ON GENERALISATION AND LEARNING

CONTRIBUTIONS TO STATISTICAL LEARNING THEORY

by Benjamin Guedj

À la tendre mémoire de mes grands-parents,
Auguste et Odette Dubois,
Michel et Georgette Guedj

FOREWORD

This manuscript highlights my contributions to the fields of machine learning and statistical learning theory, in a compact and contextualised presentation rather than a detailed treatment (for which I refer the interested reader to the corresponding research papers). This manuscript is presented as part of the requirements for the French "Habilitation à diriger des recherches" (HDR), which I publicly defended at the University of Lille (France), on the 10th of June, 2024. The jury was composed of

Prof Florence d'Alché-Buc (president)
Institut Polytechnique de Paris, France

Dr Francis Bach (reviewer)
Inria & Académie des Sciences, France

Prof Gérard Biau (examiner)
Sorbonne Université, France

Prof Christophe Biernacki (guarantor)
Université de Lille & Inria, France

Prof Arnak Dalalyan (reviewer)
ENSAE, France

Prof Gábor Lugosi (reviewer)
Universitat Pompeu Fabra, Spain

Prof Eric Moulines (examiner)
Ecole Polytechnique & Académie des Sciences, France

Université de Lille

ABSTRACT

Machine learning, as the prime workhorse of artificial intelligence, aims to build algorithms with outstanding predictive performance for a variety of tasks. A key challenge is the quest for *generalisation for intelligent systems*, *i.e.*, how that predictive performance can be assessed (both theoretically and numerically) and in some cases further promoted, in particular when algorithms are deployed on new datasets and tasks. Interestingly, studies on generalisation have led to the design of novel machine learning algorithms, which inherit solid generalisation abilities. In my academic career since 2010, I have mostly focused on one of these strategies to study and promote generalisation, the PAC-Bayes theory, which investigates the generalisation performance of randomised predictors and is now regarded as an established and principled approach to generalisation-by-design. In the past decade, it has garnered increased interest due to its flexibility and promising results, including in deep learning. Building on concepts and tools from statistical learning theory and the PAC-Bayes theory, I describe in this manuscript three sets of contributions in machine learning: (i) establishing novel generalisation guarantees for deep neural networks (ii) designing generalisation-driven learning algorithms (iii) unveiling generalisation guarantees beyond the classical learning frameworks. With this line of work, I am aiming towards a better understanding of generalisation in machine learning, with the ultimate goal to contribute to reduce the massive gap between how humans generalise with a fraction of the data and compute needed by machines, paving the way to more frugal artificial intelligence systems.

Keywords. Generalisation bounds, PAC-Bayes theory, statistical learning theory, machine learning.

ACKNOWLEDGMENTS

L'honnêteté intellectuelle commande d'entamer ces remerciements par un aveu : j'ai considérablement tardé à écrire cette habilitation à diriger des recherches, au point que ce sigle HDR, véritable arlésienne, a été ces dernières années une source inépuisable de gentilles moqueries, de questionnements interloqués, et d'incitations amicales mais fermes à l'écrire, bon sang, tu devrais être habilité depuis des années, voyons. Il est enfin temps, chère lectrice, cher lecteur, de clôre ce chapitre riche en rebondissements et de remercier ici celles et ceux qui y ont joué un rôle.

A tout seigneur, tout honneur : Gérard et Éric, depuis nos premières rencontres (respectivement en 2008 et en 2011), vous avez été des figures centrales de ma vie scientifique. Vous avez dirigé ma thèse avec chaleur et talent, combinant de façon experte supervision au plus près du vent et encouragements à prendre le large. Plus de dix ans après avoir soutenu ma thèse, je suis fier et heureux de franchir cette étape qu'est la HDR à vos côtés, et de savoir compter sur votre amitié. Je n'imaginais pas, en démarrant ma thèse en 2011, combien je m'inspirerais de vous deux comme directeur de thèse à mon tour – style qui passera à la posterité comme la supervision "audois et à l'œil" ![1]

Christophe, tu m'as fait confiance il y a dix ans, en me faisant rentrer dans ce grand institut qu'est Inria, et ton accompagnement et ta bienveillance, comme responsable d'équipe puis délégué scientifique et enfin ADS, m'ont permis de sereinement grandir scientifiquement. Alors qu'au moment d'écrire ces lignes, je me remets tout juste des célébrations de ma promotion comme directeur de recherche, ton rôle comme garant de mon HDR sonne comme la plus belle des évidences.

À Florence, Arnak, Francis et Gábor, je veux dire mon immense reconnaissance d'avoir accepté d'évaluer mes travaux : juché sur les épaules de tels géants scientifiques, on ne peut qu'aller loin. À John Shawe-Taylor, je veux témoigner ma gratitude et mon amitié : ton invitation un peu folle à "venir un peu à Londres" début 2018, à un moment où je découvrais mon nouveau rôle de père et devenais plus expert en changements de couches que je ne le croyais possible, a eu un impact décisif sur ma carrière et ma vie familiale. Merci de ta confiance.

[1] J'attends depuis beaucoup trop longtemps de faire ce jeu de mots.

Je suis reconnaissant aux très nombreux collègues (dont beaucoup sont devenus des amis), d'Inria (et de MODAL !), de l'Université de Lille, de University College London, du Turing Institute, qui ont croisé ma route – et plus largement à toutes celles et ceux rencontrés, au hasard d'une conférence ou d'une soutenance. Je ne me risquerai pas à établir de liste de peur d'en oublier – mais je suis conscient que chacune et chacun d'entre vous—vous vous reconnaîtrez—ont contribué, un peu ou beaucoup, à faire de moi le scientifique que je suis aujourd'hui. Je suis tout particulièrement reconnaissant à mes co-auteurs (plus de 120 !), avec une mention spéciale pour celles et ceux qui m'ont fait confiance et ont été mes étudiants (Bhargav, Le, Arthur, Antoine V., Felix, Florent, Valentina, Antonin, Reuben, Théophile, Maxime, Antoine P., Valentin, Fredrik : merci !). Gratitude chaque semaine renouvelée pour l'indispensable Anne. Une pensée émue pour François Laviolette, disparu beaucoup trop tôt.

Aux très nombreux amis, de tous horizons, qui font partie de ma vie, merci – vous vous reconnaîtrez. Mention spéciale à Florent et Morgane qui ont accueilli l'une de mes trop nombreuses "retraites HDR", et chez qui une version préliminaire de ce manuscrit a vu le jour (avant de retourner prendre un peu la poussière). Mention spéciale à la joyeuse et exigeante camaraderie des Young Leaders du Franco-British Council, qui tirent le meilleur de chacune et chacun de ses membres. Si mes centres de gravité professionels et personnels débordent un peu de l'Hexagone depuis quelques années, j'ai vite été intégré dans la brillante et toujours dynamique communauté française de Londres. À Minh-Hà Pham, mentor si précieuse, une pensée affectueuse et reconnaissante. Si je me risquais à un bilan comptable, je serais bien obligé d'admettre qu'une part non négligeable de mon activité se produit régulièrement dans des wagons d'Eurostar filant à 300 kilomètres-heure sous la Manche. Que d'idées, de mails et d'équations sont nés dans cet entre-deux authentiquement franco-britannique !

À ma famille : à mes parents, frères et soeur, oncles, tantes, cousines et cousins, et à la confiance qu'ils m'apportent. Au souvenir de mes grants-parents, qui m'accompagne chaque jour.

À Emeline, qui est entrée dans ma vie quelques jours après la soutenance de ma thèse en décembre 2013, et qui est devenue en un temps record ma boussole, et mon soutien le plus précieux. Cette décennie passée à tes côtés a fait de moi celui que je suis aujourd'hui : j'ai hâte de voir ce que les prochaines nous réservent. À Diane et George, nos merveilleux enfants, qui s'appliquent avec une constance qui force l'admiration à me garder les pieds sur terre quand je risque de me prendre un peu trop au sérieux. Mon rôle de papa est sans hésitation le plus exigeant de tous, mais c'est aussi le plus beau. Vous êtes une source inépuisable de joies, de nuits blanches et de néologismes *so british*.

Il est merveilleux et un peu vertigineux de fermer ainsi des chapitres importants de ma vie professionnelle – je ne doute pas un instant que ceux que je m'apprête à ouvrir soient au moins aussi passionnants. Amie lectrice, ami lecteur : je crois bien que l'aventure va continuer.[2]

[2] Voir Guedj (2013), remerciements, pages 5–6.

CONTENTS

1 INTRODUCTION

This is the manuscript of my "habilitation à diriger des recherches" (HDR), described in the official texts[1] as *validating the high scientific level of the candidate, the originality of his scientific programme, his ability to master a research strategy in a sufficiently vast scientific or technological domain, and his ability to supervise young researchers*. The typical form of a HDR manuscript is a high-level summary of one's work since their PhD, emphasising the research programme and overall directions, and I will follow that template.

Outline of this manuscript

Chapter 1 introduces my scientific background and the overview of my contributions spanning my academic career since 2010 across Denmark, France and the United Kingdom. Chapter 2 presents a selection of my works on the study of generalisation for deep neural networks. Chapter 3 illustrates the principled strategy of deriving new learning algorithms by optimising generalisation bounds. In Chapter 4, I present a selection of contributions beyond the classical learning paradigms, and I highlight future research perspectives in Chapter 5. The manuscript closes with a summary of teaching, supervision and grant management in Appendix A and my full list of publications in Appendix B.

[1] https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000298904/

1.1 SCIENTIFIC BACKGROUND

This manuscript presents a selection of my scientific contributions since 2010, when I completed my MSc in mathematics at Sorbonne Université[2] and worked as a research assistant at Danmarks Tekniske Universitet.[3] In February 2011, I started my PhD at Sorbonne Université under the joint supervision of Gérard Biau and Eric Moulines, and defended in December 2013. After a short postdoc at Sorbonne Université I was hired in November 2014 as a (tenured) research scientist at Inria (and member of the Modal team in the Lille - Nord Europe research centre, jointly with Laboratoire Paul Painlevé UMR 8524 of Université de Lille). In December 2018 I started a secondment at University College London (UCL) with the rank of Associate Professor in machine learning, to lead the Inria and UCL partnership. I have been a Turing Fellow with the Alan Turing Institute since 2021.

Since my PhD, slowly but surely, I have developed an obsession with the notion of *generalisation* in machine learning, *i.e.*, the study of how the predictive performance of machine learning algorithms can be assessed and in some cases further promoted, in particular when said algorithms are deployed on new data sets and tasks. Generalisation is arguably central to machine learning and artificial intelligence: qualitatively, lack of generalisation is commonly known as overfitting and hints at poor predictive performance. Overfitting occurs when an algorithm "copies" the training data but proves incapable of performing well on new data. This is, to use a simple analogy, the difference between memorising exam answers by heart and truly learning a subject. A crucial concept when it comes to intelligence, be it artificial or not.

Designing algorithms which are able to generalise reliably and efficiently is one of the overarching goals in machine learning. As such, *generalisation bounds* are often the hallmark of theoretical guarantees for machine learning algorithms: this forms a rich field of research known as statistical learning theory, stemming from foundational contributions starting in the 1960s.

A significant part of my work not only consists in studying the generalisation abilities of existing algorithms (therefore providing a guarantee or certificate[4] that the hypothesis performs well on new data, provided that the assumptions under which the bound was derived are

[2] Formerly known as Université Pierre et Marie Curie (UPMC).

[3] Denmark Technical University (DTU).

[4] Theoretical or numerical – ideally both.

valid), but also focuses on the design of novel algorithms by optimising generalisation-promoting training objectives – I call this strategy *generalisation-by-design*.

This led to my extensive study of different techniques to study generalisation in statistical learning theory[5], with a particular fondness for the PAC[6]-Bayes theory which has never left me[7] and ended up being a central topic in my scientific contributions with about half of my research papers with a more or less pronounced PAC-Bayes flavour. This culminated into the writing of our recent monograph Hellström et al. (2024) on generalisation through the lenses of PAC-Bayes and information theory, with which this manuscript obviously overlaps.

A learning algorithm is a (potentially stochastic) rule for selecting a hypothesis (or predictor), given a training data set. Roughly speaking, a problem is PAC learnable if there exists a learning algorithm such that, for any data distribution, the resulting predictor has a good performance with high probability. The PAC-Bayes theory investigates the generalisation performance of randomised predictors and is now regarded as an established and principled approach to study generalisation and design new learning strategies. In the past decade, it has garnered increased interest due to its flexibility and promising results, including in deep learning where understanding and controlling overfitting is crucial. PAC-Bayes learning blends concepts from Bayesian inference and PAC learning. The PAC-Bayes approach delivers bounds which are probabilistic, offering a 'probably approximately correct' guarantee of performance on unseen data. The innovative aspect of PAC-Bayes is its ability to provide tighter generalisation bounds than classical PAC learning, especially in the context of large datasets and complex models. PAC-Bayes offers a theoretically grounded way to assess how well a learning algorithm will perform in practice, not just on the training data but more importantly on new, unseen data.

The PAC-Bayesian approach was pioneered by the seminal works of Shawe-Taylor and Williamson (1997), McAllester (1998), and McAllester

[5] Which has been studied through many angles, such as the Rademacher complexity (Bartlett and Mendelson, 2002), algorithmic stability (Rogers and Wagner, 1978; Devroye and Wagner, 1979; Bousquet and Elisseeff, 2002), margins (Shawe-Taylor and Cristianini, 1999), or norms (Neyshabur et al., 2015) to name but a few. See Mohri et al., 2018 for a comprehensive treatment.

[6] Probably Approximately Correct – this acronym was coined by Valiant (1984). Nowadays this is arguably the dominant paradigm for analysing generalisation.

[7] I must here pay tribute to Pierre Alquier who introduced me to PAC-Bayes during the very first weeks of my PhD back in February 2011, with commendable enthusiasm and patience.

(1999) initially, with significant later contributions from, *e.g.*, Langford and Seeger (2001), Seeger (2002), Maurer (2004), Catoni (2003), Catoni (2004), and Catoni (2007), and started as a quest to obtain Bayesian-flavoured versions of PAC generalisation bounds, as the name implies. PAC bounds are independent of the specific learning algorithm used, as they hold uniformly over the class of possible hypotheses. In contrast, PAC-Bayesian bounds take into account the learning algorithm by explicitly incorporating a distribution over hypotheses, hence the Bayesian suffix. PAC-Bayesian bounds provide insights on how to quantify uncertainty in several machine learning problems, and further study the level of correlations between equally admissible hypotheses.

In a nutshell, the PAC-Bayes theory leads to (i) generalisation bounds which can match (minimax) optimal rates of convergence (ii) principled algorithms and models by minimising an upper bound of the generalisation error as a training objective (iii) numerically non-vacuous generalisation guarantees, provably securing future performance on unseen data. PAC-Bayes is emerging as the prime framework for analysing contemporary machine learning algorithms, which is fueled by the fact that it is one of the few strategies that deliver numerically non-vacuous generalisation bounds for some deep neural networks architectures. This manuscripts motivates my scientific vision of fostering generalisation as a central tool in machine learning, both from a theoretical and algorithmic perspective. I refer the interested reader to my primer on PAC-Bayesian learning Guedj (2019) for a quick dive into PAC-Bayes; to the excellent and recently published tutorial Alquier (2024) for a more hands-on material on PAC-Bayes; and finally to our recent monograph Hellström et al. (2024) for a broader treatment of generalisation in machine learning.

NOTATION. We consider the training examples to lie in a set $\mathcal{Z}$, referred to as the *instance space*. In supervised learning, the instance space is a product between a *feature space* $\mathcal{X}$ and a *label space* $\mathcal{Y}$, so that $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The learning algorithm has access to a *training set* $\mathbf{Z} = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$, consisting of n training examples. Usually, we assume that the training examples are independent and identically distributed (*i.i.d.*),[8] with each training example being drawn from a data distribution P_Z on $\mathcal{Z}$. We denote the distribution of $\mathbf{Z}$ as $P_{\mathbf{Z}} = P_Z^n$.

Confronted with the training data, the learner selects a hypothesis W from a set $\mathcal{W}$, called the *hypothesis space*. Again, in supervised learn-

[8] This assumption is classical in statistical learning theory, although some of my contributions focus on relaxing and even removing it.

ing, $\mathcal{W}$ is typically a subset of all functions from $\mathcal{X}$ to $\mathcal{Y}$, or the parameters of such functions. The method by which the learner chooses the hypothesis is described by a (probabilistic) mapping from the training set $\mathbf{Z}$ to the hypothesis W, denoted by $P_{W|\mathbf{Z}}$ and referred to as a *learning algorithm*. Mathematically, it can be seen as a stochastic kernel, which gives rise to a probability distribution on $\mathcal{W}$ for each instance of $\mathbf{Z}$. Note that $P_{W|\mathbf{Z}}$ is defined for a specific size n of the training set. We usually assume that the learning algorithm is flexible with respect to the size of the training set, so that the mapping $P_{W|\mathbf{Z}}$ is defined for any n.

The quality of a specific hypothesis $w \in \mathcal{W}$ with respect to a sample $z \in \mathcal{Z}$ is measured by a *loss function*, $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_+$. To give some classical examples of loss functions, consider supervised learning, where the sample is decomposed into features and labels (or inputs and outputs) as $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ and the hypotheses $w \in \mathcal{W}$ are functions $w : \mathcal{X} \to \mathcal{Y}$. For classification, where the label space $\mathcal{Y}$ is discrete, a typical loss function is the classification error $\ell(w, z) = 1\{w(x) \neq y\}$. Here, $1\{\cdot\}$ denotes the indicator function. For regression, where the label space is continuous, a standard choice is the squared loss $\ell(w, z) = (w(x) - y)^2$.

The true goal of the learner is to select a hypothesis that performs well on fresh data from the distribution P_Z, as measured by the loss function. This is formalised by the *population loss*

$$L(w) := L_Z(w) = \mathbb{E}_{P_Z} \ell(w, Z),$$

sometimes referred to as the (true) *risk* of a hypothesis. Obviously the true data distribution is unknown, which implies that the population loss cannot be computed by the learner. However, by averaging the loss function over training data, the learner obtains the *training loss* (or *empirical risk*)

$$\widehat{L}(w) := L_{\mathbf{Z}}(w) = \frac{1}{n} \sum_{z \in \mathbf{Z}} \ell(w, z),$$

which serves as an estimate of the population loss. A natural procedure for selecting a hypothesis is to minimise the training loss. This is referred to as *empirical risk minimisation* (ERM), and is successful in finding a hypothesis with low population loss if the difference between population loss and training loss is small. This is measured by the *generalisation error*

$$\text{gen}(w, \mathbf{Z}) = L(w) - \widehat{L}(w),$$

which is also called the *generalisation gap*.

In the PAC-Bayesian literature, most bounds are on the generalisation error when averaged over the learning algorithm

$$\mathbb{E}_{P_{W|\mathbf{Z}}}\mathrm{gen}(W, \mathbf{Z}),$$

and hold with probability at least $1-\delta$ under $P_{\mathbf{Z}}$ for some confidence parameter $\delta \in (0,1)$. The change in perspective in the PAC-Bayesian approach, as compared to the classical statistical learning literature, is significant. We no longer ask whether there are specific hypotheses w that perform well: instead, we ask if there are distributions $P_{W|\mathbf{Z}}$ over hypotheses that do. To highlight the conceptual connection to Bayesian statistics, the distribution $P_{W|\mathbf{Z}}$ is usually termed *posterior*. This distribution is compared, via information-theoretic metrics (such as the Kullback-Leibler divergence), to a reference measure called the *prior*. Another significant feature that is shared among many PAC-Bayesian bounds is that they hold uniformly for all choices of posterior: this opens the way to the strategy of optimising the bounds with respect to the posterior, as exemplified in Chapter 3.

To simplify notation, we will write $\underline{L} := \mathbb{E}_{P_{W|\mathbf{Z}}}L(W)$, $\underline{\widehat{L}} := \mathbb{E}_{P_{W|\mathbf{Z}}}\widehat{L}(W)$ and consequently $\underline{\mathrm{gen}} := \mathbb{E}_{P_{W|\mathbf{Z}}}\mathrm{gen}(W, \mathbf{Z})$.

In developing the PAC-Bayesian paradigm, I have focused most of my research efforts on the following three areas (with most of my research papers contributing to more than one).

AXIS 1: THEORY OF GENERALISATION. The literature on generalisation bounds has considerably enriched since the early days of PAC-Bayes in the late 1990s. Generalisation bounds often are variations of the following prototypical form. In a supervised setting, recall the population loss

$$L_{P_Z}(w) = \mathbb{E}_{P_Z}\ell(w, Z).$$

A canonical form of a PAC-Bayes bound is

$$\mathbb{P}\left[\mathbb{E}_{P_{W|\mathbf{Z}}}\mathrm{gen}(W, \mathbf{Z}) \leqslant \Phi\left(\frac{\text{Complexity} + \log\frac{1}{\delta}}{\text{Rate}}\right)\right] \geqslant 1-\delta,$$

where Complexity denotes a term characterising the difficulty of the learning problem (in most bounds, this is the Kullback-Leibler divergence between the prior and the posterior), the Rate term is a function of the sample size n, and Φ is a functional – in the vast majority of the literature, $\Phi\colon x \mapsto \sqrt{x}$. A more common form of the bound is

$$\mathbb{P}\left[\mathbb{E}_{P_{W|\mathbf{Z}}}L_{P_Z}(w) \leqslant \mathbb{E}_{P_{W|\mathbf{Z}}}L_{\mathbf{Z}}(w) + \sqrt{\frac{\text{Complexity} + \log\frac{1}{\delta}}{\text{Rate}}}\right] \geqslant 1-\delta.$$

We note that the term $\log \frac{1}{\delta}$ illustrates the tradeoff between how tight the bound can be, and with what probability it is holding true: the extreme case $\delta \to 1$ leads with probability 1 to the statement that the generalisation gap is upper bounded by $+\infty$. True, but barely useful! The typical rate is of order $\sqrt{n}$, although several papers have achieved so-called fast rates (of order n). See, *e.g.*, the works of Van Erven et al. (2015) and Grünwald and Mehta (2020).

Thus, these bounds allow controlling the generalisation gap by a term which measures the complexity of the problem that decreases with the square root of the number of points on which the algorithm is trained – all this happening with a probability arbitrarily close to 1. These bounds hold under classical assumptions, such as i.i.d. samples or bounded data distributions. The Kullback-Leibler divergence implies that the posterior is absolutely continuous with respect to the prior (which serves as a reference measure). A large part of my work has been focused on improving these bounds or extending their scope of validity. This includes obtaining fast rates [**BG-Conf3**]; [**BG-Conf21**], or substituting the Kullback-Leibler divergence with more general divergences [**BG-Journal6**]; [**BG-Preprint7**]; [**BG-Preprint10**]; [**BG-Conf23**]; [**BG-Preprint16**]. It also includes relaxing the necessary assumptions to obtain bounds [**BG-Journal6**]; [**BG-Journal14**]; [**BG-Journal15**]; [**BG-Journal21**]. We have also contributed to extending these results to the unsupervised setting [**BG-Journal9**], to ranking [**BG-Journal7**], to high-dimensional additive models [**BG-Journal2**] and to online learning [**BG-Conf17**]. We have generalised the form of the discrepancy between population and training losses (from a difference to a generic convex function, [**BG-Conf24**]), incorporated a possible hierarchical structure in the data [**BG-Preprint2**], and linked generalisation to stability properties of learning algorithms [**BG-Preprint1**]. We have developed better concentration inequalities [**BG-Conf3**]; [**BG-Preprint14**], extended the range of usable loss functions (like the error distribution [**BG-Preprint5**] or the quantiles of the loss function [**BG-Conf7**]), and studied the derandomisation of posteriors [**BG-Conf13**]; [**BG-Preprint6**], along with extending to the neural tangent kernel model [**BG-Journal20**], among others. We have also contributed to establishing bounds where none previously existed, for example, for generative models like variational autoencoders [**BG-Conf16**], unsupervised contrastive learning [**BG-Conf8**], or models of factorisation of large random matrices [**BG-Journal5**]; [**BG-Conf4**]. My contributions have therefore pushed the state-of-the-art of PAC-Bayes generalisation bounds in a large number of learning problems and with increasingly weakened assumptions.

AXIS 2: DESIGNING NEW "GENERALISATION-DRIVEN" LEARNING ALGORITHMS. A particularly remarkable aspect of the prototypical bound mentioned is that in the upper bound term

$$\mathbb{E}_{P_{W|Z}} L_{\mathbf{Z}}(w) + \sqrt{\frac{\text{Complexity} + \log \frac{1}{\delta}}{\text{Rate}}},$$

the different terms are generally calculable or approximable, and often differentiable. Since most PAC-Bayes bounds hold uniformly for all posterior, this incentivises to optimise with respect to the posterior. The major idea, therefore, is to make this upper bound the new learning objective, optimising to obtain the optimal posterior distribution over predictors. This represents a profound paradigm shift from the literature: one of the most classical learning strategies consists of optimising the empirical error (the famous Empirical Risk Minimisation, ERM principle), which often leads to overfitting – this objective has long been regularised with penalty terms that favour generalisation (such as the Lasso). I argue for using more directly generalisation-promoting objectives, *i.e.*, generalisation upper bounds (which must be as tight as possible, hence the importance of Axis 1 above).

This strategy—which I call *generalisation-by-design* has been at the core of several of my works. We have explored this strategy in frameworks as diverse as federated learning [**BG-Preprint12**], majority voting [**BG-Conf11**]; [**BG-Conf14**], Wasserstein-regularised learning [**BG-Conf23**], and the construction of ad hoc architectures by optimising a PAC-Bayes generalisation bound [**BG-Conf2**]; [**BG-Journal13**]; [**BG-Conf13**]; [**BG-Conf12**]; [**BG-Preprint4**]; [**BG-Conf10**], among others. I highlight two examples of *generalisation-by-design* in Chapter 3.

AXIS 3: FROM THEORY TO NUMERICAL GUARANTEES. Distinctly and yet complementarily to the previous axis, in certain learning problems it is possible to *numerically* calculate the value of the terms in the upper bound

$$\mathbb{E}_{P_{W|Z}} L_{\mathbf{Z}}(w) + \sqrt{\frac{\text{Complexity} + \log \frac{1}{\delta}}{\text{Rate}}},$$

and thus, it is possible to provide machine learning practitioners with guarantees that are not only theoretical but also practical (this is also referred in some papers as *algorithm certification*). Concretely, in a classification problem with a loss function bounded by 1 (such as the 0-1 loss that counts the number of errors on labels, for example of images), if the upper bound is greater than 1, the bound is said to be trivial (*vacuous*): although theoretically correct, it provides no useful information as it upper bounds a number that is bounded by 1, with a quantity that is larger. Conversely, in the case where the numerical value of the bound is less than 1 (*non-vacuous*), it can provide valuable information on the generalisation performance. This is one of the reasons behind the impressive surge in interest in PAC-Bayes strategies: Dziugaite and Roy (2017) was the first demonstration that PAC-Bayes could lead to numerically non-trivial bounds for deep learning. It is in this direction (as well as for other learning models) that I have also contributed numerical analyses demonstrating the merits of the PAC-Bayesian approach in practice [**BG-Conf2**]; [**BG-Journal13**]; [**BG-Conf13**]; [**BG-Conf12**]; [**BG-Preprint4**]; [**BG-Conf10**]; [**BG-Conf11**]; [**BG-Preprint12**]; [**BG-Conf3**]; [**BG-Conf8**]; [**BG-Conf23**]; [**BG-Preprint16**].

