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

On the Interplays between Generalisation and Optimisation: a PAC-Bayes Approach

Acknowledgements

TODO

Quote 1

Author

Quote 2

Author

Contents

List of Algorithms

List of Theorems

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Appendix

List of Notations

General

a	A scalar (integer or real)
$\mathbb{R}$	The set of real numbers
$\mathbb{R}^d$	The euclidean set of dimension d
$\|\cdot\|$	a norm of an euclidean set
$\mathrm{dist}\,(\cdot,\cdot)$	A distance on a Polish space.
$\mathbb{N}$	The set of natural numbers

Statistical Learning Theory

$\mathcal{Z}$	Data space. In supervised learning, $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ with $\mathcal{X}, \mathcal{Y}$ input and label spaces
$\mathbf{z}$	A datum of $\mathcal{Z}$, in supervised learning $\mathbf{z} = (\mathbf{x}, y)$ with $\mathbf{x}$ input and y label
$\mathcal{S}$	Learning sample $\mathcal{S} = \{\mathbf{z}_i\}_{i\geq 1}$
$\mathcal{D}_\mathcal{S}$	Distribution of $\mathcal{S}$
$\mathcal{S}_m$	Restriction of $\mathcal{S}$ to its m first data $\mathcal{S}_m = \{\mathbf{z}_i\}_{i=1\cdots m}$
$\mathcal{D}$	For *i.i.d.* $\mathcal{S}$, distribution of a single datum on $\mathcal{Z}$
$\mathcal{D}^m$	For *i.i.d.* $\mathcal{S}$, distribution of $\mathcal{S}_m$
$\mathcal{T}$	For *i.i.d.* $\mathcal{S}$, Test set drawn from $\mathcal{D}$
$\mathcal{H}$	The set of hypotheses
h	A hypothesis $h \in \mathcal{H}$
ℓ	Loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$

Probability Theory

$\mathbb{E}_{X\sim\nu}[\cdot]$	The expectation *w.r.t.* the random variable $X \sim \nu$
$\mathbb{P}_{X\sim\nu}[\cdot]$	The probability *w.r.t.* the random variable $X \sim \nu$
$\mathbb{1}[a]$	Indicator function; returns 1 if a is true and 0 otherwise
$(\mathcal{F}_i)_{i\geq 1}$	Filtration adapted to $\mathcal{S}$
$\mathbb{E}_i[\cdot]$	Conditional expectation *w.r.t.* $\mathcal{F}_i$, *i.e.* $\mathbb{E}_i[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_i]$
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution on $\mathbb{R}^d$ with mean μ and covariance matrix Σ

PAC-Bayes framework

$\mathcal{M}(\mathcal{H})$	Set of Probability densities *w.r.t.* the reference measure on $\mathcal{H}$
Q	Posterior distribution $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ on $\mathcal{H}$
P	Prior distribution $\mathrm{P} \in \mathcal{M}(\mathcal{H})$ on $\mathcal{H}$
$\mathrm{KL}(\mathrm{Q}\|\mathrm{P})$	Kullback-Leibler (KL) divergence between Q and P
$D_\alpha(\mathrm{Q}\|\mathrm{P})$	Rényi Divergence between Q and P
$\mathsf{R}_{\mathcal{D}}(h)$	Population Risk of $h \in \mathcal{H}$ *w.r.t.* $\mathcal{D}$, *i.e.* $\mathsf{R}_{\mathcal{D}}(h) := \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h, \mathbf{z})]$
$\hat{\mathsf{R}}_{\mathcal{S}_m}(h)$	Empirical Risk on $\mathcal{S}_m$, *i.e.* $\hat{\mathsf{R}}_{\mathcal{S}_m}(h) \frac{1}{m}\sum_{i=1}^{m} \ell(h, \mathbf{z}_i)$
$\mathsf{R}_{\mathcal{D}}(\mathrm{Q})$	Expected population risk *w.r.t.* Q, *i.e.* $\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) := \mathbb{E}_{h\sim\mathrm{Q}}[\mathsf{R}_{\mathcal{D}}(\mathrm{Q})]$
$\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q})$	Expected empirical risk *w.r.t.* Q, $\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) := \mathbb{E}_{h\sim\mathrm{Q}}\left[\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q})\right]$

Optimal transport

W_1	The 1-Wasserstein distance
W_2	The 2-Wasserstein distance
$\Gamma(\mathrm{Q}, \mathrm{P})$	Set of all coupling distribution on $\mathcal{H}^2$ whose marginals are Q and P.

