User-friendly Introduction to PAC-Bayes Bounds

Other titles in Foundations and Trends® in Machine Learning

A Friendly Tutorial on Mean-Field Spin Glass Techniques for Non-Physicists
Andrea Montanari and Subhabrata Sen
ISBN: 978-1-63828-212-9

Reinforcement Learning, Bit by Bit Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband and Zheng Wen
ISBN: 978-1-63828-254-9

Introduction to Riemannian Geometry and Geometric Statistics: From Basic Theory to Implementation with Geomstats
Nicolas Guigui, Nina Miolane and Xavier Pennec
ISBN: 978-1-63828-154-2

Graph Neural Networks for Natural Language Processing: A Survey
Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei and Bo Long
ISBN: 978-1-63828-142-9

Model-based Reinforcement Learning: A Survey
Thomas M. Moerland, Joost Broekens, Aske Plaat and Catholijn M. Jonker
ISBN: 978-1-63828-056-9

Divided Differences, Falling Factorials, and Discrete Splines: Another Look at Trend Filtering and Related Problems
Ryan J. Tibshirani
ISBN: 978-1-63828-036-1

User-friendly Introduction to PAC-Bayes Bounds

Pierre Alquier
ESSEC Business School
pierre.alquier.stat@gmail.com

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

P. Alquier. *User-friendly Introduction to PAC-Bayes Bounds*. Foundations and Trends® in Machine Learning, vol. 17, no. 2, pp. 174–303, 2024.

ISBN: 978-1-63828-327-0

Foundations and Trends® in Machine Learning

Volume 17, Issue 2, 2024

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2024, Volume 17, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Contents

User-friendly Introduction to PAC-Bayes Bounds

Pierre Alquier

ESSEC Business School, Asia-Pacific Campus, Singapore;
pierre.alquier.stat@gmail.com

ABSTRACT

Aggregated predictors are obtained by making a set of basic predictors vote according to some weights, that is, to some probability distribution. Randomized predictors are obtained by sampling in a set of basic predictors, according to some prescribed probability distribution.

Thus, aggregated and randomized predictors have in common that their definition rely on a probability distribution on the set of predictors. In statistical learning theory, there is a set of tools designed to understand the generalization ability of such predictors: PAC-Bayesian or PAC-Bayes bounds.

Since the original PAC-Bayes bounds (Shawe-Taylor and Williamson, 1997; McAllester, 1998), these tools have been considerably improved in many directions. We will for example describe a simplified version of the localization technique (Catoni, 2003; Catoni, 2007) that was missed by the community, and later rediscovered as “mutual information bounds”. Very recently, PAC-Bayes bounds received a considerable attention. There was workshop on PAC-Bayes at NIPS 2017, *(Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights*, organized by B. Guedj, F.

Pierre Alquier (2024), “User-friendly Introduction to PAC-Bayes Bounds”, Foundations and Trends® in Machine Learning: Vol. 17, No. 2, pp 174–303. DOI: 10.1561/2200000100.

Bach and P. Germain. One of the reasons of this recent interest is the successful application of these bounds to neural networks (Dziugaite and Roy, 2017). Since then, this is a recurring topic of workshops in the major machine learning conferences.

The objective of these notes is to provide an elementary introduction to PAC-Bayes bounds.

1

Introduction

In a supervised learning problem, such as classification or regression, we are given a data set, and we 1) *fix a set of predictors* and 2) *find a good predictor in this set.*

For example, when doing linear regression, you 1) chose to consider only linear predictors and 2) use the least-square method to chose your linear predictor.

In this tutorial, we will rather focus on "randomized" or "aggregated" predictors. By this, we mean that we will replace 2) by 2') *define weights on the predictors and make them vote according to these weights* or by 2") *draw a predictor according to some prescribed probability distribution.*

In this first section, we will introduce the main concepts of machine learning theory, and their mathematical notations. We will briefly introduce PAC bounds, that allow to control the generalization error of a predictor. These tools will allow to formalize properly the notion of "randomized" or "aggregated" predictors, and to introduce PAC-Bayes bounds.

1.1 Machine Learning and PAC Bounds

1.1.1 Machine learning: notations

In a supervised learning problem, the objective is to learn from examples to assign labels to objects. Objects can be images, videos, e-mails... The set of all possible objects will be denoted by $\mathcal{X}$. In all the examples we mentioned, it is possible to encode the objects by (large enough) vectors, and thus, we will often have $\mathcal{X} \subseteq \mathbb{R}^d$, where $\mathbb{R}$ is the set of real numbers. The set of labels will be denoted by $\mathcal{Y}$.

The most classical examples of supervised learning problems are binary classification and regression. In binary classification, $\mathcal{Y} = \{0, 1\}$. Examples includes spam detection: in this case, objects in $\mathcal{X}$ are e-mails, and the label is 1 if the e-mail is a spam, and 0 otherwise. In regression, labels can be any real number $\mathcal{Y} = \mathbb{R}$. This is the case when we try to predict a numerical quantity such as CO_2 emissions, temperature, etc.

A predictor is a function $f : \mathcal{X} \to \mathcal{Y}$: for each object x, it returns a label $f(x)$. We are usually interested in parametric sets of predictors. That is, we consider $\{f_\theta, \theta \in \Theta\}$ where Θ is any set, called the parameter set, and each f_θ is a predictor. For example, in linear regression, a common set of predictors is $f_\theta(x) = \langle x, \theta \rangle \in \mathcal{Y} = \mathbb{R}$, with $\mathcal{X} = \Theta = \mathbb{R}^d$. In classification, we can define with the same $\mathcal{X}$ and Θ,

$$f_\theta(x) = \begin{cases} 1 \text{ if } \langle x, \theta \rangle \geq 0, \\ 0 \text{ otherwise.} \end{cases}$$

Other examples include neural networks with a fixed architecture, θ being the weights of the network. Predictors are sometimes refered to as classifiers in the classification setting, and as regressors in regression.

Assume now that a pair label-object, $(x, y) \in \mathcal{X} \times \mathcal{Y}$, is given. A predictor f will propose a prediction $f(x)$ of the label y. If $f(x) = y$, the predictor f predicts the label correctly, otherwise, it makes a mistake. In order to quantify how serious a mistake is, we usually measure it by a loss function. In these notes, a loss function can be any function $\ell : \mathcal{Y}^2 \to [0, +\infty)$ such that $\ell(y, y) = 0$ for any $y \in \mathcal{Y}$; $\ell(f(x), y)$ will be interpreted as the cost of the prediction error. In classification, the most natural loss function is:

$$\ell(y',y) = \begin{cases} 1 \text{ if } y' \neq y, \\ 0 \text{ if } y' = y. \end{cases}$$

We will refer to it as the 0-1 loss function, and will use the following shorter notation: $\ell(y',y) = \mathbf{1}(y \neq y')$. For computational reasons, it is more convenient to use convex loss functions. For example, in binary classification: $\ell(y',y) = \max(1 - yy', 0)$ (the hinge loss). In regression problems, the most popular examples are $\ell(y',y) = (y'-y)^2$ the quadratic loss, or $\ell(y',y) = |y'-y|$ the absolute loss. The original PAC-Bayes bounds of McAllester (1998) were stated in the special case of the 0-1 loss, and this is also the case of most bounds published since then. However, we will see in Section 3 that their extension to any bounded loss is direct. Some PAC-Bayes bounds for regression with the quadratic loss were proven for example by Catoni (2004). **From now, and until the end of Section 4, we assume that $0 \leq \ell \leq C$.** This is typically the case in classification with the 0-1 loss, or in regression with quadratic loss under the additional assumption that $f_\theta(x)$ and y are bounded. We will discuss how to get rid of this assumption in Section 5.

Assume we want to build a machine to predict the label of objects it will encounter in the future. Of course, we don't know these objects in advance, nor their labels. A way to model this uncertainty is to assume that a future pair object-label is a random variable (X,Y) taking values in $\mathcal{X} \times \mathcal{Y}$. Let P denote the probability distribution[1] of (X,Y). The expected prediction mistake is thus $\mathbb{E}_{(X,Y)\sim P}[\ell(f(X),Y)]$. This is usually refered to as the (generalization) risk of f. As it is a very important notion in machine learning, we introduce the notation

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

As we will focus on predictors in $\{f_\theta, \theta \in \Theta\}$, we define

$$R(\theta) := R(f_\theta)$$

[1] Formally, we can only define a probability distribution on $\mathcal{X} \times \mathcal{Y}$ if it is equipped with a σ-algebra. Let $\mathcal{B}$ be such a σ-algebra. Essentially, the only thing that matters is that the loss function ℓ and the predictors $f_\theta(\cdot)$ are measurable functions, which is satisfied by all classical examples. Note that $\mathcal{B}$ will no longer appear explicitly in this tutorial.

for short. A good strategy would be to implement in our machine a predictor f_θ such that $R(\theta)$ is as small as possible – ideally, we should implement f_{θ^*} where $R(\theta^*) = \inf_{\theta\in\Theta} R(\theta)$, if this infimum is reached. Unfortunately, there is a major difficulty: we don't know the distribution P of (X,Y) in practice. Check the examples above: we are not able to describe the probability distribution of images we will see in the future, or of e-mails we will receive.