1.2 OVERVIEW OF CONTRIBUTIONS IN THIS MANUSCRIPT

The three main axes above articulate my vision of *generalisation as the driving force of learning*. Building on concepts and tools from statistical learning theory and the PAC-Bayes theory, I describe in this manuscript the following three sets of contributions in machine learning.

(I) NON-TRIVIAL GENERALISATION GUARANTEES FOR DEEP NEURAL NETWORKS. PAC-Bayes delivers state-of-the-art and/or numerically non-vacuous generalisation bounds for some specific architectures of deep networks. This is transverse to Axis 1, 2 and 3. I illustrate this in Chapter 2 through two of my recent papers on that topic: [**BG-Conf2**] and [**BG-Conf12**].

(II) GENERALISATION-DRIVEN LEARNING ALGORITHMS. PAC-Bayes yields generalisation bounds which are often computable (either directly or through proxies), which incentivises to turn these generalisation bounds into training objectives. This matches Axis 2 – and to a large extent Axis 3. Illustrated in Chapter 3 by two of my recent papers [**BG-Conf11**] and [**BG-Conf23**].

(III) GENERALISATION BEYOND THE CLASSICAL LEARNING FRAMEWORKS. PAC-Bayes allows to study and understand generalisation in new settings, where little or no generalisation results exist (including deep learning). It also allows to relax classical assumptions, extending the scope of validity of generalisation guarantees. This is transverse to Axis 1, 2 and 3. Illustrated in Chapter 4 by two of my recent papers [BG-Conf8] and [BG-Conf16].

Coherent set of contributions

Note that the frontiers between all three sets of contributions are fairly porous: several papers could have been moved accross the three chapters, and several others could have been included. I chose to focus on some of their assets to better illustrate the overall manifesto for PAC-Bayes as a driving principle of contemporary machine learning.

This forms three pairs of papers ([BG-Conf2]; [BG-Conf8]; [BG-Conf11]; [BG-Conf16]; [BG-Conf12]; [BG-Conf23]) out of the 70 (as of February 2024) documents I have written in my academic career, to highlight my view of machine learning.

With this line of work, I am aiming towards a better understanding of generalisation in machine learning, with the ultimate goal to contribute to reduce the massive gap between how humans generalise with a fraction of the data and compute needed by machines, paving the way to more frugal artificial intelligence systems.

This manuscript sums up some of the overarching guiding principles of my research since 2010, which progressively sedimented as I built my research group over the years. It is tempting to see this manuscript not only as a milestone and a collection of past work, but also as a starting point of new activities and in particular of my group, for which here is what I call the *PAC-Bayes Manifesto*, illustrating my view that PAC-Bayes is one of the rising principles in contemporary machine learning.

The PAC-Bayes Manifesto

- ◇ **Theory**: PAC-Bayes generalisation bounds are, in many cases, either the only bounds available or matching the state-of-the-art (*e.g.*, achieving the optimal rate of convergence), in a broad range of settings (batch or online, supervised or unsupervised, *etc.*).
- ◇ **Algorithms:** the PAC-Bayes theory allows generalisation-by-design. By turning the generalisation bound into a training objective, we are designing new algorithms (or recovering existing algorithms) which inherit solid generalisation guarantees.

◇ **Numerical results:** PAC-Bayes leads to numerically non-vacuous bounds (or certificates) in a broad range of settings, including deep learning for specific neuronal architectures.

1.3 OTHER CONTRIBUTIONS

For the sake of conciseness, this manuscript only covers a fraction (6 papers out of 70, about 8%) of my research since 2010. My contributions are roughly divided in two groups of papers of similar sizes. About half of my contributions are about generalisation in machine learning and PAC-Bayes: I list here the papers falling into that category, which I do not discuss in this manuscript.

- Generalisation bounds for aggregates of predictors [**BG-Academic2**].
- PAC-Bayes generalisation bounds and MCMC algorithm for the generalised additive model [**BG-Journal2**].
- An overview of Bayesian and PAC-Bayesian learning advances (back in 2015) [**BG-Journal3**].
- Design, theoretical studies, python implementation and application to image denoising of a novel non-linear algorithm for aggregating predictors called COBRA [**BG-Journal4**]; [**BG-Journal8**]; [**BG-Journal12**]; [**BG-Conf5**].
- Design and theoretical study of a new notion of stability for machine learning algorithms [**BG-Preprint1**].
- PAC-Bayesian bounds for non-negative matrix factorisation algorithms [**BG-Journal5**].
- PAC-Bayesian bounds and algorithm for binary ranking in high dimensions [**BG-Journal7**].
- PAC-Bayesian bounds holding with little to no assumptions on data distributions, and extension to f-divergences [**BG-Journal6**].
- Theoretical analysis of limitations of the PAC-Bayes approach to obtain fast rates [**BG-Journal14**].
- A primer on PAC-Bayesian learning [**BG-Conf1**] and a monograph on generalisation theory through information-theoretic and PAC-Bayes bounds [**BG-Preprint15**].

- PAC-Bayes analysis and training algorithm for deep neural networks [BG-Journal13].
- PAC-Bayes generalisation bounds for the conditional value at risk [BG-Conf7].
- PAC-Bayesian bounds for majority votes through the use of margins [BG-Conf14].
- Empirical studies of the role of data-dependent PAC-Bayes priors in deep learning, and of the performance of deep neural networks trained by optimising PAC-Bayes generalisation bounds [BG-Preprint4]; [BG-Conf10].
- Replacing the Kullback-Leibler divergence by the Wasserstein distance in PAC-Bayes generalisation bounds, for batch and online learning [BG-Preprint10]; [BG-Conf23].
- PAC-Bayes generalisation bounds and algorithms for structured prediction [BG-Preprint2].
- Upper and lower bounds on the performance of kernel PCA [BG-Preprint3].
- PAC-Bayes-inspired algorithms for federated learning [BG-Preprint12].
- PAC-Bayes bound for controlling the distribution of error rather than its expectation [BG-Preprint5].
- PAC-Bayes bounds for unbounded losses [BG-Journal15]; [BG-Journal21].
- A new change of measure inequality for f-divergences [BG-Preprint7]
- Online PAC-Bayes learning [BG-Conf17].
- Derandomised PAC-Bayes bounds [BG-Conf13]; [BG-Preprint6].
- Fast rates for PAC-Bayes bounds [BG-Conf3]; [BG-Conf21].
- Generalisation bounds exploiting flat minima [BG-Preprint14].
- Generalisation bounds with interpolation of divergences [BG-Preprint16]
- A PAC-Bayes-inspired NTK algorithm [BG-Journal20].
- Generalisation bounds for arbitrary convex functionals of the discrepancy between training and population losses [BG-Conf24].

The other half of my papers address a broader range of contributions, in machine learning and statistics.

- Sampling strategies for estimating the genetics admixture in population genetics [**BG-Academic1**]; [**BG-Journal1**].
- Online clustering algorithms and theoretical analysis [**BG-Journal9**]; [**BG-Conf9**].
- Sequential learning of principal curves [**BG-Journal16**].
- Decentralised learning with copulas [**BG-Conf6**].
- Revisiting clustering as matrix factorisation on Stiefel manifolds [**BG-Conf4**].
- Model validation using mutated training labels [**BG-Journal25**].
- Analysis of the diffusion and polarisation of opinions on (social) networks, and strategies to mitigate the emergence of echo chambers, with applications to content recommendation on social networks and forecasting of elections results [**BG-Conf20**]; [**BG-Journal17**]; [**BG-Conf20**]; [**BG-Journal26**].
- Two-sample, goodness-of-fit, and independence kernel tests based on the Maximum Mean Discrepancy (MMD), the Kernel Stein Discrepancy (KSD), and the Hilbert Schmidt Independence Criterion (HSIC) [**BG-Journal24**]; [**BG-Conf18**]; [**BG-Conf19**].
- Novel generalised Bayesian algorithm for uncertainty quantification in biology and in particular in anaerobic digestion models [**BG-Journal23**].
- On-flight learning of aerodynamics approximation for aircrafts, and algorithms for trajectory and fuel dynamic optimisation [**BG-Journal11**]; [**BG-Journal18**].
- Introduction of a novel multi-task Gaussian process series of algorithms for time series forecasting and clustering [**BG-Journal19**]; [**BG-Journal22**].
- Introduction of a novel latent space data augmentation technique for imbalanced data classification [**BG-Preprint8**].
- Introducing and analysing a diffeomorphism-invariant dissimilarity measure, and efficiently approximated through Nyström sampling [**BG-Conf15**].

- Closed-form filtering for non-linear systems [**BG-Preprint13**].

- Online learning algorithms using expert advice [**BG-Preprint11**].

I also list some non-technical documents.

- Reflections on how research software is referenced [**BG-Journal10**].

- Using virtue epistemology to evaluate AI in knowledge production, aligning it with the distinct needs of fields like social science and medicine [**BG-Preprint9**].

- Highlights from the webinar "Covid-19 and AI: unexpected challenges and lessons" that I co-organised and chaired, witht the support of the French Embassy in the United Kingdom and the Franco-British Data Society [**BG-TechReport2**].

- Reflections on the promotion of qualitative indicators to evaluate research [**BG-TechReport1**].

2 GENERALISATION FOR DEEP NEURAL NETWORKS

Outline

In this chapter, we present a selection of our contributions to the topic of generalisation bounds for deep neural networks, and we particularly emphasize that PAC-Bayes leads to numerically non-vacuous upper bounds on the risk of some architectures of deep neural networks.

The empirical successes of deep learning since the early 2010s have rightly attracted growing interest in the theory of deep learning[1]: among the numerous contributions in this field in recent years, the interest in the generalisation performance of algorithms based on deep neural networks has been an important component of my research since 2019 [**BG-Journal13**]; [**BG-Conf13**]; [**BG-Conf12**]; [**BG-Conf2**]; [**BG-Preprint4**]; [**BG-Conf10**]. In [**BG-Conf2**]; [**BG-Journal13**]; [**BG-Conf13**]; [**BG-Conf12**], we demonstrated PAC-Bayes generalisation bounds for several neural network architectures and derived algorithms to minimise these bounds to directly train the networks. We numerically evaluated the predictive performance of these networks (trained by PAC-Bayes) and the numerical value of the bounds. Doing so, we established that the PAC-Bayesian approach led to new theoretical results, which translated into (i) original neural architectures (ii) performance similar to the state

[1] As often, empirical performance has largely preceded the theoretical understanding of the algorithms at work.

of the art (iii) numerical generalisation guarantees (or *certificates*). We also conducted in [**BG-Preprint4**]; [**BG-Conf10**] empirical studies on the impact of different choices of prior distributions on predictive performance and numerical values of bounds.

The breakthrough was to establish the mathematical framework which allowed to derive a bound, and then use it as a training objective. Our big idea introduced in [**BG-Conf2**] is to use a binary activation function in the network: at first glance a curious choice since the sign function does not play nicely with gradient descent algorithms. But by adopting a PAC-Bayesian viewpoint (largely innovative in deep learning at that time), we realised that the prediction made by the network is written as the expectation of the sign function (composed as many times as there are layers), which turns out to be the error function of Gauss – and which is amenable to gradient descent. We were thus able to prove a PAC-Bayesian bound and train the binary activation network, which is not directly possible with other methods. The works [**BG-Journal13**]; [**BG-Conf13**]; [**BG-Conf12**] refine this strategy for other architectures but proceed from a similar trick: for example, in [**BG-Conf12**] we generalised the approach by constructing a shallow neural network whose activation function is the Gaussian error function itself. This leads again to a training strategy of the network by stochastic gradient descent that achieves excellent performance.

The ever-growing appetite in the scientific community and beyond for deep learning fuels our position aimed at delivering "theoretical and numerical certificates" to neural network architectures to guarantee their good performance.

We now briefly illustrate our contributions by sketching the ideas from [**BG-Conf2**] (NeurIPS 2019) and [**BG-Conf12**] (ICML 2022).

2.1 PAC-BAYESIAN BINARY ACTIVATED DEEP NEURAL NETWORKS

In the joint work [**BG-Conf2**] with Gaël Letarte, Pascal Germain and François Laviolette (Université Laval, Canada), we introduced a framework, called PBGNet (PAC-Bayesian Binary Gradient Network), to analyse and train multilayer neural networks with binary activation. The key contributions are (i) the development of an end-to-end training framework for deep neural networks with binary activations, overcoming the challenge posed by the non-differentiability of the binary activation function (ii) the establishment of empirically non-vacuous PAC-

Bayesian generalisation bounds for such networks, demonstrating their theoretical robustness and potential for practical applications.

We focus our study on deep neural networks with a sign activation function and unconstrained weights. We call such networks *binary activated multilayer* (BAM) networks. This specialisation leads to nonvacuous generalisation bounds which hold under the sole assumption that training samples are *i.i.d.*. We provide a PAC-Bayesian bound holding on the generalisation error of a continuous aggregation of BAM networks. This leads to an original approach to train BAM networks, PBGNet. The building block of PBGNet arises from the specialisation of PAC-Bayesian bounds to linear classifiers (Germain et al., 2009), that we adapt to deep neural networks (through intermediary results for shallow networks).

The output of a BAM network with L layers on an input data point x is given by

$$f_\theta(x) = \operatorname{sign}(W_L \operatorname{sign}(W_{L-1} \operatorname{sign}(\dots \operatorname{sign}(W_1 x)))) \,,$$

where W_k represent the weight matrices of each layer $k = 1, \dots, L$. We consider the averaged network with respect to a PAC-Bayes posterior Q_w (a Gaussian distribution with mean w) given by

$$F_w(x) = \mathbb{E}_{v \sim Q_w} f_v(x) \; = \operatorname{erf}\Big(\tfrac{w \cdot x}{\sqrt{2}\|x\|}\Big), \quad \text{with } \operatorname{erf}(x) = \tfrac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \,.$$

Our main result is as follows. The vector θ represents the weights of any BAM network, and F_θ represents the output of the network. We consider Gaussian distributions for the prior and posterior.

Theorem 2.1. *Given prior parameters $\mu \in \mathbb{R}^D$, with probability at least $1-\delta$, we have for all θ on $\mathbb{R}^D$*

$$\begin{aligned} L(F_\theta) &\leqslant \sup_{0 \leqslant p \leqslant 1} \Big\{ p : \mathrm{kl}(\widehat{L}(F_\theta) \| p) \leqslant \frac{1}{n} [\mathrm{KL}(\theta, \mu) + \ln \tfrac{2\sqrt{n}}{\delta}] \Big\} \\ &= \inf_{C > 0} \Big\{ \tfrac{1}{1 - e^{-C}} \Big(1 - \exp\Big(-C\, \widehat{L}(F_\theta) - \frac{1}{n} [\mathrm{KL}(\theta, \mu) + \ln \tfrac{2\sqrt{n}}{\delta}] \Big) \Big) \Big\} \,. \end{aligned}$$

In PBGNet, we optimise that bound through stochastic gradient descent. By noting P_μ the Gaussian prior with mean μ, this results in the following training objective:

$$C\, n\, \widehat{L}(F_w) + \mathrm{KL}(Q_w \| P_\mu) \; = \; C \, \frac{1}{2} \sum_{i=1}^{n} \operatorname{erf}\Big(-y_i \, \frac{w \cdot x_i}{\sqrt{2}\|x_i\|} \Big) + \frac{1}{2} \|w - \mu\|^2 \,.$$

One key aspect of PBGNet is that it can be seen as a Kullback-Leibler regularisation layer of training Bayesian neural networks. Indeed the computation of the bound relies on two crucial elements: the empirical loss on the training set and the Kullback-Leibler divergence between the prior and the posterior distributions of the parameters. This regularisation term penalises the discrepancy between the learned parameters and the prior parameters, encouraging a balance between the loss and the complexity of the model through the Kullback-Leibler divergence.

We report in the paper the results of numerical experiments on several classical datasets; a selection of which is shown in Table 2.1.

Table 2.1: Experiment results for PBGNet on classical binary classification datasets: error rates on the train and test sets (E_S and E_T), and generalisation bounds (Bnd). The PAC-Bayesian bounds hold with probability 0.95. The value of the bound is an upper bound (numerical certificate) for the test error (which is not known) with confidence 95%.

Dataset	PBGNet		
	E_S	E_T	Bound
ads	0.033	0.033	0.060
adult	0.149	0.154	0.164
mnist17	0.004	0.004	0.010
mnist49	0.016	0.017	0.028
mnist56	0.009	0.009	0.018
mnistLH	0.026	0.027	0.033

2.2 NON-VACUOUS BOUNDS FOR SHALLOW NEURAL NETWORKS

The paper **[BG-Conf12]** is mostly the work of Felix Biggs, who is doing his PhD under my supervision at UCL. We focus on a specific class of shallow neural networks with a single hidden layer, namely those with L_2-normalised data and either a sigmoid-shaped Gaussian error function ("erf") activation or a Gaussian Error Linear Unit (GELU) activation. For these networks, we derive new generalisation bounds through the PAC-Bayesian theory; unlike most existing such bounds they apply to neural networks with *deterministic* rather than randomised parameters. Our bounds are empirically non-vacuous when the network is trained with vanilla stochastic gradient descent on MNIST, Fashion-MNIST, and binary classification versions of the above.

One of the key ideas is to consider the output of neural networks as PAC-Bayes average of majority votes. Computing the average output of deep neural networks with randomised parameters is generally intractable: therefore most such works have focused on cases where the average output is simple to compute, as for example when considering linear predictors. Here, building on ideas from **[BG-Conf13]**, we show that provided our predictor structure factorises in a particular way, more complex majority votes can be constructed. In particular, we give formulations for randomised predictors whose majority vote can be expressed as a deterministic single-hidden-layer neural network. Through this, we obtain classification bounds for these *deterministic* predictors that are non-vacuous on the celebrated baselines MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al., 2017), and binarised versions of the above. We believe these are the first such results.

Our work fundamentally relates to the question: what kind of properties or structures in a trained network indicate likely generalisation to unseen data? It has been shown by Zhang et al. (2017) that neural networks trained by SGD can perfectly overfit large datasets with randomised labels, which would indicate a lack of capacity control, while simultaneously generalising well in a variety of scenarios. Thus, clearly any certification of generalisation must involve extracting additional information other than the train loss—for example, the specific final network chosen by SGD. How do the final parameters of a neural network trained on an "easy" data distribution as opposed to a pathological (*e.g.*, randomised label) one differ? A common answer to this has involved the return of capacity control and the norms of the weight matrices, often measured as a distance to the initialisation (as done, *e.g.*, in Dziugaite and Roy, 2017; Bartlett et al., 2017b; Neyshabur et al., 2018a).

We suggest, following insights from Dziugaite et al. (2020), that a better answer lies in utilising the empirically-observed stability of SGD on easy datasets. We give bounds that are tightest when a secondary run of SGD on some subset of the training set gives final weights that are close to the full-dataset derived weights. This idea combines naturally in the PAC-Bayes framework with the requirement of perturbation-robustness of the weights—related to the idea of flat-minima (Hinton and Camp, 1993; Hochreiter and Schmidhuber, 1997)—to normalise the distances between the two runs. By leveraging this commonly-observed empirical form of stability we effectively incorporate information about the inherent easiness of the dataset and how adapted our neural network architecture is to it. Although it is a deep and interesting theoretical question as to when and why such stability occurs under SGD, we believe that by making the link to generalisation explicit we solve some of the puzzle.

SETTING. We consider D-class classification on a set $\mathcal{X} \subset \mathbb{R}^d$ with "score-output" predictors returning values in $\hat{\mathcal{Y}} \subset \mathbb{R}^D$ with multi-class label space $\mathcal{Y} = [D]$, or in $\hat{\mathcal{Y}} = \mathbb{R}$ with binary label space $\mathcal{Y} = \{+1, -1\}$. The prediction is the argmaximum or sign of the output and the misclassification loss is defined as $\ell(f(x), y) = \mathbb{1}\{\arg\max_{k \in [D]} f(x)[k] \neq y\}$ or $\ell(f(x), y) = \mathbb{1}\{yf(x) \leqslant 0\}$ respectively. We write

$$L(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\ell(f(x), y)$$

and

$$\hat{L}(f) := m^{-1} \sum_{(x,y)\in S} \ell(f(x), y)$$

for the risk and empirical risk of the predictors with respect to data distribution $\mathcal{D}$ and i.i.d. m-sized sample $S \sim \mathcal{D}^m$.

OVERVIEW OF OUR CONTRIBUTIONS. We derive generalisation bounds for a single-hidden-layer neural network $F_{U,V}$ with first and second layer weights U and V respectively taking the form

$$F_{U,V}(x) = V\,\phi\left(\beta \frac{Ux}{\|x\|_2}\right)$$

with ϕ being an element-wise activation. We consider such networks with an erf activation function (we then call the network SHEL, or single hidden erf layer), or a GELU activation. The Gaussian Error Linear Unit (GELU) is a commonly-used alternative to the ReLU activation defined by $\mathrm{GELU}(t) := \Phi(t)\,t$ where $\Phi(t)$ is the standard normal CDF.

If the data is normalised to have $\|x\|_2 = \beta$ these are simply equivalent to one-hidden-layer neural networks with activation ϕ and the given data norm. We provide high-probability bounds on $L(F_{U,V})$ of the approximate form

$$2\mathbb{E}_{f\sim Q}\hat{L}(f) + \mathcal{O}\left(\frac{\beta\|U - U^n\|_F + \|V - V^n\|_F}{\sqrt{m-n}}\right),$$

where Q is a distribution over predictors f, which depends on U and V but does not necessarily take the form of a neural network. The construction of this randomised proxy Q is central to our PAC-Bayes derived proof methods. The bounds hold uniformly over any choice of weight matrices, but for many choices the bounds obtained will be vacuous; what is interesting is that they are non-vacuous for SGD-derived solutions on some real-world datasets. U^n and V^n are matrices constructed using some subset $n < m$ of the data. Since we consider SGD-derived weights, we can leverage the empirical stability of this training method (through an idea introduced by Dziugaite et al., 2020) to construct U^n, V^n which are quite close to the final true SGD-derived weights U, V, essentially by training a prior on the n-sized subset in the same way.

Our experimental validation on datasets such as Binary-MNIST and Binary-Fashion MNIST demonstrates the efficacy of our approach – see Table 2.2. Notably, we achieve test errors and PAC-Bayesian bounds that underscore the practicality of our theoretical insights for real-life applications.

Table 2.2: Results for SHEL and GELU networks trained with SGD on MNIST, Fashion-MNIST, and binarised versions of the above.

	Dataset	Test Error	Bound
SHEL	Bin-M	0.037	0.286
SHEL	Bin-F	0.085	0.300
SHEL	MNIST	0.038	0.522
SHEL	Fashion	0.136	0.844
GELU	MNIST	0.036	0.317
GELU	Fashion	0.135	0.709

3 GENERALISATION-DRIVEN ALGORITHMS

Outline
In this chapter, we illustrate the second aspect of our manifesto, on the versatility of PAC-Bayes and the use of generalisation bounds as a training objective for eliciting novel machine learning algorithms.

The initial PAC-Bayesian bounds in the late 1990s and early 2000s mainly focused on the excess risk

$$\mathbb{E}_{X,Y}\ell(Y, \phi(X)) - \mathbb{E}_{X,Y}\ell(Y, \phi^\star(X)),$$

where $\phi^\star$ denotes the optimal predictor (or Bayes predictor: it is the conditional expectation of Y given X, which is naturally unknown). These bounds allow studying the minimax learning rate but are not practically usable. It is primarily during the 2010s that most "empirical" bounds on the error

$$\mathbb{E}_{X,Y}\ell(Y, \phi(X)) - \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i, \phi(X_i))$$

appear, and under certain assumptions (for example, if the prior and posterior are Gaussian, the Kullback-Leibler divergence admits a closed form), it becomes possible to calculate (possibly at the cost of numerical approximations, like Monte Carlo) the upper bound. As the PAC-Bayes bounds are generally valid for any pair (prior, posterior), the strategy of fixing the prior and optimising over the posterior progressively became central and led to numerical results demonstrating that the algorithms obtained by minimising a PAC-Bayes bound were effective. This is the strategy I pursued in a significant number of my works since 2018, notably [BG-Journal13]; [BG-Conf13]; [BG-Conf12]; [BG-Conf2]; [BG-Preprint4]; [BG-Conf10] (with neural networks) and [BG-Conf14]; [BG-Preprint12]; [BG-Conf3]; [BG-Conf23]; [BG-Conf11].

Consider the works [BG-Preprint12]; [BG-Conf23]; [BG-Conf11]. In [BG-Preprint12], we establish a PAC-Bayes generalisation bound for learning on a network of nodes without each node having to share its data with others—a typical case of federated learning, which occurs, for example, when several hospitals wish to collaborate to predict a quantity (*e.g.*, identifying a pathology) without necessarily sharing their data for privacy reasons and/or physical constraints. The PAC-Bayes bound is then used to effectively learn a global predictor, whose performances are similar or superior to the state of the art in several situations. In [BG-Conf23], we establish the first PAC-Bayes bounds with the Wasserstein distance: this is particularly crucial as these bounds can be used with atomic posteriors, *e.g.*, a Dirac mass on a particular predictor (such as the empirical risk minimiser, or the result of a stochastic gradient descent algorithm). This is not possible with the Kullback-Leibler divergence (which requires the prior and posterior to share the same support). We extended these results in the classical (batch) and online statistical settings, with numerical results showing predictive performances similar to the state of the art and non-trivial bound values. In [BG-Conf11], we demonstrate (theoretically and empirically) that the predictive performance of stochastic majority votes (where weights are sampled from a distribution—the PAC-Bayesian posterior) is superior to votes with deterministic weights.

We now briefly illustrate our contributions by sketching the main results from [BG-Conf11] (NeurIPS 2021) and [BG-Conf23] (NeurIPS 2023).

3.1 LEARNING STOCHASTIC MAJORITY VOTES

Our paper [BG-Conf11] is a joint work led by my postdoc Valentina Zantedeschi, with Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrad and Pascal Germain. In that paper, we proved a new PAC-Bayes bound for stochastic majority votes using Dirichlet distributions on the simplex defined by initial weak predictors.

We investigate a stochastic counterpart of majority votes over finite ensembles of classifiers, and study its generalisation properties. While our approach holds for arbitrary distributions, we instantiate it with Dirichlet distributions: this allows for a closed-form and differentiable expression for the expected risk, which then turns the generalization bound into a tractable training objective. The resulting stochastic majority vote learning algorithm inherits informative generalisation properties. It achieves state-of-the-art accuracy and benefits from (non-vacuous) tight generalization bounds, in a series of numerical experiments when compared to competing algorithms which also minimize PAC-Bayes objectives – both with uninformed (data-independent) and informed (data-dependent) priors.

