Rapport soumis aux rapporteurs, dans le but de sanctionner le dossier pour l'obtention du grade de

Docteur en Mathématiques Appliquées et Applications des Mathématiques

de l'Université de Lille par

Maxime Haddouche

PAC-Bayes Learning From an Optimisation Perspective

Devant le jury composé de :

Chrétien, Stéphane	Professeur	Université Lyon 2	Président
Chazal, Frédéric	Directeur de Recherche	Inria	Rapporteur
Germain, Pascal	Professor	Univesité Laval	Rapporteur
Biau, Gérard	Professeur	Sorbonne Université	Examinateur
Boyer, Claire	Professeure	Université Paris-Saclay	Examinatrice
Guedj, Benjamin	Directeur de Recherche	Inria, University College London	Directeur

Résumés de la thèse

Résumé vulgarisé. L'apprentissage PAC-Bayésien est une branche de la théorie de l'apprentissage récemment mise en avant pour ses garanties de généralisation des réseaux de neurones profonds permettant de mieux comprendre leur performances empiriques sur des exemples jamais vus par la machine auparavant. Cette théorie a été initialement développée dans le cadre de la théorie de l'information, qui peut s'avérer limitée pour comprendre précisément la capacité de généralisation des réseaux neuronaux profonds, capacité étant acquise via un processus d'optimisation souvent non-exploité dans les bornes PAC-Bayes. Dans cette thèse, une vision optimisatoire de l'apprentissage PAC-Bayes est proposée et développée à travers de nombreux algorithmes d'apprentissage et de bornes de généralisation, mettant en évidence les différentes interactions entre les bénéfices de la phase d'apprentissage et la généralisation.

Résumé complet. L'apprentissage PAC-Bayésien est une branche de la théorie de l'apprentissage récemment mise en avant pour ses garanties de généralisation des réseaux de neurones profonds permettant de mieux comprendre leur performances empiriques sur des examples jamais vus par la machine auparavant. Cette théorie a été initialement développée dans le cadre de la théorie de l'information, qui peut s'avérer limitée pour comprendre précisément la capacité de généralisation des réseaux neuronaux profonds, capacité étant acquise via un processus d'optimisation souvent non-exploité dans les bornes PAC-Bayes. Dans cette thèse, une vision optimisatoire de l'apprentissage PAC-Bayes est proposée et développée à travers de nombreux algorithmes d'apprentissage et de bornes de généralisation, mettant en évidence les différentes interactions entre les bénéfices de la phase d'apprentissage et la généralisation.
En effet, l'apprentissage PAC-Bayes est classiquement développé via la théorie de l'information, impliquant des quantités interprétées comme bayésiennes telles que la connaissance 'a priori'. Cela peut être difficile à concilier avec l'optimisation concrète des réseaux neuronaux profonds, qui implique l'optimisation d'un grand ensemble de paramètres et ne fait pas appel à l'apprentissage Bayésien. Pour combler cette lacune, nous remettons en question les interprétations et les hypothèses du PAC-Bayes issues de la théorie de l'information et proposons une nouvelle perspective basée sur l'optimisation.

Plus précisément, nous présentons l'apprentissage PAC-Bayes au chapitre 1, ainsi que notre nouvelle vision optimisatoire. Le chapitre 2 remet en question les hypothèses statistiques du PAC-Bayes. Le chapitre 3 introduit l'apprentissage PAC-Bayesien en ligne, qui permet de réduire l'impact de l'initialisation pendant le processus d'apprentissage. Le chapitre 4 atténue l'impact de l'optimisation dans l'apprentissage par lots en exploitant les "minima plats", un certain type de minima souvent atteint par les réseaux neuronaux profonds, ce qui permet de mieux comprendre la généralisation dans de telles structures. Le chapitre 5 montre qu'il est possible de combiner l'apprentissage PAC-Bayes et le transport optimal, ce qui permet d'incorporer directement des garanties d'optimisation dans une borne PAC-Bayes. Enfin, le chapitre 6 constitue un premier pas vers la pratique en mettant en œuvre de nouveaux algorithmes PAC-Bayes en ligne et par tas pour des Diracs, ce qui n'est pas possible lorsque la théorie de l'information est utilisée.

Thesis summaries

Lay summary. PAC-Bayesian learning is a branch of learning theory recently highlighted for its tight generalisation guarantees of deep neural networks, yielding a sharper understanding of their practical efficiency on a novel, unseen example. This theory was initially developed through information theory, which may prove to be limiting for understanding precisely the generalisation capacity of deep neural networks, acquired through an optimisation process. Indeed, a large part of the PAC-Bayesian literature does not dwell on the characteristics and positive impact of the learning phase to enrich the understanding of the generalisation phenomenon observed in practice. In this thesis, an optimisation-driven vision of PAC-Bayes learning is proposed and developed via numerous learning algorithms and generalisation bounds, highlighting various interplays between the benefits of the learning phase and generalisation.

Full summary. PAC-Bayesian learning is a branch of learning theory recently highlighted for its tight generalisation guarantees of deep neural networks, yielding a sharper understanding of their practical efficiency on a novel, unseen example. This theory was initially developed through information theory, which may prove to be limiting for understanding precisely the generalisation capacity of deep neural networks, acquired through an optimisation process. Indeed, a large part of the PAC-Bayesian literature does not dwell on the characteristics and positive impact of the learning phase to enrich the understanding of the generalisation phenomenon observed in practice. In this thesis, an optimisation-driven vision of PAC-Bayes learning is proposed and developed via numerous learning algorithms and generalisation bounds, highlighting various interplays between the benefits of the learning phase and generalisation.

