



HAL
open science

A Primer on PAC-Bayesian Learning

Benjamin Guedj

► **To cite this version:**

| Benjamin Guedj. A Primer on PAC-Bayesian Learning. 2019. hal-01983732v1

HAL Id: hal-01983732

<https://inria.hal.science/hal-01983732v1>

Preprint submitted on 16 Jan 2019 (v1), last revised 7 May 2019 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A PRIMER ON PAC-BAYESIAN LEARNING

by

Benjamin Guedj

Abstract. — Generalized Bayesian learning algorithms are increasingly popular in machine learning, due to their PAC generalization properties and flexibility. The present paper aims at providing a self-contained survey on the resulting PAC-Bayes framework and some of its main theoretical and algorithmic developments.

Contents

1. Introduction.....	2
2. Notation.....	4
3. Generalized Bayesian learning.....	6
4. The PAC-Bayesian theory.....	8
5. Algorithms: PAC-Bayes in the real world.....	12
6. Some recent breakthroughs in PAC-Bayes.....	15
7. Conclusion.....	18
References.....	18

2000 Mathematics Subject Classification. — 68Q32, 62C10, 60B10.

Key words and phrases. — PAC-Bayes, machine learning, statistical learning theory.

I would like to thank Agnès Desolneux and Mylène Maïda for giving me the opportunity to contribute to the 2nd SMF Congress in Lille, France. My deepest gratitude goes to colleagues, friends and co-authors who initiated or shared my interest in PAC-Bayes, with a special mention for Pierre Alquier, Olivier Catoni, Pascal Germain, Peter Grünwald, and John Shawe-Taylor.

1. Introduction

Artificial intelligence (AI) appears as the workhorse of a striking number of revolutions in several domains. As neurosciences, robotics, or ethics—to name but a few—shape new products, ways of living or trigger new digital rights, machine learning plays a more central role than ever in the rise of AI.

In the visionary words of Arthur Samuel ([Samuel, 1959](#)), machine learning is the field of study about computers' ability to learn without being explicitly programmed. As such, a long-term goal is to mimic the inductive functioning of the human brain, and most machine learning algorithms build up on statistical models to devise automatic procedures to infer general rules from data. This effort paved the way to a mathematical theory of learning, at the crossroads of computer science, optimization and statistics ([Shalev-Shwartz and Ben-David, 2014](#)). The interest in machine learning has been considerably powered by the emergence of the so-called big data era (an abundance of data collected, and the alignment of the corresponding required computing resources), and attempts at unifying these research efforts have shaped the emerging field of *data science*.

Among several paradigms, the present paper focuses on a Bayesian perspective to machine learning. As in "classical" Bayesian statistics literature, Bayesian machine learning is a principled way of managing randomness and uncertainty in machine learning tasks. Bayes reasoning is all about the shift from inferring unknown deterministic quantities to studying distributions (of which the previous deterministic quantities are just an instance), and has proven increasingly powerful in a series of applications. We refer to the monograph [Robert \(2007\)](#) for a thorough introduction to Bayesian statistics.

Over the past years, several authors have investigated extensions of the celebrated Bayes paradigm. While these extensions no longer abide by the canonical Bayesian rules and may be harder to interpret by practitioners, they have been enjoying a growing popularity and interest from the machine learning community.

As an illustration, consider a supervised learning problem, with a regression instance: $Y = f(X) + W$ where $X \in \mathbb{R}^d$ (input), $Y \in \mathbb{R}$ (output) and $W \in \mathbb{R}^d$ (noise) are random variables. A typical Bayesian inference procedure for f (unknown – may be parametric, semiparametric or nonparametric) would focus on the posterior distribution given by

$$(1) \quad \text{posterior}(f|X, Y) \propto \text{likelihood}(X, Y|f) \times \text{prior}(f).$$

Note that when $f(X) = f_\theta(X) = \theta X$ (with $\theta \in \mathbb{R}^d$), one recovers the classical linear regression model (typically worked out under a Gaussian assumption for the noise W). To improve the model's flexibility and ability to capture a larger spectrum of phenomena, it has been suggested by [Vovk \(1990\)](#) and [Zhang \(2006a\)](#) to replace the

likelihood by its *tempered* counterpart:

$$(2) \quad \text{target}(f|X, Y) \propto \text{likelihood}(X, Y|f)^\lambda \times \text{prior}(f),$$

where $\lambda \geq 0$ is a new parameter which controls the tradeoff between the *a priori* knowledge (given by the prior) and the data-driven term (the tempered likelihood). The resulting distribution (target) is no longer a proper posterior as the tempered likelihood may not be a proper probability density function. This tempered likelihood notion is at the core of the "safe Bayesian" paradigm (Grünwald, 2011, 2012, 2018; Grünwald and Van Ommen, 2017), where the temperature λ is integrated and marginalized out to yield more robust and automatic Bayesian inference procedures.

In machine learning, the emphasis on prediction ability is usually stronger than on inference (compared to the statistical literature). With that fact in mind, it is then only natural to go even further than the tempered likelihood: one can replace it by a purely arbitrary loss term, which only serves as a measure of the quality of prediction (*i.e.*, what loss do I suffer when predicting g instead of f in the previous example). This loss term is typically driven by information-theoretic arguments and therefore, substituting a loss term to the likelihood term achieves the shift from a model-based procedure to a purely data-driven procedure (which could arguably be described as *model-free*). Purely data-driven or model-free procedures may not assume an underlying probabilistic model to be inferred, but rather focus on an agnostic measure of performance.

Extensions such as tempered likelihoods or loss terms are bundled under the term *generalized Bayes*.

PAC-Bayesian inequalities were introduced by McAllester (1999a,b) and further formalized by McAllester (2003a,b), based on earlier remarks by Shawe-Taylor and Williamson (1997). The goal was to produce PAC-type performance bounds (in the sense of a loss function) for Bayesian estimators – the term PAC-Bayes now refers to the theory delivering PAC bounds for *generalized* Bayesian algorithms (wether with a tempered likelihood or a loss term).

The acronym PAC stands for Probably Approximately Correct and may be traced back to Valiant (1984). A PAC inequality states that with an arbitrarily high probability (hence "probably"), the performance (as provided by a loss function) of a learning algorithm is upper-bounded by a term decaying to an optimal value as more data is collected (hence "approximately correct"). When applied to a Bayesian (or rather generalized Bayesian) learning algorithm, the theory is referred to as PAC-Bayesian. PAC-Bayes has proven over the past two decades to be a principled machinery to address learning problems in a striking variety of situations (sequential or batch learning, dependent or heavy-tailed data, etc.), and is now quickly re-emerging as a powerful

and relevant toolbox to derive theoretical guarantees on the most recent learning topics, such as deep learning with neural networks or domain adaptation.