Instead, we will train our machine based on examples. That is, we assume that we can access a sample of pairs object-label, that we will call the data, or the observations: $(X_1,Y_1),\dots,(X_n,Y_n)$. **From now, and until the end of Section 4, we assume that** $(X_1,Y_1),\dots,(X_n,Y_n)$ **are i.i.d. from** P. That is, they are "typical examples" of the pairs object-label the machine will have to deal with in the future. For short, we put $\ell_i(\theta) := \ell(f_\theta(X_i),Y_i) \geq 0$. We define the empirical risk:

$$r(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta).$$

Note that it satisfies

$$\mathbb{E}_{(X_1,Y_1),\dots,(X_n,Y_n)}[r(\theta)] = R(\theta).$$

The notation for the previous expectation is cumbersome. From now, we will write $\mathcal{S} = [(X_1,Y_1),\dots,(X_n,Y_n)]$ and $\mathbb{E}_{\mathcal{S}}$ (for "expectation with respect to the sample") instead of $\mathbb{E}_{(X_1,Y_1),\dots,(X_n,Y_n)}$. In the same way, we will write $\mathbb{P}_{\mathcal{S}}$ for probabilities with respect to the sample.

Finally, an estimator is a function that takes a sample of pairs object-labels of any size and returns a guess for the parameter θ we should use for future predictions. Formally,[2]

$$\hat{\theta} : \bigcup_{n=1}^{\infty} (\mathcal{X}\times\mathcal{Y})^n \to \Theta.$$

For short, we write $\hat{\theta}$ instead of $\hat{\theta}((X_1,Y_1),\dots,(X_n,Y_n))$. The most famous example is the Empirical Risk Minimizer, or ERM:

$$\hat{\theta}_{\mathrm{ERM}} = \operatorname*{argmin}_{\theta\in\Theta} r(\theta)$$

(when this minimizer exists and is unique).

[2] The proper definition also requires $\hat{\theta}$ to be a measurable function of the observations, so that probabilities of events involving $\hat{\theta}$ are well defined. This is not so important here as we will soon replace the notion of estimator with a new notion.

1.1.2 PAC bounds

Of course, our objective is to minimize R, not r. So the ERM strategy is motivated by the hope that these two functions are not so different, so that the minimizer of r almost minimizes R. Let us now discuss to what extent this is true. By doing so, we will introduce some tools that will be also useful for PAC-Bayes bounds.

Proposition 1.1. For any $\theta \in \Theta$, for any $\delta \in (0,1)$,

$$\mathbb{P}_{\mathcal{S}}\left(R(\theta) > r(\theta) + C\sqrt{\frac{\log \frac{1}{\delta}}{2n}} \right) \leq \delta. \tag{1.1}$$

The proof relies on a result that will be useful in all this tutorial.

Lemma 1.1 (Hoeffding's inequality). Let $U_1, \dots, U_n$ be independent random variables taking values in an interval $[a,b]$. Then, for any $t > 0$,

$$\mathbb{E}\left[\mathrm{e}^{t\sum_{i=1}^{n}[U_i - \mathbb{E}(U_i)]} \right] \leq \mathrm{e}^{\frac{nt^2(b-a)^2}{8}}.$$

Hoeffding's inequality is proven for example in Chapter 2 of Boucheron *et al.* (2013), which is a highly recommended reading, but it is so classical that you can as well find it on Wikipedia.

Proof of Proposition 1.1. Apply Lemma 1.1 to $U_i = \mathbb{E}[\ell_i(\theta)] - \ell_i(\theta)$:

$$\mathbb{E}_{\mathcal{S}}\left[\mathrm{e}^{tn[R(\theta) - r(\theta)]} \right] \leq \mathrm{e}^{\frac{nt^2C^2}{8}}. \tag{1.2}$$

Now, for any $s > 0$,

$$\begin{aligned} \mathbb{P}_{\mathcal{S}}(R(\theta) - r(\theta) > s) &= \mathbb{P}_{\mathcal{S}}\left(\mathrm{e}^{nt[R(\theta)-r(\theta)]} > \mathrm{e}^{nts} \right) \\ &\leq \frac{\mathbb{E}_{\mathcal{S}}\left[\mathrm{e}^{nt[R(\theta)-r(\theta)]} \right]}{\mathrm{e}^{nts}} \text{ by Markov's inequality,} \\ &\leq \mathrm{e}^{\frac{nt^2C^2}{8} - nts} \text{ by (1.2).} \end{aligned}$$

In other words,

$$\mathbb{P}_{\mathcal{S}}(R(\theta) > r(\theta) + s) \leq \mathrm{e}^{\frac{nt^2C^2}{8} - nts}.$$

We can make this bound as tight as possible, by optimizing our choice for t. Indeed, $nt^2C^2/8 - nts$ is minimized for $t = 4s/C^2$, which gives

$$\mathbb{P}_{\mathcal{S}}(R(\theta) > r(\theta) + s) \leq \mathrm{e}^{\frac{-2ns^2}{C^2}}. \tag{1.3}$$

This means that, for a given θ, the empirical risk $r(\theta)$ cannot be much smaller than the risk $R(\theta)$. The order of this "much smaller" can be better understood by introducing

$$\delta = \mathrm{e}^{\frac{-2ns^2}{C^2}}$$

and substituting δ to s in (1.3), which gives (1.1). □

Proposition 1.1 states that $R(\theta)$ will usually not exceed $r(\theta)$ by more than a term in $1/\sqrt{n}$. This is not enough, though, to justify the use of the ERM. Indeed, (1.1) is only true for the θ that was fixed above, and we cannot apply it to $\hat{\theta}_{\mathrm{ERM}}$ that is a function of the data.

The usual approach to control $R(\hat{\theta}_{\mathrm{ERM}})$ is to use the inequality

$$R(\hat{\theta}_{\mathrm{ERM}}) - r(\hat{\theta}_{\mathrm{ERM}}) \leq \sup_{\theta\in\Theta} \left[R(\theta) - r(\theta)\right], \tag{1.4}$$

and to prove a version of (1.1) that would hold uniformly on Θ. As an illustration of this method, we prove the following result.

Theorem 1.2. Assume that $\mathrm{card}(\Theta) = M < +\infty$. For any $\delta \in (0, 1)$,

$$\mathbb{P}_{\mathcal{S}}\left(R(\hat{\theta}_{\mathrm{ERM}}) \leq \inf_{\theta\in\Theta} r(\theta) + C\sqrt{\frac{\log\frac{M}{\delta}}{2n}}\right) \geq 1 - \delta.$$

Proof. As announced before the statement of the theorem, we upper bound the supremum in (1.4):

$$\begin{aligned}\mathbb{P}_{\mathcal{S}}(\sup_{\theta\in\Theta}[R(\theta) - r(\theta)] > s) &= \mathbb{P}_{\mathcal{S}}\left(\bigcup_{\theta\in\Theta}\Big\{[R(\theta) - r(\theta)] > s\Big\}\right)\\ &\leq \sum_{\theta\in\Theta} \mathbb{P}_{\mathcal{S}}(R(\theta) > r(\theta) + s)\\ &\leq M\mathrm{e}^{\frac{-2ns^2}{C^2}} \end{aligned} \tag{1.5}$$

thanks to (1.3). This time, put $\delta = M\mathrm{e}^{\frac{-2ns^2}{C^2}}$ and plug into (1.5) to get:

$$\mathbb{P}_{\mathcal{S}}\left(\sup_{\theta\in\Theta}[R(\theta) - r(\theta)] > C\sqrt{\frac{\log\frac{M}{\delta}}{2n}}\right) \leq \delta.$$

Thus, the complementary event satisfies

$$\mathbb{P}_{\mathcal{S}}\left(\sup_{\theta\in\Theta}[R(\theta) - r(\theta)] \leq C\sqrt{\frac{\log\frac{M}{\delta}}{2n}}\right) \geq 1-\delta. \tag{1.6}$$

From (1.4),

$$\mathbb{P}_{\mathcal{S}}\left(R(\hat{\theta}_{\mathrm{ERM}}) \leq r(\hat{\theta}_{\mathrm{ERM}}) + C\sqrt{\frac{\log\frac{M}{\delta}}{2n}}\right) \geq 1-\delta$$

and note that, as Θ is finite, $r(\hat{\theta}_{\mathrm{ERM}}) = \inf_{\theta\in\Theta} r(\theta)$. □

Bounds in the form of Theorem 1.2 are called Probably Approximately Correct (PAC) bounds, because $r(\hat{\theta}_{\mathrm{ERM}})$ *approximates* $R(\hat{\theta}_{\mathrm{ERM}})$ within $C\sqrt{\log(M/\delta)/2n}$ with *probability* $1-\delta$. This terminology was introduced by Valiant (1984).

Remark 1.1. The proofs of Proposition 1.1 and Theorem 1.2 used, in addition to Hoeffding's inequality, two tricks that we will reuse very often when we will prove PAC-Bayes bounds:

- given a random variable U and $s \in \mathbb{R}$, for any $t > 0$,

  $$\mathbb{P}\left(U > s\right) = \mathbb{P}\left(\mathrm{e}^{tU} > \mathrm{e}^{ts}\right) \leq \frac{\mathbb{E}\left(\mathrm{e}^{tU}\right)}{\mathrm{e}^{ts}}$$

  thanks to Markov inequality. The combo "exponential + Markov inequality" is known as **Chernoff's bounding technique**. It is is of course very useful together with exponential inequalities like Hoeffding's inequality.

