The Inria-London Programme

UCL

Inria

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

Chapter 1

Chapter 2

Chapter 3

Chapter 5

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

$\mathbb{E}_{X\sim\nu}\left[\cdot\right]$	The expectation *w.r.t.* the random variable $X \sim \nu$
$\mathbb{P}_{X\sim\nu}\left[\cdot\right]$	The probability *w.r.t.* the random variable $X \sim \nu$
$\mathbb{1}\left[a\right]$	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.* $\mathbb{E}_i[\cdot] := \mathbb{E}\left[\cdot \mid \mathcal{F}_i\right]$
$\mathcal{N}(\mu,\Sigma)$	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 $\mathcal{H}$
$\mathrm{KL}(\mathrm{Q}\Vert\mathrm{P})$	Kullback-Leibler (KL) divergence between Q and P
$D_\alpha(\mathrm{Q}\Vert\mathrm{P})$	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.* $\mathsf{R}_{\mathcal{D}}(h) := \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\ell(h,\mathbf{z})\right]$
$\hat{\mathsf{R}}_{\mathcal{S}_m}(h)$	Empirical Risk on $\mathcal{S}_m$, *i.e.* $\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\dfrac{1}{m}\sum\limits_{i=1}^{m}\ell(h,\mathbf{z}_i)$
$\Delta_{\mathcal{S}_m}(h)$	Generalisation gap $\Delta_{\mathcal{S}_m}(h) := \mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)$
$\mathsf{R}_{\mathcal{D}}(\mathrm{Q})$	Expected population risk *w.r.t.* Q, *i.e.* $\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) := \mathbb{E}_{h\sim\mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(\mathrm{Q})\right]$
$\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q})$	Expected empirical risk *w.r.t.* Q, $\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) := \mathbb{E}_{h\sim\mathrm{Q}}\left[\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q})\right]$
$\Delta_{\mathcal{S}_m}(\mathrm{Q})$	Expected generalisation gap *w.r.t.* Q, $\Delta_{\mathcal{S}_m}(\mathrm{Q}) := \mathbb{E}_{h\sim\mathrm{Q}}\left[\Delta_{\mathcal{S}_m}(h)\right]$

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].

Préambule: Apprentissage humain, Apprentissage Machine et Généralisation

Ce manuscrit étudie la question de la *généralisation* des algorithmes d'apprentissage machine, notion construite autour de celle d'*apprentissage* en apprentissage machine Pour un bref instant, prenons le luxe d'oublier les machines pour se concentrer sur l'apprentissage en ce qu'il a de plus humain.

Appréhender l'apprentissage humain. Un humain, en premier lieu, va se structurer autour d'expériences, vécues ou transmises par autrui. L'être apprenant va alors tirer bénéfice de ces vécus via diverses modalités, par exemple, en considérant une expérience médiée comme vraie (le feu brûle) et agir en fonction de ce postulat alors que la réitération ou la négation de cette même information peuvent être des symptômes d'une valeur de vérité nulle. Ces scénarios peuvent tout aussi bien apparaître pour une expérience vécue (hallucinations). Cette première dichotomie quant au traitement de l'information est intrinsèquement liée à une question clairement énoncée : est-ce que le feu brûle? Puis-je me fier à mes sens ou ai-je halluciné ? Dans ces cas de figure, l'apprentissage a eu lieu à travers l'assujettissement de l'expérience à sa valeur de vérité par rapport à une question simple (ici à deux issues). Cette vision peut facilement s'étendre à une arborescence multiple (et finie) de possibles. En effet, on peut étendre la question de la brûlure comme suit: quelle est l'intensité de la brûlure en fonction de la température du feu? On peut dès lors établir une multitude de réponses représentant divers degrés de brûlure.
Néanmoins, de nombreuses questions ne peuvent se réduire à un nombre fini de possibilités. Par exemple, qu'est-ce que le feu? Pour répondre à cette question, il est néanmoins possible d'exploiter de multiples facettes d'expériences (feu de bois, brindille, roche) pour proposer le feu comme étant la réaction chimique de l'oxygène de l'air avec un matériau combustible, un apport d'énergie servant de déclencheur.
Il est alors légitime de se demander pourquoi l'apprenant a eu besoin de comprendre la vraie nature du feu. Cette compréhension fondamentale des choses émerge de considérations pratiques : comment ne plus avoir froid? Peut-on manger de la viande autrement que crue pour diminuer les risques de maladie? Il faut alors de multiples interactions avec l'environnement pour générer des expériences et ensuite apprendre d'elles pour répondre graduellement à un besoin complexe (comment faire un feu pour se réchauffer?).

