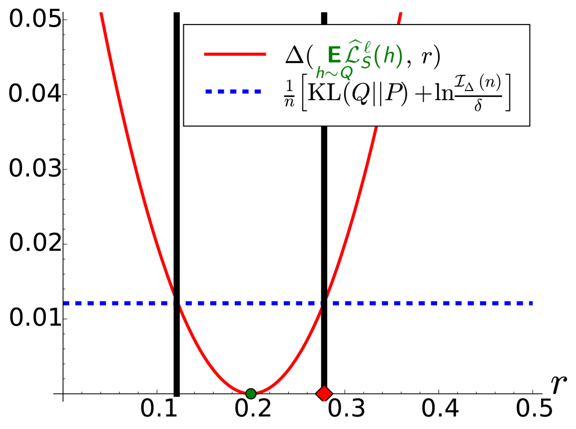
where

$$\mathcal{I}_\Delta(n) \ =\ \mathbf{E}_{h\sim P}\ \mathbf{E}_{S'\sim D^n}\ e^{n\cdot\Delta(\widehat{\mathcal{L}}_{S'}(h),\mathcal{L}_D(h))}$$

General theorem

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

Interpretation.



General theorem

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

Proof ideas.

Change of Measure Inequality (Donsker and Varadhan 1975; Csiszár 1975)

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

$$\Pr\left(X \leq \tfrac{\mathbf{E}\,X}{\delta}\right) \geq 1-\delta \quad \equiv \quad X \leq_{1-\delta} \tfrac{\mathbf{E}\,X}{\delta}\,.$$

See also the *Exponential Stochastic Inequality* $\trianglelefteq_\delta$
(proposed by Grünwald et al. 2023).

General theorem

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

Proof.

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

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

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

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

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

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

Plan

The linear case $\Delta_\lambda(q,p) := \frac{\lambda}{n}(p-q)$ (Alquier et al. 2016)

If the loss is bounded; $\forall h, z: \ \ell(h,z) \in [0,b]$:

$$\mathcal{I}_\Delta(n) \;=\; \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S'\sim D^n} e^{\lambda\cdot(\mathcal{L}_D(h)-\widehat{\mathcal{L}}_{S'}(h))} \underset{\text{(Hoeffding)}}{\leq} \mathop{\mathbf{E}}_{h\sim P} e^{\frac{\lambda^2 b^2}{2n}} = e^{\frac{\lambda^2 b^2}{2n}}$$

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

If the loss is sub-Gaussian; $\forall h, \lambda: \ \mathbf{E}_z\, e^{\lambda(\ell(h,z)-\mathcal{L}_D(h))} \leq e^{\frac{\lambda^2\sigma^2}{2n}}$:

$$\mathcal{I}_\Delta(n) \;=\; \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S'\sim D^n} e^{\lambda\cdot(\mathcal{L}_D(h)-\widehat{\mathcal{L}}_{S'}(h))} \leq \mathop{\mathbf{E}}_{h\sim P} e^{\frac{\lambda^2\sigma^2}{2n}} = e^{\frac{\lambda^2\sigma^2}{2n}}$$

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

The linear case $\Delta_\lambda(q,p) := \frac{\lambda}{n}(p-q)$

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

From an algorithm design perspective, linear "tempered bounds" promote the minimization of

$$\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S(h) + \frac{1}{\lambda}\mathrm{KL}(Q\|P)\,.$$

The *optimal Gibbs posterior* is given by (See Catoni 2007, Alquier et al. 2016,...)

$$Q^*(h) \;=\; \frac{1}{Z}P(h)\,e^{-\lambda\,\widehat{\mathcal{L}}_S(h)}\,.$$

where Z is a normalizing constant.

Tighter bounds for the $[0,1]$-loss (Classical PAC-Bayes theorems)

Corollary

With a bounded loss $\ell(h,z) \in [0,1]$:

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

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

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

$$\mathrm{kl}(q,p) = q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p} \geq 2(q-p)^2,$$

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

Tighter bounds for the [0, 1]-loss (Classical PAC-Bayes theorems)

$$\mathrm{kl}\Big(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S(h),\ \mathop{\mathbf{E}}_{h\sim Q}\mathcal{L}_D(h)\Big) \le \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{n}}{\delta}\right].$$

From an algorithm design perspective, the "kl bound" promotes the minimization of

$$\mathrm{kl}^{-1}\left(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S(h), \tfrac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\tfrac{2\sqrt{n}}{\delta}\right]\right) := \sup_{0\le p\le 1}\left\{p : \mathrm{kl}\Big(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S(h), p\Big) \le \tfrac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\tfrac{2\sqrt{n}}{\delta}\right]\right\}$$