Indeed, PAC-Bayes learning is classically developed through an information-theoretic prism involving in particular Bayesian quantities such as prior knowledge. This may be hard to fit with concrete optimisation procedure of deep neural networks, involving optimisation on large parameters set without the Bayesian paradigm. To fill this gap, we challenge the information-theoretic interpretations and assumptions disseminated within the PAC-Bayes literature by proposing an optimisation-based perspective.

More precisely, we introduce PAC-Bayes learning in Chapter 1, as well as our novel optimisation view. Chapter 2 challenges statistical assumptions of PAC-Bayes. Chapter 3, introduces Online PAC-Bayes Learning, allowing to reduce the impact of the initial-

isation during the learning process. Chapter 4 attenuates the impact of optimisation in batch learning by exploiting 'flat minima', a certain type of minima often reached by deep neural networks, providing a sharper understanding of generalisation in such structures. Chapter 5 shows that it is possible to mix up PAC-Bayes learning and optimal transport, allowing to directly incorporate optimisation guarantees in a PAC-Bayes generalisation bound. Finally, Chapter 6 is a first step towards practitioners by implementing novel batch and online PAC-Bayes algorithms for Dirac distribution, which is not possible by the information-theoretic approach of PAC-Bayes.

Acknowledgements

TODO

Contents

List of Notations

General

a	A scalar (integer or real)
$\mathbb{R}$	The set of real numbers
$\mathbb{R}^d$	The euclidean set of dimension d
$\|\cdot\|$	a norm of an euclidean set
$\operatorname{dist}(\cdot,\cdot)$	A distance on a Polish space.
$\mathbb{N}$	The set of natural numbers
∇f	the gradient of a function $f : \mathbb{R}^d \to \mathbb{R}$

Statistical Learning Theory

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

Probability Theory

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

PAC-Bayes framework

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

Optimal transport

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

List of Publications

Conference article

Paul Viallard, Maxime Haddouche, Umut Simsekli, and Benjamin Guedj. Learning via Wasserstein-Based High Probability Generalisation Bounds. (2023).

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

Journal article

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

Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, and John Shawe-Taylor. PAC-Bayes Unleashed: Generalisation Bounds with Unbounded Losses. *Entropy*. (2021).

Research Report

Maxime Haddouche, Paul Viallard, Umut Simsekli, and Benjamin Guedj. A PAC-Bayesian Link Between Generalisation and Flat Minima. (2024).

Paul Viallard, Maxime Haddouche, Umut Şimşekli, and Benjamin Guedj. Tighter Generalisation Bounds via Interpolation. (2024).

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

Maxime Haddouche, Olivier Wintenberger, and Benjamin Guedj. Optimistically Tempered Online Learning. (2023).

Pierre Jobic, Maxime Haddouche, and Benjamin Guedj. Federated Learning with Nonvacuous Generalisation Bounds. (2023).

Maxime Haddouche, Benjamin Guedj, and John Shawe-Taylor. Upper and Lower Bounds on the Performance of Kernel PCA. (2020).

Préambule: Apprentissage humain, Apprentissage Machine et Généralisation

Ce manuscrit étudie la question de la capacité de *généralisation* des algorithmes d'apprentissage machine. Pour comprendre la généralisation, il faut d'abord appréhender l'*apprentissage*, prenons donc le luxe, pour un bref instant, d'oublier les machines pour se concentrer sur l'apprentissage en ce qu'il a de plus humain.

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

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

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

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

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

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

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

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

Preamble: Human Learning, Machine Learning and Generalisation

This manuscript tackles the notion of *generalisation* a notion built upon the general notion of *learning*. For a brief moment, let's take the luxury of forgetting about machines and concentrate on learning at its most human.

Apprehending human learning A human being (here a learner) is structured around experiences, either lived or passed on by others.

The learner then benefits from these experiences in various ways, for instance, by considering a mediated experience to be true (fire burns) and acting according to this. On the contrary, reiteration or denial of this same information may be symptoms of zero truth value. These scenarios can just as easily appear for a lived experience (the question of hallucinations). This first dichotomy in information processing is intrinsically linked to a clearly stated question: does fire burn? Can I trust my senses or have I hallucinated? In these cases, learning has taken place by reducing the intrinsic complexity of an experience to its truth value *w.r.t.* a simple question (in this case with two outcomes). This vision can easily be extended to a finite tree of possibilities through multiple-choice questions. Indeed, we can extend the burning question as follows: what is the intensity of the burn as a function of the temperature of the fire? We can then establish a multitude of answers representing various degrees of burn.

However, many questions cannot be reduced to a finite number of possibilities. For example, what is fire? To answer this question, it is nevertheless possible to exploit multiple facets of experience (wood, twig, rock fire) to propose that fire is the chemical reaction of oxygen in the air with a combustible material, with a supply of energy serving as the trigger.

Then, a legitimate question is: why has mankind understood the nature of fire? This fundamental understanding emerged from practical considerations: how can we stop being cold? Can we eat meat other than raw to reduce the risk of illness? It then takes multiple interactions with the environment to generate experiences and then learn from them to gradually respond to a complex need (how to make a fire to keep yourself warm?).

Thus, through this preliminary analysis, we have found several premises of understanding human learning.

- How is learning formalised structurally? The learner must base the experience on simple questions to acquire primary certainties. These latter acquired, it is possible to reach complex questions by interweaving more and more elementary considerations.
- Where does the need to learn come from? From a practical point of view, the emergence of these complex questions often arises from a relationship between the being and its environment, making it possible to develop contextual objectives. The learner then gradually becomes capable of responding to complex needs through a succession of simple actions.