The rest of the paper is organized as follows. [Section 2](#) introduces our notation, while [Section 3](#) presents in more details the generalized Bayesian paradigm. [Section 4](#) contains a self-contained presentation of the PAC-Bayesian theory. [Section 5](#) focuses on several practical implementations of PAC-Bayes and [Section 6](#) illustrates the use of the PAC-Bayesian theory in several learning paradigms. [Section 7](#) draws perspectives and open problems.

2. Notation

Generalized Bayes and the PAC-Bayesian theory have been successfully used in a variety of topics, including sequential learning ([Gerchinovitz, 2011](#); [Li et al., 2018](#)), dependent or heavy-tailed data ([Alquier and Guedj, 2018](#); [Ralaivola et al., 2010](#); [Seldin et al., 2012](#)), classification ([Lacasse et al., 2007](#); [Langford and Shawe-Taylor, 2003](#); [Parrado-Hernández et al., 2012](#)) and many others (see also [Section 6](#)). To keep notation simple and still bear a fair amount of generality, we consider a simplified setting – let us stress however that results mentioned in this paper have been obtained in far more complex settings.

Let us assume that data comes in the form of pairs $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ where each (X_i, Y_i) is a copy of some random variable $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ whose underlying distribution is denoted by \mathbb{P} . The goal is to build a functional object $\hat{\phi}$ (which depends on \mathcal{D}_n) such that for any new query X' , $\hat{\phi}(X) \approx Y'$ in a certain sense. In other words, learning is to be able to generalize to unseen data: this remark leads to *generalization bounds*, also referred to as risk bounds, which are presented in [Section 4](#).

To assess this ability, we resort to a loss function $\ell: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$. Popular loss functions are the squared loss $\ell: (a, b) \mapsto (a - b)^2$, absolute loss $\ell: (a, b) \mapsto |a - b|$, 0-1 loss $\ell: (a, b) \mapsto \mathbb{1}[a \neq b]$, and so on. We then let

$$(3) \quad R: \hat{\phi} \mapsto \mathbb{E} \left[\ell \left(\hat{\phi}(X), Y \right) \right]$$

define the *risk* of the learning algorithm $\hat{\phi}$ (where the expectation is taken with respect to the underlying distribution of the data \mathbb{P}). As this underlying distribution is obviously unknown, the risk is not computable and is replaced by its empirical counterpart

$$(4) \quad r_n: \hat{\phi} \mapsto \frac{1}{n} \sum_{i=1}^n \ell \left(\hat{\phi}(X_i), Y_i \right).$$

As

$$\mathbb{E} \left[r_n \left(\hat{\phi} \right) \right] = R \left(\hat{\phi} \right),$$

we will see in [Section 4](#) that obtaining PAC inequalities involves a detailed study of how r_n concentrates to its mean R . Concentration inequalities such as Hoeffding's or Bernstein's are the key ingredient: we refer to the monograph [Boucheron et al. \(2013\)](#) for a thorough overview of concentration inequalities.

Let us now focus on the case where $\hat{\phi}$ is a Bayesian predictor. The learner $\hat{\phi}$ may be of parametric, semiparametric, or nonparametric nature: in any case, a Bayesian approach would consider a prior distribution on such $\hat{\phi}$: let us denote such a distribution π_0 . Let us emphasise here that this prior operates on the collection of candidate learners $\mathcal{F} = \{f: \mathbb{R}^d \rightarrow \mathbb{R}\}$, or rather on a subspace \mathcal{F}_0 of it (*e.g.*, all linear functions from \mathbb{R}^d to \mathbb{R}). A rich literature on model selection (either Bayesian or frequentist) studies refined inference techniques: see the monograph [Massart \(2007\)](#) for a solid introduction. A model assumption then allows to write a likelihood probability density function \mathcal{L} and form the posterior distribution of the model π :

$$(5) \quad \pi(\hat{\phi}|\mathcal{D}_n) \propto \mathcal{L}(\mathcal{D}_n|\hat{\phi}) \times \pi_0(\hat{\phi}).$$

Popular inference techniques would then be derived from the posterior. For example, the mean of the posterior

$$\hat{\phi}^{\text{mean}} = \mathbb{E}_\pi \phi = \int_{\mathcal{F}_0} \phi \pi(d\phi),$$

its median

$$\hat{\phi}^{\text{median}} = \text{median}(\pi),$$

the maximum a posteriori (MAP)

$$\hat{\phi}^{\text{MAP}} \in \arg \max_{\phi \in \mathcal{F}_0} \pi(\phi),$$

or a single realization

$$\hat{\phi}^{\text{draw}} \sim \pi,$$

are all popular choices. The actual implementation of such predictors is discussed in [Section 5](#). Theoretical results on Bayesian learning algorithms typically involve a thorough study of the way the posterior distribution concentrates as more data is collected. We refer the reader to the seminal papers [Ghosal et al. \(2000, iid case\)](#) and [Ghosal and Van Der Vaart \(2007, non-iid case\)](#). While Bayesian learning is a well established framework and is supported by theoretical and practical successes, a legitimate criticism is that its performance (both theoretical and practical) actually massively depends on the statistical modelling induced by the choice of the likelihood, and corresponding assumptions (additive Gaussian noise, iid data, bounded functional, etc.). As famously stated by George Box⁽¹⁾, all modelling efforts form a subjective and constrained vision of the underlying phenomenon, which may prove herself of poor quality, if any. The past few decades have thus seen an increasing

⁽¹⁾"Essentially, all models are wrong, but some are useful" (1976).

gap between the Bayesian statistical literature, and the machine learning community embracing the Bayesian paradigm – for which the Bayesian probabilistic model was too much of a constraint and had to be toned down in its influence over the learning mechanism. This movement gave rise to a series of works which laid down the extensions of Bayesian learning which are discussed in the next section.

3. Generalized Bayesian learning

A first strategy consists in modulating the influence of the likelihood term, by considering a tempered version of it: from (5), the posterior now becomes the tempered posterior π_λ :

$$(6) \quad \pi_\lambda(\hat{\phi}|\mathcal{D}_n) \propto \mathcal{L}(\mathcal{D}_n|\hat{\phi})^\lambda \times \pi_0(\hat{\phi}),$$

where $\lambda \geq 0$. The Bayesian model is now a particular case ($\lambda = 1$) of a continuum of distributions. Different values for λ will achieve different tradeoffs between the prior π_0 and the tempered likelihood \mathcal{L}^λ . Let us stress here that \mathcal{L}^λ may no longer correspond to a canonical probabilistic model.

This notion of tempered likelihood has been investigated, among others, by a series of paper (Grünwald, 2011, 2012, 2018; Grünwald and Van Ommen, 2017) which develop a "safe Bayesian" framework. As such, it is more intended as a way to "repair" the Bayesian paradigm in settings where the Bayesian model fails to capture the underlying phenomenon, rather than to replace it.

We rather focus on a second strategy which falls within generalized Bayes. We resort to an information-theoretic framework (see Csiszár and Shields, 2004, for an introduction) in which the "likelihood" of a learner $\hat{\phi}$ is no longer assessed by the probability mass from some specified model, but rather by the loss encountered when predicting $\hat{\phi}(X)$ instead of Y , the actual output value we wish to predict.