Ainsi, par cette analyse préliminaire, nous avons trouvé plusieurs prémices de compréhension de l'apprentissage chez l'homme.

- Comment l'apprentissage se formalise-t-il structurellement ? L'apprenant doit abâtardir l'expérience à des questions simples pour acquérir des certitudes primaires. Ces dernières acquises, il est possible d'atteindre des questions complexes en imbriquant de plus en plus de considérations élémentaires.

- D'où provient le besoin d'apprendre ? D'un point de vue pratique, l'émergence de ces questions complexes dérive bien souvent d'un rapport de l'être à son environnement, permettant d'élaborer des objectifs contextuels. L'apprenant devient alors graduellement capable de répondre à des besoins complexes par une succession d'actions simples.

De l'apprentissage humain à l'apprentissage machine. L'apprentissage machine s'est structuré autour de deux approches, une première symbolique qui tire profit des extrapolations humaines pour apprendre à la machine à manipuler une axiomatique et une seconde, statistique, qui consiste à fournir bon nombre d'expériences à la machine pour lui faire apprendre par de multiples exemples empiriques. Nous allons nous focaliser sur la seconde approche car, elle sous-tend une large partie de la recherche moderne. Cette méthode requiert de nombreuses expériences transmises à la machine qui en extrait les connaissances à travers des procédures optimisatoire. Plus précisément, la connaissance extraite dépend de la question posée ainsi que sa traduction mathématique. Nous pouvons alors relever des parallèles avec l'apprentissage humain décrit plus haut: il faut des expériences et une question pour réduire le réel à quelque chose d'apprenable. Pour aller plus loin, la variétés des scenarii d'apprentissages humain décrits au dessus ont une correspondance dans l'apprentssage machine moderne: à la question "Le feu brûle-t-il?" on peut associer l'apprentissage supervisé qui traite apprend sur des questions à choix multiples. A la question "qu'est-ce que le feu?", on peut associer l'apprentissage non-supervisé qui va chercher, dans le cas du clustering (ou regroupement), des similitudes non-induites par la question entre diverses expériences. Finalement, quant à l'interaction avec l'environnement et la question "puis-je faire un feu?", elle est associée à l'apprentissage par renforcement qui étudie l'apprentissage d'un agent qui interagit avec son environnement.

Comprendre la généralisation depuis l'apprentissage. La généralisation peut être vue comme la capacité d'exploiter l'apprentissage d'une expérience au delà de cette dernière. Cela englobe une compréhension théorique et axiomatique d'un phénomène, i.e. une extrapolation fructueuse ou bien la capacité à exploiter la connaissance acquise pour une situation inédite, jamais vue auparavant, *i.e.* interpoler des expériences.

Ce double aspect de la généralisation se retrouve aussi bien chez l'homme que la machine sous diverses modalités. Les réseaux de neurones profonds, qui sont le fer de lance de l'apprentissage machine moderne, se basent sur des espaces de dimension finie pour apprendre, ce qui revient à dire qu'un problème peut être appris à travers un nombre fini de principes fondateurs. Le nombre de principes pouvant être augmentés autant que les capacités numériques le permettent, nous dirons alors que les réseaux de neurones ont une puissance discrète de généralisation. Etant donné que les méthodes d'apprentissage machine sont corrêlées à leur pendantes humaines, on peut alors se demander si la puissance de généralisation (et même d'apprentissage) humaine est également discrète. Il semble cavalier d'affirmer ceci car même s'il est possible de supposer que la part consciente de l'esprit humain raisonne à horizon finie et a une puissance dénombrable (transmise d'ailleurs à la machine, apprenant selon des modalités humaines), cette dimension occulte la quantité d'information sans cesse captée et filtrée par notre cerveau ainsi que son assimilation inconsciente, relevant autant de la pensée abstraite que du biologique peut potentiellement générer une puissance de généralisation relevant d'un infini plus large et ainsi fournir une puissance de généralisation continue (relevant davantage de la ligne que du point). Dès lors, comment penser la généralisation chez l'homme alors que, mathématiquement, nos intuitions les plus simples nous font défaut lorsque cette puissance continue intervient (la boule de rayon 1 n'est pas compacte en dimension infinie, Riesz, 1955)? On peut également se demander si l'extrapolation existe dans de telles structures ou si tout revient à interpoler (Hasson *et al.*, 2020).