- given a finite number of random variables $U_1, \dots, U_M$,

$$\mathbb{P}\left(\sup_{1\leq i\leq M} U_i > s\right) = \mathbb{P}\left(\bigcup_{1\leq i\leq M}\left\{U_i > s\right\}\right)$$
$$\leq \sum_{i=1}^{M} \mathbb{P}\left(U_i > s\right).$$

  This argument is called the **union-bound argument**.

The next step in the study of the ERM would be to go beyond finite sets Θ. The union bound argument has to be modified in this case, and things become a little more complicated. We will therefore stop here the study of the ERM: it is not our objective anyway.

If the reader is interested by the study of the ERM in general: Vapnik and Chervonenkis (1968) developed the theoretical tools for this study, see the more recent monograph by Vapnik (1998). We refer the reader to Devroye *et al.* (1996) for a beautiful and very pedagogical introduction to machine learning theory. Chapters 11 and 12 in particular are dedicated to Vapnik and Chervonenkis theory. More recent references include Giraud (2014) and Wainwright (2019).

1.2 What are PAC-Bayes Bounds?

We are now in better position to explain what are PAC-Bayes bounds. A simple way to phrase things: PAC-Bayes bounds are generalization of the union bound argument, that will allow to deal with any parameter set Θ: finite or infinite, continuous... However, a byproduct of this technique is that we will have to change the notion of estimator.

Definition 1.1. Let $\mathcal{P}(\Theta)$ be the set of all probability distributions on Θ equipped with a σ-algebra $\mathcal{T}$. A data-dependent probability measure is a function:

$$\hat{\rho} : \bigcup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{P}(\Theta)$$

with a suitable measurability condition.[3] We will write $\hat{\rho}$ instead of $\hat{\rho}((X_1, Y_1), \ldots, (X_n, Y_n))$ for short.

In practice, when you have a data-dependent probability measure, and you want to build a predictor, you can:

- draw a random parameter $\tilde{\theta} \sim \hat{\rho}$, we will call this procedure "randomized estimator".
- use it to average predictors, that is, define a new predictor:

  $$f_{\hat{\rho}}(\cdot) = \mathbb{E}_{\theta \sim \hat{\rho}}[f_\theta(\cdot)]$$

  called the aggregated predictor with weights $\hat{\rho}$.

So, with PAC-Bayes bounds, we will extend the union bound argument[4] to infinite, uncountable sets Θ, but we will obtain bounds on various risks related to data-dependent probability measures, that is:

- the risk of a randomized estimator, $R(\tilde{\theta})$,
- or the average risk of randomized estimators, $\mathbb{E}_{\theta \sim \hat{\rho}}[R(\theta)]$,
- or the risk of the aggregated estimator, $R(f_{\hat{\rho}})$.

From a technical point of view, the analysis shares many similarities with the analysis of the ERM in the previous section. A key difference is that the supremum in (1.4) will be replaced by

$$\mathbb{E}_{\theta \sim \hat{\rho}}[R(\theta) - r(\theta)] \leq \sup_{\rho \in \mathcal{P}(\Theta)} \mathbb{E}_{\theta \sim \rho}[R(\theta) - r(\theta)].$$

While this might look unnecessarily complicated at first sight, PAC-Bayes bounds will actually turn out to be extremely convenient for many reasons that we hope will become clear along the next sections:

[3] I don't want to scare the reader with measurability conditions, as I will not check them in this tutorial anyway. Here, the exact condition to ensure that what follows is well defined is that for any $A \in \mathcal{T}$, the function

$$((x_1, y_1), \ldots, (x_n, y_n)) \mapsto [\hat{\rho}((x_1, y_1), \ldots, (x_n, y_n))]\,(A)$$

is measurable. That is, $\hat{\rho}$ is a regular conditional probability.

[4] See the title of van Erven's tutorial (van Erven, 2014): "PAC-Bayes mini-tutorial: a continuous union bound". Note, however, that it is argued by Catoni (2007) that PAC-Bayes bounds are actually more than that, we will come back to this in Section 4.

- first, they don't require the set of predictors to be finite, nor discrete. Of course, it is also possible to prove PAC bounds for the ERM when $\Theta \subset \mathbb{R}^p$ is not finite, but this leads to technical difficulties or strong restrictions such as the compactness of Θ. PAC-Bayes bounds do not lead to major difficulties with unbounded parameter spaces, as will be illustrated in Example 3.2.

- randomized estimators are fairly common in machine learning. This includes Bayesian estimation and related methods such as variational inference and ensemble methods. Section 2 illustrates how PAC-Bayes bounds can be applied to such estimators. Moreover, many non-randomized estimators can be derived from randomized ones: aggregation rules, majority vote classifiers, etc. The PAC-Bayes bounds on the randomized estimator often brings strong information on the de-randomized version. This will also be discussed thoroughly and illustrated in Section 2.

- Bayesian estimators incorporate prior knowledge through a prior distribution π on Θ. Even though PAC-Bayes bounds can be applied to non-Bayesian estimators, a prior π will still appear in the bound. The effect of π on the bound will be discussed thoroughly. In particular, PAC-Bayes bounds depend not only on the minimum of the empirical risk $r(\theta)$, but on the prior probability of the level sets of r: in general, this can be quantified through the so-called prior-mass condition, as described in Section 4, even though specific examples such as Example 2.1 will already illustrate this property. A consequence is that flater minima lead to tigther bounds. This is one of the reasons why PAC-Bayes bounds can be tight for deep learning (Section 3).

You will of course ask the question: if Θ is infinite, what will the $\log(M)$ term be replaced with? In PAC-Bayes bounds, this term will be replaced by the Kullback-Leibler divergence between ρ and a fixed π on Θ (the prior).

Definition 1.2. Given two probability measures μ and ν in $\mathcal{P}(\Theta)$, the Kullback-Leibler (or simply KL) divergence between μ and ν is

$$KL(\mu\|\nu) = \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\theta)\right)\mu(\mathrm{d}\theta) \in [0,+\infty]$$

if μ has a density $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ with respect to ν, and $KL(\mu\|\nu) = +\infty$ otherwise.[5]

Example 1.1. For example, if Θ is finite,

$$KL(\mu\|\nu) = \sum_{\theta\in\Theta} \log\left(\frac{\mu(\theta)}{\nu(\theta)}\right)\mu(\theta).$$

The following result is well known. You can prove it using Jensen's inequality.

Proposition 1.2. For any probability measures μ and ν, $KL(\mu\|\nu) \geq 0$ with equality if and only if $\mu = \nu$.

1.3 Why this Tutorial?

Since the "PAC analysis of a Bayesian estimator" by Shawe-Taylor and Williamson (1997) and the first PAC-Bayes bounds proven by McAllester (1998) and McAllester (1999), many new PAC-Bayes bounds appeared (we will see that some of them can be derived from a bound due to Seeger, 2002). These bounds were used in various contexts, to solve a wide range of problems. This led to hundreds of (beautiful!) papers. The consequence of this is that it's quite difficult to be aware of all the existing work on PAC-Bayes bounds. In particular, it seems that many powerful techniques in Catoni's book (Catoni, 2007) and earlier works (Catoni, 2003; 2004) are largely ignored by the community.

On the other hand, it's not easy to enter into the PAC-Bayes literature. Most papers already assume some basic knowledge on these bounds, and Catoni's book is quite technical to begin with. The objective

[5] We recall that if there is a measurable function g such that for any measurable set A,

$$\mu(A) = \int_A g(\theta)\nu(\mathrm{d}\theta),$$

then this function is essentially unique. We put $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\theta) = g(\theta)$ and refer to this function as the density of μ with respect to ν.

of these notes is thus to provide a user-friendly introduction, accessible to PhD students, that could be used as a first approach to PAC-Bayes bounds. It also provides references for more sophisticated results.

I want to mention existing short introduction to PAC-Bayes bounds, like the ones by McAllester (2013) and van Erven (2014) and the nice introductory slides of Fleuret (2011). They are very informative, and I recommend the reader to check them. However, they are focused on empirical bounds only. There are also surveys on PAC-Bayes bounds, such as Chopin *et al.* (2015, Section 5) or Guedj (2019). These papers are very useful to navigate in the ocean of publications on PAC-Bayes bounds, and they helped me a lot when I was writing this document, but might not provide enough detail for a first reading on the topic.

Finally, in order to highlight the main ideas, I will not necessarily try to present the bounds with the tightest possible constants. In particular, many oracle bounds and localized bounds in Section 4 were introduced in Catoni (2003; 2007) with better constants. Once again, this is an *introduction* to PAC-Bayes bounds. I strongly recommend the reader to check the original publications for more accurate results.

1.4 Two Types of PAC Bounds, Organization of these Notes

It is important to make a distinction between two types of PAC bounds.