In other words, the posterior from (5) or the tempered posterior from (6) now amount to the generalized posterior

$$(7) \quad \pi_\lambda(\hat{\phi}|\mathcal{D}_n) \propto \ell_{\lambda,n}(\hat{\phi}) \times \pi_0(\hat{\phi}),$$

where $\ell_{\lambda,n}$ is a loss term measuring the quality of the learner $\hat{\phi}$ on the collected data \mathcal{D}_n (the training data, on which $\hat{\phi}$ is built upon). In particular, $\ell_{\lambda,n}$ would typically involve the empirical risk r_n .

Note that the distribution defined by (7) may no longer be a proper probability distribution. In particular, the loss term may not sum up to 1, hence the term "generalized". As the loss term is merely an instrument to guide oneself towards better performing algorithms, the generalized Bayesian framework may be described as model-free, as no such assumption is required. Other terms appear in the statistical

and machine learning literature, with occurrences of "generalized posterior", "pseudo-posterior" or "quasi-posterior" succeeding one another. Similarly, the terms "prior" and "posterior" have been consistently used as they "surcharge" the existing terms in Bayesian statistics, however the distributions in (7) are now different objects. Consider for example the prior π_0 : rather than incorporating prior knowledge, π_0 serves as an exploration mechanism of the candidates space \mathcal{F}_0 .

From (7), the story goes on as in Bayesian learning: any mechanism yielding a learner from the generalized posterior is admissible. As above, the mean, median, realization or mode (MAP) are popular choices.

Among all possible loss functions $\ell_{\lambda,n}$, a most typical choice is the so-called Gibbs potential, or Gibbs measure:

$$(8) \quad \pi_\lambda \left(\widehat{\phi} | \mathcal{D}_n \right) \propto \exp \left[-\lambda r_n \left(\widehat{\phi} \right) \right] \times \pi_0 \left(\widehat{\phi} \right).$$

The loss term now exponentially penalises the performance of a learner $\widehat{\phi}$ on the training data, and the temperature parameter $\lambda \geq 0$ controls the tradeoff between the prior term and the loss term. Let us examine both extremes cases: when $\lambda = 0$, the loss term vanishes and the generalized posterior amounts to the prior: data is not taken into account. When $\lambda \rightarrow \infty$, the influence of data is overwhelming. With the normalization constant, one can see that π_λ reduces to a Dirac measure peaked in the learner(s) which minimizes the empirical risk:

$$\pi_\lambda \propto \mathbb{1}_{\{\arg \min_\phi r_n(\phi)\}}.$$

In other words, the generalized Bayesian learner reduces to the celebrated empirical risk minimizer (ERM, see [Vapnik, 1995](#), for a survey on statistical learning theory).

The Gibbs measure arises in several contexts in statistics and statistical physics: we now provide a variational perspective to its existence. Let (A, \mathcal{A}) denote a measurable space and consider μ, ν two probability measures on (A, \mathcal{A}) . We note $\mu \ll \nu$ when μ is absolutely continuous with respect to ν , and we let $\mathcal{M}_\nu(A, \mathcal{A})$ denote the space of probability measures on (A, \mathcal{A}) which are absolutely continuous with respect to ν :

$$\mathcal{M}_\nu(A, \mathcal{A}) = \{\mu: \mu \ll \nu\}.$$

We denote by \mathcal{K} the Kullback-Leibler divergence between two probability measures:

$$\mathcal{K}(\mu, \nu) = \begin{cases} \int_{\mathcal{F}_0} \log \left(\frac{d\mu}{d\nu} \right) d\mu & \text{when } \mu \ll \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

Let us consider the classical squared loss $\ell: (a, b) \mapsto (a - b)^2$, and the optimization problem

$$(9) \quad \arg \inf_{\mu \in \mathcal{M}_{\pi_0}(A, \mathcal{A})} \left\{ \int_{\mathcal{F}_0} r_n(\phi) \mu(d\phi) + \frac{\mathcal{K}(\mu, \pi_0)}{\lambda} \right\}.$$

This problem amounts to minimizing the integrated (with respect to a generalized posterior) empirical risk plus a divergence term between the generalized posterior and the prior. In other words, minimizing a criterion of performance plus a divergence from the initial distribution, which is the exact analogous of penalized regression (such as Lasso).

One can easily deduce from the Karush-Kuhn-Tucker (KKT) conditions that the Gibbs measure π_λ in (8) is the only solution to the problem (9) (see for example [Rigollet and Tsybakov, 2012](#)).

Let us also stress that the Gibbs measure arises in other domains of statistics. Consider the case where the set of candidates \mathcal{F}_0 is finite. The mean of the Gibbs measure is given by

$$\begin{aligned} \widehat{\phi}^{\text{mean}} &:= \mathbb{E}_{\pi_\lambda} \phi = \int_{\mathcal{F}_0} \phi \pi_\lambda(d\phi) \\ &= \int_{\mathcal{F}_0} \phi \exp[-\lambda r_n(\phi)] \pi_0(d\phi) \\ &= \sum_{i=1}^{\#\mathcal{F}_0} \underbrace{\frac{\exp[-\lambda r_n(\phi_i)] \pi(\phi_i)}{\sum_{j=1}^{\#\mathcal{F}_0} \exp[-\lambda r_n(\phi_j)] \pi(\phi_j)}}_{\omega_{\lambda,i}} \phi_i = \sum_{i=1}^{\#\mathcal{F}_0} \omega_{\lambda,i} \phi_i, \end{aligned}$$

which is the celebrated exponentially weighted aggregate (EWA, see for example [Leung and Barron, 2006](#)). EWA forms a convex weighted average of predictors, where each predictor has a weight which exponentially penalizes its performance on the training data. Statistical aggregation ([Nemirovski, 2000](#)) may thus be revisited as a special case of generalized Bayesian posterior distributions (as studied in [Guedj, 2013](#)).

4. The PAC-Bayesian theory

The PAC learning framework has been initiated by [Valiant \(1984\)](#) and has been at the core of a great number of breakthroughs in statistical learning theory. In its simplest form, a PAC inequality states, for any learner $\widehat{\phi}$ and any $\epsilon > 0$

$$(10) \quad \mathbb{P} \left[R(\widehat{\phi}) \leq \delta \right] \geq 1 - \epsilon,$$

where δ is a threshold usually depending on data and ϵ . These risk bounds are of central importance in statistical learning theory as they give crucial guarantee on the performance of learners, with an upper-bound and a confidence level ϵ which can be made arbitrarily small. When a matching lower bound is found, the learner $\widehat{\phi}$ is said to be minimax optimal ([Tsybakov, 2003](#), and references therein). Note that in the original definition from [Valiant \(1984\)](#), the acronym PAC was used to refer to any bound valid with arbitrarily high probability together with the constraint that the

learner must be calculable in polynomial time with respect to n and $1/\epsilon$. The acronym now has a broader meaning as it covers any risk bound holding with arbitrarily high probability.