Quid de la généralisation en apprentissage machine de nos jours? Qu'espérer alors des réseaux de neurones artificiels et de leur capacité de généralisation relativement à l'humain? Les théorèmes d'approximations universels (voir *e.g.* Lu *et al.*, 2017; Park *et al.*, 2021) assurent que les réseaux de neurones sont capables d'approximer n'importe quelle fonction vivant dans un espace à la puissance du continu (*e.g.* l'espace des fonctions continues à support compact qui n'admet pas de base dénombrable en tant qu'espace de Banach), faisant de ces structures des candidats prometteurs pour appréhender partiellement les mécanismes humains de généralisation. Les approximations prodiguées par ces machines seront, dans un avenir proche, potentiellement suffisamment puissantes pour donner l'illusion d'une capacité de généralisation humaine. Néanmoins, il demeure bon de garder en tête que, si la thèse d'une inégalité fondamentale de nature entre les puissances de généralisation humaine et machine est avérée, alors les réseaux de neurones artificiels n'atteindront jamais pleinement les capacités de compréhension du monde de leurs homologues bilogiques. Reste que leur capacité à approximer cette intelligence toute humaine font de ces structures des assistants de valeur, enrichissant les capacités de tout individu. Mieux comprendre la puissance de généralisation machine, être capable de la quantifier, d'identifier les mécanismes qui la favorisent sont les objets de ce manuscrit.

Preamble: Human Learning, Machine Learning and Generalisation

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.

Apprehending human learning A human being (here a 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.

From human to machine learning Machine learning has been structured around two approaches, the first symbolic, which takes advantage of human extrapolations to teach the machine to manipulate an axiomatic, and the second, statistical, which consists of providing the machine with a large number of experiments so that it learns from multiple empirical examples. We are going to focus on the second approach because it underpins a large part of modern research. This method requires a large number of experiments to be transmitted to the machine, which then extracts the knowledge through optimising procedures. More precisely, the knowledge extracted depends on the question posed and its mathematical translation. We can see parallels with human learning described above: you need experiments and a question to reduce reality to something learnable. To go a step further, the variety of human learning scenarios described above can be applied to modern machine learning: the question "Does fire burn?" can be associated with supervised learning, which learns from multiple-choice questions. The question "What is fire?" can be associated with unsupervised learning, which, in the case of clustering, looks for similarities between numerous experiments that are not induced by the question. Finally, the question "Can I make a fire?" can be linked to reinforcement learning which focuses on the progress of an agent learning from its interaction with the environement.

From learning to generalisation. Generalisation can be seen as the ability to exploit learning from experience beyond that experience. This encompasses a theoretical and axiomatic understanding of a phenomenon, i.e. a fruitful extrapolation, or the ability to exploit the knowledge acquired for a new, never-before-seen situation, i.e. to interpolate experiences.
This dual aspect of generalisation can be found in both humans and machines in a variety of ways. Deep neural networks, which are the spearhead of modern machine learning, are based on finite-dimensional learning spaces, which means that a problem can be learned through a finite number of founding principles. Since the number of principles can be increased as far as numerical capacity allows, we can say that neural networks have discrete generalising power. Given that machine learning methods are

correlated with their human counterparts, we might then ask whether the power of human generalisation (and even learning) is also discrete. It seems cavalier to assert this, because even if it is possible to suppose that the conscious part of the human mind reasons on a finite horizon and has a countable power (transmitted, moreover, to the machine, which learns according to human methods), this dimension obscures the quantity of information constantly captured and filtered by our brain, as well as its unconscious assimilation, In other words, the fact that our brain is as much a part of abstract thought as it is of biological thought can potentially generate a generalisation power that relates to a wider infinity and thus provide a continuous generalisation power (relating more to the line than to the point). So how can we think about generalisation in humans when, mathematically, our simplest intuitions fail us when this continuous power is involved (the ball of radius 1 is not compact in infinite dimension, RIESZ, 1955)? We might also ask whether extrapolation exists in such structures or whether it all boils down to interpolation (HASSON *et al.*, 2020).

