Université de Lille

N°d'ordre NNT : ?

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LILLE

École Doctorale MADIS N° 631

Mathématiques-Sciences du numérique et de leurs interactions

Spécialité / discipline de doctorat : Mathématique

Soutenue par :

Maxime Haddouche

A PAC-Bayes Approach of Generalisation In Machine Learning: From Heavy-Tailed Martingales To Optimisation

Acknowledgements

TODO

The further backward you look, the further forward you can see.

Winston Churchill

Science is a bit like the joke about the drunk who is looking under a lamppost for a key that he has lost on the other side of the street, because that's where the light is. It has no other choice.

Noam Chomsky

Contents

List of Figures

List of Algorithms

List of Theorems

Appendix

List of Notations

General

a	A scalar (integer or real)
$\mathbf{a}$	A vector
$\mathbf{A}$	A matrix
$\mathbb{A}, \mathbb{a}$	A set
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_*$	The set of real numbers excluding 0
$\mathbb{R}_*^+$	The set of positive real numbers excluding 0
$\mathbb{N}$	The set of natural numbers
$\mathbb{N}_*$	The set of natural numbers excluding 0
$\mathrm{card}(\cdot)$	The cardinal of a set
a_i	i-th element of the vector $\mathbf{a}$

Statistical Learning Theory

$\mathbb{X}$	Set of d-dimensional inputs ($\subseteq \mathbb{R}^d$)
$\mathbb{Y}$	Set of labels
$\mathbf{x}$	A real-valued input $\mathbf{x} \in \mathbb{X}$
y	A label $y \in \mathbb{Y}$ associated to the input $\mathbf{x}$
$\mathcal{D}$	Unknown data distribution on $\mathbb{X} \times \mathbb{Y}$
$\mathcal{D}^m$	Unknown data distribution on the m-samples, *i.e.*, on $(\mathbb{X}\times\mathbb{Y})^m$
$\mathbb{S}$	Learning sample $\mathbb{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ drawn from $\mathcal{D}^m$
$\mathcal{S}$	Uniform distribution on $\mathbb{S}$
$\mathbb{T}$	Test set drawn from $\mathcal{D}^m$

$\mathcal{T}$	Uniform distribution on $\mathbb{T}$
$\mathbb{H}$	The set of hypotheses
h	A hypothesis $h \in \mathbb{H}$
$\ell(\cdot, \cdot)$	Loss function
$\mathrm{R}^{\ell}_{\mathcal{D}'}(h)$	Risk of the hypothesis $h \in \mathbb{H}$ *w.r.t.* the loss function $\ell()$ on $\mathcal{D}'$
$\mathrm{R}_{\mathcal{D}'}(h)$	Risk of the hypothesis $h \in \mathbb{H}$ *w.r.t.* the 0-1 loss on $\mathcal{D}'$

Probability Theory

$\mathbb{E}_{X\sim\mathcal{X}}[\cdot]$	The expectation *w.r.t.* the random variable $X \sim \mathcal{X}$
$\mathbb{P}_{X\sim\mathcal{X}}[\cdot]$	The probability *w.r.t.* the random variable $X \sim \mathcal{X}$
$\mathrm{I}[a]$	Indicator function; returns 1 if a is true and 0 otherwise
$\mathbb{M}(\mathbb{H})$	Set of Probability densities *w.r.t.* the reference measure on $\mathbb{H}$
ρ	Posterior distribution $\rho \in \mathbb{M}(\mathbb{H})$ on $\mathbb{H}$
π	Prior distribution $\pi \in \mathbb{M}^*(\mathbb{H})$ on $\mathbb{H}$
$\mathrm{KL}(\rho\|\pi)$	Kullback-Leibler (KL) divergence between ρ and π
$\mathrm{D}_{\alpha}(\rho\|\pi)$	Rényi Divergence between ρ and π
$\mathrm{Uni}(\mathbb{A})$	Uniform distribution on $\mathbb{A}$
$\mathrm{Dir}(\boldsymbol{\alpha})$	Dirichlet distribution of parameters $\boldsymbol{\alpha} \in \mathbb{R}^{+}_{*}$