PAC-Bayesian inequalities date back to [Shawe-Taylor and Williamson \(1997\)](#) and [McAllester \(1999a,b\)](#). Original McAllester's PAC-Bayesian inequalities are empirical bounds, in the sense that the upper bound only depends on known quantities linked to the data.

For any measure $\mu \in \mathcal{M}_{\pi_0}(A, \mathcal{A})$, any learner $\hat{\phi}$ derived from μ and any $\epsilon > 0$, a simple form of McAllester's bound is given by

$$(11) \quad \mathbb{P} \left[R(\hat{\phi}) \leq r_n(\hat{\phi}) + \sqrt{\frac{\mathcal{K}(\mu, \pi_0) + \log \frac{1}{\epsilon}}{n}} \right] \geq 1 - \epsilon.$$

The Kullback-Leibler term $\mathcal{K}(\mu, \pi_0)$ captures the complexity of the candidates space \mathcal{F}_0 . In the simple case where \mathcal{F}_0 is the set of linear functions, this term basically reduces to the intrinsic dimension d . More favorable regimes (of order $\log d$, under a sparsity assumption) have been obtained in the recent literature (see [Guedj, 2013](#), for a survey).

This kind of bounds yield guarantees on the ("true") quality of the learner $\hat{\phi}$, with no need to evaluate or estimate its performance on some test data. This is a salient advantage of the PAC-Bayesian approach, as labelling and / or collecting test data might be cumbersome in some settings. Another key asset is that bounds of the form (11) are natural incentives to design new learning algorithms as minimizers of the right-hand side term (see [Germain, 2015](#), for a discussion). By integrating out the whole expression over $\hat{\phi}$, the constrained problem (9) appears once again and the Gibbs measure is deduced as the natural optimal generalized posterior distribution. McAllester's bounds have been improved by [Seeger \(2002, 2003\)](#).

While of great practical use, McAllester's bounds did not hint about the rate of convergence of learners, due to their empirical nature. [Catoni \(2004, 2007\)](#) therefore extended McAllester's PAC-Bayesian bounds to prove oracle-type inequalities, specifically on aggregated learners (typically the mean of the Gibbs measure - see also [Tsybakov, 2003](#), [Yang, 2003](#), and [Yang, 2004](#) for earlier works on aggregation and oracle inequalities in other settings than PAC-Bayes).

Catoni's technique consists of two ingredients:

1. A deviation inequality is used to upper bound the distance between $R(\widehat{\phi})$ and its empirical counterpart $r_n(\widehat{\phi})$ for a fixed $\widehat{\phi} \in \mathcal{F}_0$. In most of the rich PAC-Bayesian literature which followed Catoni's work, inequalities such as Bernstein's, Hoeffding's, Hoeffding-Azuma's or Bennett's have been used. More details can be found about these inequalities in the monographs [Massart \(2007\)](#) and [Boucheron et al. \(2013\)](#).
2. Then, the resulting bound is made valid for any $\widehat{\phi} \in \mathcal{F}_0$ simultaneously. Catoni suggests to consider the set of all probability distributions on \mathcal{F}_0 equipped with some suitable σ -algebra and make the deviation inequality uniform on this set with the following variational formula, presented in [Lemma 1](#).

Lemma 1 ([Csiszár, 1975](#) ; [Donsker and Varadhan, 1976](#) ; [Catoni, 2004](#))

Let (A, \mathcal{A}) be a measurable space. For any probability ν on (A, \mathcal{A}) and any measurable function $h : A \rightarrow \mathbb{R}$ such that $\int (\exp \circ h) d\nu < \infty$,

$$\log \int (\exp \circ h) d\nu = \sup_{\mu \in \mathcal{M}_\nu(A, \mathcal{A})} \left\{ \int h d\mu - \mathcal{K}(\mu, \nu) \right\},$$

with the convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of ν , the supremum with respect to μ on the right-hand side is reached for the Gibbs distribution g given by

$$\frac{dg}{d\nu}(a) = \frac{\exp \circ h(a)}{\int (\exp \circ h) d\nu}, \quad a \in A.$$

Proof. — Let $\mu \in \mathcal{M}_\nu(A, \mathcal{A})$. Since $\mathcal{K}(\cdot, \cdot)$ is non-negative, $\mu \mapsto -\mathcal{K}(\mu, g)$ reaches its supremum (equal to 0) for $\mu = g$. Then

$$\begin{aligned} -\mathcal{K}(\mu, g) &= - \int \log \left(\frac{d\mu}{d\nu} \frac{d\nu}{dg} \right) d\mu \\ &= - \int \log \left(\frac{d\mu}{d\nu} \right) d\mu + \int \log \left(\frac{dg}{d\nu} \right) d\mu \\ &= -\mathcal{K}(\mu, \nu) + \int h d\mu - \log \int (\exp \circ h) d\nu. \end{aligned}$$

Taking the supremum on all μ yields the desired result:

$$\log \int (\exp \circ h) d\nu = \sup_{\mu \in \mathcal{M}_\nu(A, \mathcal{A})} \left\{ \int h d\mu - \mathcal{K}(\mu, \nu) \right\},$$

□

Note that the second step in Catoni's technique requires to fix a reference measure ν on \mathcal{F}_0 . The reference measure is used to control the complexity of the candidates space \mathcal{F}_0 , however it kept being referred to as the "prior" to consistently extend the Bayesian setting (this has been long known in the literature, as pointed out by [Germain et al., 2016a](#)). [Catoni \(2007\)](#) also makes connections with information theory and

Rissanen's Minimum Description Length (MDL) principle (see Grünwald, 2007, for a solid introduction, and Zhang, 2006b, for the corresponding lower bounds). Other links have been studied between Catoni's bounds and generic chaining (Audibert and Bousquet, 2007) and fast rates (Audibert, 2009).

In Lemma 1, taking $\nu = \pi_0$ and $h = -\lambda r_n$ yields

$$(12) \quad \frac{1}{\lambda} \log \int \exp[-\lambda r_n(\phi)] \pi_0(d\phi) = \inf_{\mu \in \mathcal{M}_{\pi_0}(A, \mathcal{A})} \left\{ \int r_n(\phi) \mu(d\phi) + \frac{\mathcal{K}(\mu, \pi_0)}{\lambda} \right\}.$$

We therefore obtain the following empirical bound (see for example Audibert, 2004, Alquier, 2006, Alquier, 2008 and Guedj, 2013).