List of Publications

Conference article

Paul Viallard, Maxime Haddouche, Umut Simsekli, and Benjamin Guedj. Learning via Wasserstein-Based High Probability Generalisation Bounds. *Advances in Neural Information Processing Systems (NeurIPS)*. (2023).

Maxime Haddouche and Benjamin Guedj. Online PAC-Bayes Learning. *Advances in Neural Information Processing Systems (NeurIPS)*. (2022).

Journal article

Maxime Haddouche and Benjamin Guedj. PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales. *Transactions on Machine Learning Research*. (2023).

Research Report

Maxime Haddouche and Benjamin Guedj. Wasserstein PAC-Bayes Learning: A Bridge Between Generalisation and Optimisation. *arXiv*. abs/2304.07048. (2023).

Maxime Haddouche, Benjamin Guedj, and Olivier Wintenberger. Optimistic Dynamic Regret Bounds. (2023). arXiv: 2301.07530 [cs.LG].

Pierre Jobic, Maxime Haddouche, and Benjamin Guedj. Federated Learning with Nonvacuous Generalisation Bounds. (2023). arXiv: 2310.11203 [cs.LG].

Maxime Haddouche, Benjamin Guedj, and John Shawe-Taylor. Upper and Lower Bounds on the Performance of Kernel PCA. (2020). arXiv: 2012.10369 [cs.LG].

PAC-Bayes Learning, a field of many paradigms

1

This manuscript tackles the notion of *generalisation* a notion built upon the general notion of *learning*. For a brief moment, let's take the luxury of forgetting about machines and concentrate on learning at its most human. First and foremost, a human being (or learner) is structured around experiences, either lived or passed on by others. The learner then benefit from these experiences in various ways, for instance, by considering a mediated experience to be true (fire burns) and acting on this assumption, whereas reiteration or denial of this same information may be symptoms of zero truth value. These scenarios can just as easily appear for a lived experience (the question of hallucinations). This first dichotomy in information processing is intrinsically linked to a clearly stated question: does fire burn? Can I trust my senses or have I hallucinated? In these cases, learning has taken place by subjecting the experience to its truth value in relation to a simple question (in this case with two outcomes). This vision can easily be extended to a multiple (and finite) tree of possibilities. Indeed, we can extend the burning question as follows: what is the intensity of the burn as a function of the temperature of the fire? We can then establish a multitude of answers representing various degrees of burn.
However, many questions cannot be reduced to a finite number of possibilities. For example, what is fire? To answer this question, it is nevertheless possible to exploit multiple facets of experience (wood fire, twig, rock) to propose that fire is the chemical reaction of oxygen in the air with a combustible material, with a supply of energy serving as the trigger.
Then, a legitimate question is: why has mankind understood the nature of fire? This fundamental understanding emerged from practical considerations: how can we stop being cold? Can we eat meat other than raw to reduce the risk of illness? It then takes multiple interactions with the environment to generate experiences and then learn from them to gradually respond to a complex need (how to make a fire to keep yourself warm?).
Thus, through this preliminary analysis, we have found several premises of understanding human learning.

- How is learning formalised structurally? The learner must base the experience on simple questions to acquire primary certainties. These latter acquired, it is possible to reach complex questions by interweaving more and more elementary considerations.

- Where does the need to learn come from? From a practical point of view, the emergence of these complex questions often arises from a relationship between the being and its environment, making it possible to develop contextual objectives. The learner then gradually becomes capable of responding to complex needs through a succession of simple actions.