By combining the outcomes of several predictors, ensemble methods (Dietterich, 2000) have been shown to provide models that are more accurate and more robust than each predictor taken singularly. The key to their success lies in harnessing the diversity of the set of predictors (Kuncheva, 2004). Among ensemble methods, weighted Majority Votes (MV) classifiers assign a score to each base classifier (*i.e.*, voter) and output the most common prediction, given by the weighted majority. When voters have known probabilities of making an error and make independent predictions, the optimal weighting is given by the so-called Naive Bayes rule (Berend and Kontorovich, 2015). However, in most situations these assumptions are not satisfied, giving rise to the need for techniques that estimate the optimal combination of voter predictions from the data.

Among them, PAC-Bayesian based methods are well-grounded approaches for optimizing the voter weighting. Indeed, PAC-Bayes theory provides not only bounds on the true error of a MV through generalization bounds but is also suited to derive theoretically grounded learning algorithms (Germain et al., 2009; Parrado-Hernández et al., 2012; Alquier et al., 2016). PAC-Bayesian guarantees do not stand for all hypotheses (i. e.are not expressed as a worst-case analysis) but stand in expectation over the hypothesis set. The prior brings some prior knowledge on the combination of predictors, and the posterior distribution is learned (adjusted) to lead to good generalization guarantees; the deviation between the prior and the posterior distributions plays a role in generalization guarantee and is usually captured by the Kullback-Leibler (KL) divergence. In their essence, PAC-Bayesian results do not bound directly the risk of the deterministic MV, but bound the expected risk of one (or several) base voters randomly drawn according to the weight distribution of the MV (Langford and Shawe-Taylor, 2002; Lacasse et al., 2006; Lacasse et al., 2010; Germain et al., 2015; Masegosa et al., 2020).

This randomization scheme leads to upper bounds on the true risk of the MV that are then used as a proxy to derive PAC-Bayesian generalization bounds. However, the obtained risk certificates are generally not tight, as they depend on irreducible constant factors, and when optimized they can lead to sub-optimal weightings. Indeed, by considering a random subset of base predictors, state-of-the-art methods do not fully leverage the diversity of the whole set of voters. This is especially problematic when the voters are weak, and learning to combine their predictions is critical for good performance.

OUR CONTRIBUTIONS. We consider the voter weighting associated to a MV as a realization of a distribution of voter weightings. We analyze with the PAC-Bayesian framework the expected risk of a MV drawn from the posterior distribution of MVs. The main difference with the literature is that we propose a stochastic MV, while previous works aim at studying randomized evaluations of the true risk of the deterministic MV. Doing so, we are able to derive tight empirical PAC-Bayesian bounds for our model directly on its expected risk. We further propose two approaches for optimizing the generalization bounds, hence learning the optimal posterior: the first optimizes an analytical and differentiable form of the empirical risk that can be derived when considering Dirichlet distributions; the second optimizes a Monte Carlo approximation of the expected risk and can be employed with any form of posterior. In our experiments, we first compare these two approaches, highlighting in which regimes one is preferable to the other. Finally, we assess our method's performance on real benchmarks with respect to the performance of PAC-Bayesian approaches also learning MV classifiers. These results indicate that our models enjoy generalization bounds that are consistently tight and non-vacuous both when studying ensembles of data-independent predictors and when studying ensembles of data-dependent ones.

Consider the data random variable (X, Y), taking values in $\mathcal{X}\times\mathcal{Y}$ with $\mathcal{X} \subseteq \mathbb{R}^d$ a d-dimensional representation space and $\mathcal{Y}$ the set of labels. We denote $\mathcal{P}$ the (unknown) data distribution of (X, Y). We define a set (dictionary) of base classifiers $D=\{h_j : \mathcal{X} \to \mathcal{Y}\}_{j=1}^M$. The weighted majority vote classifier is a convex combination of the base classifiers from D.

Formally, a MV is parameterized by a weight vector $\theta \in [0,1]^M$, such that $\sum_{j=1}^{M} \theta_j = 1$ hence lying in the (M-1)-simplex Δ^{M-1}, as follows:

$$f_\theta(x) = \operatorname*{arg\,max}_{y \in \mathcal{Y}} \sum_{j=1}^{M} \theta_j \ \mathbb{1}(h_j(x) = y),$$

where $\mathbb{1}(\cdot)$ is the indicator function. Let $W_\theta(X,Y)$ be the random variable corresponding to the total weight assigned to base classifiers that predict an incorrect label on (X,Y), that is

$$W_\theta(X,Y) = \sum_{j=1}^{M} \theta_j \mathbb{1}(h_j(X) \neq Y).$$

We refer to Figure 3.1 for a visualisation of the density of the random weights. In binary classification with $|\mathcal{Y}|$=2, the MV errs whenever $W_\theta(X,Y) \geqslant 0.5$ (Lacasse et al., 2010; Masegosa et al., 2020). Hence the true risk (with respect to 01-loss) of the MV classifier can be expressed as

$$R(f_\theta) = \mathbb{E}_{\mathcal{P}} \ \mathbb{1}(W_\theta(X,Y) \geqslant 0.5) = \mathbb{P}(W_\theta \geqslant 0.5). \tag{3.1}$$

Similarly, the empirical risk of f_θ on a n-sample S=$\{(x_i,y_i)\sim\mathcal{P}\}_{i=1}^{n}$ is given by

$$\hat{R}(f_\theta) = \sum_{i=1}^{n} \ \mathbb{1}(W_\theta(x_i,y_i) \geqslant 0.5).$$

Note that the results we introduce in the following are stated for binary classification, but are valid also in the multi-class setting ($|\mathcal{Y}|$>2).

One of the key results in our paper is an adaptation of Seeger's bound with informed priors, which is then turned into a training objective and effectively implemented. The dataset is split into two bits, one with m points and the other one with $n-m$.

Theorem 3.1 (Seeger's bound with informed priors). *Let $\pi_{\leqslant m}$ and $\rho_{\leqslant m}$ be the prior and posterior distributions depending on $[1:m]$, and $\pi_{>m}$ and $\rho_{>m}$ the prior and posterior distributions depending on $[n-m:n]$. For any p∈$(0,1)$ and $\delta \in (0,1)$ with probability at least $1-\delta$ we have*

$$\begin{aligned} &\mathrm{kl}\big(p\hat{R}(\rho_{>m}) + (1-p)\hat{R}(\rho_{\leqslant m}) \big\| pR_{\leqslant m}(\rho_{>m}) + (1-p)_{>m}(\rho_{\leqslant m})\big) \\ &\leqslant \frac{p \ \mathrm{KL}(\rho_{>m},\pi_{>m})}{m} + \frac{(1-p) \ \mathrm{KL}(\rho_{\leqslant m},\pi_{\leqslant m})}{n-m} + \frac{\ln \frac{4\sqrt{m(n-m)}}{\delta}}{n}, \end{aligned}$$

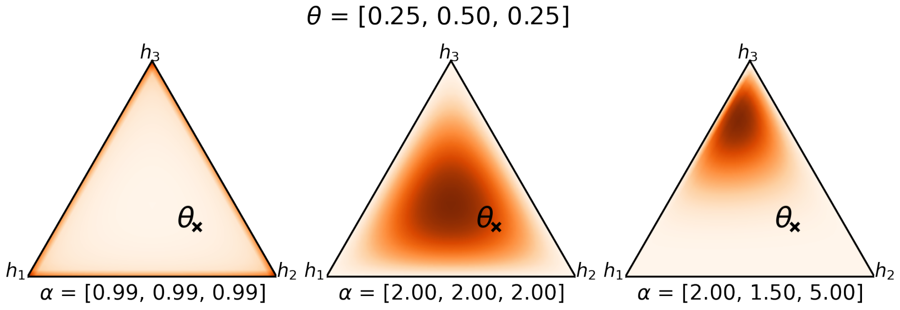

Figure 3.1: Visualization of the density measure $\rho : \Delta^2 \to \mathbb{R}_+$ taking the form of a Dirichlet distribution, with concentration parameters α. The darker the color, the higher $\rho(\theta)$. Each θ on the simplex corresponds to a majority vote classifier f_θ and has an associated probability depending on α.

with

$$R(\rho_{>m}) = \int_{\Theta_{>m}} R(f_\theta)\rho(d\theta),$$

$$R(\rho_{\leqslant m}) = \int_{\Theta_{\leqslant m}} R(f_\theta)\rho(d\theta),$$

$$\hat{R}_{\leqslant m}(\rho_{>m}) = \int_{\Theta_{>m}} \hat{R}(f_\theta)\rho(d\theta),$$

and

$$\hat{R}_{\leqslant m}(\rho_{>m}) = \int_{\Theta_{\leqslant m}} \hat{R}(f_\theta)\rho(d\theta).$$

3.2 LEARNING VIA WASSERSTEIN-BASED GENERALISATION BOUNDS

[**BG-Conf23**] is a joint work led by Paul Viallard, with Maxime Haddouche and Umut Şimşekli.

Despite its successes and unfailing surge of interest in recent years, a limitation of the PAC-Bayesian framework is that most bounds involve a Kullback-Leibler (KL) divergence term (or its variations), which might exhibit erratic behavior and fail to capture the underlying geometric structure of the learning problem – hence restricting its use in practical applications. As a remedy, recent studies have attempted to replace the KL divergence in the PAC-Bayesian bounds with the Wasserstein distance. Even though these bounds alleviated the aforementioned issues to a certain extent, they either hold in expectation, are for bounded losses, or are nontrivial to minimize. In this work, we prove novel Wasserstein distance-based PAC-Bayesian generalisation bounds for both batch learning with independent and identically distributed (*i.i.d.*) data, and online learning with potentially non-*i.i.d.* data. Contrary to previous art, our bounds are stronger in the sense that *(i)* they hold with high probability, *(ii)* they apply to unbounded (potentially heavy-tailed) losses, and *(iii)* they lead to optimizable training objectives. As a result we derive novel Wasserstein-based PAC-Bayesian learning algorithms and we illustrate their empirical advantage on a variety of experiments.

Typically, a *learning problem* is described by a tuple $(\mathcal{H}, \mathcal{Z}, \ell)$ consisting of a hypothesis (or predictor) space $\mathcal{H}$, a data space $\mathcal{Z}$, and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$. The goal is to estimate the *population risk* of a given hypothesis h, defined as $R_{\mathcal{D}}(h) = \mathbb{E}_{\mathbf{z} \sim \mathcal{D}}[\ell(h, \mathbf{z})]$, where $\mathcal{D}$ denotes the unknown *data distribution* over $\mathcal{Z}$. As $\mathcal{D}$ is not known, in practice, a hypothesis h is usually built by (approximately) minimising the *empirical risk*, given by $\hat{R}_{\mathcal{S}}(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, \mathbf{z}_i)$, where $\mathcal{S} = \{\mathbf{z}_i \in \mathcal{Z}\}_{i=1}^{m}$ is a dataset of m data points, independent and identically distributed (*i.i.d.*) from $\mathcal{D}$. We define the generalisation gap of a hypothesis h as $\hat{R}_{\mathcal{S}}(h) - R_{\mathcal{D}}(h)$.

While a plethora of techniques have been introduced, the PAC-Bayes framework has gained significant traction over the past two decades to provide non-vacuous generalisation guarantees for complex structures such as neural networks during the training phase (see Chapter 2). The bounds are also used to derive learning algorithms by minimising the right-hand side of a given bound. Beyond neural networks, the flexibility of PAC-Bayes learning makes it a useful toolbox to derive both theoretical results and practical algorithms in various learning fields such as reinforcement learning (Fard and Pineau, 2010), online learning (**[BG-Conf17]**), multi-armed bandits (Seldin et al., 2011; Seldin et al., 2012; Sakhi et al., 2023), meta-learning (Amit and Meir, 2018; Farid and Majumdar, 2021; Rothfuss et al., 2021; Rothfuss et al., 2022; Ding et al., 2021) to name but a few.

While PAC-Bayesian bounds remain nowadays of the utmost interest to explain generalisation in various learning problems they mostly rely on the KL divergence or variants which causes two main limitations: *(i)* as illustrated in the generative modeling literature, the KL divergence does not incorporate the underlying geometry or topology of the data space $\mathcal{Z}$, hence can behave in an erratic way (Arjovsky et al., 2017), *(ii)* the KL divergence and its variants require the posterior ρ to be absolutely continuous with respect to the prior π. However, recent studies

(Camuto et al., 2021) have shown that, in stochastic optimisation, the distribution of the iterates, which is the natural choice for the posterior, can converge to a *singular distribution*, which does not admit a density with respect to the Lebesgue measure. Moreover, the structure of the singularity (i. e., the *fractal dimension* of ρ) depends on the data sample $\mathcal{S}$ (Camuto et al., 2021). Hence, in such a case, it would not be possible to find a suitable prior π that can dominate ρ for almost every $\mathcal{S} \sim \mathcal{D}^m$, which will trivially make $\mathrm{KL}(\rho \| \pi) = +\infty$ and the generalisation bound vacuous.

Some works have focused on replacing the Kullback-Leibler divergence with more general divergences in PAC-Bayes (**[BG-Journal6]**; Ohnishi and Honorio, 2021; **[BG-Preprint7]**), although the problems arising from the presence of the KL divergence in the generalisation bounds are actually not specific to PAC-Bayes: information-theoretic bounds (Xu and Raginsky, 2017; Russo and Zou, 2020) also suffer from similar issues as they are based on a mutual information term, which is the KL divergence between two distributions. In this context, as a remedy to these issues introduced by the KL divergence, Zhang et al., 2018; Wang et al., 2019; Gálvez et al., 2021; Lugosi and Neu, 2022 proved analogous bounds that are based on the *Wasserstein distance*, which arises from the theory of optimal transport (Monge, 1781). As the Wasserstein distance inherits the underlying geometry of the data space and does not require absolute continuity, it circumvents the problems introduced by the KL divergence. Yet, these bounds hold only in expectation, i. e., none of these bounds is holding with high probability over the random choice of the learning sample $\mathcal{S} \sim \mathcal{D}^m$.

The recent works Amit et al., 2022; Chee and Loustau, 2021 incorporated Wasserstein distances as a complexity measure and proved generalisation bounds based on the Wasserstein distance. More precisely, Amit et al., 2022 proved a high-probability generic PAC-Bayesian bound for bounded losses depending on an integral probability metric (Müller, 1997), which contains the Wasserstein distance as a special case. On the other hand, Chee and Loustau, 2021 exploited PAC-Bayesian tools to obtain learning strategies with their associated regret bounds based on the Wasserstein distance for the *online learning* setting while requiring a finite hypothesis space and do not deal with generalisation.

CONTRIBUTIONS. The theoretical understanding of generalisation bounds based on the Wasserstein distance is still limited. The aim of this work is not only to prove generalisation bounds (for different learning settings) based on the optimal transport theory but also to propose new learning algorithms derived from our theoretical results.

(a) Using the supermartingale toolbox introduced in [**BG-Journal21**]; Chugg et al., 2023, we prove novel PAC-Bayesian bounds based on the Wasserstein distance for *i.i.d.* data. While Amit et al., 2022 proposed a McAllester-like bound for bounded losses, we propose a Catoni-like bound (see e. g., Alquier et al., 2016, Theorem 4.1) valid for heavy-tailed losses with bounded order 2 moments. This assumption is less restrictive than assuming subgaussian or bounded losses, which are at the core of many PAC-Bayes results. This assumption also covers distributions beyond subgaussian or subexponential ones (e. g., gamma distributions with a scale smaller than 1, which have an infinite exponential moment).

(b) We provide the first generalisation bounds based on Wasserstein distances for the online PAC-Bayes framework of [**BG-Conf17**]. Our results are, again, Catoni-like bounds and hold for heavy-tailed losses with bounded order 2 moments. Previous work (Chee and Loustau, 2021) already provided online strategies mixing PAC-Bayes and Wasserstein distances. However, their contributions focus on the best deterministic strategy, regularised by a Wasserstein distance, with respect to the deterministic notion of regret. Our results differ significantly as we provide the best-regularised strategy (still in the sense of a Wasserstein term) with respect to the notion of generalisation, which is new.

(c) As our bounds are linear with respect to Wasserstein terms (contrary to those of Amit et al., 2022), they are well suited for optimisation procedures. Thus, we propose the first PAC-Bayesian learning algorithms based on Wasserstein distances instead of KL divergences. For the first time, we design PAC-Bayes algorithms able to output deterministic predictors (instead of distributions over all $\mathcal{H}$) designed from deterministic priors. This is due to the ability of the Wasserstein distance to measure the discrepancy between Dirac distributions. We then instantiate those algorithms on various datasets, paving the way to promising practical developments of PAC-Bayes learning.

To sum up, we highlight two benefits of PAC-Bayes learning with Wasserstein distance. First, it ships with sound theoretical results exploiting the geometry of the predictor space, holding for heavy-tailed losses. Such a weak assumption on the loss extends the usefulness of

PAC-Bayes with Wasserstein distances to a wide range of learning problems, encompassing bounded losses. Second, it allows us to consider deterministic algorithms (i. e., sampling from Dirac measures) designed with respect to the notion of generalisation: we showcase their performance in our experiments.

NOTATION. We consider a predictor space $\mathcal{H}$ equipped with a distance d and a σ-algebra $\Sigma_{\mathcal{H}}$, a data space $\mathcal{Z}$, and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$. In this work, we consider Lipschitz functions with respect to d. We also associate a filtration $(\mathcal{F}_i)_{i \geqslant 1}$ adapted to our data $(\mathbf{z}_i)_{i=1,\dots,m}$, and we assume that the dataset $\mathcal{S}$ follows the distribution $\mathcal{D}$. In PAC-Bayes learning, we construct a data-driven posterior distribution $\rho \in \mathcal{M}(\mathcal{H})$ with respect to a prior distribution π.

For all i, we denote by $\mathbb{E}_i[\cdot]$ the conditional expectation $\mathbb{E}[\,\cdot \mid \mathcal{F}_i]$. In this work, we consider data-dependent priors. A stochastic kernel is a mapping $\pi : \cup_{m=1}^{\infty} \mathcal{Z}^m \times \Sigma_{\mathcal{H}} \to [0, 1]$ where *(i)* for any $B \in \Sigma_{\mathcal{H}}$, the function $\mathcal{S} \mapsto \pi(\mathcal{S}, B)$ is measurable, *(ii)* for any dataset $\mathcal{S}$, the function $B \mapsto \pi(\mathcal{S}, B)$ is a probability measure over $\mathcal{H}$.

In what follows, we consider two different learning paradigms: *batch learning*, where the dataset is directly available, and *online learning*, where data streams arrive sequentially.

(i) Batch setting. We assume the dataset $\mathcal{S}$ to be *i.i.d.*, so there exists a distribution $\mathcal{D}$ over $\mathcal{Z}$ such that $\mathcal{D} = \mathcal{D}^m$. We then define, for a given $h \in \mathcal{H}$, the *risk* to be $R_{\mathcal{D}} := \mathbb{E}_{\mathbf{z} \sim \mathcal{D}}[\ell(h, \mathbf{z})]$ and its empirical counterpart $\hat{R}_{\mathcal{S}} := \frac{1}{m} \sum_{i=1}^{m} \ell(h, \mathbf{z}_i)$. We aim to bound the *expected generalisation gap* defined by $\mathbb{E}_{h \sim \rho}[R_{\mathcal{D}}(h) - \hat{R}_{\mathcal{S}}(h)]$. We assume that the dataset $\mathcal{S}$ is split into K disjoint sets $\mathcal{S}_1, \dots, \mathcal{S}_K$. We consider K stochastic kernels $\pi_1, \dots, \pi_K$ such that for any $\mathcal{S}$, the distribution $\pi_i(\mathcal{S}, .)$ *does not* depend on $\mathcal{S}_i$.

(ii) Online setting. We adapt the online PAC-Bayes framework of [BG-Conf17]. We assume that we have access to a stream of data $\mathcal{S} = (\mathbf{z}_i)_{i=1,\dots,m}$, arriving sequentially, with no assumption on $\mathcal{D}$. In online PAC-Bayes, the goal is to define a posterior sequence $(\rho_i)_{i \geqslant 1}$ from a prior sequence $(\pi_i)_{i \geqslant 1}$, which can be data-dependent. We define an *online predictive sequence* $(\pi_i)_{i=1\cdots m}$ satisfying: *(i)* for all i and dataset $\mathcal{S}$, the distribution $\pi_i(S, .)$ is $\mathcal{F}_{i-1}$ measurable and *(ii)* there exists π_0 such that for all $i \geqslant 1$, we have $\pi_i(S, .) \gg \pi_0$. This last condition covers, in particular, the case where $\mathcal{H}$ is an Euclidean space and for any i, the distribution $\pi_{i,\mathcal{S}}$ is a Dirac mass. All of those measures are uniformly continuous with respect to any Gaussian distribution.

WASSERSTEIN DISTANCE. We focus on the Wasserstein distance of order 1 (Earth Mover's distance) introduced by Kantorovitch, 1960 in the optimal transport literature. Given a distance $d : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ and a Polish space $(\mathcal{A}, d)$, for any probability measures α and B_Q on $\mathcal{A}$, the Wasserstein distance is defined by

$$W(\alpha, B_Q) := \inf_{\gamma \in \Gamma(\alpha, B_Q)} \left\{ \mathop{\mathbb{E}}_{(a,b)\sim\gamma} d(a,b) \right\},$$

where $\Gamma(\alpha, B_Q)$ is the set of joint probability measures $\gamma \in \mathcal{M}(\mathcal{A}^2)$ such that the marginals are α and B_Q. The Wasserstein distance aims to find the probability measure $\gamma \in \mathcal{M}(\mathcal{A}^2)$ minimising the expected cost $\mathbb{E}_{(a,b)\sim\gamma} d(a,b)$. We refer the reader to Villani (2009) and Peyré and Cuturi (2019) for an introduction to optimal transport.

We present novel high-probability PAC-Bayesian bounds involving Wasserstein distances instead of the classical Kullback-Leibler divergence. Our bounds hold for heavy-tailed losses (instead of classical subgaussian and subexponential assumptions), extending the remits of Amit et al., 2022, Theorem 11. We exploit the supermartingale toolbox, recently introduced in PAC-Bayes framework by **[BG-Journal21]**; Chugg et al., 2023; Jang et al., 2023, to derive bounds for both batch learning and online learning. Here is one of our main results, holding for heavy-tailed non-negative losses.

Theorem 3.2. *We assume our loss ℓ to be non-negative and L-Lipschitz. We also assume that, for any $1 \leqslant i \leqslant K$, for any dataset$\mathcal{S}$, we have*

$$\mathop{\mathbb{E}}_{h\sim\pi_i(.,\mathcal{S}), z\sim\mathcal{D}} \left[\ell(h,z)^2\right] \leqslant 1,$$

bounded order 2 moments for priors. Then, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, the following holds for the distributions $\pi_{i,\mathcal{S}} := \pi_i(\mathcal{S}, .)$ and for any $\rho \in \mathcal{M}(\mathcal{H})$:

$$\mathop{\mathbb{E}}_{h\sim\rho}\left[R_{\mathcal{D}}(h) - \hat{R}_{\mathcal{S}}(h)\right] \leqslant \sum_{i=1}^{K} \frac{2|\mathcal{S}_i|L}{m} W(\rho, \pi_{i,\mathcal{S}}) + \sum_{i=1}^{K} \sqrt{\frac{2|\mathcal{S}_i| \ln\frac{K}{\delta}}{m^2}},$$

where $\pi_{i,\mathcal{S}}$ does not *depend on $\mathcal{S}_i$.*

Note that when the loss function takes values in $[0,1]$, an alternative strategy allows tightening the last term of the bound by a factor $\frac{1}{2}$.

TOWARDS NEW PAC-BAYESIAN ALGORITHMS. From that bound, we derive a new PAC-Bayesian algorithm for Lipschitz non-negative losses:

$$\operatorname*{arg\,min}_{\rho\in\mathcal{M}(\mathcal{H})} \mathop{\mathbb{E}}_{h\sim\rho}\left[\hat{R}_{\mathcal{S}}(h)\right] + \sum_{i=1}^{K} \frac{2|\mathcal{S}_i|L}{m} W(\rho, \pi_{i,\mathcal{S}}).$$

This uses Wasserstein distances as regularisers and allows the use of multiple priors.

Another key result from this paper is a bound for the online learning setting.

Theorem 3.3. *We assume our loss ℓ to be non-negative and L-Lipschitz. We also assume that, for any $i, \mathcal{S}$, $\mathbb{E}_{h\sim\pi_i(.,\mathcal{S})}\left[\mathbb{E}_{i-1}[\ell(h, \mathbf{z}_i)^2]\right] \leqslant 1$ (*bounded conditional order 2 moments for priors*). Then, for any $\delta \in (0, 1]$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, any online predictive sequence (used as priors) $(\pi_i)_{i\geqslant 1}$, we have with probability at least $1-\delta$ over the sample $S \sim \mathcal{D}$, the following, holding for the data-dependent measures $\pi_{i,\mathcal{S}} := \pi_i(S, .)$ and any posterior sequence $(\rho_i)_{i\geqslant 1}$:*

$$\frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h_i\sim\rho_i}\left[\mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] - \ell(h_i, \mathbf{z}_i)\right]$$
$$\leqslant \frac{2L}{m}\sum_{i=1}^{m} W(\rho_i, \pi_{i,\mathcal{S}}) + \sqrt{\frac{2\ln\left(\frac{1}{\delta}\right)}{m}}.$$

4 GENERALISATION IN NEW FRAMEWORKS

OUTLINE
In this chapter, we illustrate how PAC-Bayes allows for generalisation insights in models or frameworks in which no previous results were available.

PAC-Bayesian learning was long confined to a very restrictive statistical framework (bounded loss function, i.i.d. data, supervised learning, etc.) that did not match the reality of contemporary learning problems (where data is noisy, incomplete, online, heavy-tailed, unsupervised or weakly supervised, etc.), which hindered its adoption outside the (relatively small) community of theoretical machine learning researchers. The gap between theory and practice is largely explained by the mathematical arguments deployed: for example, one of the essential mechanisms of PAC-Bayesian generalisation bound proofs is a concentration argument of the empirical error $\frac{1}{n}\sum_{i=1}^{n}\ell(Y_i, \phi(X_i))$ towards its expectation $\mathbb{E}_{X,Y}\ell(Y, \phi(X))$. However, this argument requires that the learning sample $\{(X_i, Y_i)\}_{i=1}^{n}$ be indeed i.i.d. It is possible, however, to circumvent this limitation, as we did in **[BG-Journal6]** with a different proof based on Hölder's inequality. A significant part of my efforts has been devoted to extending the PAC-Bayesian theory in all directions, which can be broadly summarised into four major categories:

Rate Improving the **rate** of convergence or the potential constants appearing in the bound,

Div. Replacing the **Kullback-Leibler divergence** with other divergence measures,

Ass. Relaxing the **assumptions** necessary without altering the terms of the bound,

New Demonstrating generalisation bounds for **new learning frameworks**.