Theorem 1. — *Assume that the loss ℓ is convex and upper bounded by some constant C . Consider the Gibbs measure defined in (8) and let $\hat{\phi}^{\text{mean}} = \mathbb{E}_{\pi_\lambda} \phi$. Then for any $\epsilon > 0$,*

$$\mathbb{P} \left[R(\hat{\phi}^{\text{mean}}) \leq \inf_{\mu \in \mathcal{M}_{\pi_0}(A, \mathcal{A})} \left\{ \int r_n(\phi) \mu(d\phi) + (1 + C) \sqrt{\frac{\mathcal{K}(\mu, \pi_0) + \log \frac{1}{\epsilon}}{n}} \right\} \right] \geq 1 - \epsilon.$$

From Theorem 1, we can deduce the following oracle risk bound, which ensures that the integrated risk of learners sampled from the Gibbs measure is upper bounded, up to a complexity term which decays as $1/\sqrt{n}$, by the best (*i.e.*, smallest) integrated risk.

Theorem 2. — *Under the same assumptions as Theorem 1, for any $\epsilon > 0$,*

$$\mathbb{P} \left[\int R(\phi) \pi_\lambda(d\phi) \leq \inf_{\mu \in \mathcal{M}_{\pi_0}(A, \mathcal{A})} \left\{ \int R(\phi) \mu(d\phi) + (2 + C) \sqrt{\frac{\mathcal{K}(\mu, \pi_0) + \log \frac{1}{\epsilon}}{n}} \right\} \right] \geq 1 - \epsilon.$$

Similar results have been obtained by Dalalyan and Tsybakov (2008), Alquier and Lounici (2011), Alquier and Biau (2013), Guedj and Alquier (2013) to derive PAC-Bayesian oracle inequalities for several sparse regression models. The key is to devise a prior π_0 which enforces sparsity, *i.e.*, gives larger mass to elements $\hat{\phi} \in \mathcal{F}_0$ of small dimension (with respect to the sample size n).

The PAC-Bayesian theory consists in producing PAC risk bounds (either empirical or oracle) of generalized Bayesian learning algorithms.

A slightly different line of work has also investigated similar results, holding in expectation rather than with high probability. While obviously weaker, such results have proven important in dealing with some settings (*e.g.*, with unbounded losses). Following a method initiated by Leung and Barron (2006), Dalalyan and Tsybakov (2007, 2008) replaced the first step (the deviation inequality) with Stein's formula.

This technique was further investigated and improved in a series of papers (Alquier and Guedj, 2017; Dalalyan and Tsybakov, 2012a; Rigollet and Tsybakov, 2012).

Most refined PAC-Bayesian oracle bounds exhibit faster rates of convergence than $\mathcal{O}(1/\sqrt{n})$. Indeed, rates of magnitude $\mathcal{O}(1/n)$ have been obtained by Audibert (2009); Audibert and Catoni (2011); Dinh et al. (2016); Grünwald and Mehta (2016); van Erven et al. (2015).

5. Algorithms: PAC-Bayes in the real world

In conclusion, the PAC-Bayesian framework enjoys some of the strongest theoretical guarantees in machine learning, in the form of (often minimax optimal) oracle generalization bounds. However, the practical use of PAC-Bayes might be a computational challenge when facing complex, high-dimensional data. As a matter of fact, PAC-Bayes faces the exact same issues as Bayesian learning, as in both cases one is often required to sample from a possibly complex distribution. In Bayesian learning, sampling from the posterior; in PAC-Bayes, sampling from the generalized posterior. Let us focus on the Gibbs measure, as it is one of the most popular choices in PAC-Bayes. As in Bayesian learning, we often resort to a d -dimensional projection of the learner $\hat{\phi}$ or its development onto a functional basis (up to term K , for example). Monte Carlo Markov Chains (MCMC) are the top choice for sampling from such a distribution. We refer to Andrieu et al. (2003) and Robert (2007) for an introduction to this (rich) topic, and to Bardenet et al. (2016) for a survey on most recent techniques for massive datasets.

Let us remind the reader that the goal is to sample from the Gibbs measure, given by

$$\pi_\lambda(\hat{\phi}|\mathcal{D}_n) \propto \exp[-\lambda r_n(\hat{\phi})] \times \pi_0(\hat{\phi}).$$

The analytical form of this distribution is known (as the prior π_0 , the loss ℓ and the temperature λ are chosen). Three main techniques have been investigated in the literature.

1. The most popular one, by far, is MCMC. A naive pick is a Metropolis-Hastings algorithm (see Algorithm 1). However, due to the possibly high dimensionality of the generalized posterior π_λ , a nested model strategy coupled with a transdimensional MCMC is often a much better choice (as it could avoid sampling from a too high dimensional proposal distribution, for example). In that setting, the proposal distribution may yield states of different dimensions at each iteration. A simplified form of such a transdimensional algorithm (which was successfully applied to additive regression in Guedj and Alquier, 2013, binary ranking in Guedj and Robbiano, 2018 and online clustering in Li et al., 2018) is given by Algorithm 2. Other MCMC algorithms, such as Langevin Monte Carlo, have

Algorithm 1: Metropolis-Hastings algorithm

Input: Proposal q , target p , horizon T , initialization x_0
Output: A sequence $(x_i)_{i=0}^T$

```

1 for  $t = 1, \dots, T$  do
2    $x \sim q$  // Sample a candidate state
3    $\alpha := \min\left(1, \frac{p(x)}{p(x_{t-1})} \cdot \frac{q(x_{t-1})}{q(x)}\right)$  // acceptance ratio
4    $U \sim \mathcal{B}(\alpha)$  // draw a Bernoulli trial
5   if  $U \equiv 1$  then
6      $x_t := x$ 
7   else
8      $x_t := x_{t-1}$ 

```

also been investigated (Dalalyan and Tsybakov, 2012b). MCMC algorithms (as in Algorithm 1 and Algorithm 2) output a sequence of points, whose stationary distribution is asymptotically the target p . Whether this property is reached for a given number of iterations, or the quality of the approximation at a finite horizon, are central questions in the MCMC literature.

- When using the mode of the Gibbs measure, *i.e.*,

$$\widehat{\phi}^{\text{mode}} \in \arg \sup_{\phi \in \mathcal{F}_0} \pi_\lambda = \arg \sup_{\phi \in \mathcal{F}_0} \{\exp[-\lambda r_n(\phi)] \pi_0(\phi)\},$$

it is often more efficient to resort to stochastic optimization, such as gradient descent or its many variants (stochastic gradient descent, block gradient descent, to name a few). Gradient descent is one of the main workhorse of machine learning and we refer to Shalev-Shwartz and Ben-David (2014, Chapter 14) and references therein. A gradient-descent-based strategy has been applied in Alquier and Guedj (2017) for PAC-Bayesian-flavored non-negative matrix factorization.

- The third option which has been investigated in the literature is variational Bayes. It amounts to finding the best approximation of the Gibbs measure within a family of known measures, typically much easier to sample from. Alquier et al. (2016) propose an algorithm to find the best Gaussian approximation to the Gibbs measure (under assumptions on the prior and loss which make this approximation reasonably good).