Detail broadly what generalisation is, to what kind of structures it is applied (neural nets or linear classfier eg). Details on the other hand what optimisation is doing (ERM eg) and explain that interestingly in various methods, reaching minimisers of empirical objectives is enough to ensure a good generalisation ability. From this, discuss about the current limitations of generalisaiton: either not going so often beyond light-tailed assumptions or noticing that the interplays between generalisation (statistical arguments) and optimisation (geometric ones) remains uncharted for a vast range of cases.

Vision: after generic paragraphs on generalisation an optimisation, do a broader paragraph on PAC-Bayes and details the problem of exisiting PAC-Bayes approach:

Says that PAC-Bayes spontaneously offer a clear link from generalisation to optimisation by providing new learning algorithms: this implicitly suggests assumptions on the loss (eg convex) or on the regulariser (KL between gaussians to get a strongly convex function) to make sure the minimisation goes well and thus build a bridge with optimisation.

TODO look if there are links from optimiastion to PAC-Bayes (must have been some with dziugaite, neu with SGD).

Here, we are studying the interplays on both directions. First, we take the opposite perspective and, starting from optimisation benefits/perspectives, we want to understand generalisation, to do so we have several routes within PAC-Bayes. Second, we investigate deeper on the influence of generalisation bounds to derive novel learning algorithms

Thinking the role of the prior in PAC-Bayes: in a similar manner than initialisation/ goal to attain in optimisation: if we target a data-free posterior (eg GIbbs catoni) then ok: we target the learning objective. Otherwise, it is common to compare to a random initialisation point of a learning procedure: meaningless. Answer: Online PAC-Bayes which allows, among other, to make the prior evolve through time. (note that there is also either the data-dependent prior: its bad and the differential privacy approach, which is nice)

Switching from statistical to geometric assumptions on the loss. Most of the bounds holds for data-free light-tailed losses: do not necessarily fit the reality of losses involved in optimisation, often unbounded, and either convex, gradient lipschitz or smooth. Answer: PAC-Bayes for heavy-tailed martingales and flat minima.

Can the convergence properties of optimisation procedure play a role in generalisation? Direct answer: Wasserstein PAC-Bayes, Indirect one: flat minima.

Can we derive generalisation-based learning algorithms beyond Gaussian or Gibbs distributions? Answer: Yes as a byproduct: both in Online PAC-Bayes, Flat Minima and Paper with Paul

Precise the structure of the document: see it as a natural flow where one question implies another one:

Light-tailed losses? –¿ supermartingales!
But then : what should we do of the prior? $\rightarrow$ if you see it as an initialisation: Online PAC-Bayes with novel online algorithms or Flat Minima to attenuate the impact of the prior through fast rates.
What if it should be the optimisation goal? –¿ Wasserstein PAC-Bayes
CHALLENGE HERE: being very rigorous on the lit review.

2 PAC-Bayesian Bounds for Martingales and Heavy-Tailed losses

This chapter is based on the following paper

Maxime Haddouche and Benjamin Guedj. PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales. *Transactions on Machine Learning Research*. (2023)

Abstract

TODO: put general bounds for martingales and batch learning as corollary

3 Mitigating Initialisation Impact by Real-Time Control: Online PAC-Bayes Learning

This chapter is based on the following papers

Maxime Haddouche and Benjamin Guedj. Online PAC-Bayes Learning. *Advances in Neural Information Processing Systems (NeurIPS)*. (2022)
Maxime Haddouche and Benjamin Guedj. PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales. *Transactions on Machine Learning Research*. (2023)
Paul Viallard, Maxime Haddouche, Umut Simsekli, and Benjamin Guedj. Learning via Wasserstein-Based High Probability Generalisation Bounds. *Advances in Neural Information Processing Systems (NeurIPS)*. (2023)

Contents

Abstract

Put OPB here. Precise in the intro that the martingale bounds allow to go beyond batch learning but that this has never been made for OL. Put the supermartingale OPB bound in a supplementary section and the Online WPB bound after the main results of OPB to reach heavy-tailed losses.