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

From learning to generalisation. Generalisation can be seen as the ability to exploit learning from experience beyond that experience. This encompasses a theoretical and axiomatic understanding of a phenomenon, *i.e.* a fruitful extrapolation, or the ability to exploit the knowledge acquired for a new, yet showing similarities, situations *i.e.* to interpolate experiences.
This dual aspect of generalisation can be found in both humans and machines in a variety of ways. Deep neural networks, which are the spearhead of modern machine learning, are based on finite-dimensional learning spaces, which means that a problem can be learned through a finite number of founding principles. Since the number of principles can be increased as far as numerical capacity allows, we can say that neural networks have discrete generalising power. Given that machine learning methods are

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

What to expect from generalisation in modern machine learning? So what can we expect from artificial neural networks and their ability to generalise to humans? Universal approximation theorems (see *e.g.* Lu *et al.*, 2017; Park *et al.*, 2021) ensure that neural networks are capable of approximating any function living in a space to the power of the continuum (*e.g.* the space of continuous functions with compact support which does not admit a countable base as a Banach space), making these structures promising candidates for partially understanding human generalisation mechanisms. In the near future, machine approximations will potentially be powerful enough to give the illusion of human generalisation capacity. Nevertheless, it is worth bearing in mind that, if the thesis of a fundamental inequality in nature between the powers of human and machine generalisation is confirmed, then artificial neural networks will never fully attain the world-understanding capacities of their human counterparts. It is stll worth noticing artificial neural nets ability to approximate this human intelligence makes these structures valuable assistants, enriching the capabilities of any individual. That being said, this manuscript aims to provide a better understanding of generalisation in machine learning, quantifying and indentifying the mechanisms that promote it.

PAC-Bayes Learning, a field of many paradigms 1

Contents

1.1 A brief introduction to statistical learning

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

- First, the *population risk* $\mathsf{R}_{\mathcal{D}}(h) := \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h, \mathbf{z})]$ focus on the average performance of our learning agent *w.r.t.* any new situation $\mathbf{z} \in \mathcal{D}$, independent of $\mathcal{S}_m$, possibly faced by the agent. A small population risk ensure then efficient generalisation.

- Second, the *generalisation gap* $\Delta_{\mathcal{S}_m}(h) := \mathsf{R}_D(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)$ evaluate the coherence between the empirical risk and the population one. Having a small generalisation gap ensure that the generalisation ability of the agent has the same magnitude than its training performance.

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

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

$$\mathbb{E}_{\mathcal{S}_m}\left[\Delta_{\mathcal{S}_m}(h_{\mathcal{S}_m})\right] \leq f\left(\textsc{Complexity}, \frac{1}{m}\right). \tag{1.1}$$

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

$$\Delta_{\mathcal{S}_m}(h_{\mathcal{S}_m}) \leq f\left(\textsc{Complexity}, \frac{1}{m}, \log\frac{1}{\delta}\right). \tag{1.2}$$

The nature of f and the COMPLEXITY term depend on the facet of the complexity of the learning problem we aim to focus. Celebrated examples are for instance the dimension of $\mathcal{H}$, if euclidean, the VC dimension of $\mathcal{H}$ (VAPNIK, 2000), the Rademacher complexity (BARTLETT and MENDELSON, 2001, 2002), the stability parameter of a learning algorithm (BOUSQUET and ELISSEEFF, 2000) or the subgaussian diameter of $\mathcal{Z}$ (KONTOROVICH, 2014). Another approach relies on the Bayesian learning paradigm, deriving *posterior* knowledge from data and prior modelling of the environment. Then, the COMPLEXITY term can be borrowed from information theory (COVER and THOMAS, 2001), *e.g.* mutual information (NEAL, 2012), or from optimal transport, *e.g.* Wasserstein distances (WANG *et al.*, 2019; RODRIGUEZ-GALVEZ *et al.*, 2021).
Those two approaches have various benefits. A notable strength of expected bounds is that they may reach fast convergence rates (*i.e.* faster than $\frac{1}{\sqrt{m}}$) contrary to high-probability one, even when $\mathcal{H}$ is a singleton thanks to the central limit theorem (GRUNWALD *et al.*, 2021). However, expected bounds often involves a theoretical COMPLEXITY which cannot be estimated in practice and may be hard to interpret while high probability bounds may be fully empirical and can be considered with small confidence parameter δ as it is attenuated by a logarithm.