Several works have contributed to bridging the gap between theory and implementations for PAC-Bayes. As such, let us focus on two recent results:

- For variational Bayes (Gaussian) approximation to the Gibbs measure, Alquier et al. (2016) show that whenever a PAC-Bayesian inequality holds for the Gibbs

Algorithm 2: A transdimensional MCMC algorithm adapted from [Guedj and Alquier \(2013\)](#)

Input: Family of proposals (q_j) , target p , horizon T , initialization x_0
 /* As many proposal distributions as nested models. A model is determined by which covariates from $1, \dots, d$ are selected. Two models sharing the same number of selected covariates are said to be neighbors. */

Output: A sequence $(x_i)_{i=0}^T$

```

1 for  $t=1, \dots, T$  do
2   Dimension shift: add one, remove one, or do nothing (each with
   probability 1/3)
3   neighbors := set of models obtained from adding or subtracting one unit
   to the dimension of the current model.
4   for each  $j$  in neighbors do
5      $y_j \sim q_j$  // e.g., a Gaussian
6     Pick model  $j$  with probability  $\frac{p(y_j)/q_j(y_j)}{\sum_{k \in \text{neighbors}} p(y_k)/q_k(y_k)}$ 
7      $\alpha := \min\left(1, \frac{p(y_j)}{p(x_{t-1})} \cdot \frac{q(x_{t-1})}{q_j(y_j)}\right)$  // acceptance ratio
8      $U \sim \mathcal{B}(\alpha)$  // draw a Bernoulli trial
9     if  $U \equiv 1$  then
10       $x_t := y_j$ 
11    else
12       $x_t := x_{t-1}$ 

```

measure, a similar one (with the same rate of convergence) holds for the approximate generalized posterior (at the price of technical assumptions which control the quality of the approximation in a Kullback-Leibler sense). This breakthrough allows for PAC-Bayesian oracle generalization bounds on the actual algorithm which is implemented rather than on the theoretical object, and as such, echoes the celebrated statistical and computational tradeoff.

- MCMC has been the most used sampling scheme in the PAC-Bayes literature, however very few results were available to guarantee its validity and quality in that setting. [Li et al. \(2018\)](#) proved that the stationary distribution is indeed the Gibbs measure for a particular model (online clustering).

As a concluding remark, let us note that there is no clear heuristics as to how one should calibrate the temperature parameter λ in practice. Two strategies are possible: cross-validation (yielding interesting results, yet quite computationally demanding)

and integration and marginalization of λ similarly to what is proposed in the "safe Bayesian" framework (Grünwald, 2012).

6. Some recent breakthroughs in PAC-Bayes

Over the past years, the PAC-Bayesian approach has been applied to a large spectrum of settings. In addition to aforesaid papers, let us mention classification (Germain et al., 2009), high-dimensional sparse regression (Alquier and Biau, 2013; Alquier and Lounici, 2011; Guedj and Alquier, 2013), image denoising (Salmon and Le Pennec, 2009), completion and factorization of large random matrices (Alquier and Guedj, 2017; Mai and Alquier, 2015), recommendation systems, reinforcement learning and collaborative filtering (Ghavamzadeh et al., 2015), dependent or heavy-tailed data (Alquier and Guedj, 2018; Ralaivola et al., 2010; Seldin et al., 2012), co-clustering (Seldin and Tishby, 2010), meta-learning (Amit and Meir, 2018), binary ranking (Guedj and Robbiano, 2018; Li et al., 2013), transfer learning and domain adaptation (Germain et al., 2016b), online clustering (Li et al., 2018), multi-view learning (Sun, 2013; Zhao et al., 2017), multiple testing (Blanchard and Fleuret, 2007), tailored density estimation (Higgs and Shawe-Taylor, 2010), to name but a few.

A salient advantage of PAC-Bayes is its flexibility: the theory requires little assumptions to be applied to new topics and problems. The use of generalized Bayesian learning algorithms requires the definition of a loss, and of a prior (*i.e.*, a heuristics to navigate throughout the candidates space \mathcal{F}_0), which explains how it could have been applied to so many learning settings.

Along with its flexibility, the PAC-Bayesian paradigm often ships with powerful generalization bounds, of which many are minimax optimal and demonstrate the unbeatable performance of the corresponding learning algorithms for specific tasks.

Most recent works on PAC-Bayes have seen a growing interest in data-dependent priors (Dziugaite and Roy, 2018a,b; Rivasplata et al., 2018). This movement can be seen as an additional layer of *generalization*: since the model-based likelihood has been replaced by an agnostic data-driven loss term, why not shift from a model-constrained prior to a purely data-driven measure which captures elementary knowledge about the underlying phenomenon?

In the deep learning tide wave, the machine learning community (at large) has demonstrated the impressive empirical successes of neural networks in some tasks. However voices have risen to orient some of the research effort to obtain theoretical guarantees and bounds which would explain those successes. Very few results have been published, however a significant fraction massively relies on PAC-Bayes. Dziugaite and Roy (2017) and Neyshabur et al. (2017) prove generalization bounds for

neural networks, with computable bounds (inherited from McAllester’s initial bound) and numerical experiments proving the generalization ability of (small) networks.

Last but not least, a few research efforts in the past years have focused on more agnostic and generic perspectives to obtain PAC-Bayes bounds, and to get rid of handy yet unrealistic assumptions such as boundedness of the loss function, or independence of data. Such assumptions allow for an extensive use of powerful mathematical results, and yet are hardly met in practice. [Bégin et al. \(2016\)](#) replaced the classical Kullback-Leibler divergence by the more general Rényi divergence, allowing to derive bounds in new settings. [Alquier and Guedj \(2018\)](#) then proposed an even more general divergence class, the f -divergences (of which the Rényi divergence is a special case).

PAC bounds for heavy-tailed random variables have been studied by [Catoni \(2004\)](#) under strong exponential moments assumptions. In a striking series of papers, several authors have taken over and improved those bounds with different tools: the small ball property ([Grünwald and Mehta, 2016](#); [Mendelson, 2015](#)), robust loss functions ([Catoni, 2016](#)) and median-of-means tournaments ([Devroye et al., 2016](#); [Lugosi and Mendelson, 2016](#)). However those papers mostly focus on linear regression, and the ERM as a learner (or intricated functionals of the median of samples). [Alquier and Guedj \(2018\)](#) derived PAC bounds with similar rates of convergence, holding for generalized Bayesian learning algorithms. As for dependent data, several PAC or PAC-Bayesian bounds have been proven ([Agarwal and Duchi, 2013](#); [Ralaivola et al., 2010](#); [Seldin et al., 2012](#)) under boundedness or exponential moments assumptions.

Let us conclude this section by sketching the main arguments of the scheme of proof. Note that data points are not required to be independent nor identically distributed. For the sake of concision we shall now omit the argument ϕ when no confusion can arise.