I illustrate these four types of contributions with four articles [BG-Journal6]; [BG-Journal21]; [BG-Conf3]; [BG-Conf8] among my related works [BG-Conf3]; [BG-Conf21]; [BG-Preprint7]; [BG-Preprint10]; [BG-Conf23]; [BG-Preprint16]; [BG-Journal6]; [BG-Journal14]; [BG-Journal15]; [BG-Journal21]; [BG-Journal9]; [BG-Journal7]; [BG-Journal2]; [BG-Conf17]; [BG-Conf24]; [BG-Preprint2]; [BG-Preprint1]; [BG-Conf3]; [BG-Preprint14]; [BG-Preprint5]; [BG-Conf7]; [BG-Conf13]; [BG-Preprint6]; [BG-Journal20]; [BG-Conf16]; [BG-Conf8]; [BG-Journal5]; [BG-Conf4].

Rate. In [BG-Conf3], we demonstrate that it is possible to achieve fast rates in $\mathcal{O}(n)$ (rather than $\mathcal{O}(\sqrt{n})$), which is a significant advancement for attesting to the generalisation of algorithms when large samples are available. The central argument is a new concentration inequality (which we named *unexpected Bernstein*).

Div. In [BG-Journal6], we proposed the first PAC-Bayes generalisation bound with f-divergences, a class significantly broader than what was previously used. The vast majority of available bounds use the Kullback-Leibler divergence, which is a specific case of f-divergence (with the choice $f\colon x \mapsto x \log x$).

Ass. In [BG-Journal21], we demonstrate that it is possible to drop the assumption that the loss function is bounded (as assumed in the vast majority of the literature, even though a simple and popular loss such as the squared loss does not satisfy this). The key idea is a new flavour of Markov inequality, applied to supermartingales.

New. In [BG-Conf8], we demonstrate the first PAC-Bayes generalisation bounds for contrastive learning, which is widely used in industry but for which no theoretical generalisation result existed before 2019.

We now briefly illustrate our contributions by sketching the main results from [BG-Conf8] (UAI 2020) and [BG-Conf16] (AISTATS 2022).

The paper **[BG-Conf8]** is a joint work with Pascal Germain and Kento Nozawa, on contrastive unsupervised representation learning (CURL).

Contrastive unsupervised representation learning (CURL) is the state-of-the-art technique to learn representations (as a set of features) from unlabelled data. While CURL has collected several empirical successes recently, theoretical understanding of its performance was still missing. Arora et al. (2019) provided the first generalisation bounds for CURL, relying on a Rademacher complexity. We extend their framework to the flexible PAC-Bayes setting, allowing us to deal with the non-iid setting. We present PAC-Bayesian generalisation bounds for CURL, which are then used to derive a new representation learning algorithm. Numerical experiments on real-life datasets illustrate that our algorithm achieves competitive accuracy, and yields non-vacuous generalisation bounds.

Arora et al. (2019) introduced the first theoretical results on CURL, using Rademacher complexity. In a nutshell, for any predictor f and $\widehat{f}$ an ERM, w.p. $\geqslant 1-\delta$,

$$\mathrm{Loss}_{\mathrm{sup}}(\widehat{f}) \leqslant C_1 \mathrm{Loss}_{\mathrm{uns}}(f) + C_2 \left(\frac{\mathrm{Rad}}{m} + \sqrt{\frac{\log(1/\delta)}{m}} \right).$$

We proposed a PAC-Bayes generalisation which improves on their results by removing the iid assumption, and by deriving a SOTA learning algorithm. For any prior P, any posterior Q, any $\lambda > 0$, w.p. $\geqslant 1-\delta$

$$\mathrm{Loss}_{\mathrm{sup}}(Q) \leqslant C \left(\frac{1 - \exp\left(-\lambda \widehat{\mathrm{Loss}}_{\mathrm{uns}}(Q) - \frac{\mathrm{KL}(Q,P)+\log(1/\delta)}{m} \right)}{1-\exp(-\lambda)} \right).$$

Unsupervised representation learning (Bengio et al., 2013) aims at extracting features representation from an unlabelled dataset for downstream tasks such as classification and clustering (see Mikolov et al., 2013; Noroozi and Favaro, 2016; Zhang et al., 2016; Caron et al., 2018; Devlin et al., 2019). An unsupervised representation learning model is typically learnt by solving a pretext task without supervised information. Trained model work as a feature extractor for supervised tasks.

In unsupervised representation learning, contrastive loss is a widely used objective function class. Contrastive loss uses two types of data pair, namely, similar pair and dissimilar pair. Their similarity is defined without label information of a supervised task. For example, in word representation learning, Mikolov et al. (2013) define a similar pair as co-occurrence words in the same context, while dissimilar pairs are randomly sampled from a fixed distribution. Intuitively, by minimising a contrastive loss, similar data samples are mapped to similar representations in feature space in terms of some underlying metric (as the inner product), and dissimilar samples are not mapped to similar representations.

Contrastive unsupervised representation learning improves the performance of supervised models in practice, and has attracted a lot of research interest lately (see Chen et al., 2020, and references therein), although usage is still quite far ahead of theoretical understanding. Arora et al. (2019) introduced a theoretical framework for contrastive unsupervised representation learning and derived the first generalisation bounds for CURL. In parallel, PAC-Bayes is emerging as a principled framework to understand and quantify the generalisation ability of many machine learning algorithms.

OUR CONTRIBUTIONS. We extend the framework introduced by Arora et al. (2019), by adopting a PAC-Bayes approach to contrastive unsupervised representation learning. We derive the first PAC-Bayes generalisation bounds for CURL, both in iid and non-iid settings. Our bounds are then used to derive new CURL algorithms, for which we provide a complete implementation. The paper closes with numerical experiments on two real-life datasets (`CIFAR-100` and `AUSLAN`) showing that our bounds are non-vacuous in the iid setting.

Inputs are denoted $\mathbf{x} \in X = \mathbb{R}^{d_0}$, and outputs are denoted $y \in Y$, where Y is a discrete and finite set. The *representation* is learnt from a (large) unlabelled dataset $U = \{\mathbf{z}_i\}_{i=1}^m$, where $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_{i1}^-, \dots, \mathbf{x}_{ik}^-)$ is a tuple of $k+2$ elements; $\mathbf{x}_i$ being *similar* to $\mathbf{x}_i^+$ and *dissimilar* to every element of the *negative sample set* $\{\mathbf{x}_{ij}^-\}_{j=1}^k$. The predictor is learnt from a labelled dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$.

In the following, we present the contrastive framework proposed by Arora et al. (2019) in a simplified scenario in order to highlight the key ideas, where the supervised prediction task is binary and the negative sample sets for unsupervised representation learning contain one element. Thus, we choose the label set to be $Y = \{-1, 1\}$, and the unsupervised set U contains triplets $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)$. The extension to a more generic setting (for $|Y| > 2$ and $k > 1$) bears no particular difficulty. It is important to note at this stage that both U and S are assumed to be

iid (independent, identically distributed) collections, as also assumed by Arora et al. (2019).

Latent classes and data distributions. The main assumption is the existence of a set of *latent classes* $\mathcal{C}$. Let us denote by ρ a probability distribution over $\mathcal{C}$. Moreover, with each class $c \in \mathcal{C}$, comes a class distribution $\mathcal{D}_c$ over the input space X. A similar pair $(\mathbf{x}, \mathbf{x}^+)$ is such that both $\mathbf{x}$ and $\mathbf{x}^+$ are generated by the same class distribution. Note that an input $\mathbf{x}$ possibly belongs to multiple classes: take the example of $\mathbf{x}$ being an image and $\mathcal{C}$ a set of latent classes including "the image depicts a dog" and "the image depicts a cat" (both classes are not mutually exclusive).

Definition 4.1. *Let ρ^2 be a shorthand for the joint distribution (ρ, ρ). We refer to the* unsupervised data distribution $\mathcal{U}$ *as the process that generates an unlabelled sample $\mathbf{z} = (\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-)$ according to the following scheme:*
1. Draw two latent classes $(c^+, c^-) \sim \rho^2$;
2. Draw two similar samples $(\mathbf{x}, \mathbf{x}^+) \sim (\mathcal{D}_{c^+})^2$;
3. Draw a negative sample $\mathbf{x}^- \sim \mathcal{D}_{c^-}$.

The labelled sample S is obtained by fixing two classes $c^{\pm} = \{c^-, c^+\} \in \mathcal{C}^2$ (from now on, the shorthand notation $c^{\pm}$ is used to refer to a pair of latent classes). Each class is then mapped on a label of Y. We fix $y_{c^-} = -1$ and $y_{c^+} = 1$; Thus we can write $Y = \{y_{c^-}, y_{c^+}\}$ as an ordered set. The label is obtained from the latent class distribution restricted to two values $\rho_{c^{\pm}}$:

$$\rho_{c^{\pm}}(c^-) = \frac{\rho(c^-)}{\rho(c^-)+\rho(c^+)}, \ \rho_{c^{\pm}}(c^+) = \frac{\rho(c^+)}{\rho(c^-)+\rho(c^+)} .$$

Definition 4.2. *We refer to the* supervised data distribution $\mathcal{S}$ *as the process that generates a labelled sample $(\mathbf{x}, y)$ according to the following scheme:*
1. Draw a class $c \sim \rho_{c^{\pm}}$ and set label $y = y_c$;
2. Draw a sample $\mathbf{x} \sim \mathcal{D}_c$.

Loss function. The learning process is divided in two sequential steps, the unsupervised and supervised steps. In order to relate these two steps, the key is to express them in terms of a common convex loss function $\ell : \mathbb{R} \rightarrow \mathbb{R}^+$. Typical choices are

$$\ell_{\log}(v) = \log_2(1 + e^{-v}) \, , \text{(logistic loss)}$$
$$\ell_{\text{hinge}}(v) = \max\{0, 1 - v\} \, , \text{(hinge loss)}$$

where the loss argument v expresses a notion of *margin*.

In the first step, an unsupervised representation learning algorithm produces a feature map $\mathbf{f} : X \to \mathbb{R}^d$. The *contrastive loss* associated with $\mathbf{f}$ is defined as

$$
\begin{aligned}
L_{\mathrm{un}}(\mathbf{f}) &= \mathop{\mathbf{E}}_{(c^+,c^-)\sim\rho^2} \mathop{\mathbf{E}}_{\substack{(\mathbf{x},\mathbf{x}^+)\sim\mathcal{D}^2_{c^+}\\ \mathbf{x}^-\sim\mathcal{D}_{c^-}}} \ell\Big(\mathbf{f}(\mathbf{x})\cdot\big[\mathbf{f}(\mathbf{x}^+)-\mathbf{f}(\mathbf{x}^-)\big]\Big)\\
&= \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{x}^+,\mathbf{x}^-)\sim\mathcal{U}} \ell\Big(\mathbf{f}(\mathbf{x})\cdot\big[\mathbf{f}(\mathbf{x}^+)-\mathbf{f}(\mathbf{x}^-)\big]\Big)\,.
\end{aligned}
$$

More precisely, from the unsupervised training dataset

$$
U = \{(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)\}_{i=1}^m \sim \mathcal{U}^m\,,
$$

we are interested in learning the feature map $\mathbf{f}$ that minimises the following empirical contrastive loss:

$$
\widehat{L}_{\mathrm{un}}(\mathbf{f}) = \frac{1}{m}\sum_{i=1}^{m} \ell\Big(\mathbf{f}(\mathbf{x}_i)\cdot\big[\mathbf{f}(\mathbf{x}_i^+)-\mathbf{f}(\mathbf{x}_i^-)\big]\Big)\,. \tag{4.1}
$$

In the second step, a supervised learning algorithm is given the *mapped* dataset $\widehat{S}=\{(\hat{\mathbf{x}}_i, y_i)\}_{i=1}^n$, with $\hat{\mathbf{x}}_i = \mathbf{f}(\mathbf{x}_i)$, and returns a predictor $g : \mathbb{R}^d \to \mathbb{R}$. For a fixed pair $c^\pm = \{c^-, c^+\}$, the predicted label on an input $\mathbf{x}$ is then obtained from $\hat{y} = \mathrm{sgn}[g(\hat{\mathbf{x}})]$ (recall that $Y = \{-1, 1\}$), and we aim to minimise the supervised loss

$$
\begin{aligned}
L_{\mathrm{sup}}(g\circ\mathbf{f}) &= \mathop{\mathbf{E}}_{c\sim\rho_{c^\pm}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_c} \ell\Big(y_c\, g(\mathbf{f}(\mathbf{x}))\Big)\\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \ell\Big(y\, g(\mathbf{f}(\mathbf{x}))\Big)\,.
\end{aligned}
$$

Given a labelled dataset $S \sim \mathcal{S}^n$, the empirical counterpart of the above supervised loss is

$$
\widehat{L}_{\mathrm{sup}}(g\circ\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(y_i\, g(\mathbf{f}(\mathbf{x}_i))\Big)\,.
$$

Mean classifier. Following Arora et al. (2019), we study the mean classifier defined by the linear function

$$
g_{c^\pm}(\hat{\mathbf{x}}) = \mathbf{w}_{c^\pm}\cdot\hat{\mathbf{x}}\,,
$$

where $\mathbf{w}_{c^\pm} = \boldsymbol{\mu}_{c^+} - \boldsymbol{\mu}_{c^-}$, and $\boldsymbol{\mu}_c = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_c}\mathbf{f}(\mathbf{x})$. Then, the *supervised average loss* of the mean classifier is the expected loss on a dataset whose pair of labels is sampled from the latent class distribution ρ.

$$
L^{\mu}_{\mathrm{sup}}(\mathbf{f}) = \mathop{\mathbf{E}}_{c^\pm\sim\rho^2_{\mathrm{w/o}}} L_{\mathrm{sup}}(g_{c^\pm}\circ\mathbf{f})\,, \tag{4.2}
$$

with $\rho^2_{\mathrm{w/o}}$ being a shorthand notation for the sampling *without replacement* of two classes among $\mathcal{C}$. Indeed, we want positive and negative samples that are generated by distinct latent class distributions, i. e., $c^- \neq c^+$.

A major contribution of the framework introduced by Arora et al. (2019) is that it rigorously links the unsupervised representation learning task and the subsequent prediction task: it provides generalisation guarantees on the supervised average loss of Equation (4.2) in terms of the empirical contrastive loss in Equation (4.1).

Theorem 4.1 (Arora et al., 2019, Theorem 4.1)**.** *Let* $B \in \mathbb{R}_+$ *be such that* $\|\mathbf{f}(\cdot)\| \leqslant B$*, with probability* $1-\delta$ *over training samples* $U \sim \mathcal{U}^m$, $\forall \mathbf{f} \in \mathcal{F}$

$$L^{\mu}_{\sup}(\widehat{\mathbf{f}}) \leqslant \frac{1}{1-\tau}\left(L_{\mathrm{un}}(\mathbf{f})-\tau\right)+\frac{1}{1-\tau}\,\mathcal{O}\left(B\frac{\mathcal{R}_U(\mathcal{F})}{m}+B^2\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right),$$

where $\hat{\mathbf{f}} = \underset{\mathbf{f}\in\mathcal{F}}{\arg\min}\, \widehat{L}_{\mathrm{un}}(\mathbf{f})\,.$

The bound focuses on a class of feature map functions $\mathcal{F}$ through its empirical Rademacher complexity on a training dataset U, defined by

$$\mathcal{R}_U(\mathcal{F}) = \underset{\boldsymbol{\sigma}\sim\{\pm 1\}^{3dm}}{\mathbf{E}}\left(\sup_{\mathbf{f}\in\mathcal{F}}\left[\boldsymbol{\sigma}\cdot\mathbf{f}_{|U}\right]\right),$$

where $\mathbf{f}_{|U} = \mathrm{vec}(\{\mathbf{f}(\mathbf{x}_i), \mathbf{f}(\mathbf{x}_i^+), \mathbf{f}(\mathbf{x}_i^-)\}_{i=1}^m) \in \mathbb{R}^{3dm}$ is the concatenation of all feature mapping given by $\mathbf{f}$ on U, and $\boldsymbol{\sigma}\sim\{\pm1\}^{3dm}$ denotes the uniformly sampled Rademacher variables over that "representation" space. We proved the following PAC-Bayesian doppelgänger of that result.

Theorem 4.2. *Let* $B \in \mathbb{R}_+$ *such that* $\|\mathbf{f}(\cdot)\| \leqslant B$ *for all* $\mathbf{f} \in \mathcal{F}$*. Given* $\lambda > 0$ *and a prior* $\mathcal{P}$ *over* $\mathcal{F}$*, with probability at least* $1-\delta$ *over training samples* $U \sim \mathcal{U}^m$, $\forall \mathcal{Q}$ *over* $\mathcal{F}$,

$$L^{\mu}_{\sup}(\mathcal{Q}) \leqslant \frac{1}{1-\tau}\left(B_\ell \frac{1-\exp\left(-\frac{\lambda}{B_\ell}\widehat{L}_{\mathrm{un}}(\mathcal{Q}) - \frac{\mathrm{KL}(\mathcal{Q}\|\mathcal{P})+\ln\frac{1}{\delta}}{m}\right)}{1-\exp(-\lambda)} - \tau\right).$$

An interesting byproduct of Arora et al. (2019)'s approach is that the proof of the main bound (Theorem 4.1) is modular: we mean that in the proof of Theorem 4.2, instead of plugging in Catoni's bound, we can use any relevant bound. We therefore leverage the work of Alquier and Guedj (2018) who proved a PAC-Bayes generalisation bound which no longer needs to assume that data are iid, and even holds when the data-generating distribution is heavy-tailed. We can therefore cast our results onto the non-iid setting. We believe removing the iid assumption is especially relevant for contrastive unsupervised learning, as we deal with triplets of data points governed by a relational causal link (similar and dissimilar examples). In fact, several contrastive representation learning algorithms violate the iid assumption (Goroshin et al., 2015; Logeswaran and Lee, 2018).

Alquier and Guedj (2018)'s framework generalises the Kullback-Leibler divergence in the PAC-Bayes bound with the class of f-divergences (see Csiszár and Shields, 2004, for an introduction). Given a convex function f such that $f(1) = 0$, the f-divergence between two probability distributions is given by

$$D_f(\mathcal{P}\|\mathcal{Q}) = \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} f\left(\frac{\mathcal{P}(h)}{\mathcal{Q}(h)}\right).$$

Moreover, PAC-Bayes provides bounds on the expected loss of the predictors under the distribution $\mathcal{Q}$. We now present the classical supervised setup, where the zero-one loss is used.[1] We refer to this loss as the classification risk, denoted by $r(y,\hat{y}) = \mathbb{1}[y\,\hat{y} < 0]$. Given a data-generating distribution $\mathcal{S}$ on $X \times Y$, the expected $\mathcal{Q}$-risk is

$$R(\mathcal{Q}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{S}}\ \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} r(y, h(\mathbf{x}))\,,$$

and the empirical counterpart, i. e., the $\mathcal{Q}$-weighted empirical risk on a training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \sim \mathcal{S}^n$, is given by

$$\widehat{R}(\mathcal{Q}) = \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{h\sim\mathcal{Q}} r(y_i, h(\mathbf{x}_i))\,.$$

Theorem 4.3. *Given $p > 1, q = \frac{p}{p-1}$ and a prior $\mathcal{P}$ over $\mathcal{F}$, with probability at least $1-\delta, \forall\mathcal{Q}$ over $\mathcal{F}$,*

$$L^{\mu}_{\text{sup}}(\mathcal{Q}) \leqslant \frac{1}{1-\tau}\left(\widehat{L}_{\text{un}}(\mathcal{Q}) - \tau\right) + \frac{1}{1-\tau}\left(\frac{\mathcal{M}_q}{\delta}\right)^{\frac{1}{q}}\left(D_{\phi_p - 1}(\mathcal{Q}\|\mathcal{P}) + 1\right)^{\frac{1}{p}},$$

where $\mathcal{M}_q = \mathbb{E}_{\mathbf{f}\sim\mathcal{P}}\,\mathbb{E}_{\mathcal{U}\sim\mathcal{U}^m}(|L_{\text{un}}(\mathbf{f}) - \widehat{L}_{\text{un}}(\mathbf{f})|^q)$ and $\phi_p(x) = x^p$.

Up to our knowledge, Theorem 4.3 is the first generalisation bound for contrastive unsupervised representation learning that holds without the iid assumption, therefore extending the framework introduced by Arora et al. (2019) in a non-trivial and promising direction. Note that Theorem 4.3 does not require iid assumption for both unsupervised and supervised steps.

[1] Classical PAC-Bayes analyses consider the supervised learning setting, but non-supervised learning approaches exist (e. g., Seldin and Tishby, 2010; Higgs and Shawe-Taylor, 2010; Germain et al., 2013).

[BG-Conf16] is a joint work, led by Badr-Eddine Chérief-Abdellatif, with Yuyang Shi and Arnaud Doucet.

Despite its wide use and empirical successes, the theoretical understanding and study of the behaviour and performance of the variational autoencoder (VAE) have only emerged in the past few years. We contribute to this recent line of work by analysing the VAE's reconstruction ability for unseen test data, leveraging arguments from the PAC-Bayes theory. We provide generalisation bounds on the theoretical reconstruction error, and provide insights on the regularisation effect of VAE objectives. We illustrate our theoretical results with supporting experiments on classical benchmark datasets.

Since its introduction by Kingma and Welling, 2014, the Variational AutoEncoder (VAE) has attracted considerable interest and is now widely used for learning low dimensional representations of high dimensional complex data, such as images. The VAE provides a probabilistic view on the autoencoder, a structure which trains an encoder that maps a high dimensional input to a low dimensional latent code, which is then reconstructed using a decoder. The probabilistic encoder $q_\phi(\mathbf{z}|\mathbf{x})$ is a distribution over the possible values of the code $\mathbf{z}$ given a datapoint $\mathbf{x}$, while the probabilistic decoder $p_\theta(\mathbf{x}|\mathbf{z})$ is a distribution over the possible corresponding values of $\mathbf{x}$ given the code $\mathbf{z}$.

As any autoencoder, the VAE offers a powerful framework for learning compressed representations by encoding the information required to reconstruct the original signal accurately. Beyond this simple coding theory perspective, the VAE is more generally presented as a deep generative model. Assuming that the latent code $\mathbf{z}$ is distributed according to a prior $p(\mathbf{z})$ (typically an isotropic Gaussian distribution) and that the decoder is defined via a likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ parameterised by a neural network, the encoder $q_\phi(\mathbf{z}|\mathbf{x})$ is then represented as a variational approximation of the posterior $p_\theta(\mathbf{z}|\mathbf{x})$ which is also parameterised by a neural network. Both the encoder and the decoder networks weights are jointly learnt by minimising a variational objective:

$$\underbrace{-\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction loss}} + \underbrace{\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))}_{\text{rate}}$$

averaged over the dataset, where the first term is a reconstruction loss and the second term is the Kullback–Leibler (KL) divergence between the encoder and the prior $p(\mathbf{z})$. This can alternatively be seen as maximising the celebrated Evidence Lower Bound (ELBO) on the (intractable) log-evidence of this latent variable model. In information-theoretic words, the reconstruction loss is the distortion measured through the encoder-decoder channel, while the KL term is called the rate, an upper bound on the mutual information between the input and the code (a quantity which is usually interpreted as a regulariser controlling the degree of compression through the autoencoder). The celebrated β-VAE variant (Higgins et al., 2017) corresponds to the case where the KL rate has a multiplicative factor β in the variational objective.

Related work. There is a growing body of works aiming to understand the empirical success of the VAE, and a number of reasons have been put forward to explain its apparent good generalisation properties. A large part of the literature has addressed generalisation through the lens of generative modeling, by measuring the sampling ability of the VAE and of its variants using either the log-marginal likelihood, the ELBO, or relative quantitative metrics such as FID (Heusel et al., 2017; Kumar and Poole, 2020) and precision and recall scores (Sajjadi et al., 2018). Rate-distortion curves rather than the log-likelihood have also been proposed to obtain more information about the model (Alemi et al., 2018; Huang et al., 2020). Probably the most remarkable phenomenon is the now widely acknowledged fact that an infinite capacity VAE memorises the training data and interpolates: in other words the VAE predicts novel data points between given training samples by decoding a convex combination of the latent codes (Alemi et al., 2018; Rezende and Viola, 2018; Shu et al., 2018). In light of these findings, Shu et al., 2018 investigated the impact of the encoder capacity on the memorisation property, while Zhao et al., 2018 focused on the question of sampling out-of-domain data from the learnt representation. Another interesting line of research lies in evaluating the quality of the representation via different semantic notions such as disentanglement (achieving interpretability via the decomposability of the latent representation, with the hope to ultimately generalise to new combinations of factors, as explored by Higgins et al., 2017; Chen et al., 2018; Kim and Mnih, 2018; Mathieu et al., 2019; Esmaeili et al., 2019; Locatello et al., 2019) or robustness (by exploring metrics that capture some effects that rare events from multiple generative factors can have on feature encodings, see Suter et al., 2019). While this is out of scope of the present paper, we also acknowledge recent works on other generative models, such as Generative Adversarial Networks (Biau et al., 2021; Schreuder et al., 2021).

Nevertheless, the notion of generalisation is intrinsically subjective. Indeed, a given VAE objective can lead to good reconstruction on unseen test data while being poor for sampling, and can conversely lead to poor reconstruction while being able to generate realistic images. Furthermore, while the ELBO objective naturally defines a proper generative model as a lower bound on the log-marginal likelihood, this is no longer true for the β-VAE when $\beta < 1$. Hence, the approach consisting in evaluating any VAE objective from the sampling perspective is not always appropriate. We do not focus here on the generative abilities of the VAE. Similarly, although disentanglement potentially induces generalisation, we are not focusing on that notion per se.

In this paper, we study the VAE from a reconstruction perspective: we consider the VAE as a model that learns a lossy encoder and decoder, with the belief that a model generalising well should capture a meaningful representation of the data. To better understand the generalisation ability of the VAE in terms of reconstruction, Bozkurt et al., 2021 have investigated the regularisation properties of the VAE objective. Somewhat counter-intuitively, they demonstrated through extensive experiments that the KL term neither actually acts as a regulariser nor improves generalisation when focusing on reconstruction. They also showed that reducing β always decreases the generalisation gap when test data deviates substantially from the training data in pixel space. Their work differentiates between test data that can be reconstructed easily by taking the most similar memorised training data points, and more complicated test data with out-of-domain samples. Hence, the influence of the KL term on the generalisation ability of the VAE is not the same depending on the difficulty of the generalisation task, although the impact of the KL is always monotonic in β. This sheds additional light on the observations made by Alemi et al., 2018 and Rezende and Viola, 2018 which were conducted on training data only. This somewhat surprising behaviour calls for a further study of the regularisation effect of VAE objectives – a contribution of our paper.