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

> **Example 1.1.1** (VC dimension of multilayer perceptrons)**.** Consider a supervised learning problem where $\mathcal{Z} = \mathbb{R}^k \times \mathcal{Y}$ with $\mathcal{Y} = \{0, 1\}$, k smaller than m and with loss $\ell(h, (x, y)) = \mathbb{1}\{h(x) \neq y\}$. First, assume that $\mathcal{H}$ is the set of linear classifiers; *i.e.* $\mathcal{H}_1 := \{h_\theta(x) = sgn(\langle \theta, x \rangle)\}$, where $sgn(a)$ denotes the sign of a. In this case, using the VC dimension may lead to non-vacuous generalisation bounds (VAPNIK, 2000).
> However, in modern machine learning, deep neural networks are often considered, let us first define a celebrated class of deep neural networks.
>
> > **Definition 1.1.1** (Multlilayer perceptron)**.** A multilayer perceptron with depth K and architecture $\{N_1, \cdots, N_K\}$, denoted as $h_{\mathbf{w}}(\mathbf{x}) := Wh^K(\cdots h^1(\mathbf{x})) + b$, is composed of K layers $h^1(\cdot), \ldots, h^K(\cdot)$. $W \in \mathbb{R}^{|\mathcal{Y}| \times N_K}$ and $b \in \mathbb{R}^{N_K}$ are the weight matrix and the bias of the last layer, and the i-th layer h^i, composed of N_i nodes, is defined by $h^i(\mathbf{x}) := \sigma_i(W_i\mathbf{x} + b_i)$, where $W_i \in \mathbb{R}^{N_i \times N_{i-1}}$ and the bias $b_i \in \mathbb{R}^{N_i}$ are its weight matrix and bias respectively; $\sigma_i : \mathbb{R}^{N_i} \to \mathbb{R}^{N_i}$ is an activation function. The weights $\mathbf{w} = \mathrm{vec}(\{W, W_K, \ldots, W_1, b, b_K, \ldots, b_1\})$ represent the vectorisation of all parameters of the network.
>
> Now, consider the learning problem with the same $\mathcal{Z}, \ell$ as above, but with $\mathcal{H}_2$ being the set of multilayer perceptrons *w.r.t.* a fixed depth K and architecture $\{N_1, \cdots, N_K\}$. To be consistent with modern practice, assume also that we are in the *overparametrised setting*, meaning that the space $\mathcal{H}_2$ has a dimension d far greater than m. In this case, VC dimension fails to explain the good generalisation ability (seen in practice) of multilayer perceptrons (BARTLETT and MAASS, 2003).

Understanding the generalisation ability of deep neural networks remains nowadays a major challenge and in what follows, we focus on a modern branch of learning theory which provided non-vacuous bounds of the generalisation ability of deep neural networks: PAC-Bayes learning.

1.2 An information-theoretic exposure of PAC-Bayes learning

PAC-Bayes learning is a recent branch of learning theory which emerged in the late 90s via the seminal work of (SHAWE-TAYLOR and WILLIAMSON, 1997; MCALLESTER, 1998, 1999, 2003b) and later pursued by (CATONI, 2003, 2007). Modern surveys are available to describe the various advances in the field (GUEDJ, 2019; HELLSTRÖM *et al.*, 2023; ALQUIER, 2024). Similarly to subfields of statistical learning described in Section 1.1, PAC-Bayes provide generalisation bounds involving a COMPLEXITY term, inspired here from the Bayesian learning paradigm of designing a *posterior* knowledge of the learning problem based on both training data and a *prior* knowledge of the considered situation.

A concrete example of Bayesian learning would be an explorer mapping an ill-known territory. The explorer has to adapt the existing maps at its disposal before exploration to its discoveries. Doing so, he creates an *a posteriori* map imbricating the benefits of both the prior knowledge alongside its findings.

From a mathematical perspective, the Bayes approach relies on the Bayes formula, providing an update recipe from a prior distribution $\mathrm{P} \in \mathcal{M}(\mathcal{H})$ over the predictor space $\mathcal{H}$ to a posterior $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ through a likelihood. On the contrary, PAC-Bayes, while inspired from the Bayesian philosophy, relies historically on tools from information theory. This general approach benefits from additional flexibility as PAC-Bayes can be linked and applied to Bayesian learning (see GUEDJ, 2019) but also blurs the notion of prior and posterior distributions, now independent of the fundamental Bayes formula. We further develop those points through two celebrated high-probability bounds: the McAllester and Catoni ones.

Two fundamental results

The McAllester's bound (MCALLESTER, 2003b) enriched with Maurer's trick (MAURER, 2004) and Catoni's bound (ALQUIER *et al.*, 2016, Theorem 4.1, being a relaxation of CATONI, 2007, Theorem 1.2.6) are probably the most known high-probability PAC-Bayes bounds. We recall them in Proposition 1.2.1.

Proposition 1.2.1 (McAllester and Catoni's bounds). Assume $\mathcal{S}_m$ to be *i.i.d.*.

McAllester's bound, (Maurer, 2004, Theorem 5). For any $\delta \in (0,1), \ell \in [0,1]$, any data-free prior $\mathrm{P} \in \mathcal{M}(\mathcal{H})$, with probability at least $1-\delta$, for any posterior $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$,

$$\Delta_{\mathcal{S}_m}(\mathrm{Q}) \leq \sqrt{\frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \ln \frac{2\sqrt{m}}{\delta}}{2m}}. \tag{1.3}$$

Catoni's bound, (Alquier *et al.*, 2016, Theorem 4.1). For any $\lambda \in \mathbb{R}/\{0\}, \delta \in (0,1), \ell$ being σ^2-subgaussian and a data-free prior P, with probability at least $1-\delta$ over $\mathcal{S}$, for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$,

$$\Delta_{\mathcal{S}_m}(\mathrm{Q}) \leq \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda} + \frac{\lambda\sigma^2}{2m}. \tag{1.4}$$

For both results, $\Delta_{\mathcal{S}_m}(\mathrm{Q})$ denotes the expected generalisation gap *w.r.t.* Q and KL denotes the Kullback-Leibler divergence.