The function kl^{-1} is differentiable (see Reeb et al. 2018)

pyTorch implementation (Viallard et al. 2021):
https://github.com/paulviallard/ECML21-PB-CBound/blob/master/core/kl_inv.py

Lemma (see Letarte, Germain, et al. 2019)

$$\mathrm{kl}^{-1}\left(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S(h), \tfrac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\tfrac{2\sqrt{n}}{\delta}\right]\right) = \inf_{c>0}\left\{\frac{1}{1-e^{-c}}\left(c\cdot \mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S(h) + \tfrac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\tfrac{2\sqrt{n}}{\delta}\right]\right)\right\}$$

Plan

Plan

Distribution over parameters

Given a model / predictor h_θ, where θ are parameters.

Consider P and Q as distributions over the set of parameters Θ.

$$\forall\, Q \text{ on } \Theta: \quad \mathrm{kl}\Big(\mathop{\mathbf{E}}_{\theta\sim Q} \widehat{\mathcal{L}}_S(h_\theta), \mathop{\mathbf{E}}_{\theta\sim Q} \mathcal{L}_D(h_\theta) \Big) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln \frac{2\sqrt{n}}{\delta} \right].$$

Typical approach for (stochastics) neural networks
(Dziugaite and D. M. Roy 2017; Neyshabur et al. 2018; Nozawa et al. 2020; Pérez-Ortiz et al. 2021, among many others.)

- $P = \mathcal{N}(\mathbf{W}_p, \sigma_p \mathbf{I})$ where $\mathbf{W}_p$ are the random/pre-learned weights initialization.
- $Q = \mathcal{N}(\mathbf{W}, \sigma \mathbf{I})$, where $\mathbf{W}$ are the learned/fine-tuned neural network weights.

Then, $\mathrm{KL}(Q\|P) = \frac{1}{2}\|\mathbf{W} - \mathbf{W}_p\|^2$.

Self-certified learning of neural networks (Pérez-Ortiz et al. 2021)

Pérez-Ortiz, Rivasplata, Shawe-Taylor and Szepesvári

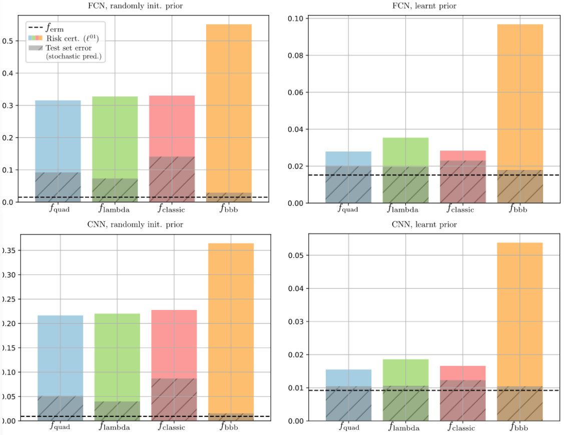


Figure 3: Tightness of the risk certificates for MNIST across different architectures, priors and training objectives. The bottom shaded areas correspond to the test set

- Build on the pioneer work of Dziugaite and D. M. Roy 2017.
- Tight guarantees!

risk $\leq 1.55\%$ **on MNIST (CNN)** with probability $\geq 95\%$.

- Easy to train.

Source code (pyTorch):
https://github.com/mperezortiz/PBB

Plan

Bayesian Learning (Zhang 2006, Grünwald 2012, Germain, Bach, et al. 2016, Masegosa et al. 2020)

Negative log-likelihood loss function

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

Bayesian Rule

For each $\theta \in \Theta$:

$$p(\theta|X,Y) \;=\; \frac{p(\theta)\, p(Y|X,\theta)}{p(Y|X)} \qquad \text{with} \quad \begin{array}{l} X = \{x_1,\ldots,x_n\} \\ Y = \{y_1,\ldots,y_n\} \end{array}$$

- $p(\theta|X,Y)$ is the *posterior* given X, Y (similar Q over $\mathcal{H}$)
- $p(\theta)$ is the *prior* (similar to P over $\mathcal{H}$)
- $p(Y|X,\theta)$ is the *likelihood* of the parameter θ given X, Y
- $p(Y|X) = \int_\Theta p(\theta)\, p(Y|X,\theta) d\theta$ is the *marginal likelihood* of the model at hand.

Then,

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

Plan

Mutual Information

Consider a learning algorithm that returns a distribution $Q(S)$ on $\mathcal{H}$ given $S{\sim}D^n$.