Majority Vote

$\mathrm{MV}_{\rho}(\cdot)$	The majority vote classifier
$m_{\rho}(\mathbf{x}, y)$	Majority vote's margin for the example $(\mathbf{x}, y) \in \mathbb{X}\times\mathbb{Y}$
$\widehat{m_{\rho}}(\mathbf{x}, y)$	$\frac{1}{2}$-Margin for the example $(\mathbf{x}, y) \in \mathbb{X}\times\mathbb{Y}$
$\mathrm{sign}(a)$	Sign function; returns $+1$ if $a \geq 0$ and -1 otherwise
$r_{\mathcal{D}'}(\rho)$	Gibbs risk on the distribution $\mathcal{D}'$ associated to the majority vote $\mathrm{MV}_{\rho}()$
$e_{\mathcal{D}'}(\rho)$	Joint Error on the distribution $\mathcal{D}'$ associated to the majority vote $\mathrm{MV}_{\rho}()$
$d_{\mathcal{D}'}(\rho)$	Disagreement on the distribution $\mathcal{D}'$ associated to the majority vote $\mathrm{MV}_{\rho}()$

List of Publications

International Conference

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. Self-bounding Majority Vote Learning Algorithms by the Direct Minimization of a Tight PAC-Bayesian C-Bound. *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. (2021a).

Paul Viallard, Guillaume Vidot, Amaury Habrard, and Emilie Morvant. A PAC-Bayes Analysis of Adversarial Robustness. *Advances in Neural Information Processing Systems (NeurIPS)*. (2021d).

Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrard, Pascal Germain, and Benjamin Guedj. Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound. *Advances in Neural Information Processing Systems (NeurIPS)*. (2021).

International Workshop

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. Interpreting Neural Networks as Majority Votes through the PAC-Bayesian Theory. *NeurIPS 2019 Workshop on Machine Learning with Guarantees*. (2019).

National Conference

Paul Viallard, Rémi Emonet, Pascal Germain, Amaury Habrard, Emilie Morvant, and Valentina Zantedeschi. Intérêt des bornes désintégrées pour la généralisation avec des mesures de complexité. *Conférence sur l'Apprentissage automatique (CAp)*. (2022a).

Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrard, Pascal Germain, and Benjamin Guedj. Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound. *Conférence sur l'Apprentissage automatique (CAp)*. (2022).

Paul Viallard, Pascal Germain, and Emilie Morvant. Apprentissage de Vote de Majorité par Minimisation d'une C-Borne PAC-Bayésienne. *Conférence sur l'Apprentissage automatique (CAp)*. (2021b).

Paul Viallard, Pascal Germain, and Emilie Morvant. Dérandomisation des Bornes PAC-Bayésiennes. *Conférence sur l'Apprentissage automatique (CAp)*. (2021c).

Guillaume Vidot, Paul Viallard, and Emilie Morvant. Une Analyse PAC-Bayésienne de la Robustesse Adversariale. *Conférence sur l'Apprentissage automatique (CAp)*. (2021).

Paul Viallard, Rémi Emonet, Amaury Habrard, Emilie Morvant, and Pascal Germain. Théorie PAC-Bayésienne pour l'apprentissage en deux étapes de réseaux de neurones. *Conférence sur l'Apprentissage automatique (CAp)*. (2020).

Research Report

Paul Viallard, Rémi Emonet, Amaury Habrard, Emilie Morvant, and Valentina Zantedeschi. Generalization Bounds with Arbitrary Complexity Measures. *Submitted to ICLR 2023*. (2022b).

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A General Framework for the Disintegration of PAC-Bayesian Bounds. *Submitted to Machine Learning Journal*. (2022c).

Part I

Background

Part II

PAC-Bayesian Majority Vote:
Theory and Self-bounding Algorithms

Part III

Beyond PAC-Bayesian Bounds:
From Disintegration to Novel Bounds

Part IV

Conclusion and Perspectives

Part V

Appendix

Some Mathematical Tools

A

A.1 Jensen's Inequality

Theorem A.1.1 (Jensen's Inequality)**.** Let $X \in \mathbb{X}$ a random variable following a probability distribution $\mathcal{X}$ with $f : \mathbb{X} \to \mathbb{R}$ a measurable convex function, we have

$$f\left(\mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right) \leq \mathop{\mathbb{E}}_{X\sim\mathcal{X}}\Big[f\left(X\right)\Big].$$