Recall that a random variable X is σ^2-subgaussian if for any $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$ and that any loss $\ell \in [0, C]$ is C-subgaussian. Both McAllester and Catoni bounds fit the general shape of (1.2). In both cases, $\textsc{Complexity} = \mathrm{KL}(\mathrm{Q},\mathrm{P})$ and f varies. The immediate link with the Bayesian philosophy of learning is that the prior has to be data-free. However, (1.3) and (1.4) are both valid simultaneously for any posterior, which is strictly more general than considering the Bayesian posterior. Note that if λ is optimised, then Catoni's bound would boil down to an upgraded McAllester bound without the $\log(\sqrt{m})$ term, but such an optimisation is not feasible as λ has to be chosen independently of the dataset $\mathcal{S}_m$. Note that this gap has been recently filled by Dupuis and Şimşekli (2024, Theorem 33). While the theoretical links between those two bounds are clear, they involve two different toolboxes: McAllester's bound heavily relies on the KL divergence between Bernoullis alongisde calculation tricks exploiting the boundedness of the loss while the original Catoni's bound (Catoni, 2007, Theorem 1.2.6) exploits tools from statistical physics. The relaxation (1.4) proposed here is reachable by a few key arguments, involved in a vast majority of PAC-Bayes proofs. We propose it below for pedagogical purpose.

Proof of Equation (1.4). Note that the first part of the proof holds for a large part of PAC-Bayes literature.

A generic pattern for PAC-Bayes bounds. This part is designed upon two cornerstones, retrievable in many existing results: the change of measure inequality (Csiszár, 1975; Donsker and Varadhan, 1976 – see also Banerjee, 2006; Guedj, 2019 for a proof) and Markov's inequality.

Lemma 1.2.1 (Change of measure inequality)**.** For any measurable function $\psi : \mathcal{H} \to \mathbb{R}$ and any distributions Q,P on $\mathcal{H}$:

$$\mathbb{E}_{h\sim\mathrm{Q}}[\psi(h)] \leq \mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log\left(\mathbb{E}_{h\sim\mathrm{P}}[\exp(\psi(h))]\right).$$

For a given $\lambda > 0$, the change of measure inequality is then applied to a certain function $f_m : \mathcal{H} \to \mathbb{R}$, possibly involving $\mathcal{S}_m$: for all posteriors Q,

$$\mathbb{E}_{h\sim \mathrm{Q}}[f_m(h)] \leq \mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log\left(\mathbb{E}_{h\sim \mathrm{P}}[\exp(f_m(h))]\right). \tag{1.5}$$

To deal with the random variable $X(\mathcal{S}_m) := \mathbb{E}_{h\sim\mathrm{P}}[\exp(f_m(h))]$, our second building block is Markov's inequality $\left(\mathbb{P}(X > a) \leq \frac{\mathbb{E}[X]}{a}\right)$ which we apply for a fixed $\delta \in (0,1)$ on $X(\mathcal{S}_m)$ with $a = \mathbb{E}_{\mathcal{S}_m}[X(\mathcal{S}_m)]/\delta$. Taking the complementary event gives that for any m, with probability at least $1-\delta$ over the sample $\mathcal{S}_m$, $X(\mathcal{S}_m) \leq \mathbb{E}_{\mathcal{S}_m}[X(\mathcal{S}_m)]/\delta$, thus:

$$\mathbb{E}_{h\sim \mathrm{Q}}[f_m(h)] \leq \mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta) + \log\left(\mathbb{E}_{h\sim P}\mathbb{E}_{\mathcal{S}_m}[\exp(f_m(h))]\right). \tag{1.6}$$

Note that in (1.6), we swapped the two expectations in the last term thanks to Fubini's theorem and the fact that P is data-free.

Proving Catoni's bound. Now, we take $f_m(h) = \lambda\Delta_{\mathcal{S}_m}$ and consider for any $h \in \mathcal{H}, A(h) = \mathbb{E}_{\mathcal{S}_m}[\exp(f_m(h))]$.

Note that, given $\mathcal{S}_m$ is iid,

$$A(h) = \prod_{i=1}^{m} \mathbb{E}_{\mathcal{S}_m}\left[\exp\left(\frac{\lambda}{m}(\mathsf{R}_{\mathcal{D}}(h) - \ell(h,\mathbf{z}_i))\right)\right],$$

and thanks to Heoffding's lemma alongside ℓ being σ^2-subgaussian,

$$A(h) \leq \prod_{i=1}^{m} \exp\left(\frac{\lambda^2\sigma^2}{2m^2}\right) = \exp\left(\frac{\lambda^2\sigma^2}{2m}\right).$$

Plugging this upper bound in (1.6) and dividing by λ concludes the proof. ■

The generic pattern (1.6), allows to retrieve many PAC-Bayes bounds, starting with McAllester's one, where $f_m = kl(\mathsf{R}_{\mathcal{D}}(h), \hat{\mathsf{R}}_{\mathcal{S}_m}(h)), kl$ being the KL divergence between Bernoullis and completing with the subtle calculations of Maurer (2004). This pattern is also valid, for instance, for the results of Germain *et al.* (2009), the Bernstein PAC-Bayesian bounds of Tolstikhin and Seldin (2013) and Mhammedi *et al.* (2019) and many other results, *e.g.* Thiemann *et al.* (2017), Guedj and Robbiano (2018), Holland (2019), and Wu and Seldin (2022). This then pins two major points for a large part of PAC-Bayes literature:

1. Interpreting PAC-Bayes from a Bayesian point of view is legitimated by the change of measure inequality, yet the KL divergence. More generally, this property allows interpreting PAC-Bayes under a more general information-theoretic paradigm, where relevant prior information is transferred to the posterior (here

by absolute continuity to keep the KL finite). This information-theoretic vision is also retrieved in in-expectation PAC-Bayes bounds, where mutual information can be considered instead of KL divergence (Russo and Zou, 2016; Xu and Raginsky, 2017; Hellström and Durisi, 2020; Steinke and Zakynthinou, 2020; Grunwald *et al.*, 2021; Hellström and Durisi, 2022).