Our approach. A natural way to study regularisation is to derive statistical guarantees to quantify the risk of overfitting. We address this by computing generalisation bounds on the reconstruction loss using PAC-Bayes theory. PAC-Bayes has been extensively and successfully used in many settings in machine learning and statistics – however, to the best of our knowledge, it has never been leveraged in the VAE literature. The inference model represented by the encoder is a stochastic function of the inputs that is learnt using amortised inference to improve computational efficiency for huge datasets.

We first formulate PAC-Bayes bounds for the VAE structure. We then show that minimising directly the PAC-Bayes bound over the reconstruction error for amortised variational inference not only provides non-vacuous generalisation bounds, but also significantly decreases the generalisation gap between the test and the training reconstruction errors. The idea of using PAC-Bayes to evaluate the generalisation ability of autoencoders has appeared in the past few years. Epstein and Meir, 2019 have indeed recently adapted margin- and norm-based results for deep neural networks (Bartlett et al., 2017a; Neyshabur et al., 2018b; Arora et al., 2018) to obtain a generalisation bound for deterministic autoencoders. However, there are two substantial differences between their work and ours. First, their bound on the generalisation gap can only be obtained up to a large constant independent of the network parameters, the sample size and the margin, while our bound can be computed and used as an objective for designing an alternative learning algorithm. Second, their bound is deterministic. This is due to the fact that PAC-Bayes inequalities only appear as an artefact in their proofs, in which the networks parameters are artificially perturbed using a Gaussian noise. The deterministic generalisation gap is then controlled by the means of the perturbed network parameters at the price of a looser inequality involving different margin levels. In contrast, our bound is a genuine PAC-Bayes bound on the probabilistic structure of the VAE whose stochasticity is used as a way to inject noise during the learning phase.

Summary of our contributions. Hence, the primary motivation for this work is to complement the findings on the role of the rate as a regulariser Bozkurt et al., 2021 by providing the first theoretical results on the generalisation ability of the VAE and the regularising property of the KL in terms of reconstruction. We choose to derive statistical guarantees by computing generalisation bounds on the reconstruction loss. Leveraging PAC-Bayes theory, we provide bounds that can not only be computed empirically, but can also be used as new learning objectives with good generalisation properties and strong theoretical guarantees. Consequently, our contribution is two-fold: i) we formulate a derandomised PAC-Bayes generalisation bound for the VAE structure which is the first such bound in the VAE literature; ii) we use a non-derandomised variant of this bound to propose a novel PAC-Bayes objective for the VAE structure that will generalise well while achieving tight risk certificates. We provide empirical evidence on real-world datasets, evaluate the generalisation ability of both the β-VAE (including the original VAE with $\beta = 1$) and PAC-Bayes objectives, and compute generalisation bounds for these strategies.

We consider a dataset $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of independent copies of a random variable $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^D$ sampled from an unknown probability distribution $\mathcal{D}$, with D a (potentially large) positive integer. We assume a generative model $p_\theta(\mathbf{x}, \mathbf{z})$ involving a latent random variable $\mathbf{z}$ in a lower dimensional space $\mathcal{Z} \subset \mathbb{R}^d$: the model is composed of a prior $p(\mathbf{z})$ (*e.g.* a standard Gaussian distribution), and of a conditional likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ from a parametric family indexed by $\theta \in \Theta$ (*e.g.* the weights of a neural network). The marginal likelihood over the observed variables, given by $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}$, is typically intractable.

The VAE (Kingma and Welling, 2014) adopts a variational approach to turn the intractable posterior inference and learning problem into a tractable one, which results in the maximisation of a lower bound on the log-evidence (ELBO). The encoder and the decoder, respectively parameterised by ϕ and θ, attempt to learn: (i) a variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$ that approximates the intractable posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$, and (ii) the conditional likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ that approximates the data generating distribution. We recall the variational objective minimised by the β-VAE (Higgins et al., 2017), an extension of the VAE that reweights the KL term in the variational objective:

$$\begin{aligned}\mathcal{L}_\beta(\phi, \theta) = \sum_{i=1}^{n} -\mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{x}_i)} [\log p_\theta(\mathbf{x}_i|\mathbf{z}_i)] \\ + \beta \cdot \sum_{i=1}^{n} \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i) \| p(\mathbf{z})).\end{aligned}$$

The standard VAE framework corresponds to $\beta = 1$, in which case the variational objective can be rewritten as the (opposite of the) ELBO:

$$\begin{aligned}\mathcal{L}_1(\phi, \theta) &= \sum_{i=1}^{n} -\log p_\theta(\mathbf{x}_i) + \sum_{i=1}^{n} \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i) \| p_\theta(\mathbf{z}|\mathbf{x}_i)) \\ &\geqslant \sum_{i=1}^{n} -\log p_\theta(\mathbf{x}_i).\end{aligned}$$

The prior over the latent variables is typically set to be the isotropic multivariate Gaussian $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, while the conditional likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ is generally defined as a Gaussian (in case of real-valued data) or Bernoulli (in case of binary data) whose distribution parameters are computed from $\mathbf{z}$ using a neural network. For binary data $\mathbf{x}$ for instance, the shape of the variational and likelihood distributions can be taken as a Gaussian latent distribution and a factorised Bernoulli observation likelihood:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(\mathbf{z}))\right), \tag{4.3}$$

$$\log p_\theta(\mathbf{x}|\mathbf{z}) = \sum_{j=1}^{D} \{x_j \log \omega_\theta(\mathbf{z})_j + (1 - x_j)\log(1 - \omega_\theta(\mathbf{z})_j)\}, \tag{4.4}$$

where both the encoder distribution parameters $(\boldsymbol{\mu}_\phi(\mathbf{x}), \log \boldsymbol{\sigma}_\phi(\mathbf{x})) = \mathrm{NN}_\phi(\mathbf{x})$ and the decoder distribution parameter $\boldsymbol{\omega}_\theta(\mathbf{z}) = \mathrm{NN}_\theta(\mathbf{z})$ are outputs of neural networks, with $0 < \omega_\theta(\mathbf{z})_j < 1$ for any j, which can be obtained for example via a sigmoid nonlinearity as the last layer of the neural network. Here, ϕ and θ are the weights of the corresponding neural networks.

Let us stress that we focus on the generalisation properties of the VAE and its variants in terms of reconstruction, and mainly interpret the structure as a model for learning representations using an encoder and a decoder. Note that we recover the case of a deterministic autoencoder in the limit of infinite capacity when $\beta = 0$ as the reconstruction loss alone is minimised when the encoder $q_\phi(\mathbf{z}|\mathbf{x})$ is a Dirac mass at $\arg\max_{\mathbf{z}} \log p_\theta(\mathbf{x}|\mathbf{z})$ and when θ minimises the corresponding log-likelihood. This partly explains the fact that the KL term is often interpreted as a regulariser that smoothes the representation and makes the VAE less prone to overfitting.

We adopt in this section a PAC-Bayesian approach on the VAE structure, both for computing generalisation bounds and for learning the related learning objective. The term pseudo-VAE refers to the structure learnt by any objective, whether that of the exact VAE, that of a β-VAE or that of a PAC-Bayes objective which we present in this section.

We consider a dataset $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ composed of binary data, typically images, from an unknown distribution $\mathcal{D}$. We use a Gaussian encoder and a standard Bernoulli conditional likelihood in the decoder as detailed in (4.3) and (4.4). Here, $\omega = (\phi, \theta)$, and the reconstruction loss $\ell(\phi, \theta, \mathbf{x})$ is obtained via rescaling $-\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$ where $p_\theta(\mathbf{x}|\mathbf{z})$ is a truncated version of the conditional likelihood, so that the loss is bounded with range $[0, 1]$. Then, the theoretical $R(\phi, \theta) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\ell(\phi, \theta, \mathbf{x})]$ and empirical $\widehat{R}_{\mathcal{S}}(\phi, \theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\phi, \theta, \mathbf{x}_i)$ reconstruction losses measure the quality of the reconstruction of any variant of the VAE whose encoder and decoder weights are respectively ϕ and θ.

We now present our main results, which are generalisation bounds on the reconstruction gap. The first one, dedicated to designing the learning objective, is a bound on average over the VAE parameters, while the second one is a derandomised PAC-Bayes bound used to evaluate the upper bound. The first bound is given in the following theorem.

Theorem 4.4. *Let $\delta \in (0,1)$, (ϕ^0, θ^0) and $\sigma_\theta^2 > 0$, $\sigma_\phi^2 > 0$. With probability at least $1-\delta$ over $\{\mathbf{x}_1,\dots,\mathbf{x}_n\} \sim \mathcal{D}^n$, we have for any (ϕ, θ), for any (s_ϕ^2, s_θ^2):*

$$\mathbb{E}_{\mathcal{N}(\phi, s_\phi^2 I), \mathcal{N}(\theta, s_\theta^2 I)}[R(\tilde{\phi}, \tilde{\theta})] \leqslant \mathbb{E}_{\mathcal{N}(\phi, s_\phi^2 I), \mathcal{N}(\theta, s_\theta^2 I)}[\widehat{R}_{\mathcal{S}}(\tilde{\phi}, \tilde{\theta})]$$
$$+ \sqrt{\frac{\|\phi - \phi^0\|_2^2}{4\sigma_\phi^2 n} + \frac{N_\phi \left(\frac{s_\phi^2}{\sigma_\phi^2} + \log(\frac{\sigma_\phi^2}{s_\phi^2}) - 1 \right)}{4n} + \frac{\log(\frac{2\sqrt{n}}{\delta})}{2n}}$$
$$+ \sqrt{\frac{\|\theta - \theta^0\|_2^2}{4\sigma_\theta^2 n} + \frac{N_\theta \left(\frac{s_\theta^2}{\sigma_\theta^2} + \log(\frac{\sigma_\theta^2}{s_\theta^2}) - 1 \right)}{4n} + \frac{\log(\frac{2\sqrt{n}}{\delta})}{2n}},$$

where N_ϕ and N_θ are respectively the encoder and decoder neural networks size.

The risks $R(\tilde{\phi}, \tilde{\theta})$ and $\widehat{R}_{\mathcal{S}}(\tilde{\phi}, \tilde{\theta})$ in the inequality are averaged over the random parameters $\tilde{\phi}$ and $\tilde{\theta}$ following Gaussians centered at the VAE parameters ϕ, θ with respective variances s_ϕ^2, s_θ^2. The bound serves as a learning objective for both the VAE parameters ϕ, θ and the corresponding variance levels s_ϕ^2, s_θ^2, and contains separate terms involving each of them.

We present now a derandomised bound, which is a key point in order to evaluate the performance of the learnt strategy itself, in contrast to the previous standard averaged PAC-Bayes bound.

Theorem 4.5. *Let $\delta \in (0,1)$, (ϕ^0, θ^0) and $\sigma_\theta^2 > 0$, $\sigma_\phi^2 > 0$. Then with probability at least $1-\delta$ over both $\mathcal{S} = \{\mathbf{x}_1,\dots,\mathbf{x}_n\} \sim \mathcal{D}^n$, and $\varepsilon_\phi \sim \mathcal{N}(0, \sigma_\phi^2 I)$, $\varepsilon_\theta \sim \mathcal{N}(0, \sigma_\theta^2 I)$:*

$$kl\left(R(\phi + \varepsilon_\phi, \theta + \varepsilon_\theta) \| \widehat{R}_{\mathcal{S}}(\phi + \varepsilon_\phi, \theta + \varepsilon_\theta) \right)$$
$$\leqslant \frac{\|\phi - \phi^0 + \varepsilon_\phi\|_2^2 - \|\varepsilon_\phi\|_2^2}{2\sigma_\phi^2 n} + \frac{\|\theta - \theta^0 + \varepsilon_\theta\|_2^2 - \|\varepsilon_\theta\|_2^2}{2\sigma_\theta^2 n}$$
$$+ \frac{\log(2\sqrt{n}/\delta)}{n}, \quad (4.5)$$

where (ϕ, θ) *is the output of the algorithm minimising the bound in* 4.4 *given the dataset* $\mathcal{S}$, *and where kl is the binary Kullback-Leibler divergence*

$$kl(q \| p) = q \log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right)$$

for any $q, p \in (0, 1)$.

To summarise, we considered two different bounds with two different perspectives. Theorem 4.4 is a randomised inequality designed to learn the noise level so that it encourages the learning process to end in flat regions with small generalisation error, while Theorem 4.5 is a high probability bound over the noise which is used to evaluate the final bound for small fixed values of the noise so that the bound can be considered as deterministic. Note that learning is no longer self-certified, but the principles driving this approach could still lead to good guarantees.

Our PAC-Bayes approach can be summarised as:

(i) First learn (ϕ^0, θ^0);

(ii) Then learn (ϕ, θ) along with (s_ϕ^2, s_θ^2). This can be done by minimising the bound in Theorem 4.4 using SGD on the entire dataset;

(iii) Finally evaluate the bound at (ϕ, θ). This can be done by inverting the kl bound in Theorem 4.5.

5 PERSPECTIVES

OUTLINE

Towards a new generalisation-centric paradigm: from generalisation to frugal AI.

My scientific activity to date shows that my major interest in generalisation in machine learning structures my vision of the field. I am convinced that this is one of the keys to designing and deploying future intelligent systems. This long-term goal resonates remarkably with the emergence, beyond academic walls, of large models (termed *foundation models*) in language over the past eighteen months—it is, however, striking to note that the most powerful models require astonishing volumes of data, to train and fine-tune an also prodigious number of parameters (this article estimates GPT-4's parameters at over 1.8 trillion). This inflationary trend will significantly concentrate the capacity to train such models within an increasingly limited number of actors, already excluding most academic players, hindering the accessibility of future systems. This trend seems neither sustainable nor desirable to me. The quest for algorithms capable of generalising well from **limited data** will thus become increasingly pressing, at a time when large-scale data acquisition (necessary for training **foundation models**) can prove costly for a large part of research or industry players, and legitimate privacy concerns limit its availability. It will become increasingly important to design algorithms capable of learning **from a fraction of the data** and/or **computational power** currently necessary.

I will carry on my work on generalisation and will investigate many follow-up leads of results mentioned in this manuscript, with the group of outstandingly talented students I am fortunate to supervise. However, I feel there is now an opportunity for not only work on generalisation for the sake of it, but to translate this knowledge towards shifting the way we build and deploy learning algorithms.

The high-level ideas I detail here are the product of three convictions:

(a) **generalisation theory** (especially its PAC-Bayes component) is today sufficiently mature and developed to tackle ambitious problems in machine learning,

(b) one of the major topics will not be to collect **always more** data, but **to learn better from less**,

(c) the most advanced systems will be capable of **continuous learning** and flexibly evolving their representation of the world.

On top of these three scientific convictions that I have forged over my career, I am acutely aware of the growing environmental footprint of artificial intelligence systems, and its unsustainable nature. I will thus aim to design learning algorithms that are more **frugal**, and more efficient.

If I ever need an elevator pitch, here's what I'd use

The guiding principle of my research is the quest for generalisation for intelligent systems, that is, the ability for an artificial entity to generalise notions or abstract concepts from specific examples (data), and the long-term goal is to drastically reduce the amount of data and computation necessary for generalisation, leading to frugal AI systems.

In the realm of machine learning and artificial intelligence, the quest for algorithms that can generalize well from limited data is paramount, especially in an era where data acquisition can be costly or privacy concerns limit data availability. I will aim my future reseach efforts towards the intersection of PAC-Bayes learning, frugal AI principles, and the innovative design of algorithms capable of learning from a fraction of data and/or computational resources. PAC-Bayes, an influential theory that provides a statistical framework to understand the generalization ability of machine learning models, offers a robust foundation for exploring new frontiers in algorithmic efficiency and effectiveness. By leveraging insights from PAC-Bayes, we aim to pioneer frugal AI algorithms that not only require minimal data for training but also optimize computational overhead, addressing critical challenges in scalability and accessibility. The envisioned algorithms have the potential to revolutionize the landscape of AI and ML by enabling more sustainable, inclusive, and efficient learning models, thereby broadening the applicability of AI technologies across various environments constrained by ecological imperatives.

Through fundamental advances towards stronger principles, smaller models, and reduced data sets, my research ideas will enable tomorrow's best AI systems to operate on yesterday's devices, thus offering a remedy against obsolescence, and contribute to more frugal AI systems.

I close this manuscript by echoing the PAC-Bayes manifesto introduced in Chapter 1. The previous chapters support my claim that PAC-Bayes is to play a pivotal role in the study and promotion of generalisation, in particular for designing new frugal intelligent systems. This is evidenced by the steady growth of scientific works submitted to arXiv relating to PAC-Bayes depicted in Figure 5.1.

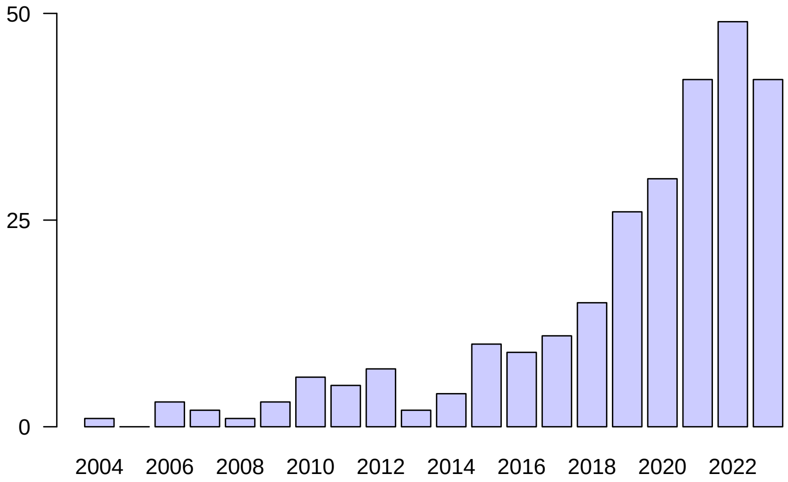

Figure 5.1: Number of submissions on arXiv whose title or abstract contains the terms "PAC-Bayes" or "PAC-Bayesian", from 2004 to 2023.

The PAC-Bayes Manifesto

- ◇ **Theory**: PAC-Bayes generalization bounds are the most precise or the only ones existing for many learning problems.
- ◇ **Algorithms**: The *generalisation-by-design* strategy allows constructing *ad hoc* algorithms with the best generalization performance.
- ◇ **Numerical Results**: PAC-Bayes leads to numerically non-trivial bounds (or certificates) for a wide range of problems.

BIBLIOGRAPHY

Adams, R., Shawe-Taylor, J., and Guedj, B. (2022). "Controlling Multiple Errors Simultaneously with a PAC-Bayes Bound". Submitted. arXiv: 2202.05560 [stat.ML]. URL: https://arxiv.org/abs/2202.05560 (pp. 7, 12, 36, 75).

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. (2018). "Fixing a broken ELBO". *International Conference on Machine Learning* (pp. 44, 45).

Alliez, P., Di Cosmo, R., Guedj, B., Girault, A., Hacid, M.-S., Legrand, A., and Rougier, N. (Jan. 2020). "Attributing and Referencing (Research) Software: Best Practices and Outlook From Inria". *Computing in Science & Engineering* 22.1, pp. 39–52. ISSN: 1558-366X. DOI: 10.1109/mcse.2019.2949413. URL: http://dx.doi.org/10.1109/MCSE.2019.2949413 (p. 14).

Alquier, P. (2024). "User-friendly introduction to PAC-Bayes bounds". *Foundations and Trends® in Machine Learning* 17.2, pp. 174–303. ISSN: 1935-8237. DOI: 10.1561/2200000100. URL: http://dx.doi.org/10.1561/2200000100 (p. 4).

Alquier, P. and Guedj, B. (2017). "An oracle inequality for quasi-Bayesian nonnegative matrix factorization". *Mathematical Methods of Statistics* 26.1, pp. 55–67. ISSN: 1934-8045. DOI: 10.3103/S1066530717010045. URL: https://link.springer.com/article/10.3103%2FS1066530717010045 (pp. 7, 11, 36).

Alquier, P. and Guedj, B. (2018). "Simpler PAC-Bayesian bounds for hostile data". *Machine Learning* 107.5, pp. 887–902. ISSN: 1573-0565. DOI: 10.1007/s10994-017-5690-0. URL: https://doi.org/10.1007/s10994-017-5690-0 (pp. 7, 11, 30, 35, 36, 41, 42).

Alquier, P., Ridgway, J., and Chopin, N. (Dec. 2016). "On the properties of variational approximations of Gibbs posteriors". *Journal of Machine Learning Research (JMLR)* 17.236, pp. 1–41 (pp. 25, 31).

Amit, R., Epstein, B., Moran, S., and Meir, R. (2022). "Integral Probability Metrics PAC-Bayes Bounds". *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 30, 31, 33).

Amit, R. and Meir, R. (2018). "Meta-Learning by Adjusting Priors Based on Extended PAC-Bayes Theory". *International Conference on Machine Learning (ICML)* (p. 29).

Arjovsky, M., Chintala, S., and Bottou, L. (2017). "Wasserstein Generative Adversarial Networks". *International Conference on Machine Learning (ICML)* (p. 29).

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. (2018). "Stronger generalization bounds for deep nets via a compression approach". *International Conference on Machine Learning* (p. 46).

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. (2019). "A Theoretical Analysis of Contrastive Unsupervised Representation Learning". *ICML*, pp. 5628–5637 (pp. 37–42).

Bartlett, P., Foster, D. J., and Telgarsky, M. (2017a). "Spectrally-normalized margin bounds for neural networks". *Advances in Neural Information Processing Systems* (p. 46).

Bartlett, P. L., Foster, D. J., and Telgarsky, M. (2017b). "Spectrally-normalized margin bounds for neural networks". *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, pp. 6240–6249. URL: https://proceedings.neurips.cc/paper/2017/hash/b22b257ad0519d4500539da3c8bcf4dd-Abstract.html (p. 19).

Bartlett, P. L. and Mendelson, S. (Nov. 2002). "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results". *Journal of Machine Learning Research (JMLR)* 3, pp. 463–482. ISSN: 1532-4435. DOI: 10.1007/3-540-44581-1_15 (p. 3).

Bengio, Y., Courville, A., and Vincent, P. (2013). "Representation Learning: A Review and New Perspectives". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8, pp. 1798–1828 (p. 37).

Berend, D. and Kontorovich, A. (2015). "A finite sample analysis of the Naive Bayes classifier". *JMLR* (p. 25).

Biau, G., Fischer, A., Guedj, B., and Malley, J. D. (2016). "COBRA: A combined regression strategy". *Journal of Multivariate Analysis* 146. Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces, pp. 18–28. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2015.04.007. URL: http://www.sciencedirect.com/science/article/pii/S0047259X15000950 (pp. 11, 77).

Biau, G., Sangnier, M., and Tanielian, U. (2021). "Some Theoretical Insights into Wasserstein GANs". *Journal of Machine Learning Research* 22.119, pp. 1–45 (p. 44).

Biggs, F. and Guedj, B. (2021). "Differentiable PAC-Bayes Objectives with Partially Aggregated Neural Networks". *Entropy* 23.10: *Approx-*

imate Bayesian Inference. ISSN: 1099-4300. DOI: 10.3390/e23101280. arXiv: 2006.12228 [cs.LG]. URL: https://www.mdpi.com/1099-4300/23/10/1280 (pp. 8, 9, 12, 15, 16, 24, 76).

Biggs, F. and Guedj, B. (July 2022a). "Non-Vacuous Generalisation Bounds for Shallow Neural Networks". *Proceedings of the 39th International Conference on Machine Learning [ICML]*. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 1963–1981. arXiv: 2202.01627 [cs.LG]. URL: https://proceedings.mlr.press/v162/biggs22a.html (pp. 8–10, 15, 16, 18, 24, 76).

Biggs, F. and Guedj, B. (28–30 Mar 2022b). "On Margins and Derandomisation in PAC-Bayes". *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics [AISTATS]*. Ed. by G. Camps-Valls, F. J. R. Ruiz, and I. Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, pp. 3709–3731. arXiv: 2107.03955 [cs.LG]. URL: https://proceedings.mlr.press/v151/biggs22a.html (pp. 7–9, 12, 15, 16, 19, 24, 36, 76).

Biggs, F. and Guedj, B. (2023). "Tighter PAC-Bayes Generalisation Bounds by Leveraging Example Difficulty". *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics [AISTATS]*. Vol. 206. PMLR, pp. 8165–8182. DOI: 10.48550/arXiv.2210.11289. arXiv: 2210.11289 [cs.LG]. URL: https://proceedings.mlr.press/v206/biggs23a.html (pp. 7, 12, 36, 76).

Biggs, F., Zantedeschi, V., and Guedj, B. (2022). "On Margins and Generalisation for Voting Classifiers". *Advances in Neural Information Processing Systems [NeurIPS]*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 9713–9726. arXiv: 2206.04607 [cs.LG]. URL: https://papers.nips.cc/paper_files/paper/2022/hash/3f8675af3da6da231c9e75b889b7f047-Abstract-Conference.html (pp. 8, 12, 24, 74, 76).

Bousquet, O. and Elisseeff, A. (Mar. 2002). "Stability and Generalization". *Journal of Machine Learning Research (JMLR)* 2, pp. 499–526 (p. 3).

Bozkurt, A., Esmaeili, B., Tristan, J.-B., Brooks, D., Dy, J., and Meent, J.-W. van de (2021). "Rate-Regularization and Generalization in Variational Autoencoders". *International Conference on Artificial Intelligence and Statistics* (pp. 45, 46).

Brotcorne, L., Canteaut, A., Carneiro Viana, A., Grandmont, C., Guedj, B., Huot, S., Issarny, V., Pallez, G., Perrier, V., Quema, V., Pomet, J.-B., Rival, X., Salvati, S., and Thomé, E. (Dec. 2020). *Indicateurs de suivi de l'activité scientifique de l'Inria*. Research Report. Inria. URL: https://hal.inria.fr/hal-03033764 (p. 14).

Camuto, A., Deligiannidis, G., Erdogdu, M. A., Gürbüzbalaban, M., Şimşekli, U., and Zhu, L. (2021). "Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms". *Advances in Neural Information Processing Systems (NeurIPS)* (p. 30).

Candiotto, L., Cayco-Gajic, N. A., Soarez, P. C. de, Luca, M. D., Frassinelli, D., Fujita, A., Growiec, J., Guedj, B., Kasturi, S. N., Kawanishi, Y., Kellmeyer, P., Livermore, M. A., Moodley, D., Rabinowitch, I., Raliteera, T., Stalnov, O., Taylor, H., and Wevers, M. (2023). "Nurturing Knowledge: A Virtue Epistemic Approach to Explainable AI". Submitted. DOI: 10.48550/arXiv.xxxx.xxxxx. arXiv: xxxx.xxxxx [cs.LG]. URL: https://arxiv.org/abs/xxxx.xxxxx (p. 14).