Definition 1. — For any $p \in \mathbb{N}$, let

$$\mathcal{M}_{\psi_p, n} := \int_{\mathcal{F}_0} \mathbb{E} (|r_n(\phi) - R(\phi)|^p) \pi_0(d\phi).$$

Definition 2. — Let f be a convex function with $f(1) = 0$. Csiszár’s f -divergence between two measures μ and ν is given by

$$D_f(\mu, \nu) = \begin{cases} \int f\left(\frac{d\mu}{d\nu}\right) d\nu & \text{when } \mu \ll \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

Note that $\mathcal{K}(\mu, \nu) = D_{x \log(x)}(\mu, \nu)$.

Theorem 3 ([Alquier and Guedj, 2018](#)). — Fix $p > 1$, $q = \frac{p}{p-1}$ and $\epsilon \in (0, 1)$. With probability at least $1 - \epsilon$ we have for any distribution μ

$$\left| \int R d\mu - \int r_n d\mu \right| \leq \left(\frac{\mathcal{M}_{\psi_q, n}}{\epsilon} \right)^{\frac{1}{q}} (D_{\psi_{p-1}}(\mu, \pi_0) + 1)^{\frac{1}{p}}.$$

Proof. — Let $\Delta_n(\phi) := |r_n(\phi) - R(\phi)|$. From [Bégin et al. \(2016\)](#), we derive:

$$\begin{aligned} \left| \int R d\mu - \int r_n d\mu \right| &\leq \int \Delta_n d\mu = \int \Delta_n \frac{d\mu}{d\pi_0} d\pi_0 \\ &\leq \left(\int \Delta_n^q d\pi_0 \right)^{\frac{1}{q}} \left(\int \left(\frac{d\mu}{d\pi_0} \right)^p d\pi_0 \right)^{\frac{1}{p}} && \text{(Hölder ineq.)} \\ &\leq \left(\frac{\mathbb{E} \int \Delta_n^q d\pi_0}{\epsilon} \right)^{\frac{1}{q}} \left(\int \left(\frac{d\mu}{d\pi_0} \right)^p d\pi_0 \right)^{\frac{1}{p}} && \text{(Markov, w.p. } 1 - \epsilon) \\ &= \left(\frac{\mathcal{M}_{\psi_q, n}}{\epsilon} \right)^{\frac{1}{q}} (D_{\psi_{p-1}}(\mu, \pi_0) + 1)^{\frac{1}{p}}. \end{aligned}$$

□

The message from [Theorem 3](#) is that we can compare $\int r_n d\mu$ (observable) to $\int R d\mu$ (unknown, the objective) in terms of

- the moment $\mathcal{M}_{\psi_q, n}$ (which depends on the distribution of the data)
- the divergence $D_{\psi_{p-1}}(\mu, \pi_0)$ (which measures the complexity of the set \mathcal{F}_0).

As a straightforward consequence, we have with probability at least $1 - \epsilon$, for any μ ,

$$\int R d\mu \leq \int r_n d\mu + \left(\frac{\mathcal{M}_{\psi_q, n}}{\epsilon} \right)^{\frac{1}{q}} (D_{\psi_{p-1}}(\mu, \pi_0) + 1)^{\frac{1}{p}},$$

which is a strong incitement to deduce the optimal generalized posterior as the minimizer of the right-hand side.

Definition 3. — Define $\bar{r}_n = \bar{r}_n(\epsilon, p)$ as

$$\bar{r}_n = \min \left\{ u \in \mathbb{R}, \int [u - r_n(\phi)]_+^q \pi_0(d\phi) = \frac{\mathcal{M}_{\psi_q, n}}{\epsilon} \right\}.$$

The minimum always exists as the integral is a continuous function of u , is equal to 0 when $u = 0$ and $\rightarrow \infty$ when $u \rightarrow \infty$. We then define the optimal generalized posterior $\hat{\mu}_n$ as

$$\frac{d\hat{\mu}_n}{d\pi_0}(\phi) = \frac{[\bar{r}_n - r_n(\phi)]_+^{\frac{1}{p-1}}}{\int [\bar{r}_n - r_n(\psi)]_+^{\frac{1}{p-1}} \pi_0(d\psi)}.$$

[Alquier and Guedj \(2018\)](#) then focus on the explicit computation of the two terms $\mathcal{M}_{\psi_q, n}$ and $D_{\psi_{p-1}}(\mu, \pi_0)$ in several cases: bounded and unbounded losses, iid or dependent observations, and prove the first PAC-Bayesian bound for a time series without any boundedness nor exponential moment assumption. As [Theorem 3](#) is a completely generic result and holds under no assumption whatsoever, it may serve as an elementary brick to build up new PAC-Bayesian generalization bounds.

7. Conclusion

As developed throughout the present paper, PAC-Bayesian learning is a flexible and powerful machinery, as it yields state-of-the-art oracle generalization bounds under little assumptions.

A NIPS 2017 workshop⁽²⁾ and the "PAC-Bayes" query on arXiv⁽³⁾ both illustrate how PAC-Bayes is quickly re-emerging as a principled theory to efficiently address modern machine learning topics, such as leaning with heavy-tailed and dependent data, or deep neural networks generalization abilities.

References

- A. Agarwal and J. C. Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2013.
- P. Alquier. *Transductive and Inductive Adaptive Inference for Regression and Density Estimation*. PhD thesis, Université Paris 6, 2006.
- P. Alquier. PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.
- P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14:243–280, 2013.
- P. Alquier and B. Guedj. An oracle inequality for quasi-Bayesian nonnegative matrix factorization. *Mathematical Methods of Statistics*, 26(1):55–67, 2017. ISSN 1934-8045. doi: 10.3103/S1066530717010045. URL <https://doi.org/10.3103/S1066530717010045>.
- P. Alquier and B. Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018. ISSN 1573-0565. doi: 10.1007/s10994-017-5690-0. URL <https://doi.org/10.1007/s10994-017-5690-0>.
- P. Alquier and K. Lounici. PAC-Bayesian theorems for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.
- P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.
- R. Amit and R. Meir. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *International Conference on Machine Learning*, pages 205–214, 2018.
- C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.
- J.-Y. Audibert. *Une approche PAC-bayésienne de la théorie statistique de l'apprentissage*. PhD thesis, Université Paris 6, 2004.
- J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.

⁽²⁾<https://bguedj.github.io/nips2017/50shadesbayesian.html> – slides and videos are available.

⁽³⁾<https://arxiv.org/search/?query=PAC-Bayes&searchtype=all&source=header>

- J.-Y. Audibert and O. Bousquet. Combining PAC-Bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8(Apr):863–889, 2007.
- J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.
- R. Bardenet, A. Doucet, and C. Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 2016.
- L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. Pac-bayesian bounds based on the rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444, 2016.
- G. Blanchard and F. Fleuret. Occam’s hammer. In *International Conference on Computational Learning Theory*, pages 112–126. Springer, 2007.
- S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. École d’Été de Probabilités de Saint-Flour 2001. Springer, 2004.
- O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Lecture notes – Monograph Series*. Institute of Mathematical Statistics, 2007.
- O. Catoni. PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*, 2016.
- I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975.
- I. Csiszár and P. C. Shields. *Information theory and statistics: a tutorial*. Now Publishers Inc, 2004.
- A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory (COLT 2007), Lecture Notes in Computer Science*, pages 97–111, 2007.
- A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.
- A. S. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 2012a.
- A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012b.
- L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.
- V. C. Dinh, L. S. Ho, B. Nguyen, and D. Nguyen. Fast learning rates with heavy-tailed losses. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 505–513. Curran Associates, Inc., 2016.
- M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time — III. *Communications on pure and applied Mathematics*, 29(4):389–461, 1976.
- G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data.