- Let $\theta \sim Q(S)$. Xu and Ragingsky (2017) showed that for sub-Gaussian losses:

$$\mathop{\mathbf{E}}_{S\sim D}\left|\mathop{\mathbf{E}}_{h\sim Q}\mathcal{L}_D(h) - \mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}_S(h)\right| \le \sqrt{\frac{2\sigma I(\theta,S)}{n}}\,,$$

  where $I(\theta, S)$ is the *mutual information* between the parameters and the train data.

- This is equivalent to a PAC-Bayesian bound *in expectation* (e.g., Alquier 2021):

$$\begin{aligned} I(\theta,S) &= \mathop{\mathbf{E}}_{S\sim D}\mathrm{KL}\left(Q(S)\middle\|P_D^*\right) && \text{for the } \textit{data-dependent prior } P_D^* := \mathop{\mathbf{E}}_{S\sim D} Q(S) \\ &\le \mathop{\mathbf{E}}_{S\sim D}\mathrm{KL}\left(Q(S)\middle\|P\right) && \text{for any prior } P. \end{aligned}$$

- Negrea et al. (2019) showed that *Stochastic Gradient Langevin Dynamics* (SGLD) minimizes a PAC-Bayes bound with a data-dependant prior P_D^*.

Plan

Some of our recent work

PAC-Bayesian learning of:

- *Aggregated* binary-activated neural networks (Letarte, Germain, et al. 2019; Biggs and Guedj 2021; Fortier-Dubois et al. 2023).
- Kernels, via a posterior distribution over random Fourier features (Letarte, Morvant, et al. 2019), and extension to contrastive learning (Letarte 2023, chapter 3).
- Wassertein GANs (Mbacke et al. 2023)

Plan

Other Recorded Video Tutorials

- Laviolette 2017: Tutorial on PAC-Bayesian Theory. https://youtu.be/GnRX9Pvw6Xw

  Part of the NeurIPS workshop "*(Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights*". https://bguedj.github.io/nips2017/

- Shawe-Taylor & Rivasplata 2018: Statistical Learning Theory - a Hitchhiker's Guide, https://youtu.be/m8PLzDmW-TY (NeurIPS tutorial)
- Guedj & Shawe-Taylor 2019: A Primer on PAC-Bayesian Learning. https://bguedj.github.io/icml2019/ (ICML tutorial)

Other Monographs

- Langford 2005: Tutorial on Practical Prediction Theory for Classification. http://www.jmlr.org/papers/v6/langford05a.html
- Catoni 2007: Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. https://arxiv.org/abs/0712.0248
- McAllester 2013: A PAC-Bayesian Tutorial with A Dropout Bound. https://arxiv.org/abs/1307.2118
- Van Erven 2014: PAC-Bayes Mini-tutorial: A Continuous Union Bound. https://arxiv.org/abs/1405.1580
- Germain, Lacasse, Laviolette, Marchand, and J.-F. Roy 2015: Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm http://jmlr.org/papers/v16/germain15a.html
- Guedj 2019: A Primer on PAC-Bayesian Learning. https://arxiv.org/abs/1901.05353
- Alquier 2021: User-friendly introduction to PAC-Bayes bounds. https://arxiv.org/abs/2110.11216
- Letarte 2023: PAC-Bayesian representation learning (PhD thesis). http://hdl.handle.net/20.500.11794/120163

Thank you!

References I

Alquier, Pierre (2021). "User-friendly introduction to PAC-Bayes bounds". In: *CoRR* abs/2110.11216.

Alquier, Pierre, James Ridgway, and Nicolas Chopin (2016). "On the properties of variational approximations of Gibbs posteriors". In: *J. Mach. Learn. Res.* 17, 239:1–239:41.

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*.

Biggs, Felix and Benjamin Guedj (2021). "Differentiable PAC–Bayes Objectives with Partially Aggregated Neural Networks". In: *Entropy* 23.10.

Breiman, Leo (2001). "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". In: *Statistical Science* 16.3, pp. 199–231.

Catoni, Olivier (2003). "A PAC-Bayesian approach to adaptive classification". In: *preprint LPMA* 840.

— (2004). *Statistical Learning Theory and Stochastic Optimization*. École d'Été de Probabilités de Saint-Flour 2001. Springer.

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

Csiszár, I. (1975). "I-divergence geometry of probability distributions and minimization problems". In: *Annals of Probability* 3, pp. 146–158.

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

References II

Dziugaite, Gintare Karolina and Daniel M. Roy (2017). "Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data". In: *UAI*. AUAI Press.