Cantelobre, T., Ciliberto, C., Guedj, B., and Rudi, A. (17–23 Jul 2022). "Measuring dissimilarity with diffeomorphism invariance". *Proceedings of the 39th International Conference on Machine Learning [ICML]*. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 2572–2596. arXiv: 2202.05614 [stat.ML]. URL: https://proceedings.mlr.press/v162/cantelobre22a.html (pp. 13, 75).

Cantelobre, T., Ciliberto, C., Guedj, B., and Rudi, A. (2024). "Closed-form filtering for non-linear systems". Submitted. DOI: 10.48550/ARXIV.2402.09796. arXiv: 2402.09796 [stat.ML]. URL: https://arxiv.org/abs/2402.09796 (pp. 14, 75).

Cantelobre, T., Guedj, B., Pérez-Ortiz, M., and Shawe-Taylor, J. (2020). "A PAC-Bayesian Perspective on Structured Prediction with Implicit Loss Embeddings". Submitted. arXiv: 2012.03780 [cs.LG]. URL: https://arxiv.org/abs/2012.03780 (pp. 7, 12, 36, 75).

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). "Deep Clustering for Unsupervised Learning of Visual Features". *ECCV* (p. 37).

Catoni, O. (2007). *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Vol. 56. IMS Lecture Notes Monogr. Ser., pp. 1–163 (p. 4).

Catoni, O. (Oct. 2003). *A PAC-Bayesian approach to adaptive classification*. URL: yaroslavvb.com/papers/notes/catoni-pac.pdf (p. 4).

Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization*. Ed. by J. Picard. Lecture Notes in Mathematics: Saint-Flour Summer School on Probability Theory XXXI 2001. DOI: 10.1007/b99352 (p. 4).

Celisse, A. and Guedj, B. (2016). "Stability revisited: new generalisation bounds for the Leave-one-Out". Preprint. URL: https://arxiv.org/abs/1608.06412 (pp. 7, 11, 36).

Chee, A. and Loustau, S. (2021). "Learning with BOT - Bregman and Optimal Transport divergences" (pp. 30, 31).

Chen, R. T., Li, X., Grosse, R., and Duvenaud, D. (2018). "Isolating sources of disentanglement in variational autoencoders". *Advances in Neural Information Processing Systems* (p. 44).

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. (2020). "A Simple Framework for Contrastive Learning of Visual Representations". *arXiv preprint arXiv:2002.05709v1* (p. 38).

Chérief-Abdellatif, B.-E., Shi, Y., Doucet, A., and Guedj, B. (28–30 Mar 2022). "On PAC-Bayesian reconstruction guarantees for VAEs". *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics [AISTATS]*. Ed. by G. Camps-Valls, F. J. R. Ruiz, and I. Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, pp. 3066–3079. arXiv: 2202.11455 [cs.LG]. URL: https://proceedings.mlr.press/v151/cherief-abdellatif22a.html (pp. 7, 10, 36, 43).

Chopin, N., Gadat, S., Guedj, B., Guyader, A., and Vernet, E. (2015). "On some recent advances on high dimensional Bayesian statistics". *ESAIM: Proceedings and Surveys* 51, pp. 293–319. DOI: 10.1051/proc/201551016. URL: https://doi.org/10.1051/proc/201551016 (p. 11).

Chrétien, S. and Guedj, B. (2020). "Revisiting clustering as matrix factorisation on the Stiefel manifold". *LOD – The Sixth International Conference on Machine Learning, Optimization, and Data Science*. Ed. by G. Nicosia, V. Ojha, E. La Malfa, G. Jansen, V. Sciacca, P. Pardalos, G. Giuffrida, and R. Umeton. Springer International Publishing, pp. 1–12. ISBN: 978-3-030-64583-0. DOI: 10.1007/978-3-030-64583-0_1. URL: https://link.springer.com/chapter/10.1007%2F978-3-030-64583-0_1 (pp. 7, 13, 36).

Chugg, B., Wang, H., and Ramdas, A. (2023). "A unified recipe for deriving (time-uniform) PAC-Bayes bounds". *arXiv* abs/2302.03421 (pp. 31, 33).

Clerico, E., Farghly, T., Deligiannidis, G., Guedj, B., and Doucet, A. (2022). "Generalisation under gradient descent via deterministic PAC-Bayes". Submitted. DOI: 10.48550/ARXIV.2209.02525. arXiv: 2209.02525 [stat.ML]. URL: https://arxiv.org/abs/2209.02525 (pp. 7, 12, 36, 77).

Clerico, E. and Guedj, B. (2023). "A note on regularised NTK dynamics with an application to PAC-Bayesian training". *Transactions on Machine Learning Research [TMLR]*. ISSN: 2835-8856. DOI: 10.48550/ARXIV.2312.13259. arXiv: 2312.13259 [stat.ML]. URL: https://openreview.net/forum?id=2la55BeWwy (pp. 7, 12, 36, 77).

Cohen-Addad, V., Guedj, B., Kanade, V., and Rom, G. (Apr. 2021). "Online k-means Clustering". *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics [AISTATS]*. Ed. by A. Banerjee and K. Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 1126–1134. URL: http://proceedings.mlr.press/v130/cohen-addad21a.html (p. 13).

Csiszár, I. and Shields, P. C. (2004). "Information Theory and Statistics: A Tutorial". *Foundations and Trends® in Communications and Information Theory* 1.4, pp. 417–528 (p. 42).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *NAACL-HLT*, pp. 4171–4186 (p. 37).

Devroye, L. and Wagner, T. (Sept. 1979). "Distribution-free performance bounds for potential function rules". *IEEE Trans. Inf. Theory* 25.5, pp. 601–604. DOI: 10.1109/TIT.1979.1056087 (p. 3).

Dewez, F., Guedj, B., Talpaert, A., and Vandewalle, V. (2022). "An end-to-end data-driven optimisation framework for constrained trajectories". *Data-Centric Engineering* 3. DOI: 10.1017/dce.2022.6. arXiv: 2011.11820 [stat.AP]. URL: https://www.cambridge.org/core/journals/data-centric-engineering/article/an-endtoend-datadriven-optimization-framework-for-constrained-trajectories/B4BF729E1681F795FFE79A1F4B544CE5 (pp. 13, 74, 77, 79).

Dewez, F., Guedj, B., and Vandewalle, V. (2020). "From industry-wide parameters to aircraft-centric on-flight inference: improving aeronautics performance prediction with machine learning". *Data-Centric Engineering* 1. DOI: 10.1017/dce.2020.12. arXiv: 2005.05286 [stat.AP]. URL: https://www.cambridge.org/core/journals/data-centric-engineering/article/from-industrywide-parameters-to-aircraftcentric-onflight-inference-improving-aeronautics-performance-prediction-with-machine-learning/7A5662351D23A3D855E7FBC58B45AB6D (pp. 13, 74, 79).

Dietterich, T. G. (2000). "Ensemble Methods in Machine Learning". *Multiple Classifier Systems*. Lecture Notes in Computer Science. Springer (p. 25).

Ding, N., Chen, X., Levinboim, T., Goodman, S., and Soricut, R. (2021). "Bridging the Gap Between Practice and PAC-Bayes Theory in Few-Shot Meta-Learning". *Advances in Neural Information Processing Systems (NeurIPS)* (p. 29).

Dziugaite, G. K. and Roy, D. M. (Aug. 2017). "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data". *Proc. Conf. Uncertainty in Artif. Intell. (UAI)*. Sydney, Australia (pp. 9, 19).

Dziugaite, G. K., Hsu, K., Gharbieh, W., and Roy, D. M. (2020). "On the role of data in PAC-Bayes bounds". *CoRR* abs/2006.10929. arXiv: 2006.10929. URL: https://arxiv.org/abs/2006.10929 (pp. 19, 21).

Epstein, B. and Meir, R. (2019). "Generalization bounds for unsupervised and semi-supervised learning with autoencoders". *arXiv preprint arXiv:1902.01449* (p. 46).

Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D. H., Dy, J., and Meent, J.-W. (2019). "Structured disentangled representations". *International Conference on Artificial Intelligence and Statistics* (p. 44).

Fard, M. M. and Pineau, J. (2010). "PAC-Bayesian Model Selection for Reinforcement Learning". *Advances in Neural Information Processing Systems (NIPS)* (p. 29).

Farid, A. and Majumdar, A. (2021). "Generalization Bounds for Meta-Learning via PAC-Bayes and Uniform Stability". *Advances in Neural Information Processing Systems (NeurIPS)* (p. 29).

Gálvez, B. R., Bassi, G., Thobaben, R., and Skoglund, M. (2021). "Tighter Expected Generalization Error Bounds via Wasserstein Distance". *Advances in Neural Information Processing Systems (NeurIPS)* (p. 30).

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. (2013). "A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers". *ICML*, pp. 738–746 (p. 42).

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. (June 2009). "PAC-Bayesian Learning of Linear Classifiers". *Proc. Int. Conf. Mach. Learning (ICML)*. Montreal, Canada. DOI: 10.1145/1553374.1553419 (pp. 17, 25).

Germain, P., Lacasse, A., Laviolette, F., Marchand, M., and Roy, J. (2015). "Risk bounds for the majority vote: from a PAC-Bayesian analysis to a learning algorithm". *JMLR* (p. 26).

Goroshin, R., Bruna, J., Tompson, J., Eigen, D., and LeCun, Y. (2015). "Unsupervised Learning of Spatiotemporally Coherent Metrics". *ICCV*, pp. 4086–4093 (p. 42).

Grünwald, P. and Mehta, N. A. (Mar. 2020). "Fast Rates for General Unbounded Loss Functions: From ERM to Generalized Bayes". *Journal of Machine Learning Research (JMLR)* 21, pp. 1–80 (p. 7).

Guedj, B. (2010). "A Bayesian modelling of the hybridization mechanism". MA thesis. Sorbonne Université and Danmarks Tekniske Universitet (p. 13).

Guedj, B. (Dec. 2013). "Aggregation of estimators and classifiers: theory and methods". Theses. Université Pierre et Marie Curie - Paris VI.

URL: https://tel.archives-ouvertes.fr/tel-00922353 (pp. xi, 11).

Guedj, B. (2019). “A Primer on PAC-Bayesian Learning”. *Proceedings of the second congress of the French Mathematical Society*. Vol. 33. URL: https://arxiv.org/abs/1901.05353 (pp. 4, 11).

Guedj, B. (July 2021). *Covid-19 and AI: Unexpected Challenges and Lessons*. Research Report. URL: https://hal.inria.fr/hal-03277494 (p. 14).

Guedj, B. and Alquier, P. (2013). “PAC-Bayesian estimation and prediction in sparse additive models”. *Electron. J. Statist.* 7, pp. 264–291. DOI: 10.1214/13-EJS771. URL: https://doi.org/10.1214/13-EJS771 (pp. 7, 11, 36).

Guedj, B. and Guillot, G. (2011). “Estimating the location and shape of hybrid zones”. *Molecular Ecology Resources* 11.6, pp. 1119–1123. DOI: 10.1111/j.1755-0998.2011.03045.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2011.03045.x (p. 13).

Guedj, B. and Pujol, L. (2021). “Still no free lunches: the price to pay for tighter PAC-Bayes bounds”. *Entropy* 23.11: *Approximate Bayesian Inference*. ISSN: 1099-4300. DOI: 10.3390/e23111529. arXiv: 1910.04460 [cs.LG]. URL: https://www.mdpi.com/1099-4300/23/11/1529 (pp. 7, 11, 36, 78).

Guedj, B. and Rengot, J. (2020). “Non-linear aggregation of filters to improve image denoising”. *SAI: Intelligent Computing*. Ed. by K. Arai, S. Kapoor, and R. Bhatia. Springer International Publishing, pp. 314–327. ISBN: 978-3-030-52246-9. DOI: 10.1007/978-3-030-52246-9_22. arXiv: 1904.00865. URL: https://link.springer.com/chapter/10.1007%2F978-3-030-52246-9_22 (pp. 11, 78).

Guedj, B. and Robbiano, S. (2018). “PAC-Bayesian high dimensional bipartite ranking”. *Journal of Statistical Planning and Inference* 196, pp. 70–86. ISSN: 0378-3758. DOI: https://doi.org/10.1016/j.jspi.2017.10.010. URL: http://www.sciencedirect.com/science/article/pii/S0378375817301945 (pp. 7, 11, 36).

Guedj, B. and Srinivasa Desikan, B. (2018). “Pycobra: A Python Toolbox for Ensemble Learning and Visualisation”. *Journal of Machine Learning Research* 18.190, pp. 1–5. URL: http://jmlr.org/beta/papers/v18/17-228.html (pp. 11, 78).

Guedj, B. and Srinivasa Desikan, B. (Jan. 2020). “Kernel-Based Ensemble Learning in Python”. *Information* 11.2, p. 63. ISSN: 2078-2489. DOI: 10.3390/info11020063. URL: http://dx.doi.org/10.3390/info11020063 (pp. 11, 78).

Haddouche, M. and Guedj, B. (2022). “Online PAC-Bayesian Learning”. *Advances in Neural Information Processing Systems [NeurIPS]*. Ed. by S.

Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 25725–25738. arXiv: 2206.00024 [cs.LG]. URL: https://papers.nips.cc/paper_files/paper/2022/hash/a4d991d581accd2955a1e1928f4e6965-Abstract-Conference.html (pp. 7, 12, 29, 31, 32, 36, 75).

Haddouche, M. and Guedj, B. (2023a). "PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales". *Transactions on Machine Learning Research [TMLR]*. ISSN: 2835-8856. DOI: 10.48550/ARXIV.2210.00928. arXiv: 2210.00928 [stat.ML]. URL: https://openreview.net/forum?id=qxrwt6F3sf (pp. 7, 12, 31, 33, 36, 75).

Haddouche, M. and Guedj, B. (2023b). "Wasserstein PAC-Bayes Learning: Exploiting Optimisation Guarantees to Explain Generalisation". Submitted. DOI: 10.48550/arXiv.2304.07048. arXiv: 2304.07048 [stat.ML]. URL: https://arxiv.org/abs/2304.07048 (pp. 7, 12, 36, 75).

Haddouche, M., Guedj, B., Rivasplata, O., and Shawe-Taylor, J. (2021). "PAC-Bayes unleashed: generalisation bounds with unbounded losses". *Entropy* 23.10: *Approximate Bayesian Inference*. ISSN: 1099-4300. DOI: 10.3390/e23101330. arXiv: 2006.07279 [stat.ML]. URL: https://www.mdpi.com/1099-4300/23/10/1330 (pp. 7, 12, 36, 75).

Haddouche, M., Guedj, B., and Shawe-Taylor, J. (2020). "Upper and Lower Bounds on the Performance of Kernel PCA". Submitted. arXiv: 2012.10369 [cs.LG]. URL: https://arxiv.org/abs/2012.10369 (pp. 12, 75).

Haddouche, M., Guedj, B., and Wintenberger, O. (2023). "Optimistically Tempered Online Learning". Submitted. DOI: 10.48550/arXiv.2301.07530. arXiv: 2301.07530 [cs.LG]. URL: https://arxiv.org/abs/2301.07530 (pp. 14, 75).

Haddouche, M., Viallard, P., Şimşekli, U., and Guedj, B. (2024). "A PAC-Bayesian Link Between Generalisation and Flat Minima". Submitted. DOI: 10.48550/ARXIV.2402.08508. arXiv: 2402.08508 [stat.ML]. URL: https://arxiv.org/abs/2402.08508 (pp. 7, 12, 36, 75).

Hellström, F., Durisi, G., Guedj, B., and Raginsky, M. (2024). "Generalization Bounds: Perspectives from Information Theory and PAC-Bayes". Accepted for publication. DOI: 10.48550/arXiv.2309.04381. arXiv: 2309.04381 [cs.LG]. URL: https://arxiv.org/abs/2309.04381 (pp. 3, 4, 11, 74).

Hellström, F. and Guedj, B. (2024). "Comparing Comparators in Generalization Bounds". *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics [AISTATS]*. DOI: 10.48550/arXiv.

2310.10534. arXiv: 2310.10534 [cs.LG]. URL: https://arxiv.org/abs/2310.10534 (pp. 7, 12, 36, 74).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). “GANs trained by a two time-scale update rule converge to a local Nash equilibrium”. *Advances in Neural Information Processing Systems* (p. 44).

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). “Beta-VAE: Learning basic visual concepts with a constrained variational framework”. *International Conference on Learning Representations* (pp. 44, 47).

Higgs, M. and Shawe-Taylor, J. (2010). “A PAC-Bayes Bound for Tailored Density Estimation”. *ALT*, pp. 148–162 (p. 42).

Hinton, G. E. and Camp, D. van (1993). “Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights”. *Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory, COLT 1993, Santa Cruz, CA, USA, July 26-28, 1993*. Ed. by L. Pitt. ACM, pp. 5–13. DOI: 10.1145/168304.168306. URL: https://doi.org/10.1145/168304.168306 (p. 19).

Hochreiter, S. and Schmidhuber, J. (1997). “Flat Minima”. *Neural Comput.* 9.1, pp. 1–42. DOI: 10.1162/neco.1997.9.1.1. URL: https://doi.org/10.1162/neco.1997.9.1.1 (p. 19).

Huang, S., Makhzani, A., Cao, Y., and Grosse, R. (2020). “Evaluating lossy compression rates of deep generative models”. *International Conference on Machine Learning* (p. 44).

Jang, K., Jun, K., Kuzborskij, I., and Orabona, F. (2023). “Tighter PAC-Bayes Bounds Through Coin-Betting”. *Conference on Learning Theory (COLT)* (p. 33).

Jobic, P., Haddouche, M., and Guedj, B. (2023). “Federated Learning with Nonvacuous Generalisation Bounds”. Submitted. DOI: 10.48550/arXiv.2310.11203. arXiv: 2310.11203 [cs.LG]. URL: https://arxiv.org/abs/2310.11203 (pp. 8, 9, 12, 24, 75).

Kantorovitch, L. V. (1960). “Mathematical Methods of Organizing and Planning Production”. *Management Science* (p. 33).

Kim, H. and Mnih, A. (2018). “Disentangling by factorising”. *International Conference on Machine Learning* (p. 44).

Kingma, D. P. and Welling, M. (2014). “Auto-encoding variational Bayes”. *International Conference on Learning Representations* (pp. 43, 47).

Klein, J., Albardan, M., Guedj, B., and Colot, O. (2020). “Decentralized Learning with Budgeted Network Load Using Gaussian Copulas and Classifier Ensembles”. *ECML-PKDD 2019: Machine Learning and Knowledge Discovery in Databases*. Ed. by P. Cellier and K.

Driessens. Springer International Publishing, pp. 301–316. ISBN: 978-3-030-43823-4. DOI: 10.1007/978-3-030-43823-4_26. URL: https://link.springer.com/chapter/10.1007%2F978-3-030-43823-4_26 (p. 13).

Kumar, A. and Poole, B. (2020). "On Implicit Regularization in β -VAEs". *International Conference on Machine Learning* (p. 44).

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley (p. 25).

Lacasse, A., Laviolette, F., Marchand, M., Germain, P., and Usunier, N. (Dec. 2006). "PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier". *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada. DOI: 10.7551/mitpress/7503.003.0101 (p. 26).

Lacasse, A., Laviolette, F., Marchand, M., and Turgeon-Boutin, F. (2010). "Learning with Randomized Majority Votes". *ECML/PKDD (2)*. Springer (pp. 26, 27).

Langford, J. and Seeger, M. (2001). "Bounds for Averaging Classifiers". *CMU Technical report* CMU-CS-01-102 (p. 4).

Langford, J. and Shawe-Taylor, J. (Dec. 2002). "PAC-Bayes & margins". *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada (p. 26).

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). "Gradient-based learning applied to document recognition". *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791 (p. 19).

Leroy, A., Latouche, P., Guedj, B., and Gey, S. (2022). "MAGMA: Inference and Prediction with Multi-Task Gaussian Processes". *Machine Learning*. DOI: 10.1007/s10994-022-06172-1. arXiv: 2007.10731 [stat.CO]. URL: https://arxiv.org/abs/2007.10731 (pp. 13, 76).

Leroy, A., Latouche, P., Guedj, B., and Gey, S. (2023). "Cluster-Specific Predictions with Multi-Task Gaussian Processes". *Journal of Machine Learning Research [JMLR]* 24.5, pp. 1–49. arXiv: 2011.07866 [cs.LG]. URL: https://jmlr.org/papers/v24/20-1321.html (pp. 13, 76).

Letarte, G., Germain, P., Guedj, B., and Laviolette, F. (2019). "Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks". *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems [NeurIPS] 2019, 8-14 December 2019, Vancouver, BC, Canada*. Ed. by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, pp. 6869–6879. eprint: https://arxiv.org/abs/1905.10259. URL: https://papers.nips.cc/paper/8911-dichotomize-and-generalize-pac-bayesian-binary-activated-deep-neural-networks (pp. 8–10, 15, 16, 24).

Li, L. and Guedj, B. (2021). "Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly". *Entropy* 23.11: *Approximate Bayesian Inference*. ISSN: 1099-4300. DOI: 10.3390/e23111534. arXiv: 1805.07418 [stat.ML]. URL: https://www.mdpi.com/1099-4300/23/11/1534 (pp. 13, 76).

Li, L., Guedj, B., and Loustau, S. (2018). "A quasi-Bayesian perspective to online clustering". *Electron. J. Statist.* 12.2, pp. 3071–3113. DOI: 10.1214/18-EJS1479. URL: https://doi.org/10.1214/18-EJS1479 (pp. 7, 13, 36, 76).

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). "Challenging common assumptions in the unsupervised learning of disentangled representations". *International Conference on Machine Learning* (p. 44).

Logeswaran, L. and Lee, H. (2018). "An Efficient Framework for Learning Sentence Representations". *ICLR* (p. 42).

Lugosi, G. and Neu, G. (2022). "Generalization Bounds via Convex Analysis". *Conference on Learning Theory (COLT)* (p. 30).

Masegosa, A. R., Lorenzen, S. S., Igel, C., and Seldin, Y. (2020). "Second Order PAC-Bayesian Bounds for the Weighted Majority Vote". *NeurIPS* (pp. 26, 27).

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. (2019). "Disentangling disentanglement in variational autoencoders". *International Conference on Machine Learning* (p. 44).

Maurer, A. (Nov. 2004). "A Note on the PAC Bayesian Theorem". *arXiv*. DOI: 10.48550/arxiv.cs/0411099 (p. 4).

McAllester, D. A. (July 1998). "Some PAC-Bayesian Theorems". *Proc. Conf. Learn. Theory (COLT)*. Madison, WI, USA (p. 3).

McAllester, D. A. (July 1999). "PAC-Bayesian Model Averaging". *Proc. Conf. Comp. Learn. Theory (COLT)*. Santa Cruz, CA, USA. DOI: 10.1145/307400.307435 (p. 3).

Mhammedi, Z., Grünwald, P., and Guedj, B. (2019). "PAC-Bayes Un-Expected Bernstein Inequality". *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems [NeurIPS] 2019, 8-14 December 2019, Vancouver, BC, Canada*. Ed. by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, pp. 12180–12191. URL: https://papers.nips.cc/paper/9387-pac-bayes-un-expected-bernstein-inequality (pp. 7, 9, 12, 24, 36).

Mhammedi, Z., Guedj, B., and Williamson, R. C. (2020). “PAC-Bayesian Bound for the Conditional Value at Risk”. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems [NeurIPS] 2020, December 6-12, 2020, virtual*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. arXiv: 2006.14763 [cs.LG]. URL: https://proceedings.neurips.cc/paper/2020/hash/d02e9bdc27a894e882fa0c9055c99722-Abstract.html (pp. 7, 12, 36).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). “Distributed Representations of Words and Phrases and their Compositionality”. *NeurIPS* (pp. 37, 38).

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. 2nd ed. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press (p. 3).

Monge, G. (1781). “Mémoire sur la théorie des déblais et des remblais”. *Histoire de l’Académie Royale des Sciences de Paris* (p. 30).

Müller, A. (1997). “Integral Probability Metrics and Their Generating Classes of Functions”. *Advances in Applied Probability* 29.2 (p. 30).

Neyshabur, B., Bhojanapalli, S., and Srebro, N. (2018a). “A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks”. *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=Skz%5C_WfbCZ (p. 19).

Neyshabur, B., Bhojanapalli, S., and Srebro, N. (2018b). “A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks”. *International Conference on Learning Representations* (p. 46).

Neyshabur, B., Tomioka, R., and Srebro, N. (July 2015). “Norm-Based Capacity Control in Neural Networks”. *Proc. Conf. Learn. Theory (COLT)*. Paris, France (p. 3).

Noroozi, M. and Favaro, P. (2016). “Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles”. *ECCV* (p. 37).

Nozawa, K., Germain, P., and Guedj, B. (2020). “PAC-Bayesian Contrastive Unsupervised Representation Learning”. *Conference on Uncertainty in Artificial Intelligence [UAI]*. arXiv: 1910.04464 [cs.LG]. URL: https://proceedings.mlr.press/v124/nozawa20a.html (pp. 7, 9, 10, 36, 37, 77).

Ohnishi, Y. and Honorio, J. (2021). “Novel Change of Measure Inequalities with Applications to PAC-Bayesian Bounds and Monte Carlo Estimation”. *International Conference on Artificial Intelligence and Statistics (AISTATS)* (p. 30).

Parrado-Hernández, E., Ambroladze, A., Shawe-Taylor, J., and Sun, S. (Dec. 2012). "PAC-Bayes Bounds with Data Dependent Priors". *Journal of Machine Learning Research (JMLR)* 13.112, pp. 3507–3531. DOI: 10.1007/978-3-7908-2604-3_21 (p. 25).

Pérez-Ortiz, M., Rivasplata, O., Guedj, B., Gleeson, M., Zhang, J., Shawe-Taylor, J., Bober, M., and Kittler, J. (2021a). "Learning PAC-Bayes Priors for Probabilistic Neural Networks". Submitted. arXiv: 2109.10304 [cs.LG]. URL: https://arxiv.org/abs/2109.10304 (pp. 8, 9, 12, 15, 16, 24).

Pérez-Ortiz, M., Rivasplata, O., Parrado-Hernandez, E., Guedj, B., and Shawe-Taylor, J. (2021b). "Progress in self-certified neural networks". *NeurIPS 2021 workshop Bayesian Deep Learning [BDL]*. arXiv: 2111.07737 [cs.LG]. URL: http://bayesiandeeplearning.org/2021/papers/38.pdf (pp. 8, 9, 12, 15, 16, 24).

Peyré, G. and Cuturi, M. (2019). "Computational Optimal Transport". *Foundations and Trends in Machine Learning* 11.5-6 (p. 33).

Picard-Weibel, A., Capson-Tojo, G., Guedj, B., and Moscoviz, R. (2023). "Bayesian Uncertainty Quantification for Anaerobic Digestion models". *Bioresource Technology* 394, pp. 130–147. ISSN: 0960-8524. DOI: 10.1016/j.biortech.2023.130147. arXiv: 2312.xxxxx [cs.LG]. URL: https://www.sciencedirect.com/science/article/pii/S0960852423015754 (pp. 13, 75).