4 Mitigating Initialisation Impact through Flat Minima: Fast Rates for Small Gradients

This chapter is based on the following paper

TODO

Contents

Abstract

This is the PLS paper, precise that the supermartingales bounds are richer than simply recovering classical batch guarantees: we can incorporate gradient norms, which explains generalisation when a flat minima is reached.

4.1 Introduction

5 Wasserstein PAC-Bayes Learning: Exploiting Optimisation Guarantees to Explain Generalisation

This chapter is based on the following papers

Maxime Haddouche and Benjamin Guedj. Wasserstein PAC-Bayes Learning: A Bridge Between Generalisation and Optimisation. *arXiv*. abs/2304.07048. (2023)
Paul Viallard, Maxime Haddouche, Umut Simsekli, and Benjamin Guedj. Learning via Wasserstein-Based High Probability Generalisation Bounds. *Advances in Neural Information Processing Systems (NeurIPS)*. (2023)

Contents

Abstract

Put WPB here, precise that, when the prior is seen as the learning goal, it is possible for a certain optimisation algorithm to directly incorporate sound geometric optimisation guarantee into a generalisation bound, trading the hope to reach a flat minima with a sound convergence guarantees. However, this comes at the cost of the explicit impact of the dimension. Also put the paper with Paul(batch bounds) as a supplementary content.

5.1 Introduction

Some Mathematical Tools

A

A.1 Jensen's Inequality

Theorem A.1.1 (Jensen's Inequality)**.** Let X a random variable following a probability distribution ν with f a real-valued measurable convex function, we have

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

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

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

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

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

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

A.2 Markov's Inequality

Theorem A.2.1 (Markov's Inequality)**.** Let X a non-negative random variable following a probability distribution ν and $\delta > 0$, we have

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

Proof. First of all, remark that we have the following inequality for any X

$$\delta \mathbb{1}[X \geq \delta] \;\leq\; X \mathbb{1}[X \geq \delta] \;\leq\; X. \tag{A.1}$$

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

$$\mathop{\mathbb{E}}_{X \sim \nu}\Big[\delta \mathbb{1}[X \geq \delta]\Big] \leq \mathop{\mathbb{E}}_{X \sim \nu}\Big[X\Big].$$

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

$$\delta \mathop{\mathbb{P}}_{X \sim \nu}[X \geq \delta] \leq \mathop{\mathbb{E}}_{X \sim \nu}\Big[X\Big] \quad \iff \quad \mathop{\mathbb{P}}_{X \sim \nu}[X \geq \delta] \leq \frac{\mathbb{E}_{X \sim \nu}[X]}{\delta},$$

which is the desired result. ■

A.3 Ville's Inequality

Lemma A.3.1 (Ville's maximal inequality for supermartingales)**.** Let $(\mathcal{F}_t)$ be a filtration and (Z_t) a non-negative super-martingale satisfying $Z_0 = 1$ a.s. If Z_t is adapted to $\mathcal{F}_t$ and $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \leq Z_{t-1}$ a.s., $t \geq 1$, then, for any $0 < \delta < 1$, it holds

$$\mathbb{P}\left(\exists T \geq 1 : Z_T > \delta^{-1}\right) \leq \delta.$$

Proof. We apply the optional stopping theorem (Durrett, 2019, Thm 4.8.4) with Markov's inequality defining the stopping time $i = \inf\{t > 1 :\; Z_t > \delta^{-1}\}$ so that

$$\mathbb{P}\left(\exists t \geq 1 : Z_t > \delta^{-1}\right) = \mathbb{P}\left(Z_i > \delta^{-1}\right) \leq \mathbb{E}[Z_i]\,\delta \leq \mathbb{E}[Z_0]\,\delta \leq \delta.$$

■

Additional Background

B

B.1 Online Learning

B.2 Wasserstein distances and optimisation on distributions spaces

B.3 Differential privacy

B.4 Log-Sobolev and Poincaré Inequalities

References

Rick Durrett. Probability: theory and examples. Vol. 49. *Cambridge university press*. (2019)
——— **Cited on page 26**.

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.