Theorem 1.2 is usually refered to as an *empirical bound*. It means that, for any θ, $R(\theta)$ is upper bounded by an empirical quantity, that is, by something that we can compute when we observe the data. This allows to study the ERM as the minimizer of this bound. It also provides a numerical certificate of the generalization error of the ERM. You will really end up with something like

$$\mathbb{P}_{\mathcal{S}}\left(R(\hat{\theta}_{\mathrm{ERM}}) \leq 0.12\right) \geq 0.99.$$

However, a numerical certificate on the generalization error does not tell you one thing. Can this 0.12 be improved using a larger sample size? Or is it the best that can be done with our set of predictors? The right tools to answer these questions are excess risk bounds, also known as oracle PAC bounds. In these bounds, you have a control of the form

$$\mathbb{P}_{\mathcal{S}}\left(R(\hat{\theta}_{\mathrm{ERM}}) \leq \inf_{\theta\in\Theta} R(\theta) + r_n(\delta)\right) \geq 1-\delta,$$

where the remainder $r_n(\delta)$ should be as small as possible and satisfy $r_n(\delta) \to 0$ when $n \to \infty$. Of course, the upper bound on $R(\hat{\theta}_{\mathrm{ERM}})$ cannot be computed because R is unknown in practice, so it doesn't lead to a numerical certificate on $R(\hat{\theta}_{\mathrm{ERM}})$. Still, these bounds are very interesting, because they tell you how close you can expect $R(\hat{\theta}_{\mathrm{ERM}})$ to be to the smallest possible value of R.

In the same way, there are empirical PAC-Bayes bounds, and oracle PAC-Bayes bounds (also known as excess-risk PAC-Bayes bounds). The very first PAC-Bayes bounds by McAllester (1998) and McAllester (1999) were empirical bounds. The first oracle PAC-Bayes bounds came later (Catoni, 2003; Catoni, 2004; Zhang, 2006; Catoni, 2007).

In some sense, empirical PAC-Bayes bounds are more useful in practice, and oracle PAC-Bayes bounds are theoretical objects. But this might be an oversimplification. We will see that empirical bounds are tools used to prove some oracle bounds, so they are also useful in theory. On the other hand, when we design a data-dependent probability measure, we don't know if it will lead to large or small empirical bounds. A preliminary study of its theoretical properties through an oracle bound is the best way to ensure that it is efficient, and so that it has a chance to lead to small empirical bounds.

In Section 2, we will study an example of empirical PAC-Bayes bound, essentially taken from a preprint by Catoni (2003). We will prove it together, play with it and modify it in many ways. In Section 3, we cover many empirical PAC-Bayes bounds, and explain the race to tighter bounds. This led to bounds that are tight enough to provide good generalization certificates for deep learning, we will discuss this based on Dziugaite and Roy's paper (Dziugaite and Roy, 2017) and a more recent work by Pérez-Ortiz, Rivasplata, Shawe-Taylor, and Szepesvàri (Pérez-Ortiz *et al.*, 2021).

In Section 4, we will turn to oracle PAC-Bayes bounds. We will see how to derive these bounds from empirical bounds, and apply them to some classical set of predictors. We will examine the assumptions leading to fast rates in these inequalities. Section 5 will be devoted to the various attempts to extend PAC-Bayes bounds beyond the setting introduced in this introduction, that is: bounded loss, and i.i.d. observations. Finally, in Section 6 we will discuss briefly the connection between PAC-Bayes bounds and many other approaches in machine learning and statistics, including regret bounds and Mutual Information bounds (MI).

References

Alquier, P. (2006). “Transductive and inductive adaptative inference for regression and density estimation”. *PhD thesis, University Paris 6.*

Alquier, P. (2008). “PAC-Bayesian bounds for randomized empirical risk minimizers”. *Mathematical Methods of Statistics*. 17(4): 279–304.

Alquier, P. (2013). “Bayesian methods for low-rank matrix estimation: short survey and theoretical study”. In: *International Conference on Algorithmic Learning Theory*. Springer. 309–323.

Alquier, P. and G. Biau. (2013). “Sparse single-index model.” *Journal of Machine Learning Research*. 14(1).

Alquier, P. and B. Guedj. (2018). “Simpler PAC-Bayesian bounds for hostile data”. *Machine Learning*. 107(5): 887–902.

Alquier, P., X. Li, and O. Wintenberger. (2013). “Prediction of time series by statistical learning: general losses and fast rates”. *Dependence Modeling*. 1(2013): 65–93.

Alquier, P. and K. Lounici. (2011). “PAC-Bayesian bounds for sparse regression estimation with exponential weights”. *Electronic Journal of Statistics*. 5: 127–145.

Alquier, P. and J. Ridgway. (2020). “Concentration of tempered posteriors and of their variational approximations”. *Annals of Statistics*. 48(3): 1475–1497.

Alquier, P., J. Ridgway, and N. Chopin. (2016). “On the properties of variational approximations of Gibbs posteriors”. *Journal of Machine Learning Research*. 17(239): 1–41.

Alquier, P. and O. Wintenberger. (2012). “Model selection for weakly dependent time series forecasting”. *Bernoulli*. 18(3): 883–913.

Ambroladze, A., E. Parrado-hernández, and J. Shawe-taylor. (2006). “Tighter PAC-Bayes bounds”. In: *Advances in Neural Information Processing Systems*. Ed. by B. Schölkopf, J. Platt, and T. Hoffman. Vol. 19. MIT Press.

Aminian, G., Y. Bu, L. Toni, M. R. D. Rodrigues, and G. Wornell. (2021). “Characterizing the deneralization error of Gibbs algorithm with symmetrized KL information”. *arXiv preprint arXiv:2107.13656*.

Amit, R. and R. Meir. (2018). “Meta-learning by adjusting priors based on extended PAC-Bayes theory”. In: *International Conference on Machine Learning*. PMLR. 205–214.

Aouali, I., V.-E. Brunel, D. Rohde, and A. Korba. (2023). “Exponential Smoothing for Off-Policy Learning”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett. Vol. 202. *Proceedings of Machine Learning Research*. PMLR. 984–1017.

Appert, G. and O. Catoni. (2021). “New bounds for k-means and information k-means”. *arXiv preprint arXiv:2101.05728*.

Asadi, A., E. Abbe, and S. Verdu. (2018). “Chaining mutual information and tightening generalization bounds”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc.

Audibert, J.-Y. (2004). “PAC-Bayesian statistical learning theory”. *PhD thesis, Université Paris VI*.

Audibert, J.-Y. (2009). “Fast learning rates in statistical inference through aggregation”. *The Annals of Statistics*. 37(4): 1591–1646.

Audibert, J.-Y. and O. Bousquet. (2007). “Combining PAC-Bayesian and generic chaining bounds”. *Journal of Machine Learning Research*. 8(4).

Audibert, J.-Y. and O. Catoni. (2011). “Robust linear least squares regression”. *The Annals of Statistics*. 39(5): 2766–2794.

Avena Medina, M., J. L. Montiel Olea, C. Rush, and A. Velez. (2021). “On the robustness to misspecification of α-posteriors and their variational approximations”. *arXiv preprint arXiv:2104.08324*.

Banerjee, A. (2006). “On Bayesian bounds”. In: *Proceedings of ICML*. ACM. 81–88.

Banerjee, I., V. A. Rao, and H. Honnappa. (2021a). “PAC-Bayes bounds on variational tempered posteriors for Markov models”. *Entropy*. 23(3): 313.

Banerjee, P. K. and G. Montúfar. (2021). “Information complexity and generalization bounds”. In: *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 676–681.

Banerjee, S., I. Castillo, and S. Ghosal. (2021b). “Bayesian inference in high-dimensional models”. *arXiv preprint arXiv:2101.04491*.

Barron, A., J. Rissanen, and B. Yu. (1998). “The minimum description length principle in coding and modeling”. *IEEE Transactions on Information Theory*. 44(6): 2743–2760.

Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe. (2006). “Convexity, classification, and risk bounds”. *Journal of the American Statistical Association*. 101(473): 138–156.

Bartlett, P. L. and S. Mendelson. (2006). “Empirical minimization”. *Probability theory and related fields*. 135(3): 311–334.

Barzdinš, J. and R. Freivalds. (1974). “Prediction and limiting synthesis of recursively enumerable classes of functions”. *Latvijas Valsts Univ. Zimatm. Raksti*. 210: 101–111.

Bassily, R., S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff. (2018). “Learners that use little information”. In: *Proceedings of Algorithmic Learning Theory*. Ed. by F. Janoos, M. Mohri, and K. Sridharan. Vol. 83. *Proceedings of Machine Learning Research*. PMLR. 25–55.

Bégin, L., P. Germain, F. Laviolette, and J.-F. Roy. (2016). “PAC-Bayesian bounds based on the Rényi divergence”. In: *Artificial Intelligence and Statistics*. PMLR. 435–444.

Bhattacharya, A., D. Pati, and Y. Yang. (2019). “Bayesian fractional posteriors”. *The Annals of Statistics*. 47(1): 39–66.

Biggs, F. and B. Guedj. (2021). "Differentiable PAC–Bayes objectives with partially aggregated neural networks". *Entropy*. 23(10).

Biggs, F. and B. Guedj. (2022). "On margins and derandomisation in PAC-Bayes". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 3709–3731.

Bilodeau, B., D. J. Foster, and D. M. Roy. (2021). "Minimax rates for conditional density estimation via empirical entropy". *arXiv preprint arXiv:2109.10461, to appear in the Annals of Statistics.*

Bissiri, P. G., C. C. Holmes, and S. G. Walker. (2016). "A general framework for updating belief distributions". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 78(5): 1103–1130.