Proof. Since $f()$ is a convex function, the following inequality holds, *i.e.*, we have

$$\forall X' \in \mathbb{X}, \quad a\left(X' - \mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right) \leq f(X') - f\left(\mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right),$$

where a is the tangent's slope. By taking the expectation to both sides of the inequality, we have

$$\underbrace{a\left(\mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X] - \mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right)}_{=0} \leq \mathop{\mathbb{E}}_{X\sim\mathcal{X}}[f(X)] - f\left(\mathop{\mathbb{E}}_{X\sim\mathcal{X}}[X]\right).$$

Hence, by rearranging the terms, we prove the claimed result. ■

A.2 Markov's Inequality

Theorem A.2.1 (Markov's Inequality)**.** Let $X \in \mathbb{X}$ a non-negative random variable following a probability distribution $\mathcal{X}$ and $\tau > 0$, we have

$$\mathop{\mathbb{P}}_{X\sim\mathcal{X}}[X \geq \tau] \leq \frac{\mathbb{E}_{X\sim\mathcal{X}}[X]}{\tau}.$$

Proof. First of all, remark that we have the following inequality for any $X \in \mathbb{X}$

$$\tau \, \mathrm{I}[X \geq \tau] \;\leq\; X \, \mathrm{I}[X \geq \tau] \;\leq\; X. \tag{A.1}$$

Indeed, on the one hand, if $X < \tau$, $\mathrm{I}[X \geq \tau] = 0$, the inequality holds trivially. On the other hand, if $X \geq \tau$, $\mathrm{I}[X \geq \tau] = 1$ and the inequality becomes $\tau \leq X$, which is true. By taking the expectation of Equation (A.1), we have

$$\mathop{\mathbb{E}}_{X \sim \mathcal{X}} \Big[\tau \, \mathrm{I}[X \geq \tau] \Big] \leq \mathop{\mathbb{E}}_{X \sim \mathcal{X}} \Big[X \Big].$$

From the fact that the expectation of a constant is the constant and by definition of the probability, we have

$$\tau \mathop{\mathbb{P}}_{X \sim \mathcal{X}} [X \geq \tau] \leq \mathop{\mathbb{E}}_{X \sim \mathcal{X}} \Big[X \Big] \quad \Longleftrightarrow \quad \mathop{\mathbb{P}}_{X \sim \mathcal{X}} [X \geq \tau] \leq \frac{\mathbb{E}_{X \sim \mathcal{X}} [X]}{\tau},$$

which is the desired result. ■

A.3 2nd Order Markov's Inequality

Theorem A.3.1 (2nd Order MARKOV's Inequality)**.** Let X a non-negative random variable following a probability distribution $\mathcal{X}$ and $\tau > 0$, we have

$$\mathop{\mathbb{P}}_{X \sim \mathcal{X}} [X \geq \tau] \leq \frac{\mathbb{E}_{X \sim \mathcal{X}} [X^2]}{\tau^2}.$$

Proof. We apply MARKOV's inequality (Theorem A.2.1) to have

$$\mathop{\mathbb{P}}_{X \sim \mathcal{X}} \Big[X^2 \geq \tau^2 \Big] \leq \frac{\mathbb{E}_{X \sim \mathcal{X}} [X^2]}{\tau^2}.$$

Moreover, since $\mathrm{I}[X \geq \tau] = \mathrm{I}[X^2 \geq \tau^2]$, we have

$$\mathop{\mathbb{P}}_{X \sim \mathcal{X}} [X \geq \tau] = \mathop{\mathbb{P}}_{X \sim \mathcal{X}} \Big[X^2 \geq \tau^2 \Big],$$

which proves the desired result. ■

A.4 Chebyshev-Cantelli Inequality

Theorem A.4.1 (Chebyshev-Cantelli Inequality)**.** Let X a random variable following a probability distribution $\mathcal{X}$ and $\tau > 0$, we have

$$\mathop{\mathbb{P}}_{X\sim\mathcal{X}}\left[X - \mathop{\mathbb{E}}_{X'\sim\mathcal{X}} X' \geq \tau\right] \leq \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\mathbb{V}_{X'\sim\mathcal{X}}\, X' + \tau^2}.$$