- In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 2017.
- G. K. Dziugaite and D. M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In *NeurIPS*, 2018a.
- G. K. Dziugaite and D. M. Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors. In *International Conference on Machine Learning*, pages 1376–1385, 2018b.
- S. Gerchinovitz. *Prédiction de suites individuelles et cadre statistique classique : étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation*. PhD thesis, Université Paris-Sud, 2011.
- P. Germain. *Généralisations de la théorie PAC-bayésienne pour l'apprentissage inductif, l'apprentissage transductif et l'adaptation de domaine*. PhD thesis, Université Laval, 2015.
- P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML, 2009*.
- P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016a.
- P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A new PAC-Bayesian perspective on domain adaptation. In *Proceedings of International Conference on Machine Learning*, volume 48, 2016b.
- M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015.
- S. Ghosal and A. Van Der Vaart. Convergence rates of posterior distributions for non-iid observations. *The Annals of Statistics*, 35(1):192–223, 2007.
- S. Ghosal, J. K. Ghosh, and A. Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.
- P. D. Grünwald. *The minimum description length principle*. MIT press, 2007.
- P. D. Grünwald. Safe learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 397–420, 2011.
- P. D. Grünwald. The Safe Bayesian. In *Algorithmic Learning Theory*, pages 169–183. Springer Berlin Heidelberg, 2012.
- P. D. Grünwald. Safe probability. *Journal of Statistical Planning and Inference*, 195: 47–63, 2018.
- P. D. Grünwald and N. A. Mehta. Fast rates for general unbounded loss functions: from erm to generalized bayes. *arXiv preprint arXiv:1605.00252*, 2016.
- P. D. Grünwald and T. Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4): 1069–1103, 2017.
- B. Guedj. *Aggregation of estimators and classifiers: theory and methods*. Theses, Université Pierre et Marie Curie - Paris VI, December 2013. URL <https://tel.archives-ouvertes.fr/tel-00922353>.

- B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electron. J. Statist.*, 7:264–291, 2013. doi: 10.1214/13-EJS771. URL <https://doi.org/10.1214/13-EJS771>.
- B. Guedj and S. Robbiano. PAC-Bayesian high dimensional bipartite ranking. *Journal of Statistical Planning and Inference*, 196:70 – 86, 2018. ISSN 0378-3758. doi: <https://doi.org/10.1016/j.jspi.2017.10.010>. URL <http://www.sciencedirect.com/science/article/pii/S0378375817301945>.
- M. Higgs and J. Shawe-Taylor. A PAC-Bayes bound for tailored density estimation. In *International Conference on Algorithmic Learning Theory*, pages 148–162. Springer, 2010.
- A. Lacasse, F. Laviolette, M. Marchand, P. Germain, and N. Usunier. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *Advances in Neural information processing systems*, pages 769–776, 2007.
- J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *Advances in neural information processing systems*, pages 439–446, 2003.
- G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006.
- C. Li, W. Jiang, and M. Tanner. General oracle inequalities for Gibbs posterior with application to ranking. In *Conference on Learning Theory*, pages 512–521, 2013.
- L. Li, B. Guedj, and S. Loustau. A quasi-Bayesian perspective to online clustering. *Electron. J. Statist.*, 12(2):3071–3113, 2018. doi: 10.1214/18-EJS1479. URL <https://doi.org/10.1214/18-EJS1479>.
- G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society (to appear)*, 2016.
- T. T. Mai and P. Alquier. A Bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 9(1):823–841, 2015. doi: 10.1214/15-EJS1020. URL <https://doi.org/10.1214/15-EJS1020>.
- P. Massart. *Concentration inequalities and model selection*. Springer, 2007.
- D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999a.
- D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th annual conference on Computational Learning Theory*, pages 164–170. ACM, 1999b.
- D. A. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003a.
- D. A. McAllester. Simplified PAC-Bayesian margin bounds. In *COLT*, pages 203–215, 2003b.
- S. Mendelson. Learning without concentration. *Journal of the ACM*, 62(3):1–25, 2015. doi: 10.1145/2699439.
- A. Nemirovski. Topics in non-parametric statistics. *Ecole d’Eté de Probabilités de Saint-Flour*, 28:85, 2000.
- B. Neyshabur, S. Bhojanapalli, D. A. McAllester, and N. Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

- E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. PAC-Bayes bounds with data dependent priors. *Journal of Machine Learning Research*, 13: 3507–3531, 2012.
- L. Ralaivola, M. Szafranski, and G. Stempfel. Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β -mixing processes. *Journal of Machine Learning Research*, 11(Jul):1927–1956, 2010.
- P. Rigollet and A.B. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 4(27):558–575, 2012.
- O. Rivasplata, E. Parrado-Hernandez, J. Shawe-Taylor, S. Sun, and C. Szepesvari. PAC-Bayes bounds for stable algorithms with instance-dependent priors. *arXiv preprint arXiv:1806.06827*, 2018.
- C. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.
- J. Salmon and E. Le Pennec. An aggregator point of view on NL-Means. In *Wavelets XIII*, volume 7446, page 74461E. International Society for Optics and Photonics, 2009.
- A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959. ISSN 0018-8646. doi: 10.1147/rd.33.0210.
- M. Seeger. PAC-Bayesian generalization bounds for gaussian processes. *Journal of Machine Learning Research*, 3:233–269, 2002.
- M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, 2003.
- Y. Seldin and N. Tishby. PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11:3595–3646, 2010.
- Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.
- S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*, pages 2–9. ACM, 1997. doi: 10.1145/267460.267466.
- S. Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.
- A. B. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pages 303–313. Springer, 2003.
- L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11): 1134–1142, 1984.
- T. van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16: 1793–1861, 2015. Special issue in Memory of Alexey Chervonenkis.
- V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.
- V. G. Vovk. Aggregating strategies. *Proc. of Computational Learning Theory*, 1990.

- Y. Yang. Regression with multiple candidate models: selecting or mixing? *Statistica Sinica*, pages 783–809, 2003.
- Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004.
- T. Zhang. From ε -entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 10 2006a.
- T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006b.
- J. Zhao, X. Xie, X. Xu, and S. Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

BENJAMIN GUEDJ, Inria, Lille - Nord Europe research center, Modal project-team & University College London, Computer Science department, Machine Learning group
E-mail : benjamin.guedj@inria.fr • *Url* : <https://bguedj.github.io>