Blanchard, G. and F. Fleuret. (2007). "Occam's hammer". In: *International Conference on Computational Learning Theory*. Springer. 112–126.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe. (2017). "Variational inference: A review for statisticians". *Journal of the American statistical Association*. 112(518): 859–877.

Boucheron, S., G. Lugosi, and P. Massart. (2013). *Concentration inequalities*. Oxford University Press.

Bu, Y., S. Zou, and V. V. Veeravalli. (2020). "Tightening mutual information-based bounds on generalization error". *IEEE Journal on Selected Areas in Information Theory*. 1(1): 121–130.

Bubeck, S. and N. Cesa-Bianchi. (2012). "Regret analysis of stochastic and nonstochastic multi-armed bandit problems". *Foundations and Trends® in Machine Learning*. 5(1): 1–122. DOI: 10.1561/2200000024.

Bunea, F. and A. Nobel. (2008). "Sequential procedures for aggregating arbitrary estimators of a conditional mean". *IEEE Transactions on Information Theory*. 54(4): 1725–1735.

Catoni, O. (2007). *PAC-Bayesian supervised classification: The thermodynamics of statistical learning. Institute of Mathematical Statistics Lecture Notes – Monograph Series, 56.* Institute of Mathematical Statistics, Beachwood, OH.

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

Catoni, O. (2004). *Statistical learning theory and stochastic optimization. Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics.* Springer. 1–269.

Catoni, O. (2012). "Challenging the empirical mean and empirical variance: a deviation study". In: *Annales de l'IHP Probabilités et statistiques*. Vol. 48. No. 4. 1148–1185.

Catoni, O. and I. Giulini. (2017). "Dimension free PAC-Bayesian bounds for the estimation of the mean of a random vector". In: *NIPS-2017 Workshop (Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights.*

Cesa-Bianchi, N., Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. (1997). "How to use expert advice". *Journal of the ACM*. 44(3): 427–485.

Cesa-Bianchi, N. and G. Lugosi. (2006). *Prediction, learning, and games.* Cambridge university press.

Chafaï, D., O. Guédon, G. Lecué, and A. Pajor. (2012). *Interactions between compressed sensing random matrices and high dimensional geometry*. Société Mathématique de France (SMF).

Chee, A. and S. Loustau. (2021). "Learning with BOT-Bregman and Optimal Transport divergences". *Preprint hal-03262687*.

Cherief-Abdellatif, B.-E. (2019). "Consistency of ELBO maximization for model selection". In: *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*. Ed. by F. Ruiz, C. Zhang, D. Liang, and T. Bui. Vol. 96. *Proceedings of Machine Learning Research*. PMLR. 11–31.

Chérief-Abdellatif, B.-E. (2020). "Convergence rates of variational inference in sparse deep learning". In: *International Conference on Machine Learning*. PMLR. 1831–1842.

Chérief-Abdellatif, B.-E. and P. Alquier. (2018). "Consistency of variational Bayes inference for estimation and model selection in mixtures". *Electronic Journal of Statistics*. 12(2): 2995–3035.

Chérief-Abdellatif, B.-E., P. Alquier, and M. E. Khan. (2019). "A generalization bound for online variational inference". In: *Proceedings of The Eleventh Asian Conference on Machine Learning*. Ed. by W. S. Lee and T. Suzuki. Vol. 101. *Proceedings of Machine Learning Research*. Nagoya, Japan. 662–677.

Chérief-Abdellatif, B.-E., Y. Shi, A. Doucet, and B. Guedj. (2022). "On PAC-Bayesian reconstruction guarantees for VAEs". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 3066–3079.

Chopin, N., S. Gadat, B. Guedj, A. Guyader, and E. Vernet. (2015). "On some recent advances on high dimensional Bayesian statistics". *ESAIM: Proceedings and Surveys*. 51: 293–319.

Chugg, B., H. Wang, and A. Ramdas. (2023). "A unified recipe for deriving (time-uniform) PAC-Bayes bounds". *arXiv preprint arXiv:2302.03421*.

Clerico, E., G. Deligiannidis, and A. Doucet. (2022a). "Conditionally Gaussian PAC-Bayes". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2311–2329.

Clerico, E., G. Deligiannidis, B. Guedj, and A. Doucet. (2022b). "A PAC-Bayes bound for deterministic classifiers". *arXiv preprint arXiv:2209.02525*.

Clerico, E., A. Shidani, G. Deligiannidis, and A. Doucet. (2022c). "Chained generalisation bounds". In: *Conference on Learning Theory*. PMLR. 4212–4257.

Clerico, E., A. Shidani, G. Deligiannidis, and A. Doucet. (2023). "Wide stochastic networks: Gaussian limit and PAC-Bayesian training". In: *International Conference on Algorithmic Learning Theory*. PMLR. 447–470.

Cottet, V. and P. Alquier. (2018). "1-bit matrix completion: PAC-Bayesian analysis of a variational approximation". *Machine Learning*. 107(3): 579–603.

Dai, D., P. Rigollet, L. Xia, and T. Zhang. (2014). "Aggregation of affine estimators". *Electronic Journal of Statistics*. 8(1): 302–327.

Dai, D., P. Rigollet, and T. Zhang. (2012). "Deviation optimal learning using greedy Q-aggregation". *The Annals of Statistics*. 40(3): 1878–1905.

Dalalyan, A. S., E. Grappin, and Q. Paris. (2018). "On the exponentially weighted aggregate with the Laplace prior". *Annals of Statistics*. 46(5): 2452–2478.

Dalalyan, A. S. and J. Salmon. (2012). “Sharp oracle inequalities for aggregation of affine estimators”. *The Annals of Statistics*. 40(4): 2327–2355.

Dalalyan, A. S. and A. B. Tsybakov. (2008). “Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity”. *Machine Learning*. 72(1-2): 39–61.

Dalalyan, A. S. and A. B. Tsybakov. (2012). “Sparse regression learning by aggregation and Langevin Monte-Carlo”. *Journal of Computer and System Sciences*. 78(5): 1423–1443.

Dalalyan, A. S. (2020). “Exponential weights in multivariate regression and a low-rankness favoring prior”. *Ann. Inst. H. Poincaré Probab. Statist.* 56(2): 1465–1483.

Dedecker, J., P. Doukhan, G. Lang, L. R. J. Rafael, S. Louhichi, and C. Prieur. (2007). “Weak dependence”. In: *Weak dependence: With examples and applications*. Springer. 9–20.

Devroye, L., L. Györfi, and G. Lugosi. (1996). *A probabilistic theory of pattern recognition*. Springer Science & Business Media.

Donsker, M. D. and S. S. Varadhan. (1976). “Asymptotic evaluation of certain Markov process expectations for large time. III.” *Communications on Pure and Applied Mathematics*. 28: 389–461.

Dziugaite, G. K., K. Hsu, W. Gharbieh, G. Arpino, and D. Roy. (2021a). “On the role of data in PAC-Bayes bounds”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 604–612.

Dziugaite, G. K. and D. M. Roy. (2017). “Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data”. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Dziugaite, G. K. and D. M. Roy. (2018). “Data-dependent PAC-Bayes priors via differential privacy”. In: *Advances in Neural Information Processing Systems*. 8430–8441.

Dziugaite, G. K., K. Hsu, W. Gharbieh, G. Arpino, and D. Roy. (2021b). “On the role of data in PAC-Bayes”. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Banerjee and K. Fukumizu. Vol. 130. *Proceedings of Machine Learning Research*. PMLR. 604–612.

Eringis, D., J. Leth, Z.-H. Tan, R. Wisniewski, and M. Petreczky. (2021). "PAC-Bayesian theory for stochastic LTI systems". In: IEEE. 6626–6633.

Fleuret, F. (2011). "Machine learning, PAC-learning". *Slides available on the author's website*. URL: https://fleuret.org/public/EN_20110511-pac/pac-fleuret-2011.pdf.

Foong, A. Y. K., W. P. Bruinsma, D. R. Burt, and R. E. Turner. (2021). "How tight can PAC-Bayes be in the small data regime?" In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc. 4093–4105.

Foret, P., A. Kleiner, H. Mobahi, and B. Neyshabur. (2020). "Sharpness-aware minimization for efficiently improving generalization". *arXiv preprint arXiv:2010.01412*.

Frazier, D. T., R. Loaiza-Maya, G. M. Martin, and B. Koo. (2021). "Loss-based variational Bayes prediction". *arXiv preprint arXiv:2104.14054*.

Gaïffas, S. and G. Lecué. (2007). "Optimal rates and adaptation in the single-index model using aggregation". *Electronic journal of statistics*. 1: 538–573.

Geffner, T. and J. Domke. (2020). "On the difficulty of unbiased alpha divergence minimization". *arXiv preprint arXiv:2010.09541*.

Germain, P., F. Bach, A. Lacoste, and S. Lacoste-Julien. (2016a). "PAC-Bayesian theory meets Bayesian inference". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 1884–1892.

Germain, P., A. Habrard, F. Laviolette, and E. Morvant. (2016b). "A new PAC-Bayesian perspective on domain adaptation". In: *International conference on machine learning*. PMLR. 859–868.

Germain, P., A. Lacasse, F. Laviolette, M. March, and J.-F. Roy. (2015). "Risk bounds for the majority vote: from a PAC-Bayesian analysis to a learning algorithm". *Journal of Machine Learning Research*. 16(26): 787–860.