Proof. First of all, remark that we have

$$\begin{aligned}\mathop{\mathbb{P}}_{X\sim\mathcal{X}}\left[X - \mathop{\mathbb{E}}_{X'\sim\mathcal{X}} X' \geq \tau\right] &= \mathop{\mathbb{P}}_{X\sim\mathcal{X}}\left[X - \mathop{\mathbb{E}}_{X'\sim\mathcal{X}} X' + \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\tau} \geq \tau + \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\tau}\right] \\ &\leq \mathop{\mathbb{P}}_{X\sim\mathcal{X}}\left[\left[X - \mathop{\mathbb{E}}_{X'\sim\mathcal{X}} X' + \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\tau}\right]^2 \geq \left[\tau + \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\tau}\right]^2\right], \end{aligned} \tag{A.2}$$

where $\mathbb{V}_{X\sim\mathcal{X}}\, X$ is the variance of the random variable $X \sim \mathcal{X}$. From Equation (A.2) and Markov's Inequality (Theorem A.2.1), we can deduce that

$$\begin{aligned}\mathop{\mathbb{P}}_{X\sim\mathcal{X}}\left[X - \mathop{\mathbb{E}}_{X'\sim\mathcal{X}} X' \geq \tau\right] &\leq \frac{\mathbb{E}_{X\sim\mathcal{X}}\left[X - \mathbb{E}_{X'\sim\mathcal{X}}\, X' + \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\tau}\right]^2}{\left[\tau + \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\tau}\right]^2} \\ &= \frac{\mathbb{V}_{X'\sim\mathcal{X}}\, X'}{\mathbb{V}_{X'\sim\mathcal{X}}\, X' + \tau^2}.\end{aligned}$$

■

A.5 Hölder's Inequality

In order to prove Hölder's inequality, we first prove the following lemma.

Lemma A.5.1 (Young's Inequality)**.** For any $\alpha > 1$ and $\beta > 1$ such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, for any $a \geq 0$ and $b \geq 0$, we have

$$ab \leq \frac{a^\alpha}{\alpha} + \frac{b^\beta}{\beta}.$$

Proof. We first develop $\ln [ab]$ and we apply JENSEN's inequality (Theorem A.1.1) since the logarithm is concave and $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. Indeed, we have

$$\ln [ab] = \ln a + \ln b = \frac{\alpha}{\alpha} \ln a + \frac{\beta}{\beta} \ln b = \frac{\ln a^\alpha}{\alpha} + \frac{\ln b^\beta}{\beta} \leq \ln \left[\frac{a^\alpha}{\alpha} + \frac{b^\beta}{\beta} \right].$$

Then, we take the exponential to both sides of the inequality and we are done. ■

We are now ready to prove HÖLDER's inequality.

Theorem A.5.1 (HÖLDER's Inequality)**.** For any measurable function $f()$ and $g()$, for any $\alpha > 1$ and $\beta > 1$ such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, we have

$$\mathop{\mathbb{E}}_{X \sim \mathcal{X}} |f(X)g(X)| \leq \left[\mathop{\mathbb{E}}_{X \sim \mathcal{X}} |f(X)|^\alpha \right]^{\frac{1}{\alpha}} \left[\mathop{\mathbb{E}}_{X \sim \mathcal{X}} |g(X)|^\beta \right]^{\frac{1}{\beta}}.$$

Proof. For convenience of notation, let $\|f\|_\alpha = [\mathbb{E}_{X \sim \mathcal{X}} |f(X)|^\alpha]^{\frac{1}{\alpha}}$ and $\|g\|_\beta = \left[\mathbb{E}_{X \sim \mathcal{X}} |g(X)|^\beta\right]^{\frac{1}{\beta}}$. If $\|f\|_\alpha = 0$ or $\|g\|_\beta = 0$, then $\mathbb{E}_{X \sim \mathcal{X}} |f(X)g(X)| = 0$, hence, the inequality holds in this case. Then for $\|f\|_\alpha > 0$ and $\|g\|_\beta > 0$, we upper-bound the term $\frac{|f(X)g(X)|}{\|f\|_\alpha \|g\|_\beta}$ with YOUNG's inequality (Lemma A.5.1), *i.e.*, we have

$$\frac{|f(X)g(X)|}{\|f\|_\alpha \|g\|_\beta} \leq \frac{|f(X)|^\alpha}{\alpha \|f\|_\alpha^\alpha} + \frac{|f(X)|^\beta}{\beta \|f\|_\beta^\beta}.$$