Picard-Weibel, A. and Guedj, B. (2022). "On change of measure inequalities for f-divergences". Submitted. arXiv: 2202.05568 [stat.ML]. URL: https://arxiv.org/abs/2202.05568 (pp. 7, 12, 30, 36, 75).

Rezende, D. J. and Viola, F. (2018). "Taming VAEs". *arXiv preprint arXiv:1810.00597* (pp. 44, 45).

Rogers, W. H. and Wagner, T. J. (May 1978). "A Finite Sample Distribution-Free Performance Bound for Local Discrimination Rules". *The Annals of Statistics* 6.3, pp. 506–514. DOI: 10.1214/aos/1176344196 (p. 3).

Rothfuss, J., Fortuin, V., Josifoski, M., and Krause, A. (2021). "PACOH: Bayes-optimal meta-learning with PAC-guarantees". *International Conference on Machine Learning (ICML)* (p. 29).

Rothfuss, J., Josifoski, M., Fortuin, V., and Krause, A. (2022). "PAC-Bayesian Meta-Learning: From Theory to Practice". *arXiv* abs/2211.07206 (p. 29).

Russo, D. and Zou, J. (2020). "How Much Does Your Data Exploration Overfit? Controlling Bias via Information Usage". *IEEE Transactions on Information Theory* 66.1 (p. 30).

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. (2018). "Assessing generative models via precision and recall". *Advances in Neural Information Processing Systems* (p. 44).

Sakhi, O., Alquier, P., and Chopin, N. (2023). "PAC-Bayesian Offline Contextual Bandits With Guarantees". *International Conference on Machine Learning (ICML)* (p. 29).

Schrab, A., Guedj, B., and Gretton, A. (2022a). "KSD Aggregated Goodness-of-fit Test". *Advances in Neural Information Processing Systems [NeurIPS]*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 32624–32638. arXiv: 2202.00824 [stat.ML]. URL: https://papers.nips.cc/paper_files/paper/2022/hash/d241a7b1499cee1bf40769ceade2444d-Abstract-Conference.html (pp. 13, 75).

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2023). "MMD Aggregated Two-Sample Test". *Journal of Machine Learning Research [JMLR]* 24.194, pp. 1–81. arXiv: 2110.15073 [stat.ML]. URL: https://jmlr.org/papers/v24/21-1289.html (pp. 13, 75).

Schrab, A., Kim, I., Guedj, B., and Gretton, A. (2022b). "Efficient Aggregated Kernel Tests using Incomplete U-statistics". *Advances in Neural Information Processing Systems [NeurIPS]*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 18793–18807. arXiv: 2206.09194 [stat.ML]. URL: https://papers.nips.cc/paper_files/paper/2022/hash/774164b966cc277c82a960934445140d-Abstract-Conference.html (pp. 13, 75).

Schreuder, N., Brunel, V., and Dalalyan, A. S. (2021). "Statistical guarantees for generative models without domination". *International Conference on Algorithmic Learning Theory* (p. 44).

Seeger, M. (Oct. 2002). "PAC-Bayesian Generalisation Error Bounds for Gaussian Process Classification". *Journal of Machine Learning Research (JMLR)* 3, pp. 233–269 (p. 4).

Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J., and Auer, P. (2012). "PAC-Bayesian Inequalities for Martingales". *IEEE Transactions on Information Theory* 58.12 (p. 29).

Seldin, Y., Laviolette, F., Shawe-Taylor, J., Peters, J., and Auer, P. (2011). "PAC-Bayesian Analysis of Martingales and Multiarmed Bandits". *arXiv* abs/1105.2416 (p. 29).

Seldin, Y. and Tishby, N. (2010). "PAC-Bayesian Analysis of Co-clustering and Beyond". *Journal of Machine Learning Research* 11, pp. 3595–3646 (p. 42).

Shawe-Taylor, J. and Cristianini, N. (Mar. 1999). "Margin Distribution Bounds on Generalization". *Proc. European Conf. Comp. Learn. Theory (EuroCOLT)*. Nordkirchen, Germany. doi: 10.1007/3-540-49097-3_21 (p. 3).

Shawe-Taylor, J. and Williamson, R. C. (July 1997). "A PAC Analysis of a Bayesian Estimator". *Proc. Conf. Learn. Theory (COLT)*. Nashville, TN, USA. doi: 10.1145/267460.267466 (p. 3).

Shu, R., Bui, H. H., Zhao, S., Kochenderfer, M. J., and Ermon, S. (2018). "Amortized inference regularization". *Advances in Neural Information Processing Systems* (p. 44).

Suter, R., Miladinovic, D., Bauer, S., and Schölkopf, B. (2019). "Interventional robustness of deep latent variable models". *International Conference on Machine Learning* (p. 44).

Valiant, L. G. (Nov. 1984). "A Theory of the Learnable". *Commun. ACM* 27.11, pp. 1134–1142. doi: 10.1145/1968.1972 (p. 3).

Van Erven, T., Grünwald, P., Mehta, N., Reid, M., and Williamson, R. (Sept. 2015). "Fast rates in statistical and online learning". *Journal of Machine Learning Research (JMLR)* 16, pp. 1793–1861 (p. 7).

Vendeville, A., Giovanidis, A., Papanastasiou, E., and Guedj, B. (2023). "Opening up echo chambers via optimal content recommendation". *Complex Networks and Their Applications XI*. Ed. by H. Cherifi, R. N. Mantegna, L. M. Rocha, C. Cherifi, and S. Miccichè. Cham: Springer International Publishing, pp. 74–85. isbn: 978-3-031-21127-0. doi: 10.1007/978-3-031-21127-0_7. arXiv: 2206.03859 [cs.SI]. url: https://link.springer.com/chapter/10.1007/978-3-031-21127-0_7 (p. 76).

Vendeville, A., Guedj, B., and Zhou, S. (2021). "Forecasting elections results via the voter model with stubborn nodes". *Applied Network Science* 6. doi: 10.1007/s41109-020-00342-7. arXiv: 2009.10627 [cs.SI]. url: https://appliednetsci.springeropen.com/articles/10.1007/s41109-020-00342-7 (pp. 13, 76).

Vendeville, A., Guedj, B., and Zhou, S. (2022). "Towards control of opinion diversity by introducing zealots into a polarised social group". *Complex Networks & Their Applications X*. Ed. by R. M. Benito, C. Cherifi, H. Cherifi, E. Moro, L. M. Rocha, and M. Sales-Pardo. Springer International Publishing, pp. 341–352. isbn: 978-3-030-93413-2. doi: 10.1007/978-3-030-93413-2_29. arXiv: 2006.07265 [cs.SI]. url: https://link.springer.com/chapter/10.1007%2F978-3-030-93413-2_29 (pp. 13, 76).

Vendeville, A., Zhou, S., and Guedj, B. (2024). "Discord in the voter model for complex networks". *Physical Review E* 109 (2). doi: 10.

1103/PhysRevE.109.024312. arXiv: 2203.02002 [cs.SI]. URL: https://link.aps.org/doi/10.1103/PhysRevE.109.024312 (pp. 13, 76).

Viallard, P., Haddouche, M., Şimşekli, U., and Guedj, B. (2023). "Learning via Wasserstein-Based High Probability Generalisation Bounds". *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems [NeurIPS] 2023*. DOI: 10.48550/arXiv.2306.04375. arXiv: 2306.04375 [stat.ML]. URL: https://arxiv.org/abs/2306.04375 (pp. 7–10, 12, 24, 28, 36, 75).

Viallard, P., Haddouche, M., Şimşekli, U., and Guedj, B. (2024). "Tighter Generalisation Bounds via Interpolation". Submitted. DOI: 10.48550/ARXIV.2402.05101. arXiv: 2402.05101 [stat.ML]. URL: https://arxiv.org/abs/2402.05101 (pp. 7, 9, 12, 36, 75).

Villani, C. (2009). *Optimal transport: old and new*. Grundlehren der mathematischen Wissenschaften 338. Springer (p. 33).

Wang, H., Díaz, M., Filho, J. C. S. S., and Calmon, F. P. (2019). "An Information-Theoretic View of Generalization via Wasserstein Distance". *IEEE International Symposium on Information Theory (ISIT)* (p. 30).

Wei, J., Chen, Q., Peng, P., Guedj, B., and Li, L. (2022). "Reprint: a randomized extrapolation based on principal components for data augmentation". Submitted. arXiv: 2204.12024 [cs.CL]. URL: https://arxiv.org/abs/2204.12024 (p. 13).

Xiao, H., Rasul, K., and Vollgraf, R. (2017). "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms". *CoRR* abs/1708.07747. arXiv: 1708.07747. URL: https://arxiv.org/abs/1708.07747 (p. 19).

Xu, A. and Raginsky, M. (2017). "Information-theoretic analysis of generalization capability of learning algorithms". *Advances in Neural Information Processing Systems (NeurIPS)* (p. 30).

Zantedeschi, V., Viallard, P., Morvant, E., Emonet, R., Habrard, A., Germain, P., and Guedj, B. (2021). "Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound". *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems [NeurIPS] 2021*. Ed. by A. Beygelzimer, P. Liang, J. W. Vaughan, and Y. Dauphin. arXiv: 2106.12535 [cs.LG]. URL: https://proceedings.neurips.cc/paper/2021/hash/0415740eaa4d9decbc8da001d3fd805
Abstract.html (pp. 8–10, 24, 74).

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). "Understanding deep learning requires rethinking generalization". *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=Sy8gdB9xx (p. 19).

Zhang, J. M., Harman, M., Guedj, B., Barr, E. T., and Shawe-Taylor, J. (2023). “Model Validation Using Mutated Training Labels: An Exploratory Study”. *Neurocomputing* 539. DOI: 10.1016/j.neucom.2023.02.042. URL: https://www.sciencedirect.com/science/article/abs/pii/S0925231223001911 (p. 13).

Zhang, J., Liu, T., and Tao, D. (2018). “An Optimal Transport View on Generalization”. *arXiv* abs/1811.03270 (p. 30).

Zhang, R., Isola, P., and Efros, A. A. (2016). “Colorful Image Colorization”. *ECCV*, pp. 649–666 (p. 37).

Zhao, S., Ren, H., Yuan, A., Song, J., Goodman, N., and Ermon, S. (2018). “Bias and generalization in deep generative models: An empirical study”. *arXiv preprint arXiv:1811.03259* (p. 44).

A TEACHING, SUPERVISION, GRANT MANAGEMENT

To complement the main body of this manuscript which highlights my scientific contributions, I gather in this appendix details on my teaching contributions, supervision of students and grant management.

A.1 TEACHING

Inria positions come with no teaching duty. Out of personal inclination, I maintain since 2014 a (moderate) teaching activity at the master level. I find the live interaction with students extremely stimulating, and a great complement to research. On top of this, having to teach a topic is probably the fastest way to become an expert. More prosaically, teaching at the master-level is also an incredibly efficient way to recruit PhD students.

I have taught approximately 700 hours since 2010 (about 350 hours since joining Inria in 2014): as a research assistant in Denmark in 2010-2011 (100 hours), as a teaching assistant and "khôlleur" during my PhD at Sorbonne University (2011-2013, 250 hours), at the University of Lille, Centrale Lille, the University of Le Mans, Sorbonne University, ENSAE (2014-2019, 280 hours). I chose not to teach between 2019 and 2022 (due to a parental leave and the Covid-19 pandemic) and waited to be able to come back in front of students in person.

Since 2022, I have been teaching in the master's program (which I helped to establish) *MSc AI for sustainable development and biomedicine and healthcare* at UCL (30 hours in 2022, 30 hours in 2023). Since 2014, I have taught the following subjects: statistical learning theory, PAC-Bayes, computational statistics, deep learning, probabilistic modelling, exclusively at the master's level.

A.2 SUPERVISION OF STUDENTS

Since joining Inria in 2014, I have (co-)supervised **3 postdocs**, **10 PhD students** (including two on industrial CIFRE contracts), **2 research engineers**, and **23 master's interns**. The names of the current members of my team are underlined. For postdocs, PhD students, and engineers, I specify the source of funding: note that for PhD students, those funded on the French side have three years to complete their PhDs (with little flexibility to extend) to complete their thesis, while those funded on the British side have four years (with some flexibility to extend). I mention the prestigious journals and conferences where some of the work has been published.

Postdocs (3; 1 ongoing)

(a) **Fredrik Hellström** [web], since May 2023 (funded until May 2025), with a **100%** supervision share, funded by a WASP Fellowship (a prestigious Swedish government scholarship). Fredrik and I work on the theory of generalisation, mainly *via* PAC-Bayesian theory (of which I am an expert) and information theory (of which Fredrik is an expert). We have already written the monograph [**BG-Preprint15**] (accepted at Foundations and Trends in Machine Learning) and the paper [**BG-Conf24**] which has just been accepted at AISTATS 2024.

(b) **Valentina Zantedeschi** [web], January 2021 – July 2022, with a **100%** supervision share, funded by our ANR APRIORI. Valentina and I primarily worked on the design, theoretical study, and implementation of majority voting strategies in learning, with two papers [**BG-Conf11**]; [**BG-Conf14**] published at NeurIPS 2021 and 2022. Since August 2022, Valentina is a Senior Research Scientist at ServiceNow Research in Montreal (Canada).

(c) **Florent Dewez** [web], February 2019 – August 2021, with a **50%** supervision share (with Vincent Vandewalle, Inria), funded by our European project PERF-AI. With Florent, we primarily worked on the design and deployment of algorithms to optimise fuel consumption (notably by optimising trajectory) in aeronautics using real-time flight data, leading to savings in the order of 5% to 10% [**BG-Journal11**]; [**BG-Journal18**]. Florent is now a Senior Data Scientist at DiagRAMS Technologies, in Lille (France).

PhD students (10; 7 ongoing)

(a) **Valentin Kilian** [web], since October 2023, with a **40%** supervision share (with François Caron, University of Oxford), funded by the University of Oxford. Valentin works on learning on graphs, particularly the study of community detection algorithm properties, and the links with *geometric deep learning*. Viva planned for around Autumn 2027.

(b) **Théophile Cantelobre** [web], since September 2021, with a **50%** supervision share (with Alessandro Rudi, Inria), funded by my ANR JCJC BEAGLE. Théophile works on kernel methods learning, particularly on using similarity measures to promote robustness to invariants. Viva planned for the second half of 2024. Works [**BG-Preprint2**]; [**BG-Conf15**]; [**BG-Preprint13**] (notable: ICML).

(c) **Maxime Haddouche** [web], since September 2021, with a **100%** supervision share, funded by an ENS Cachan PhD scholarship. Maxime mainly works on PAC-Bayesian learning, on relaxing assumptions, extending to the online framework and measures like the Wasserstein distance, and also contributes to online optimisation and federated learning. Viva planned for the second half of 2024. Works [**BG-Preprint11**]; [**BG-Preprint10**]; [**BG-Conf23**]; [**BG-Preprint12**]; [**BG-Preprint16**]; [**BG-Preprint14**]; [**BG-Journal15**]; [**BG-Preprint3**]; [**BG-Conf17**]; [**BG-Journal21**] (notable: TMLR, NeurIPS x2).

(d) **Antoine Picard** [web], since September 2021, with a **50%** supervision share (with Roman Moscoviz, SUEZ), funded by the ANRT and SUEZ (**CIFRE thesis**). Antoine works on modelling and forecasting of anaerobic digestion for organic compounds, *via* PAC-Bayesian theory and particularly develops an online meta-learning strategy. Viva planned for the first semester of 2025. Works [**BG-Preprint7**]; [**BG-Journal23**].

(e) **Reuben Adams** [web], since September 2020, with a **90%** supervision share (with John Shawe-Taylor, UCL and UNESCO), funded by UCL CDT in Foundational AI. Reuben works on generalisation theory, mainly with PAC-Bayesian theory. Viva planned for the first semester of 2025. Works [**BG-Preprint5**].

(f) **Antonin Schrab** [web], since September 2020, with a **50%** supervision share (with Arthur Gretton, UCL and DeepMind), funded by UCL CDT in Foundational AI. Antonin works on adaptive statistical test algorithms in high dimensions. Viva planned for the second semester of 2024. Works [**BG-Journal24**]; [**BG-Conf18**]; [**BG-Conf19**] (notable: NeurIPS x2, JMLR).

(g) **Felix Biggs** [web], since September 2019, with a **90%** supervision share (with John Shawe-Taylor, UCL and UNESCO), funded by UCL CDT in Foundational AI. Felix works on the theory of generalisation, especially the design, analysis, and deployment of

PAC-Bayesian strategies for deep learning. Viva planned for March 2024. Works [**BG-Journal13**]; [**BG-Conf13**]; [**BG-Conf12**]; [**BG-Conf14**]; [**BG-Conf21**] (notable: NeurIPS, ICML, AISTATS x2).

(h) **Antoine Vendeville** [web], September 2019 – October 2023, with a **75%** supervision share (with Shi Zhou, UCL), funded by UCL CDT in Cybersecurity. Antoine worked on the analysis of information diffusion on interaction graphs (typically social networks), the emergence of the polarisation phenomenon, and on designing and analysing strategies to prevent and reduce the emergence of echo chambers. Antoine applied his work to content recommendation on social networks, and to forecasting election outcomes in the United States and France. His thesis is available here `https://theses.hal.science/tel-04431872`. Antoine is now a postdoctoral researcher at the MédiaLab of Sciences Po, Paris (France). Works [**BG-Conf22**]; [**BG-Conf20**]; [**BG-Journal17**]; [**BG-Journal26**]; [**BG-Conf22**] (notable: Physical Review E).

(i) **Arthur Leroy** [web], October 2017 – December 2020, with a **33%** supervision share (with Servane Gey and Pierre Latouche, Université Paris Descartes), funded by a doctoral contract from Université Paris Descartes. Arthur worked on the design, analysis, and implementation of forecasting and clustering algorithms for time series, with a model of multi-task Gaussian processes. His work was motivated by applications to sports performance and early detection of future athletes: the results of his thesis were passed on to the French Swimming Federation in view of the Paris 2024 Olympic Games. His thesis is available here: `https://arthur-leroy.netlify.app/files/Thesis-Arthur_LEROY.pdf`. Arthur is now a postdoctoral researcher at University of Manchester, UK under the supervision of Mauricio Alvarez. Works [**BG-Journal19**]; [**BG-Journal22**] (notable: JMLR).

(j) **Le Li** [web], October 2014 – November 2018, with a **50%** supervision share (with Sébastien Loustau, Université d'Angers), funded by the ANRT and iAdvize (**CIFRE thesis**). Le worked on the design, theoretical analysis, implementation, and deployment (for customer clustering on the iAdvize platform) of online clustering algorithms, partly through PAC-Bayesian theory. His thesis is accessible here: `https://tel.archives-ouvertes.fr/tel-01970795/`. Le is a lecturer at Central China Normal University, China. Works [**BG-Journal9**]; [**BG-Journal16**] (notable: EJS).

VISITING PHD STUDENTS (4; 1 ONGOING. I DO NOT HAVE ANY OFFICIAL SUPERVISION INVOLVEMENT)

(a) **Szilvia Ujvary** [web], since October 2023 (expected to finish in September 2024), PhD student at University of Cambridge (supervised by Ferenc Huszár), visiting to work on sampling methods for algorithms inspired by PAC-Bayes bounds.

(b) **Ludovic Arnould** [web], May 2022 – July 2022, PhD student at Sorbonne Université (supervised by Claire Boyer and Erwan Scornet), visiting to work on training neural networks via PAC-Bayes bounds.

(c) **Eugenio Clerico** [web], April 2022 – July 2022, PhD student at University of Oxford (supervised by Arnaud Doucet and George Deligiannidis), visiting to work on the Neural Tangent Kernel model and the derandomisation of PAC-Bayes bounds. Works [**BG-Preprint6**]; [**BG-Journal20**].

(d) **Kento Nozawa** [web], May 2019 – October 2019, PhD student at University of Tokyo (supervised by Issei Sato), visiting to work on contrastive learning and the first PAC-Bayes view of this problem. Works [**BG-Conf8**] (notable: UAI).

RESEARCH ENGINEERS (2)

(a) **Arthur Talpaert** [web], October 2019 - September 2020, with a **20%** supervision share (with Florent Dewez and Vincent Vandewalle, Inria), funded by our European project PERF-AI. Arthur implemented the Python library pyrotor, accompanying the paper [**BG-Journal18**] on aircraft trajectory optimisation. My role was to guide him and direct the project (jointly with Vincent Vandewalle). The library allows for the optimisation of a physical trajectory under constraint (we applied it to aeronautical and sailing navigation data). Arthur is a data science engineer at Data Prisme, France.

(b) **Bhargav Srinivasa Desikan** [web], October 2016 – September 2018, with a **100%** supervision share, funded by an engineering scholarship from the Hauts-de-France region (SLAP-ME project). Bhargav worked on extending, and implementing in Python, the COBRA algorithm [**BG-Journal4**] that I designed during my thesis. My role was to guide him and lead the project. The Python library pycobra automatically constructs a non-linear aggregate of preliminary predictors. Bhargav also worked on visualising psychomotor principles in human-machine interactions (with Stéphane Huot, Fanny Chevalier, Pierre Dragicevic from Inria). Works [**BG-Journal8**]; [**BG-Journal12**] (notable: JMLR MLOSS). Bhargav then completed a master's at the University of Chicago and started a PhD at EPFL (Switzerland).

MASTER'S INTERNS. I have supervised, for durations ranging from 3 months to 6 months, 23 Master's interns, including Valentin Kilian, Théophile Cantelobre, and Maxime Haddouche before taking them on as PhD students, as well as students from the University of Lille (Mostafa Bouziane, Wilfried Heyse in 2017-2018), from India (Astha Gupta in 2016), from the MVA of ENS Cachan (between 2017 and 2019, Adèle Gillier, Victor Sanh, Adrien Doumergue, Kawisorn Kamtue, Marc Etheve, Jean-Baptiste Remy; Juliette Rengot with whom I wrote [**BG-Conf5**]), from Université Paris-Saclay (in 2019, Louis Pujol with whom I wrote [**BG-Journal14**]), and from University College London (since 2022, Rita Kurban, Alexandra Udaltsova, Shoujing Zhu, Bjorn Kischelewski, Chloé Hashimoto-Cullen, Shahar Pelles, Naomi Fuchs, Kenza Benkirane, Ilai Bachrach). I supervised these students on a variety of topics in machine learning, generally more applied than my research, such as image inpainting, forecasting of landmines, automatic annotation of medical imaging for the prostate, survival modelling in statistics, modelling patient pathways in geriatrics in the Lille region, short-term forecasting of electricity consumption using few-shot learning, semantic characterisation of invariants in handwritten digit classification, and many others.

A.3 GRANT MANAGEMENT

Regardless of what one might think about this evolution, securing individual or project-based grants is now an essential component to research management, and is likely to remain so in the foreseeable future. I have been incredibly fortunate to progressively secure access to considerable funding, including the grants below.

AS A PI

- Inria INSIGHT Associated Team (2016-2018) with University College Dublin (Ireland), members were Christophe Biernacki from Inria, Brendan Murphy, and Nial Friel from UCD. Budget of

€10k for funding bilateral visits. We explored new model selection strategies in statistical learning.

- SLAP-ME (2017-2018), funded by the Hauts-de-France region (other members Stéphane Huot, Fanny Chevalier, Pierre Dragicevic). Budget of **€55k** for missions and the engineering position of Bhargav Srinivasa Desikan.

- Inria Associated Team 6PAC (2018-2021) with CWI (Netherlands). Members Emilie Kaufmann (Inria), Wooter Koolen, and Peter Grünwald from CWI. Budget of **€30k** to fund bilateral visits, working on extending the PAC-Bayesian framework to achieve fast rates.

- ANR JCJC BEAGLE (2018-2023). The success rate for obtaining funding in 2018 was 12%. Budget of **€180k** (equipment, missions, interns, and the thesis of Théophile Cantelobre), to work on PAC-Bayesian theory.

AS A CO-I

- ANR PRC APRIORI (2019-2024). The success rate for obtaining funding in 2018 was 12%. Members were Emilie Morvant (PI), Amaury Habrard, Rémi Emonet (University of Saint-Etienne), and Pascal Germain (Inria). Budget of **€300k**, of which **€120k** managed by me (missions, equipment, and the postdoctoral position of Valentina Zantedeschi), to work on learning representations via PAC-Bayesian theory.

- PERF-AI (2018-2020, European Commission, CleanSky 2 programme). The other members were Vincent Vandewalle (Inria) and the startup Safety Line (PI). Budget of **€700k**, with **€250k** managed by Vincent and me (equipment, the postdoctoral position of Florent Dewez, and the engineering position of Arthur Talpaert). Works [**BG-Journal11**]; [**BG-Journal18**].

- PEPR AI, SHARP project (2024-2028). Consortium of 8 co-Is (with PI Rémi Gribonval, Inria) to work on frugal learning. Total budget of **€7M**, with **€950k** managed by me, with the first recruitments of postdocs and PhD students planned for the end of 2024.

B COMPLETE LIST OF CONTRIBUTIONS

This appendix gathers all my publications, grouped by type. My complete list of publications is available on https://bguedj.github.io/publications/. I have (co-)authored **26** articles published in international journals, **24** articles published in international conferences, **16** preprints currently submitted to journals or conferences, **2** technical reports and **2** theses (MSc and PhD).

Articles in International Journals

[BG-Journal1] Guedj, B. and Guillot, G. (2011). "Estimating the location and shape of hybrid zones". *Molecular Ecology Resources* 11.6, pp. 1119–1123. DOI: 10.1111/j.1755-0998.2011.03045.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2011.03045.x (p. 13).

[BG-Journal2] Guedj, B. and Alquier, P. (2013). "PAC-Bayesian estimation and prediction in sparse additive models". *Electron. J. Statist.* 7, pp. 264–291. DOI: 10.1214/13-EJS771. URL: https://doi.org/10.1214/13-EJS771 (pp. 7, 11, 36).

[BG-Journal3] Chopin, N., Gadat, S., Guedj, B., Guyader, A., and Vernet, E. (2015). "On some recent advances on high dimensional Bayesian statistics". *ESAIM: Proceedings and Surveys* 51, pp. 293–319. DOI: 10.1051/proc/201551016. URL: https://doi.org/10.1051/proc/201551016 (p. 11).

[BG-Journal4] Biau, G., Fischer, A., Guedj, B., and Malley, J. D. (2016). "COBRA: A combined regression strategy". *Journal of Multivariate Analysis* 146. Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces, pp. 18–28. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2015.04.007. URL: http://www.sciencedirect.

com/science/article/pii/S0047259X15000950 (pp. 11, 77).

[BG-Journal5] Alquier, P. and Guedj, B. (2017). “An oracle inequality for quasi-Bayesian nonnegative matrix factorization”. *Mathematical Methods of Statistics* 26.1, pp. 55–67. ISSN: 1934-8045. DOI: 10.3103/S1066530717010045. URL: https://link.springer.com/article/10.3103%2FS1066530717010045 (pp. 7, 11, 36).