Germain, P., A. Lacasse, F. Laviolette, and M. Marchand. (2009). "PAC-Bayesian learning of linear classifiers". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 353–360.

Ghosal, S. and A. Van der Vaart. (2017). *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press.

Giraud, C. (2014). *Introduction to high-dimensional statistics*. CRC Press.

Giulini, I. (2018). “Robust dimension-free Gram operator estimates”. *Bernoulli*. 24(4B): 3864–3923.

Głuch, G. and R. Urbanke. (2023). “Bayes complexity of learners vs overfitting”. *arXiv preprint arXiv:2303.07874*.

Grünwald, P., T. Steinke, and L. Zakynthinou. (2021). “PAC-Bayes, MAC-Bayes and conditional mutual information: fast rate bounds that handle general VC classes”. In: *Conference on Learning Theory*. PMLR. 2217–2247.

Grünwald, P. and T. Van Ommen. (2017). “Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it”. *Bayesian Analysis*. 12(4): 1069–1103.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.

Grünwald, P. D. and N. A. Mehta. (2020). “Fast rates for general unbounded loss functions: From ERM to Generalized Bayes”. *Journal of Machine Learning Research*. 21(56): 1–80.

Guedj, B. (2019). “A primer on PAC-Bayesian learning”. In: *Proceedings of the second congress of the French Mathematical Society*.

Guedj, B. and P. Alquier. (2013). “PAC-Bayesian estimation and prediction in sparse additive models”. *Electronic Journal of Statistics*. 7: 264–291.

Haddouche, M. and B. Guedj. (2022). “Online PAC-Bayes Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc. 25725–25738.

Haddouche, M. and B. Guedj. (2023). “PAC-Bayes generalisation bounds for heavy-tailed losses through supermartingales”. *Transactions on Machine Learning Research*.

Haddouche, M., B. Guedj, O. Rivasplata, and J. Shawe-Taylor. (2020). “Upper and lower bounds on the performance of kernel PCA”. *arXiv preprint arXiv:2012.10369*.

Haddouche, M., B. Guedj, O. Rivasplata, and J. Shawe-Taylor. (2021). "PAC-Bayes unleashed: generalisation bounds with unbounded losses". *Entropy*. 23(10).

Haghifam, M., G. K. Dziugaite, S. Moran, and D. Roy. (2021). "Towards a unified information-theoretic framework for generalization". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc. 26370–26381.

Haghifam, M., S. Moran, D. M. Roy, and G. K. Dziugiate. (2022). "Understanding generalization via leave-one-out conditional mutual information". In: *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2487–2492.

Haghifam, M., J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite. (2020). "Sharpened generalization bounds based on conditional mutual information and anapplication to noisy, iterative algorithms". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc. 9925–9935.

Haghifam, M., B. Rodríguez-Gálvez, R. Thobaben, M. Skoglund, D. M. Roy, and G. K. Dziugaite. (2023). "Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization". In: *International Conference on Algorithmic Learning Theory*. PMLR. 663–706.

Haußmann, M., S. Gerwinn, A. Look, B. Rakitsch, and M. Kandemir. (2021). "Learning partially known stochastic dynamics with empirical PAC-Bayes". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 478–486.

Hellström, F. and G. Durisi. (2020). "Generalization bounds via information density and conditional information density". *IEEE Journal on Selected Areas in Information Theory*. 1(3): 824–839.

Herbrich, R. and T. Graepel. (2002). "A PAC-Bayesian margin bound for linear classifiers". *IEEE Transactions on Information Theory*. 48(12): 3140–3150.

Higgs, M. and J. Shawe-Taylor. (2010). "A PAC-Bayes bound for tailored density estimation". In: *International Conference on Algorithmic Learning Theory*. Springer. 148–162.

Hinton, G. E. and D. Van Camp. (1993). "Keeping the neural networks simple by minimizing the description length of the weights". In: *Proceedings of the sixth annual conference on Computational learning theory*. 5–13.

Hoeven, D., T. Erven, and W. Kotłowski. (2018). "The many faces of exponential weights in online learning". In: *Conference On Learning Theory*. PMLR. 2067–2092.

Holland, M. (2019). "PAC-Bayes under potentially heavy tails". *Advances in Neural Information Processing Systems*. 32: 2715–2724.

Honorio, J. and T. Jaakkola. (2014). "Tight bounds for the expected risk of linear classifiers and PAC-Bayes finite-sample guarantees". In: *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*. 384–392.

Huggins, J. H., T. Campbell, M. Kasprzak, and T. Broderick. (2018). "Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach". *arXiv preprint arXiv:1809.09505*.

Jaiswal, P., V. Rao, and H. Honnappa. (2020). "Asymptotic consistency of α-Rényi-approximate posteriors". *Journal of Machine Learning Research*. 21(156): 1–42.

Jang, K., K.-S. Jun, I. Kuzborskij, and F. Orabona. (2023). "Tighter PAC-Bayes Bounds Through Coin-Betting". In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Ed. by G. Neu and L. Rosasco. Vol. 195. *Proceedings of Machine Learning Research*. PMLR. 2240–2264.

Jiang, W. and M. A. Tanner. (2008). "Gibbs posterior for variable selection in high-dimensional classification and data mining". *The Annals of Statistics*: 2207–2231.

Jin, G., Y. X., P. Yang, L. Zhang, S. Schewe, and X. Huang. (2022). "Weight expansion: a new perspective on dropout and generalization". *arXiv preprint arXiv:2201.09209*.

Jose, S. T. and O. Simeone. (2021). "Transfer meta-learning: Information-theoretic bounds and information meta-risk minimization". *IEEE Transactions on Information Theory*. 68(1): 474–501.

Juditsky, A. and A. Nemirovski. (2000). "Functional aggregation for nonparametric regression". *Annals of Statistics*: 681–712.

Juditsky, A., P. Rigollet, and A. B. Tsybakov. (2008). "Learning by mirror averaging". *The Annals of Statistics*. 36(5): 2183–2206.

Kakade, S. M., K. Sridharan, and A. Tewari. (2008). "On the complexity of linear prediction: Risk bounds, margin bounds, and regularization". In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc.

Khan, M. E. and H. Rue. (2021). "The Bayesian Learning Rule". *arXiv preprint arXiv:2107.04562*.

Kivinen, J. and M. K. Warmuth. (1999). "Averaging expert predictions". In: *European Conference on Computational Learning Theory*. Springer, Berlin. 153–167.

Knoblauch, J., J. Jewson, and T. Damoulas. (2022). "An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference". *Journal of Machine Learning Research*. 23(132): 1–109.

Kullback, S. (1959). *Information theory and statistics*. John Wiley & Sons.

Lacasse, A., F. Laviolette, M. Marchand, P. Germain, and N. Usunier. (2006). "PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier". In: *Advances in Neural Information Processing Systems*. Ed. by B. Schölkopf, J. Platt, and T. Hoffman. Vol. 19. MIT Press.

Lan, X., X. Guo, and K. E. Barner. (2020). "PAC-Bayesian generalization bounds for multiLayer perceptrons". *Preprint arXiv:2006.08888*.

Langford, J. and A. Blum. (2003). "Microchoice bounds and self-bounding learning algorithms". *Machine Learning*. 51(2): 165–179.

Langford, J. and R. Caruana. (2002). "(Not) bounding the true error". *Advances in Neural Information Processing Systems*. 2: 809–816.

Langford, J. and M. Seeger. (2001). "Bounds for averaging classifiers". *Technical Report CMU-CS-01-102, Carnegie Mellon University*.

Langford, J. and J. Shawe-Taylor. (2002). "PAC-Bayes & margins". In: *Proceedings of the 15th International Conference on Neural Information Processing Systems*. MIT Press. 439–446.

Laviolette, F., M. Marchand, and J.-F. Roy. (2011). "From PAC-Bayes bounds to quadratic programs for majority votes". In: *Proceedings of International Conference on Machine Learning*. Citeseer. 5–59.

Lecué, G. (2007). "Aggregation procedures: optimality and fast rates". *PhD thesis*. Université Pierre et Marie Curie-Paris VI.

Lecué, G. and S. Mendelson. (2013). "On the optimality of the aggregate with exponential weights for low temperatures". *Bernoulli*. 19(2): 646–675.

Lepski, O. (1992). "Asymptotically minimax adaptive estimation I: upper bounds". *Theory of Probability and its Applications*. 36(4): 682–697.

Letarte, G., P. Germain, B. Guedj, and F. Laviolette. (2019). "Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks". In: *Advances in Neural Information Processing Systems*. 6872–6882.

Leung, G. and A. R. Barron. (2006). "Information theory and mixing least-squares regressions". *IEEE Trans. Inform. Theory*. 52(8): 3396–3410.

Lever, G., F. Laviolette, and J. Shawe-Taylor. (2010). "Distribution-dependent PAC-Bayes Priors". In: *Proceedings of the 15th International Conference on Algorithmic Learning Theory*. Berlin, Heidelberg: Springer. 119–133.

Lever, G., F. Laviolette, and J. Shawe-Taylor. (2013). "Tighter PAC-Bayes bounds through distribution-dependent priors". *Theoretical Computer Science*. 473: 4–28.