2. The statistical properties of the learning problem are linked to the exponential moment coming from the change of measure inequality, this often implies the strong assumptions of Proposition 1.2.1: data-free prior, bounded or subgaussian losses (sometimes attenuated to subexponentiality Catoni, 2004).

A theory suited for Example 1.1.1? The two previous points show that Proposition 1.2.1 holds for learning problem with light-tailed losses (often bounded), *i.i.d.* data, encompassing classification tasks for instance. Then, PAC-Bayes learning seems suited to understand, on such problems, the McAllester and Catoni bounds are suited to the learning problem $(\mathcal{H}_2, \mathcal{Z}, \ell)$ of Example 1.1.1.
However, the question of their tightness is unsolved as we do not know the behavior of the KL term in practice. Furthermore the question of which distribution Q should be taken in Proposition 1.2.1 remains open. Hopefully, PAC-Bayes bounds can be transformed into learning algorithms.

1.3 From theory to learning algorithms

Algorithms associated to McAllester and Catoni bounds

A shared particularity of McAllester and Catoni bounds is that they are both fully empirical. Then it is possible to minimise them in practice and thus, deriving new theory-driven learning algorithms which are expected to have at worse, a small generalisation gap and at best, a small population risk. More precisely, learning algorithms associated to Proposition 1.2.1 are stated below:

$$\mathrm{Q}_M := \operatorname*{argmin}_{\mathrm{Q}\in\mathcal{C}} \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \sqrt{\frac{\mathrm{KL}(\mathrm{Q},\mathrm{P})}{2m}}. \tag{1.7}$$

For any $\lambda > 0$,

$$\mathrm{Q}_C := \operatorname*{argmin}_{\mathrm{Q}\in\mathcal{C}} \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P})}{\lambda}. \tag{1.8}$$

In both cases, $\mathcal{C} \subseteq \mathcal{M}(\mathcal{H})$ is the class of distributions on which we optimise. The choice of $\mathcal{C}$ may come from prior knowledge of the problem or from optimisation concerns to make the KL divergence tractable.

Knowing Catoni's bound is a relaxation of McAllester's one, it seems more natural to consider Q_M over Q_C. However, the presence of a square root in (1.7) can be challenging for practical optimisation. We illustrate this below.

> **Example 1.3.1** (A celebrated class of measures for PAC-Bayes algorithms)**.** Consider the case where, for a given $\sigma > 0$, $\mathcal{C} = \{\mathcal{N}(\mu, \sigma^2 \mathrm{Id}) \mid \mu \in \mathbb{R}^d\}$. Then the for any $\mathrm{P} = \mathcal{N}(\mu_1, \sigma^2\mathrm{Id}), \mathrm{Q} = \mathcal{N}(\mu_2, \sigma^2\mathrm{Id})$, $\mathrm{KL}(\mathrm{Q},\mathrm{P}) = \frac{\|\mu_1 - \mu_2\|^2}{2\sigma^2}$. Then, optimising (1.7) in this case implies to lose the strong convexity of the KL divergence while it is retained for (1.8).

Another practical advantage of (1.8) over (1.7) emerges when $\mathcal{C} = \mathcal{M}(\mathcal{H})$. In this case, Catoni's bound admits a closed form solution, while McAllester's one should be numerically optimised on all the space of distributions, which is not feasible. This closed form, extracted from Catoni (2003, Section 5.1), is recalled below.

$$\text{When } \mathcal{C} = \mathcal{M}(\mathcal{H}),\ d\mathrm{Q}_C(h) = \frac{\exp(-\lambda \hat{\mathsf{R}}_{\mathcal{S}_m}(h))}{\mathbb{E}_{h\sim\mathrm{P}}[\exp(-\lambda\hat{\mathsf{R}}_{\mathcal{S}_m}(h))]} d\mathrm{P}(h) \tag{1.9}$$

Then, $\mathrm{Q}_C = \mathrm{P}_{-\lambda\hat{\mathsf{R}}_{\mathcal{S}_m}}$ is the *Gibbs posterior* associated to $\mathrm{P}, \lambda\hat{\mathsf{R}}_{\mathcal{S}_m}$. By introducing Gibbs posterior in statistical learning, Catoni (2007) draws a theoretical link between statistical physics and learning theory. Unfortunately, Gibbs posteriors often require Monte Carlo methods to be implemented, which can be time-consuming. Below, we then focus on PAC-Bayes algorithms working on a subset $\mathcal{C}$ of $\mathcal{M}(\mathcal{H})$.

Instantiation and efficiency of PAC-Bayesian algorithms

A general pattern for PAC-Bayesian algorithms The introductory examples (1.7),(1.8) unveil a general design for any KL-based PAC-Bayesian algorithm, satisfying a trade-off between *(i)* the empirical risk, showing that the learner has to fit the training dataset, and *(ii)* a *regulariser* being a function of $\mathrm{KL}(\mathrm{Q},\mathrm{P})$. This regulariser ensures that, during training, the learner will not overfit on training data. This training ensures a good generalisation ability as long as the associated generalisation bound is small.

While the conceptual ins and outs of PAC-Bayes algorithms are getting clearer, two unanswered questions remains:

1. How are those algorithms instantiated in practice?
2. Are these algorithms efficient and do they come with non-vacuous theoretical guarantees?