[BG-Journal6] Alquier, P. and Guedj, B. (2018). “Simpler PAC-Bayesian bounds for hostile data”. *Machine Learning* 107.5, pp. 887–902. ISSN: 1573-0565. DOI: 10.1007/s10994-017-5690-0. URL: https://doi.org/10.1007/s10994-017-5690-0 (pp. 7, 11, 30, 35, 36, 41, 42).

[BG-Journal7] Guedj, B. and Robbiano, S. (2018). “PAC-Bayesian high dimensional bipartite ranking”. *Journal of Statistical Planning and Inference* 196, pp. 70–86. ISSN: 0378-3758. DOI: https://doi.org/10.1016/j.jspi.2017.10.010. URL: http://www.sciencedirect.com/science/article/pii/S0378375817301945 (pp. 7, 11, 36).

[BG-Journal8] Guedj, B. and Srinivasa Desikan, B. (2018). “Pycobra: A Python Toolbox for Ensemble Learning and Visualisation”. *Journal of Machine Learning Research* 18.190, pp. 1–5. URL: http://jmlr.org/beta/papers/v18/17-228.html (pp. 11, 78).

[BG-Journal9] Li, L., Guedj, B., and Loustau, S. (2018). “A quasi-Bayesian perspective to online clustering”. *Electron. J. Statist.* 12.2, pp. 3071–3113. DOI: 10.1214/18-EJS1479. URL: https://doi.org/10.1214/18-EJS1479 (pp. 7, 13, 36, 76).

[BG-Journal10] Alliez, P., Di Cosmo, R., Guedj, B., Girault, A., Hacid, M.-S., Legrand, A., and Rougier, N. (Jan. 2020). “Attributing and Referencing (Research) Software: Best Practices and Outlook From Inria”. *Computing in Science & Engineering* 22.1, pp. 39–52. ISSN: 1558-366X. DOI: 10.1109/mcse.2019.2949413. URL: http://dx.doi.org/10.1109/MCSE.2019.2949413 (p. 14).

[BG-Journal11] Dewez, F., Guedj, B., and Vandewalle, V. (2020). “From industry-wide parameters to aircraft-centric on-flight inference: improving aeronautics performance prediction with machine learning”. *Data-Centric Engineering* 1. DOI: 10.1017/dce.2020.12. arXiv: 2005.05286 [stat.AP]. URL: https://www.cambridge.org/core/journals/data-centric-engineering/article/from-industrywide-parameters-to-aircraftcentric-onflight-inference-improving-aeronautics-performance-prediction-with-machine-learning/7A5662351D23A3D855E7FBC58B45AB6D (pp. 13, 74, 79).

[BG-Journal12] Guedj, B. and Srinivasa Desikan, B. (Jan. 2020). “Kernel-Based Ensemble Learning in Python”. *Information* 11.2, p. 63. ISSN: 2078-2489. DOI: 10.3390/info11020063. URL: http://dx.doi.org/10.3390/info11020063 (pp. 11, 78).

[BG-Journal13] Biggs, F. and Guedj, B. (2021). “Differentiable PAC-Bayes Objectives with Partially Aggregated Neural Networks”. *Entropy* 23.10: *Approximate Bayesian Inference*. ISSN: 1099-4300. DOI: 10.3390/e23101280. arXiv: 2006.12228 [cs.LG]. URL: https://www.mdpi.com/1099-4300/23/10/1280 (pp. 8, 9, 12, 15, 16, 24, 76).

[BG-Journal14] Guedj, B. and Pujol, L. (2021). “Still no free lunches: the price to pay for tighter PAC-Bayes bounds”. *Entropy* 23.11: *Approximate Bayesian Inference*. ISSN: 1099-4300. DOI: 10.3390/e23111529. arXiv: 1910.04460 [cs.LG]. URL: https://www.mdpi.com/1099-4300/23/11/1529 (pp. 7, 11, 36, 78).

[BG-Journal15] Haddouche, M., Guedj, B., Rivasplata, O., and Shawe-Taylor, J. (2021). “PAC-Bayes unleashed: generalisation bounds with unbounded losses”. *Entropy* 23.10: *Approximate Bayesian Inference*. ISSN: 1099-4300. DOI: 10.3390/e23101330. arXiv: 2006.07279 [stat.ML]. URL: https://www.mdpi.com/1099-4300/23/10/1330 (pp. 7, 12, 36, 75).

[BG-Journal16] Li, L. and Guedj, B. (2021). “Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly”. *Entropy* 23.11: *Approximate Bayesian Inference*. ISSN: 1099-4300. DOI: 10.3390/e23111534. arXiv:

1805.07418 [stat.ML]. URL: https://www.mdpi.com/1099-4300/23/11/1534 (pp. 13, 76).

[BG-Journal17] Vendeville, A., Guedj, B., and Zhou, S. (2021). “Forecasting elections results via the voter model with stubborn nodes”. *Applied Network Science* 6. DOI: 10.1007/s41109-020-00342-7. arXiv: 2009.10627 [cs.SI]. URL: https://appliednetsci.springeropen.com/articles/10.1007/s41109-020-00342-7 (pp. 13, 76).

[BG-Journal18] Dewez, F., Guedj, B., Talpaert, A., and Vandewalle, V. (2022). “An end-to-end data-driven optimisation framework for constrained trajectories”. *Data-Centric Engineering* 3. DOI: 10.1017/dce.2022.6. arXiv: 2011.11820 [stat.AP]. URL: https://www.cambridge.org/core/journals/data-centric-engineering/article/an-endtoend-datadriven-optimization-framework-for-constrained-trajectories/B4BF729E1681F795FFE79A1F4B5 (pp. 13, 74, 77, 79).

[BG-Journal19] Leroy, A., Latouche, P., Guedj, B., and Gey, S. (2022). “MAGMA: Inference and Prediction with Multi-Task Gaussian Processes”. *Machine Learning*. DOI: 10.1007/s10994-022-06172-1. arXiv: 2007.10731 [stat.CO]. URL: https://arxiv.org/abs/2007.10731 (pp. 13, 76).

[BG-Journal20] Clerico, E. and Guedj, B. (2023). “A note on regularised NTK dynamics with an application to PAC-Bayesian training”. *Transactions on Machine Learning Research [TMLR]*. ISSN: 2835-8856. DOI: 10.48550/ARXIV.2312.13259. arXiv: 2312.13259 [stat.ML]. URL: https://openreview.net/forum?id=2la55BeWwy (pp. 7, 12, 36, 77).

[BG-Journal21] Haddouche, M. and Guedj, B. (2023a). “PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales”. *Transactions on Machine Learning Research [TMLR]*. ISSN: 2835-8856. DOI: 10.48550/ARXIV.2210.00928. arXiv: 2210.00928 [stat.ML]. URL: https://openreview.net/forum?id=qxrwt6F3sf (pp. 7, 12, 31, 33, 36, 75).

[BG-Journal22] Leroy, A., Latouche, P., Guedj, B., and Gey, S. (2023). "Cluster-Specific Predictions with Multi-Task Gaussian Processes". *Journal of Machine Learning Research [JMLR]* 24.5, pp. 1–49. arXiv: 2011.07866 [cs.LG]. URL: https://jmlr.org/papers/v24/20-1321.html (pp. 13, 76).

[BG-Journal23] Picard-Weibel, A., Capson-Tojo, G., Guedj, B., and Moscoviz, R. (2023). "Bayesian Uncertainty Quantification for Anaerobic Digestion models". *Bioresource Technology* 394, pp. 130–147. ISSN: 0960-8524. DOI: 10.1016/j.biortech.2023.130147. arXiv: 2312.xxxxx [cs.LG]. URL: https://www.sciencedirect.com/science/article/pii/S0960852423015754 (pp. 13, 75).

[BG-Journal24] Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2023). "MMD Aggregated Two-Sample Test". *Journal of Machine Learning Research [JMLR]* 24.194, pp. 1–81. arXiv: 2110.15073 [stat.ML]. URL: https://jmlr.org/papers/v24/21-1289.html (pp. 13, 75).

[BG-Journal25] Zhang, J. M., Harman, M., Guedj, B., Barr, E. T., and Shawe-Taylor, J. (2023). "Model Validation Using Mutated Training Labels: An Exploratory Study". *Neurocomputing* 539. DOI: 10.1016/j.neucom.2023.02.042. URL: https://www.sciencedirect.com/science/article/abs/pii/S0925231223001911 (p. 13).

[BG-Journal26] Vendeville, A., Zhou, S., and Guedj, B. (2024). "Discord in the voter model for complex networks". *Physical Review E* 109 (2). DOI: 10.1103/PhysRevE.109.024312. arXiv: 2203.02002 [cs.SI]. URL: https://link.aps.org/doi/10.1103/PhysRevE.109.024312 (pp. 13, 76).

Articles in Peer-Reviewed International Conferences

[BG-Conf1] Guedj, B. (2019). "A Primer on PAC-Bayesian Learning". *Proceedings of the second congress of the French Mathematical Society*. Vol. 33. URL: https://arxiv.org/abs/1901.05353 (pp. 4, 11).

[BG-Conf2] Letarte, G., Germain, P., Guedj, B., and Laviolette, F. (2019). “Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks”. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems [NeurIPS] 2019, 8-14 December 2019, Vancouver, BC, Canada*. Ed. by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, pp. 6869–6879. eprint: https://arxiv.org/abs/1905.10259. URL: https://papers.nips.cc/paper/8911-dichotomize-and-generalize-pac-bayesian-binary-activated-deep-neural-networks (pp. 8–10, 15, 16, 24).

[BG-Conf3] Mhammedi, Z., Grünwald, P., and Guedj, B. (2019). “PAC-Bayes Un-Expected Bernstein Inequality”. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems [NeurIPS] 2019, 8-14 December 2019, Vancouver, BC, Canada*. Ed. by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, pp. 12180–12191. URL: https://papers.nips.cc/paper/9387-pac-bayes-un-expected-bernstein-inequality (pp. 7, 9, 12, 24, 36).

[BG-Conf4] Chrétien, S. and Guedj, B. (2020). “Revisiting clustering as matrix factorisation on the Stiefel manifold”. *LOD – The Sixth International Conference on Machine Learning, Optimization, and Data Science*. Ed. by G. Nicosia, V. Ojha, E. La Malfa, G. Jansen, V. Sciacca, P. Pardalos, G. Giuffrida, and R. Umeton. Springer International Publishing, pp. 1–12. ISBN: 978-3-030-64583-0. DOI: 10.1007/978-3-030-64583-0_1. URL: https://link.springer.com/chapter/10.1007%2F978-3-030-64583-0_1 (pp. 7, 13, 36).

[BG-Conf5] Guedj, B. and Rengot, J. (2020). “Non-linear aggregation of filters to improve image denoising”. *SAI: Intelligent Computing*. Ed. by K. Arai, S. Kapoor, and R. Bhatia. Springer International Publishing, pp. 314–327. ISBN: 978-3-030-52246-9. DOI: 10.1007/978-3-030-52246-9_22. arXiv: 1904.00865. URL:

https://link.springer.com/chapter/10.1007%2F978-3-030-52246-9_22 (pp. 11, 78).

[BG-Conf6] Klein, J., Albardan, M., Guedj, B., and Colot, O. (2020). "Decentralized Learning with Budgeted Network Load Using Gaussian Copulas and Classifier Ensembles". *ECML-PKDD 2019: Machine Learning and Knowledge Discovery in Databases*. Ed. by P. Cellier and K. Driessens. Springer International Publishing, pp. 301–316. ISBN: 978-3-030-43823-4. DOI: 10.1007/978-3-030-43823-4_26. URL: https://link.springer.com/chapter/10.1007%2F978-3-030-43823-4_26 (p. 13).

[BG-Conf7] Mhammedi, Z., Guedj, B., and Williamson, R. C. (2020). "PAC-Bayesian Bound for the Conditional Value at Risk". *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems [NeurIPS] 2020, December 6-12, 2020, virtual*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. arXiv: 2006.14763 [cs.LG]. URL: https://proceedings.neurips.cc/paper/2020/hash/d02e9bdc27a894e882fa0c9055c99722-Abstract.html (pp. 7, 12, 36).

[BG-Conf8] Nozawa, K., Germain, P., and Guedj, B. (2020). "PAC-Bayesian Contrastive Unsupervised Representation Learning". *Conference on Uncertainty in Artificial Intelligence [UAI]*. arXiv: 1910.04464 [cs.LG]. URL: https://proceedings.mlr.press/v124/nozawa20a.html (pp. 7, 9, 10, 36, 37, 77).

[BG-Conf9] Cohen-Addad, V., Guedj, B., Kanade, V., and Rom, G. (Apr. 2021). "Online k-means Clustering". *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics [AISTATS]*. Ed. by A. Banerjee and K. Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 1126–1134. URL: http://proceedings.mlr.press/v130/cohen-addad21a.html (p. 13).

[BG-Conf10] Pérez-Ortiz, M., Rivasplata, O., Parrado-Hernandez, E., Guedj, B., and Shawe-Taylor, J. (2021b). "Progress in self-certified neural networks". *NeurIPS 2021 workshop Bayesian Deep Learning [BDL]*. arXiv: 2111.07737 [cs.LG]. URL: http://bayesiandeeplearning.org/2021/papers/38.pdf (pp. 8, 9, 12, 15, 16, 24).

[BG-Conf11] Zantedeschi, V., Viallard, P., Morvant, E., Emonet, R., Habrard, A., Germain, P., and Guedj, B. (2021). "Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound". *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems [NeurIPS] 2021*. Ed. by A. Beygelzimer, P. Liang, J. W. Vaughan, and Y. Dauphin. arXiv: 2106.12535 [cs.LG]. URL: https://proceedings.neurips.cc/paper/2021/hash/0415740eaa4d9decbc8da001d3fd805f-Abstract.html (pp. 8–10, 24, 74).

[BG-Conf12] Biggs, F. and Guedj, B. (July 2022a). "Non-Vacuous Generalisation Bounds for Shallow Neural Networks". *Proceedings of the 39th International Conference on Machine Learning [ICML]*. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 1963–1981. arXiv: 2202.01627 [cs.LG]. URL: https://proceedings.mlr.press/v162/biggs22a.html (pp. 8–10, 15, 16, 18, 24, 76).

[BG-Conf13] Biggs, F. and Guedj, B. (28–30 Mar 2022b). "On Margins and Derandomisation in PAC-Bayes". *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics [AISTATS]*. Ed. by G. Camps-Valls, F. J. R. Ruiz, and I. Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, pp. 3709–3731. arXiv: 2107.03955 [cs.LG]. URL: https://proceedings.mlr.press/v151/biggs22a.html (pp. 7–9, 12, 15, 16, 19, 24, 36, 76).

[BG-Conf14] Biggs, F., Zantedeschi, V., and Guedj, B. (2022). "On Margins and Generalisation for Voting Classifiers". *Advances in Neural Information Processing Systems [NeurIPS]*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 9713–9726. arXiv: 2206.04607 [cs.LG]. URL: https://papers.nips.cc/paper_files/paper/2022/hash/3f8675af3da6da231c9e75b889b7f047-Abstract-Conference.html (pp. 8, 12, 24, 74, 76).

[BG-Conf15] Cantelobre, T., Ciliberto, C., Guedj, B., and Rudi, A. (17–23 Jul 2022). "Measuring dissimilarity with diffeomorphism invariance". *Proceedings of the 39th International Conference on Machine Learning [ICML]*. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 2572–2596. arXiv: 2202.05614 [stat.ML]. URL: https://proceedings.mlr.press/v162/cantelobre22a.html (pp. 13, 75).

[BG-Conf16] Chérief-Abdellatif, B.-E., Shi, Y., Doucet, A., and Guedj, B. (28–30 Mar 2022). "On PAC-Bayesian reconstruction guarantees for VAEs". *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics [AISTATS]*. Ed. by G. Camps-Valls, F. J. R. Ruiz, and I. Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, pp. 3066–3079. arXiv: 2202.11455 [cs.LG]. URL: https://proceedings.mlr.press/v151/cherief-abdellatif22a.html (pp. 7, 10, 36, 43).

[BG-Conf17] Haddouche, M. and Guedj, B. (2022). "Online PAC-Bayesian Learning". *Advances in Neural Information Processing Systems [NeurIPS]*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 25725–25738. arXiv: 2206.00024 [cs.LG]. URL: https://papers.nips.cc/paper_files/paper/2022/hash/a4d991d581accd2955a1e1928f4e6965-Abstract-Conference.html (pp. 7, 12, 29, 31, 32, 36, 75).

[BG-Conf18] Schrab, A., Guedj, B., and Gretton, A. (2022a). "KSD Aggregated Goodness-of-fit Test". *Advances in Neural Information Processing Systems [NeurIPS]*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 32624–32638. arXiv: 2202.00824 [stat.ML]. URL: https://papers.nips.cc/paper_files/paper/2022/hash/d241a7b1499cee1bf40769ceade2444d-Abstract-Conference.html (pp. 13, 75).

[BG-Conf19] Schrab, A., Kim, I., Guedj, B., and Gretton, A. (2022b). "Efficient Aggregated Kernel Tests using Incomplete

U-statistics". *Advances in Neural Information Processing Systems [NeurIPS]*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 18793–18807. arXiv: 2206.09194 [stat.ML]. URL: https://papers.nips.cc/paper_files/paper/2022/hash/774164b966cc277c82a960934445140d-Abstract-Conference.html (pp. 13, 75).

[BG-Conf20] Vendeville, A., Guedj, B., and Zhou, S. (2022). "Towards control of opinion diversity by introducing zealots into a polarised social group". *Complex Networks & Their Applications X*. Ed. by R. M. Benito, C. Cherifi, H. Cherifi, E. Moro, L. M. Rocha, and M. Sales-Pardo. Springer International Publishing, pp. 341–352. ISBN: 978-3-030-93413-2. DOI: 10.1007/978-3-030-93413-2_29. arXiv: 2006.07265 [cs.SI]. URL: https://link.springer.com/chapter/10.1007%2F978-3-030-93413-2_29 (pp. 13, 76).

[BG-Conf21] Biggs, F. and Guedj, B. (2023). "Tighter PAC-Bayes Generalisation Bounds by Leveraging Example Difficulty". *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics [AISTATS]*. Vol. 206. PMLR, pp. 8165–8182. DOI: 10.48550/arXiv.2210.11289. arXiv: 2210.11289 [cs.LG]. URL: https://proceedings.mlr.press/v206/biggs23a.html (pp. 7, 12, 36, 76).

[BG-Conf22] Vendeville, A., Giovanidis, A., Papanastasiou, E., and Guedj, B. (2023). "Opening up echo chambers via optimal content recommendation". *Complex Networks and Their Applications XI*. Ed. by H. Cherifi, R. N. Mantegna, L. M. Rocha, C. Cherifi, and S. Miccichè. Cham: Springer International Publishing, pp. 74–85. ISBN: 978-3-031-21127-0. DOI: 10.1007/978-3-031-21127-0_7. arXiv: 2206.03859 [cs.SI]. URL: https://link.springer.com/chapter/10.1007/978-3-031-21127-0_7 (p. 76).

[BG-Conf23] Viallard, P., Haddouche, M., Şimşekli, U., and Guedj, B. (2023). "Learning via Wasserstein-Based High Probability Generalisation Bounds". *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems [NeurIPS]*

2023. DOI: 10.48550/arXiv.2306.04375. arXiv: 2306.04375 [stat.ML]. URL: https://arxiv.org/abs/2306.04375 (pp. 7–10, 12, 24, 28, 36, 75).

[BG-Conf24] Hellström, F. and Guedj, B. (2024). "Comparing Comparators in Generalization Bounds". *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics [AISTATS]*. DOI: 10.48550/arXiv.2310.10534. arXiv: 2310.10534 [cs.LG]. URL: https://arxiv.org/abs/2310.10534 (pp. 7, 12, 36, 74).

Preprints

[BG-Preprint1] Celisse, A. and Guedj, B. (2016). "Stability revisited: new generalisation bounds for the Leave-one-Out". Preprint. URL: https://arxiv.org/abs/1608.06412 (pp. 7, 11, 36).

[BG-Preprint2] Cantelobre, T., Guedj, B., Pérez-Ortiz, M., and Shawe-Taylor, J. (2020). "A PAC-Bayesian Perspective on Structured Prediction with Implicit Loss Embeddings". Submitted. arXiv: 2012.03780 [cs.LG]. URL: https://arxiv.org/abs/2012.03780 (pp. 7, 12, 36, 75).

[BG-Preprint3] Haddouche, M., Guedj, B., and Shawe-Taylor, J. (2020). "Upper and Lower Bounds on the Performance of Kernel PCA". Submitted. arXiv: 2012.10369 [cs.LG]. URL: https://arxiv.org/abs/2012.10369 (pp. 12, 75).

[BG-Preprint4] Pérez-Ortiz, M., Rivasplata, O., Guedj, B., Gleeson, M., Zhang, J., Shawe-Taylor, J., Bober, M., and Kittler, J. (2021a). "Learning PAC-Bayes Priors for Probabilistic Neural Networks". Submitted. arXiv: 2109.10304 [cs.LG]. URL: https://arxiv.org/abs/2109.10304 (pp. 8, 9, 12, 15, 16, 24).

[BG-Preprint5] Adams, R., Shawe-Taylor, J., and Guedj, B. (2022). "Controlling Multiple Errors Simultaneously with a PAC-Bayes Bound". Submitted. arXiv: 2202.05560 [stat.ML]. URL: https://arxiv.org/abs/2202.05560 (pp. 7, 12, 36, 75).

[BG-Preprint6] Clerico, E., Farghly, T., Deligiannidis, G., Guedj, B., and Doucet, A. (2022). "Generalisation under gradient descent via deterministic PAC-Bayes". Submitted. DOI: 10.48550/ARXIV.2209.02525. arXiv: 2209.02525 [stat.ML]. URL: https://arxiv.org/abs/2209.02525 (pp. 7, 12, 36, 77).

[BG-Preprint7] Picard-Weibel, A. and Guedj, B. (2022). "On change of measure inequalities for f-divergences". Submitted. arXiv: 2202.05568 [stat.ML]. URL: https://arxiv.org/abs/2202.05568 (pp. 7, 12, 30, 36, 75).

[BG-Preprint8] Wei, J., Chen, Q., Peng, P., Guedj, B., and Li, L. (2022). "Reprint: a randomized extrapolation based on principal components for data augmentation". Submitted. arXiv: 2204.12024 [cs.CL]. URL: https://arxiv.org/abs/2204.12024 (p. 13).

[BG-Preprint9] Candiotto, L., Cayco-Gajic, N. A., Soarez, P. C. de, Luca, M. D., Frassinelli, D., Fujita, A., Growiec, J., Guedj, B., Kasturi, S. N., Kawanishi, Y., Kellmeyer, P., Livermore, M. A., Moodley, D., Rabinowitch, I., Ralitera, T., Stalnov, O., Taylor, H., and Wevers, M. (2023). "Nurturing Knowledge: A Virtue Epistemic Approach to Explainable AI". Submitted. DOI: 10.48550/arXiv.xxxx.xxxxx. arXiv: xxxx.xxxxx [cs.LG]. URL: https://arxiv.org/abs/xxxx.xxxxx (p. 14).

[BG-Preprint10] Haddouche, M. and Guedj, B. (2023b). "Wasserstein PAC-Bayes Learning: Exploiting Optimisation Guarantees to Explain Generalisation". Submitted. DOI: 10.48550/arXiv.2304.07048. arXiv: 2304.07048 [stat.ML]. URL: https://arxiv.org/abs/2304.07048 (pp. 7, 12, 36, 75).

[BG-Preprint11] Haddouche, M., Guedj, B., and Wintenberger, O. (2023). "Optimistically Tempered Online Learning". Submitted. DOI: 10.48550/arXiv.2301.07530. arXiv: 2301.07530 [cs.LG]. URL: https://arxiv.org/abs/2301.07530 (pp. 14, 75).

[BG-Preprint12] Jobic, P., Haddouche, M., and Guedj, B. (2023). "Federated Learning with Nonvacuous Generalisation Bounds". Submitted. DOI: 10.48550/arXiv.2310.11203. arXiv: 2310.11203 [cs.LG]. URL: https://arxiv.org/abs/2310.11203 (pp. 8, 9, 12, 24, 75).

[BG-Preprint13] Cantelobre, T., Ciliberto, C., Guedj, B., and Rudi, A. (2024). "Closed-form filtering for non-linear systems". Submitted. DOI: 10.48550/ARXIV.2402.09796. arXiv: 2402.09796 [stat.ML]. URL: https://arxiv.org/abs/2402.09796 (pp. 14, 75).

[BG-Preprint14] Haddouche, M., Viallard, P., Şimşekli, U., and Guedj, B. (2024). "A PAC-Bayesian Link Between Generalisation and Flat Minima". Submitted. DOI: 10.48550/ARXIV.2402.08508. arXiv: 2402.08508 [stat.ML]. URL: https://arxiv.org/abs/2402.08508 (pp. 7, 12, 36, 75).

[BG-Preprint15] Hellström, F., Durisi, G., Guedj, B., and Raginsky, M. (2024). "Generalization Bounds: Perspectives from Information Theory and PAC-Bayes". Accepted for publication. DOI: 10.48550/arXiv.2309.04381. arXiv: 2309.04381 [cs.LG]. URL: https://arxiv.org/abs/2309.04381 (pp. 3, 4, 11, 74).

[BG-Preprint16] Viallard, P., Haddouche, M., Şimşekli, U., and Guedj, B. (2024). "Tighter Generalisation Bounds via Interpolation". Submitted. DOI: 10.48550/ARXIV.2402.05101. arXiv: 2402.05101 [stat.ML]. URL: https://arxiv.org/abs/2402.05101 (pp. 7, 9, 12, 36, 75).

Technical Reports

[BG-TechReport1] Brotcorne, L., Canteaut, A., Carneiro Viana, A., Grandmont, C., Guedj, B., Huot, S., Issarny, V., Pallez, G., Perrier, V., Quema, V., Pomet, J.-B., Rival, X., Salvati, S., and Thomé, E. (Dec. 2020). *Indicateurs de suivi de l'activité scientifique de l'Inria*. Research Report. Inria. URL: https://hal.inria.fr/hal-03033764 (p. 14).

[BG-TechReport2] Guedj, B. (July 2021). *Covid-19 and AI: Unexpected Challenges and Lessons*. Research Report. URL: https://hal.inria.fr/hal-03277494 (p. 14).

Academic works

[BG-Academic1] Guedj, B. (2010). "A Bayesian modelling of the hybridization mechanism". MA thesis. Sorbonne Université and Danmarks Tekniske Universitet (p. 13).

[BG-Academic2] Guedj, B. (Dec. 2013). "Aggregation of estimators and classifiers: theory and methods". Theses. Université Pierre et Marie Curie - Paris VI. URL: https://tel.archives-ouvertes.fr/tel-00922353 (pp. xi, 11).