Littlestone, N. and M. K. Warmuth. (1989). "The weighted majority algorithm". In: *Proceedings of the 30th Annual Symposium on the Foundations of Computer Science*. IEEE. 256–261.

Liu, T., J. Lu, Z. Yan, and G. Zhang. (2021a). "PAC-Bayes bounds for meta-learning with data-dependent prior". *arXiv preprint arXiv:2102. 03748*.

Liu, T., J. Lu, Z. Yan, and G. Zhang. (2021b). "Statistical generalization performance guarantee for meta-learning with data dependent prior". *Neurocomputing*. 465: 391–405.

Livni, R. and S. Moran. (2020). "A limitation of the PAC-Bayes framework". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc. 20543–20553.

London, B. (2017). “A PAC-Bayesian analysis of randomized learning with application to stochastic gradient descent”. In: *Advances in Neural Information Processing Systems*. 2931–2940.

London, B. and T. Sandler. (2019). “Bayesian counterfactual risk minimization”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. *Proceedings of Machine Learning Research*. PMLR. 4125–4133.

Lorenzen, S. S., C. Igel, and Y. Seldin. (2019). “On PAC-Bayesian bounds for random forests”. *Machine Learning*. 108(8-9): 1503–1522.

Lugosi, G. and G. Neu. (2021). “Online-to-PAC conversions: Generalization bounds via regret analysis”. *arXiv preprint arXiv:2305.19674*.

Lugosi, G. and G. Neu. (2022). “Generalization bounds via convex analysis”. In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by P.-L. Loh and M. Raginsky. Vol. 178. *Proceedings of Machine Learning Research*. PMLR. 3524–3546.

Luu, T. D., J. Fadili, and C. Chesneau. (2019). “PAC-Bayesian risk bounds for group-analysis sparse regression by exponential weighting”. *Journal of Multivariate Analysis*. 171: 209–233.

Mai, T. T. (2017). “PAC-Bayesian estimation of low-rank matrices”. *PhD thesis, Université Paris Saclay*.

Mai, T. T. (2023a). “From bilinear regression to inductive matrix completion: a quasi-Bayesian analysis”. *Entropy*. 25(2): 333.

Mai, T. T. (2023b). “Simulation comparisons between Bayesian and de-biased estimators in low-rank matrix completion”. *METRON*: 1–22.

Mai, T. T. and P. Alquier. (2015). “A Bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution”. *Electronic Journal of Statistics*. 9(1): 823–841.

Mai, T. T. and P. Alquier. (2017). “Pseudo-Bayesian quantum tomography with rank-adaptation”. *Journal of Statistical Planning and Inference*. 184: 62–76.

Mammen, E. and A. B. Tsybakov. (1999). “Smooth discrimination analysis”. *The Annals of Statistics*. 27(6): 1808–1829.

Masegosa, A., S. Lorenzen, C. Igel, and Y. Seldin. (2020). "Second order PAC-Bayesian bounds for the weighted majority vote". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc. 5263–5273.

Masegosa, A. R. (2020). "Learning under model misspecification: Applications to variational and ensemble methods". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc. 5479–5491.

Maurer, A. (2004). "A note on the PAC Bayesian theorem". *arXiv preprint cs/0411099*.

Mbacke, S. D., F. Clerc, and P. Germain. (2023). "PAC-Bayesian generalization bounds for adversarial generative models". *arXiv preprint arXiv:2302.08942*.

McAllester, D. A. (1998). "Some PAC-Bayesian theorems". In: *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. New York: ACM. 230–234.

McAllester, D. A. (1999). "PAC-Bayesian model averaging". In: *Proceedings of the twelfth annual conference on Computational learning theory*. 164–170.

McAllester, D. A. (2003). "PAC-Bayesian stochastic model selection". *Machine Learning*. 51(1): 5–21.

McAllester, D. A. (2013). "A PAC-Bayesian tutorial with a dropout bound". *arXiv preprint arXiv:1307.2118*.

McDiarmid, C. (1998). "Concentration". In: *Probabilistic methods for algorithmic discrete mathematics*. Ed. by M. Habib, C. McDiarmid, and B. Reed. Springer. 195–248.

Meir, R. and T. Zhang. (2003). "Generalization error bounds for Bayesian mixture algorithms". *Journal of Machine Learning Research*. 4(Oct): 839–860.

Meunier, D. and P. Alquier. (2021). "Meta-strategy for learning tuning parameters with guarantees". *Entropy*. 23(10).

Mhammedi, Z., P. D. Grünwald, and B. Guedj. (2019). "PAC-Bayes unexpected Bernstein inequality". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.

Mourtada, J., T. Vaškevičius, and N. Zhivotovskiy. (2023). "Local Risk Bounds for Statistical Aggregation". *arXiv preprint arXiv:2306.17151*.

Nachum, I., J. Shafer, and A. Yehudayoff. (2018). "A direct sum result for the information complexity of learning". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. *Proceedings of Machine Learning Research*. PMLR. 1547–1568.

Nakakita, S., P. Alquier, and M. Imaizumi. (2022). "Dimension-free bounds for sum of dependent matrices and operators with heavy-tailed distribution". *arXiv preprint arXiv:2210.09756*.

Negrea, J., M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy. (2019). "Information-theoretic generalization bounds for SGLD via data-dependent estimates". *Advances in Neural Information Processing Systems*. 32: 11015–11025.

Nemirovski, A. (2000). "Topics in non-parametric statistics". *Ecole d'Eté de Probabilités de Saint-Flour*. 28: 85.

Neu, G., G. K. Dziugaite, M. Haghifam, and D. M. Roy. (2021). "Information-theoretic generalization bounds for stochastic gradient descent". In: *Conference on Learning Theory*. PMLR. 3526–3545.

Neyshabur, B., S. Bhojanapalli, D. McAllester, and N. Srebro. (2017). "A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks". *NIPS 2017 Workshop: (Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights*.

Nozawa, K., P. Germain, and B. Guedj. (2020). "PAC-Bayesian contrastive unsupervised representation learning". In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by J. Peters and D. Sontag. Vol. 124. *Proceedings of Machine Learning Research*. PMLR. 21–30.

Nozawa, K. and I. Sato. (2019). "PAC-Bayes analysis of sentence representation". *arXiv preprint arXiv:1902.04247*.

Ohn, I. and L. Lin. (2021). “Adaptive variational Bayes: Optimality, computation and applications”. *arXiv preprint arXiv:2109.03204*.

Ohnishi, Y. and J. Honorio. (2021). “Novel change of measure inequalities with applications to PAC-Bayesian bounds and Monte Carlo estimation”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 1711–1719.

Oneto, L., M. Donini, M. Pontil, and J. Shawe-Taylor. (2020). “Randomized learning and generalization of fair and private classifiers: From PAC-Bayes to stability and differential privacy”. *Neurocomputing*. 416: 231–243.

Orabona, F. (2019). “A modern introduction to online learning”. *arXiv preprint arXiv:1912.13213*.

Parrado-Hernández, E., A. Ambroladze, J. Shawe-Taylor, and S. Sun. (2012). “PAC-Bayes bounds with data dependent priors”. *The Journal of Machine Learning Research*. 13(1): 3507–3531.

Pentina, A. and C. Lampert. (2014). “A PAC-Bayesian bound for lifelong learning”. In: *International Conference on Machine Learning*. PMLR. 991–999.

Pérez-Ortiz, M., O. Rivasplata, J. Shawe-Taylor, and C. Szepesvári. (2021). “Tighter risk certificates for neural networks”. *The Journal of Machine Learning Research*. 22(1): 10326–10365.

Pitas, K. (2020). “Dissecting non-vacuous generalization bounds based on the mean-field approximation”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. *Proceedings of Machine Learning Research*. PMLR. 7739–7749.

Plummer, S., D. Pati, and A. Bhattacharya. (2020). “Dynamics of coordinate ascent variational inference: A case study in 2D Ising models”. *Entropy*. 22(11).

Raginsky, M., A. Rakhlin, M. Tsao, Y. Wu, and A. Xu. (2016). “Information-theoretic analysis of stability and bias of learning algorithms”. In: *2016 IEEE Information Theory Workshop (ITW)*. IEEE. 26–30.

Ralaivola, L., M. Szafranski, and G. Stempfel. (2010). “Chromatic PAC-Bayes bounds for non-i.i.d. data: Applications to ranking and stationary β-mixing processes”. *Journal of Machine Learning Research*. 11(Jul): 1927–1956.

Rezazadeh, A. (2022). "A general framework for PAC-Bayes bounds for meta-learning". *arXiv preprint arXiv:2206.05454*.

Ridgway, J., P. Alquier, N. Chopin, and F. Liang. (2014). "PAC-Bayesian AUC classification and scoring". *Advances in Neural Information Processing Systems*. 1(January): 658–666.

Rio, E. (2000). "Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes". *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*. 330(10): 905–908.

Riou, C., P. Alquier, and B.-E. Chérief-Abdellatif. (2023). "Bayes meets Bernstein at the meta level: an analysis of fast rates in meta-learning with PAC-Bayes". *arXiv preprint arXiv:2302.11709*.

Rissanen, J. (1978). "Modeling by shortest data description". *Automatica*. 14(5): 465–471.

Rivasplata, O., I. Kuzborskij, C. Szepesvári, and J. Shawe-Taylor. (2020). "PAC-Bayes analysis beyond the usual bounds". In: *Advances in Neural Information Processing Systems*.