By taking the expectation *w.r.t.* $X \sim \mathcal{X}$, we have

$$\begin{aligned}
\frac{\mathbb{E}_{X \sim \mathcal{X}} |f(X)g(X)|}{\|f\|_\alpha \|g\|_\beta} &\leq \frac{\mathbb{E}_{X \sim \mathcal{X}} |f(X)|^\alpha}{\alpha \|f\|_\alpha^\alpha} + \frac{\mathbb{E}_{X \sim \mathcal{X}} |f(X)|^\beta}{\beta \|f\|_\beta^\beta} \\
&= \frac{\|f\|_\alpha^\alpha}{\alpha \|f\|_\alpha^\alpha} + \frac{\|f\|_\beta^\beta}{\beta \|f\|_\beta^\beta} \\
&= \frac{1}{\alpha} + \frac{1}{\beta} \\
&= 1.
\end{aligned}$$

This concludes the proof since

$$\frac{\mathbb{E}_{X\sim\mathcal{X}}\,|f(X)g(X)|}{\|f\|_\alpha\|g\|_\beta} \leq 1 \iff \mathop{\mathbb{E}}_{X\sim\mathcal{X}}|f(X)g(X)| \leq \|f\|_\alpha\|g\|_\beta.$$

■

Abstract. In machine learning, a model is learned from data to solve a task automatically. In the supervised classification setting, the model aims to predict the label associated with an input. The model is learned using a limited number of examples, each consisting of an input and its associated label. However, the model's performance on the examples, computed by the empirical risk, does not necessarily reflect the performance on the task, which is represented by the true risk. Moreover, since it is not computable, the true risk is upper-bounded by a generalization bound that mainly depends on two quantities: the empirical risk and a complexity measure. One way to learn a model is to minimize a bound by a type of algorithm called self-bounding. PAC-Bayesian bounds are well suited to the derivation of this type of algorithm. In this context, the first contribution consists in developing self-bounding algorithms that minimize PAC-Bayesian bounds to learn majority votes. If these bounds are well adapted to majority votes, their use for other models becomes less natural. To overcome this difficulty, a second contribution focuses on the disintegrated PAC-Bayesian bounds that are natural for more general models. In this framework, we provide the first empirical study of these bounds. In a third contribution, we derive bounds that allow us to incorporate complexity measures defined by the user.

Keywords. Machine Learning, Generalization, PAC-Bayesian Bound, Disintegrated PAC-Bayesian Bound, Self-Bounding Algorithm, Majority Vote, Neural Network, Complexity Measure.

Résumé. En apprentissage automatique, un modèle est appris à partir de données pour résoudre une tâche de manière automatique. Dans le cadre de la classification supervisée, le modèle vise à prédire la classe associée à une entrée. Le modèle est appris à l'aide d'un nombre limité d'exemples, chacun étant constitué d'une entrée et de sa classe associée. Cependant, la performance du modèle sur les exemples, calculée par le risque empirique, ne reflète pas nécessairement la performance sur la tâche qui est représentée par le risque réel. De plus, n'étant pas calculable, le risque réel est majoré pour obtenir une borne en généralisation qui dépend principalement de deux quantités : le risque empirique et une mesure de complexité. Une façon d'apprendre un modèle est de minimiser une borne par un type d'algorithme appelé auto-certifié (ou auto-limitatif). Les bornes PAC-Bayésiennes sont bien adaptées à la dérivation de ce type d'algorithmes. Dans ce contexte, la première contribution consiste à développer des algorithmes auto-certifiés qui minimisent des bornes PAC-Bayésiennes pour apprendre des votes de majorité. Si ces bornes sont bien adaptées aux votes de majorité, leur utilisation pour d'autres modèles devient moins naturelle. Pour pallier cette difficulté, une seconde contribution se concentre sur les bornes PAC-Bayésiennes désintégrées qui sont naturelles pour des modèles plus généraux. Dans ce cadre, nous apportons la première étude empirique de ces bornes. Dans une troisième contribution, nous dérivons des bornes permettant d'incorporer des mesures de complexité pouvant être définies par l'utilisateur.

Mot-clés. Apprentissage Automatique, Généralisation, Borne PAC-Bayésienne, Borne PAC-Bayésienne Désintégrée, Algorithme Auto-certifié, Algorithme Auto-limitatif, Vote de Majorité, Réseau de Neurones, Mesure de Complexité.