What to expect from generalisation in modern machine learning? So what can we expect from artificial neural networks and their ability to generalise to humans? Universal approximation theorems (see *e.g.* LU *et al.*, 2017; PARK *et al.*, 2021) ensure that neural networks are capable of approximating any function living in a space to the power of the continuum (*e.g.* the space of continuous functions with compact support which does not admit a countable base as a Banach space), making these structures promising candidates for partially understanding human generalisation mechanisms. In the near future, machine approximations will potentially be powerful enough to give the illusion of human generalisation capacity. Nevertheless, it is worth bearing in mind that, if the thesis of a fundamental inequality in nature between the powers of human and machine generalisation is confirmed, then artificial neural networks will never fully attain the world-understanding capacities of their two-generation counterparts. It is stll worth noticing artificial neural nets ability to approximate this human intelligence makes these structures valuable assistants, enriching the capabilities of any individual. That being said, this manuscript aims to provide a better understanding of generalisation in machine learning, to be able to quantify it and to identify the mechanisms that promote it.
JURY:
Rapporteurs: Alquier (sur)/Chopin (moins)
Membres: Seldin (rapporteur mais peut etre pas fou pour le rapport)/ Pascal Germain (rapporteur) Gérard (si Pierre pas dispo), John Shawe-Taylor (examinateur), Emilie Morvant (présidente) + le Maitre + Arnak Dalalyan (trop proche?), Alessandro Rudi
CHALLENGE HERE: being very rigorous on the lit review.

PAC-Bayes Learning, a field of many paradigms 1

1.1 A brief introduction to statistical learning

Statistical learning (Vapnik, 1999; James *et al.*, 2013) quantifies and identifies how learning algorithms, trained on a specific task using a finite training dataset, generalise to novel, unseen datum. More precisely, a learning agent has to learn how to answer a question, formalised as a *learning problem* being a tuple $(\mathcal{H}, \mathcal{Z}, \ell)$ composed of a *predictor space* on which evolves the agent during the learning process, a *data space* $\mathcal{Z}$ and a *loss function* being the mathematical formulation of the question. Such a minimalistic structure is convenient to encompass a broad range of real-life learning scenarii. To learn, the agent has access to a *training dataset* $\mathcal{S}_m = (\mathbf{z}_i)_{i=1\cdots m}$. The most classical way to learn from $\mathcal{S}_m$ is the empirical risk minimisation (ERM), minimising the *empirical risk* $\hat{\mathsf{R}}_{\mathcal{S}_m} := \frac{1}{m}\sum_{i=1}^m \ell(h, \mathbf{z}_i)$. In this setting, when $\mathcal{S}_m$ is *i.i.d.* (following the distribution $\mathcal{D}$), two facets of generalisation are commonly studied in statistical learning for a given agent $h \in \mathcal{H}$.

- First, the *population risk* $\mathsf{R}_{\mathcal{D}}(h) := \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h, \mathbf{z})]$ focus on the average performance of our learning agent *w.r.t.* any new situation $\mathbf{z} \in \mathcal{D}$, independent of $\mathcal{S}_m$, possibly faced by the agent. A small population risk ensure then efficient generalisation.
- Second, the *generalisation gap* $\Delta_{\mathcal{S}_m}(h) := \mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}\mathcal{S}_m(h)$ evaluate the coherence between the empirical risk and the population one. Having a small generalisation gap ensure that the generalisation ability of the agent has the same magnitude than its training performance.

Note that the population risk is a stronger notion of generalisation than the generalisation gap. However, a small generalisation gap (in absolute value) as well as a small empirical risk is enough to ensure a good population risk. Given that modern optimisation algorithm are often enough to ensure a small empirical risk, the generalisation gap has received a particular attention in statistical learning.

Generalisation bounds. Generalisation bounds are inequalities often controlling the generalisation gap by various quantities depending either on $\mathcal{H}, \mathcal{Z}$ or $\mathcal{S}_m$. We propose below general patterns usually involved in generalisation bounds for an agent $h_{\mathcal{S}_m} \in \mathcal{H}$ depending on $\mathcal{S}_m$ (for instance the output of the ERM).