Instantiating a PAC-Bayes algorithm In practice, using a single prior P usually does not work, but it remains theoretically possible to consider a finite set of priors. Indeed, if one wants to consider k priors, then it is possible to consider k PAC-Bayes bounds holding for each of those priors with probability at least $1-\frac{\delta}{k}$ and then consider a union bound, such a set of priors is called a grid. This method has been widely used in many PAC-Bayes work with clever grids, deteriorating initial bounds at the cost of supplementary $\log(n)$ or $\log\log(n)$ (divided by m), see *e.g.* ALQUIER (2024). This can also be used, for Catoni-typed algorithms, to the parameter λ. In both cases, considering grids allows optimising on both the prior, the posterior and possibly λ when involved. Then, taking the closest value of those optimised parameters on the grid to still obtain theoretical guarantees. Another technique to ensure a good prior is to sacrifice a part of the training set to pre-train P. Doing so, the prior is then data-dependent and yields tighter bounds alongside increased performance (PEREZ-ORTIZ *et al.*, 2021a,c).

Efficiency of PAC-Bayes algorithms on supervised learning problems. The work of DZIUGAITE and ROY (2017) showed that optimising (1.7) when $\mathcal{C}$ is a class of Gaussian measures for the weights of a deep neural network yields non-vacuous generalisation bound, meaning that the generalisation benefits of PAC-Bayesian training on deep nets can be theoretically ensured. Note that PAC-Bayesian bounds can also be used to quantify the generalisation ability of other learning algorithms, but the bound value is then suboptimal. DZIUGAITE and ROY (2017) used the toolbox described in the 'instantiation' paragraph, alongside a preliminary use of Stochastic Gradient Descent (SGD) to update Q before the PAC-Bayes training algorithm. This promising work paved the way to various extensions, providing non-vacuous guarantees for a wide range PAC-Bayes algorithms (LETARTE *et al.*, 2019; RIVASPLATA *et al.*, 2019; DZIUGAITE *et al.*, 2021; PEREZ-ORTIZ *et al.*, 2021a,b,c; BIGGS and GUEDJ, 2022a, 2023), showing that the PAC-Bayes toolbox provides elements of answer to understand the generalisation ability of neural networks. Beyond generalisation guarantees, PAC-Bayes bounds are also useful to propose original training methods, even if the associated guarantees are vacuous (BIGGS and GUEDJ, 2021, 2022b). Another important empirical use is to exploit PAC-Bayes bounds as correlation measures, to see whether a decrease of the bound is related to an increased generalisation ability of the learner. For instance NEYSHABUR *et al.* (2017) used McAllester's bound (1.3) as a 'flatness' measure and showed that it correlates well with a good generalisation ability for a few learning problems. This conclusion has been extended to a wider range of problems in DZIUGAITE *et al.* (2020) and JIANG *et al.* (2020).

PAC-Bayes algorithms beyond supervised learning. While supervised learning is a widely used to perform experiments in PAC-Bayes (often involving celebrated datasets such as MNIST or CIFAR-10), the McAllester bound holds for any learning problem with bounded loss, going beyond this setting. This theoretical flexibility has been exploited to derive PAC-Bayesian algorithm for various learning settings reinforcement learning (Fard and Pineau, 2010), multi-armed bandits (Seldin *et al.*, 2011, 2012b; Sakhi *et al.*, 2023), meta-learning (Amit and Meir, 2018; Ding *et al.*, 2021; Farid and Majumdar, 2021; Rothfuss *et al.*, 2021, 2022) to name but a few.

Is Example 1.1.1 tackled now? (Dziugaite and Roy, 2017) and following works have provided a positive answer by obtaining non-vacuous guarantees (sometimes tight) for $(\mathcal{H}_2, \mathcal{Z}, \ell)$ of Example 1.1.1 for various $\mathcal{Z}$ (being, *e.g.*, set of images for MNIST CIFAR-10 etc...). To obtain such guarantees, a PAC-Bayesian training needs to be performed to minimise its associated theoretical bound. That being said, several questions then legitimately emerge.

- Modern machine learning often implies learning problems where assumptions such as bounded (or subgaussian) losses or *i.i.d.* data do not hold. Is PAC-Bayes theory extendable beyond those assumptions?
- As shown in Dziugaite and Roy (2017), the PAC-Bayesian training is often combined to another procedure (*e.g.* ERM) to yield non-vacuous bounds. However, PAC-Bayes bounds do not bring the theoretical understanding of such additional methods, often outputting deterministic predictors (*i.e.* Dirac distributions). This kind of predictor is not allowed in (1.3), (1.4). Is it possible to obtain PAC-Bayes bounds valid for such methods?

1.4 An optimisation perspective of PAC-Bayes

The questions raised at the end of the previous part are important as they underline a gap between the information-theoretic approach of PAC-Bayes bounds and practical optimisation. A supplementary example of this is the grid required in practice to optimise the prior (and/or λ in Catoni's bound). Indeed, this hybrid solution is required to roughly fit theory,(exploiting a single prior) and practice (optimising freely the prior on a continuous space), while not being truly adapted to any of these settings. This then raises the following fundamental question:

Can we think PAC-Bayes learning from an optimisation perspective?

The elements of answer to this question are multiple. First, one can mix up PAC-Bayes argument with geometric properties of optimisation procedures to obtain generalisation bounds designed for specific algorithms including but not limited to, SGD, Langevin dynamics (London, 2017; Dziugaite and Roy, 2018a; Neu *et al.*, 2021; Clerico

et al., 2022; HAGHIFAM *et al.*, 2023; ZHOU *et al.*, 2023). Those works shows both convergence properties as well as minimax rates, showing the impact of PAC-Bayes learning to provide a better theoretical understanding of the generalisation ability of concrete algorithms.
A second approach consists in describe general principles that should be satisfied by the various terms and assumptions in PAC-Bayes when looking at this through the prism of optimisation. We propose such an analysis below.