Fortier-Dubois, Louis, Gaël Letarte, Benjamin Leblanc, François Laviolette, and Pascal Germain (2023). "Learning Aggregations of Binary Activated Neural Networks with Probabilities over Representations". In: *AI*. Canadian Artificial Intelligence Association.

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://hdl.handle.net/20.500.11794/26130.

Germain, Pascal, Francis R. Bach, Alexandre Lacoste, and Simon Lacoste-Julien (2016). "PAC-Bayesian Theory Meets Bayesian Inference". In: *NIPS*, pp. 1876–1884.

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.

Grünwald, Peter (2012). "The Safe Bayesian - Learning the Learning Rate via the Mixability Gap". In: *ALT*.

Grünwald, Peter, Muriel Felipe Pérez-Ortiz, and Zakaria Mhammedi (2023). *Exponential Stochastic Inequality*. arXiv: 2304.14217 [math.ST].

Guedj, Benjamin (2019). "A Primer on PAC-Bayesian Learning". In: *CoRR* abs/1901.05353.

Langford, John (2005). "Tutorial on Practical Prediction Theory for Classification". In: *JMLR* 6.

Langford, John and Rich Caruana (2001). "(Not) Bounding the True Error". In: *NIPS*. MIT Press, pp. 809–816.

References III

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

Langford, John and John Shawe-Taylor (2002). "PAC-Bayes & Margins". In: *NIPS*.

Letarte, Gaël (2023). "PAC-Bayesian representation learning". PhD thesis. Université Laval. URL: http://hdl.handle.net/20.500.11794/120163.

Letarte, Gaël, Pascal Germain, Benjamin Guedj, and François Laviolette (2019). "Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks". In: *NeurIPS*, pp. 6869–6879.

Letarte, Gaël, Emilie Morvant, and Pascal Germain (2019). "Pseudo-Bayesian Learning with Kernel Fourier Transform as Prior". In: *AISTATS*. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 768–776.

Masegosa, Andrés R., Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin (2020). "Second Order PAC-Bayesian Bounds for the Weighted Majority Vote". In: *NeurIPS*.

Mbacke, Sokhna Diarra, Florence Clerc, and Pascal Germain (2023). "PAC-Bayesian Generalization Bounds for Adversarial Generative Models". In: *ICML*. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 1–9.

McAllester, David (1998). "Some PAC-Bayesian Theorems". In: *COLT*. ACM, pp. 230–234.

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

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

— (2013). "A PAC-Bayesian Tutorial with A Dropout Bound". In: *CoRR* abs/1307.2118.

Negrea, Jeffrey, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M. Roy (2019). "Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates". In: *NeurIPS*, pp. 11013–11023.

References IV

Neyshabur, Behnam, Srinadh Bhojanapalli, and Nathan Srebro (2018). "A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks". In: *ICLR (Poster)*. OpenReview.net.

Nozawa, Kento, Pascal Germain, and Benjamin Guedj (2020). "PAC-Bayesian Contrastive Unsupervised Representation Learning". In: *UAI*. Vol. 124. Proceedings of Machine Learning Research. AUAI Press, pp. 21–30.

Pérez-Ortiz, María, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári (2021). "Tighter Risk Certificates for Neural Networks". In: *J. Mach. Learn. Res.* 22, 227:1–227:40.

Reeb, David, Andreas Doerr, Sebastian Gerwinn, and Barbara Rakitsch (2018). "Learning Gaussian Processes by Minimizing PAC-Bayesian Generalization Bounds". In: *NeurIPS*, pp. 3341–3351.

Shawe-Taylor, John and Robert C. Williamson (1997). "A PAC Analysis of a Bayesian Estimator". In: *COLT*.

Van Erven, Tim (2014). "PAC-Bayes Mini-tutorial: A Continuous Union Bound". In: arXiv: `1405.1580 [stat.ML]`.

Viallard, Paul, Pascal Germain, Amaury Habrard, and Emilie Morvant (2021). "Self-bounding Majority Vote Learning Algorithms by the Direct Minimization of a Tight PAC-Bayesian C-Bound". In: *ECML/PKDD (2)*. Vol. 12976. Lecture Notes in Computer Science. Springer, pp. 167–183.

Xu, Aolin and Maxim Raginsky (2017). "Information-theoretic analysis of generalization capability of learning algorithms". In: *NIPS*, pp. 2524–2533.

Zhang, Tong (2006). "Information-theoretic upper and lower bounds for statistical estimation". In: *IEEE Trans. Information Theory* 52.4.