Expected generalisation bound. For any training set $\mathcal{S}_m$:

$$\mathbb{E}_{\mathcal{S}_m}\left[\Delta_{\mathcal{S}_m}(h_{\mathcal{S}_m})\right] \leq f\left(\textsc{Complexity}, \frac{1}{m}\right).$$

High-probability generalisation bounds. For any training set $\mathcal{S}_m$, with probability $1-\delta$ pver the draw of $\mathcal{S}_m$:

$$\Delta_{\mathcal{S}_m}(h_{\mathcal{S}_m}) \leq f\left(\textsc{Complexity}, \frac{1}{m}, \log\frac{1}{\delta}\right).$$

The nature of f and the COMPLEXITY term depend on the facet of the complexity of the learning problem we aim to focus. Celebrated examples are for instance the dimension of $\mathcal{H}$, if euclidean, the VC dimension of $\mathcal{H}$ (VAPNIK, 2000), the Rademacher complexity (BARTLETT and MENDELSON, 2001, 2002), the stability parameter of a learning algorithm (BOUSQUET and ELISSEEFF, 2000) or the subgaussian diameter of $\mathcal{Z}$ (KONTOROVICH, 2014). Another approach relies on the Bayesian learning paradigm, deriving *posterior* knowledge from data and prior modelling of the environment. Then, the COMPLEXITY can be borrowed from information theory (COVER and THOMAS, 2001), *e.g.* mutual information (NEAL, 2012), or from optimal transport, *e.g.* Wasserstein distances (WANG *et al.*, 2019; RODRIGUEZ-GALVEZ *et al.*, 2021).

Those two approaches have various benefits. A notable strength of expected bounds is that they may reach fast convergence rates (*i.e.* faster than $\frac{1}{\sqrt{m}}$) contrary to high-probability one, even when $\mathcal{H}$ is a singleton thanks to the central limit theorem (GRUNWALD *et al.*, 2021). However, expected bounds often involves a theoretical COMPLEXITY which cannot be estimated in practice and may be hard to interpret while high probability bounds may be fully empirical and can be considered with small confidence parameter δ as it is attenuated by a logarithm.

How to choose the complexity term ? An introductory example. There is no evidence proving that a certain notion of complexity is preferrable to another. The choice of COMPLEXITY may however be driven by practical considerations, emerging from the learning problem of interest. To illustrate this point, let us focus on the following example, providing two learning problems which differs only from the predictor space $\mathcal{H}$ and which have very different interactions with the VC dimension.

First, consider a supervised learning problem with loss ℓ where $\mathcal{Z} = \mathbb{R}^k \times \mathcal{Y}$ with k smaller than m and assume that $\mathcal{H}$ is the set of linear classifiers; *i.e.* $\mathcal{H}_1 := \{h_\theta(x) = sgn(\langle\theta, x\rangle)\}$, where $sgn(a)$ denotes the sign of a. In this case, using the VC dimension may lead to non-vacuous generalisation bounds (VAPNIK, 2000).

However, in modern machine learning, deep neural networks are often considered, let us first define a celebrated class of deep neural networks.

Definition 1.1.1. A multilayer perceptron with depth K and architecture $\{N_1, \cdots, N_K\}$, denoted as $h_{\mathbf{w}}(\mathbf{x}) := Wh^K(\cdots h^1(\mathbf{x})) + b$, is composed of K layers $h^1(\cdot), \ldots, h^K(\cdot)$. $W \in \mathbb{R}^{|\mathcal{Y}| \times N_K}$ and $b \in \mathbb{R}^{N_K}$ are the weight matrix and the bias of the last layer, and the i-th layer h^i, composed of N_i nodes, is defined by $h^i(\mathbf{x}) := \sigma_i(W_i\mathbf{x} + b_i)$, where $W_i \in \mathbb{R}^{N_i \times N_{i-1}}$ and the bias $b_i \in \mathbb{R}^{N_i}$ are its weight matrix and bias respectively; $\sigma_i : \mathbb{R}^{N_i} \to \mathbb{R}^{N_i}$ is an activation function. The weights $\mathbf{w} = \mathrm{vec}(\{W, W_K, \ldots, W_1, b, b_K, \ldots, b_1\})$ represent the vectorisation of all parameters of the network.