Rivasplata, O., V. M. Tankasali, and C. Szepesvari. (2019). "PAC-Bayes with backprop". *arXiv preprint arXiv:1908.07380*.

Rodrígues-Gálvez, B., R. Thobaden, and M. Skoglund. (2023). "More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime-validity". *arXiv preprint arXiv:2306.12214*.

Rodríguez-Gálvez, B., G. Bassi, R. Thobaben, and M. Skoglund. (2021). "Tighter expected generalization error bounds via Wasserstein distance". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc. 19109–19121.

Rothfuss, J., V. Fortuin, M. Josifoski, and A. Krause. (2021). "PACOH: Bayes-optimal meta-learning with PAC-guarantees". In: *International Conference on Machine Learning*. PMLR. 9116–9126.

Rousseau, J. (2016). "On the frequentist properties of Bayesian nonparametric methods". *Annual Review of Statistics and Its Application*. 3: 211–231.

Russo, D. and J. Zou. (2019). "How much does your data exploration overfit? Controlling bias via information usage". *IEEE Transactions on Information Theory*. 66(1): 302–323.

Sakhi, O., P. Alquier, and N. Chopin. (2023). “PAC-Bayesian Offline Contextual Bandits With Guarantees”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett. Vol. 202. *Proceedings of Machine Learning Research*. PMLR. 29777–29799.

Samson, P.-M. (2000). “Concentration of measure inequalities for Markov chains and Φ-mixing processes”. *The Annals of Probability*. 28(1): 416–461.

Seeger, M. (2002). “PAC-Bayesian generalisation error bounds for Gaussian process classification”. *Journal of machine learning research*. 3(Oct): 233–269.

Seeger, M. (2003). “Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations”. *Tech. rep.* University of Edinburgh.

Seldin, Y., P. Auer, J. Shawe-Taylor, R. Ortner, and F. Laviolette. (2011). “PAC-Bayesian analysis of contextual bandits”. In: *Advances in Neural Information Processing Systems*. 1683–1691.

Seldin, Y., N. Cesa-Bianchi, P. Auer, F. Laviolette, and J. Shawe-Taylor. (2012a). “PAC-Bayes-Bernstein inequality for martingales and its application to multiarmed bandits”. In: *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*. Ed. by D. Glowacka, L. Dorard, and J. Shawe-Taylor. Vol. 26. *Proceedings of Machine Learning Research*. Bellevue, Washington, USA: PMLR. 98–111.

Seldin, Y., F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. (2012b). “PAC-Bayesian inequalities for martingales”. *IEEE Transactions on Information Theory*. 58(12): 7086–7093.

Seldin, Y. and N. Tishby. (2010). “PAC-Bayesian analysis of co-clustering and beyond.” *Journal of Machine Learning Research*. 11(12).

Shalev-Shwartz, S. (2011). “Online learning and online convex optimization”. *Foundations and Trends® in Machine Learning*. 4(2): 107–194.

Shawe-Taylor, J. and R. Williamson. (1997). “A PAC analysis of a Bayes estimator”. In: *Proceedings of the Tenth Annual Conference on Computational Learning Theory*. New York: ACM. 2–9.

Sheth, R. and R. Khardon. (2017). “Excess risk bounds for the Bayes risk using variational inference in latent Gaussian models”. In: *Advances in Neural Information Processing Systems*. 5151–5161.

Steffen, M. F. and M. Trabs. (2022). “PAC-Bayes training for neural networks: sparsity and uncertainty quantification”. *arXiv preprint arXiv:2204.12392*.

Steinke, T. and L. Zakynthinou. (2020). “Reasoning about generalization via conditional mutual information”. In: *Conference on Learning Theory*. PMLR. 3437–3452.

Sucker, M. and P. Ochs. (2023). “PAC-Bayesian learning of optimization algorithms”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 8145–8164.

Suzuki, T. (2012). “PAC-Bayesian bound for Gaussian process regression and multiple kernel additive model”. In: *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings. 8–1.

Suzuki, T. (2015). “Convergence rate of Bayesian tensor estimator and its minimax optimality”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. *Proceedings of Machine Learning Research*. Lille, France: PMLR. 1273–1282.

Suzuki, T. (2020). “Generalization bound of globally optimal non-convex neural network training: Transportation map estimation by infinite dimensional Langevin dynamics”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc. 19224–19237.

Syring, N. and R. Martin. (2019). “Calibrating general posterior credible regions”. *Biometrika*. 106(2): 479–486.

Syring, N. and R. Martin. (2023). “Gibbs posterior concentration rates under sub-exponential type losses”. *Bernoulli*. 29(2): 1080–1108.

Tasdighi, B., A. Akgül, K. K. Brink, and M. Kandemir. (2023). “PAC-Bayesian soft actor-critic learning”. *arXiv preprint arXiv:2301.12776*.

Thiemann, N., C. Igel, O. Wintenberger, and Y. Seldin. (2017). "A strongly quasiconvex PAC-Bayesian bound". In: *International Conference on Algorithmic Learning Theory*. 466–492.

Tolstikhin, I. and Y. Seldin. (2013). "PAC-Bayes-empirical-Bernstein inequality". *Advances in Neural Information Processing Systems 26 (NIPS 2013)*: 1–9.

Tsuzuku, Y., I. Sato, and M. Sugiyama. (2020). "Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis". In: *International Conference on Machine Learning*. PMLR. 9636–9647.

Tsybakov, A. B. (2003). "Optimal rates of aggregation". In: *Computational Learning Theory and Kernel Machines. Proc. 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines*. Ed. by B. Schölkopf and M. Warmuth. Springer Lecture Notes in Artificial Intelligence. 303–313.

Valiant, L. (1984). "A theory of the learnable". *Communications of the ACM*. 27(11): 1134–1142.

van Erven, T. (2014). "PAC-Bayes mini-tutorial: a continuous union bound". *arXiv preprint arXiv:1405.1580*.

Vapnik, V. (1998). *Statistical learning theory*. Wiley–Blackwell.

Vapnik, V. N. and A. Y. Chervonenkis. (1968). "The uniform convergence of frequencies of the appearance of events to their probabilities". *Doklady Akademii Nauk*. 181(4): 781–783.

Viallard, P., R. Emonet, P. Germain, A. Habrard, and E. Morvant. (2019). " Interpreting neural networks as majority votes through the PAC-Bayesian theory". *NeurIPS 2019 Workshop on Machine Learning with Guarantees*.

Vovk, V. G. (1990). "Aggregating strategies". *Proceedings of Computational Learning Theory, 1990*.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press.

Wang, H., S. Zheng, C. Xiong, and R. Socher. (2019). "On the generalization gap inreparameterizable reinforcement learning". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. *Proceedings of Machine Learning Research*. PMLR. 6648–6658.

Wang, Z., S.-L. Huang, E. E. Kuruoglu, J. Sun, X. Chen, and Y. Zheng. (2021). “PAC-Bayes information bottleneck”. *arXiv preprint arXiv:2109.14509*.

Wintenberger, O. (2010). “Deviation inequalities for sums of weakly dependent time series”. *Electronic Communications in Probability*. (15): 489–503.

Wu, Y.-S., A. Masegosa, S. Lorenzen, C. Igel, and Y. Seldin. (2021). “Chebyshev-Cantelli PAC-Bayes-Bennett inequality for the weighted majority vote”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc. 12625–12636.

Wu, Y.-S. and Y. Seldin. (2022). “Split-kl and PAC-Bayes-split-kl inequalities for ternary random variables”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc. 11369–11381.

Xu, A. and M. Raginsky. (2017). “Information-theoretic analysis of generalization capability of learning algorithms”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS’17*. Long Beach, California, USA: Curran Associates Inc. 2521–2530.

Yang, J., S. Sun, and D. M. Roy. (2019). “Fast-rate PAC-Bayes generalization bounds via shifted Rademacher processes”. *Advances in Neural Information Processing Systems*. 32: 10803–10813.

Yang, Y. (2001). “Adaptive regression by mixing”. *Journal of the American Statistical Association*. 96(454): 574–588.

Yang, Y. (2004). “Aggregating regression procedures to improve performance”. *Bernoulli*. 10(1): 25–47.

Yang, Y., D. Pati, and A. Bhattacharya. (2020). “α-variational inference with statistical guarantees”. *Annals of Statistics*. 48(2): 886–905.

Zhang, F. and C. Gao. (2020). “Convergence rates of variational posterior distributions”. *Annals of Statistics*. 48(4): 2180–2207.

Zhang, T. (2006). “Information-theoretic upper and lower bounds for statistical estimation”. *IEEE Transactions on Information Theory*. 52(4): 1307–1321.

Zhang, X., A. Ghosh, G. Liu, and R. Wang. (2023). “Auto-tune: PAC-Bayes optimization over prior and posterior for neural networks”. *arXiv preprint arXiv:2305.19243*.

Zhivotovskiy, N. (2021). “Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle”. *arXiv preprint arXiv:2108.08198*.

Zhou, W., V. Veitch, M. Austern, R. P. Adams, and P. Orbanz. (2018). “Non-vacuous generalization bounds at the imagenet scale: a PAC-Bayesian compression approach”. *arXiv preprint arXiv:1804.05862*.