Now, consider the learning problem with the same $\mathcal{Z}, \ell$ as above, but with $\mathcal{H}_2$ being the set of multilayer perceptrons *w.r.t.* a fixed depth K and architecture $\{N_1, \cdots, N_K\}$. To be consistent with modern practice, assume also that we are in the *overparametrised setting*, meaning that the space $\mathcal{H}_2$ has a dimension d far greater than m. In this case, VC dimension fails to explain the good generalisation ability (seen in practice) of multilayer perceptrons (Bartlett and Maass, 2003). Understanding the generalisation ability of deep neural networks remains nowadays a major challenge and in what follows, we focus on a modern branch of learning theory which provided non-vacuous bounds of the generalisation ability of deep neural networks: PAC-Bayes learning.

1.2 PAC-Bayes learning

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 generalisation: 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? → 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

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

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

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

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

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

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

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

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

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

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

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

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

References

Peter Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Conference on Computational Learning Theory (COLT)*. (2001)
—— **Cited on page 22.**

Peter Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*. (2002)
—— **Cited on page 22.**

Peter L Bartlett and Wolfgang Maass. Vapnik-Chervonenkis dimension of neural nets. *The handbook of brain theory and neural networks*. (2003)
—— **Cited on page 23.**

Olivier Bousquet and Andre Elisseeff. Algorithmic Stability and Generalization Performance. *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*. (2000). URL: https://proceedings.neurips.cc/paper/2000/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html
—— **Cited on page 22.**

Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. *Wiley*. (2001)
—— **Cited on page 22.**

Rick Durrett. Probability: theory and examples. Vol. 49. *Cambridge university press*. (2019)
—— **Cited on page 36.**

Peter Grunwald, Thomas Steinke, and Lydia Zakynthinou. PAC-Bayes, MAC-Bayes and Conditional Mutual Information: Fast rate bounds that handle general VC classes. *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. *Proceedings of Machine Learning Research. PMLR*. (15–19 Aug 2021). URL: https://proceedings.mlr.press/v134/grunwald21a.html
—— **Cited on page 22.**

Uri Hasson, Samuel A Nastase, and Ariel Goldstein. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*. 105.3. (2020)
——— **Cited on pages 15, 19.**

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *et al.* An introduction to statistical learning. Vol. 112. *Springer*. (2013)
——— **Cited on page 21.**

Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. Vol. 32. *JMLR Workshop and Conference Proceedings*. *JMLR.org*. (2014). URL: http://proceedings.mlr.press/v32/kontorovicha14.html
——— **Cited on page 22.**

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The Expressive Power of Neural Networks: A View from the Width. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett. (2017). URL: https://proceedings.neurips.cc/paper/2017/hash/32cbf687880eb1674a07bf717761dd3a-Abstract.html
——— **Cited on pages 15, 19.**

Radford M. Neal. Bayesian learning for neural networks. *Springer Science & Business Media*. (2012)
——— **Cited on page 22.**

Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum Width for Universal Approximation. *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. (2021). URL: https://openreview.net/forum?id=O-XJwyoIF-k
——— **Cited on pages 15, 19.**

Frigyes Riesz. Leçons d'Analyse Fonctionelle. (1955)
——— **Cited on pages 15, 19.**

Borja Rodriguez-Galvez, German Bassi, Ragnar Thobaben, and Mikael Skoglund. Tighter Expected Generalization Error Bounds via Wasserstein Distance. *Advances in Neural Information Processing Systems 34: Annual Conference on Neural*

Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. (2021)
——— **Cited on page 22.**

Vladimir Vapnik. An overview of statistical learning theory. *IEEE Trans. Neural Networks*. (1999). url: https://doi.org/10.1109/72.788640
——— **Cited on page 21.**

Vladimir Naumovich Vapnik. The Nature of Statistical Learning Theory, Second Edition. *Statistics for Engineering and Information Science*. *Springer*. (2000)
——— **Cited on page 22.**

Hao Wang, Mario Diaz, Jose Czndido Silveira Santos Filho, and Flavio P. Calmon. An Information-Theoretic View of Generalization via Wasserstein Distance. *IEEE*. (2019). url: https://doi.org/10.1109/ISIT.2019.8849359
——— **Cited on page 22.**

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.