An optimisation-driven view of PAC-Bayes

- **Statistical assumptions.** While ℓ satisfies desirable geometric properties (convexity, gradient lipschitz ...), no statistical assumption is needed to have optimisation algorithms with convergence properties, on then may wonder about the generalisation ability of the reached empirical minima. It happens that the output of two runs of a stochastic optimisation algorithm on the same training set may vary a lot, for instance, the specific case of SGD shows that heavy-tailed behaviour (see *e.g.* ŞIMŞEKLI *et al.*, 2019; ZHANG *et al.*, 2020; GÜRBÜZBALABAN *et al.*, 2021) may emerge in practice. Given such behaviours, generalisation bounds, from an optimisation point of view, should hold with weak statistical assumptions on the dataset, possibly at the cost of additional geometric assumptions on the loss.

- **The role of the prior.** The information-theoretic approach justifies the Bayesian view of the prior, as discussed earlier. In this spirit it is also possible to sacrifice a part of the training set (*i.e.* of the available information) to enrich P. Doing so, we accept to not understand what happens during the training of P and thus, to explain only partially the efficiency of an information-theoretic training. Those two visions (Bayesian prior or data-dependent prior) are not easily linked to optimisation concerns as the first one would be linked to a 'good' initialisation, something we cannot know in advance, while the second makes little sense as P is obtained through a first, unexplained, optimisation process which is necessary to understand the efficiency of the second part of training, outputting Q. From an optimisation stance, we suggest assigning only two possible roles to P: *(i)* the initialisation of the optimisation algorithm, then its impact should be attenuated through the learning process and *(ii)* a minimum we aim to reach through optimisation. In this case, its impact is crucial as it translates the speed of convergence of our learning algorithms.

- **The place of stochastic predictors.** Involving a KL divergence as a complexity brings a particular focus on stochastic predictors, drawn from a distribution Q. Classical PAC-Bayes bounds usually focus on the average performance of such a predictor (hence the expectation over Q in (1.3),(1.4)), but recent extensions directly proposed guarantees for a single draw over Q (Rivasplata *et al.*, 2020; Viallard *et al.*, 2023a). However, involving a KL implies that Q has to be absolutely continuous *w.r.t.* P, meaning that the support of Q cannot go beyond the one of P: this excludes the case of Dirac distributions, *i.e.* deterministic predictors. This is a clear limitation of the information-theoretic approach, as many learning algorithms outputs a deterministic predictor and thus, should be avoided to be in line with common practice in optimisation.

Those three points, while not necessarily considered explicitly through the lens of optimisation have been recently challenged.

PAC-Bayes beyond the usual setting

Recall that according to what we saw in McAllester's bound (1.3) and Catoni's one (1.4), we denote by usual setting a bound holding for *i.i.d.* $\mathcal{S}_m$, with bounded or subgaussian losses and involving a KL divergence as Complexity term. Many works overcame at least one of this assumption as precised below.

Beyond *i.i.d.* data The work of Fard and Pineau (2010) established links between reinforcement learning and PAC-Bayes theory. This naturally led to the study of PAC-Bayesian bound for martingales instead of *i.i.d.* data (Seldin *et al.*, 2011, 2012a,b). Also, PAC-Bayesian bound for lifelong learning (Pentina and Lampert, 2014; Flynn *et al.*, 2022) challenged also the *i.i.d.* assumption. We also denote that the PAC-Bayes bound for meta learning (Amit and Meir, 2018; Ding *et al.*, 2021; Farid and Majumdar, 2021; Rothfuss *et al.*, 2021, 2022) consider independent but non-identically distributed datasets (corresponding to different tasks).

Avoiding light-tailed losses. Light-tailed losses encompass bounded, subgaussian, subexponential losses. Deriving PAC-Bayes bound for heavy-tailed losses, starting from Audibert and Catoni (2011) which provided PAC-Bayes bounds for least square estimators with heavy-tailed random variables. Their results was suboptimal with respect to the intrinsic dimension and was followed by further works from Catoni (2016) and Catoni and Giulini (2017). More recently, this question has been addressed in the works of Alquier and Guedj (2018), Holland (2019), Kuzborskij and Szepesvári (2019), and Haddouche *et al.* (2021), extending PAC-Bayes to heavy-tailed losses under additional technical assumptions.

Towards data-dependent priors. The work of (Catoni, 2007; Lever *et al.*, 2010, 2013) proposed priors, not directly data-dependent, but depending of the data distribution $\mathcal{D}$ when *i.i.d.* data are considered. can be informed by the data-generating distribution, Parrado-Hernández *et al.* (2012), Oneto *et al.* (2016), Dziugaite and Roy (2017), and Mhammedi *et al.* (2019) also obtained PAC-Bayes bound with data-dependent priors by infusing directly data in the prior (and sacrificing a part of the dataset). The drawback of this method is that, in practice, such a prior allows tighter bounds, but at the cost of a reduced theoretical understanding as the prior is in practice often learned via ERM, and the PAC-Bayes bound hardly gives insights on what happens during this pre-training. Furthermore, if this pre-training has already made the bound converge to a minimum generalising well, then the PAC-Bayes training has no effect and the associated bound is no more than a test bound (the case $\mathrm{Q} = \mathrm{P}$). It has been shown for instance in (Perez-Ortiz *et al.*, 2021a) that when P is trained with a consequent fraction of data, then the generalisation performance of the pre-trained P was roughly the same than Q, obtained from P after a PAC-Bayesian training. It is then unclear how impacting are PAC-Bayes methods compared to a test bound in this case. To alleviate this issue, another original route (Dziugaite and Roy, 2018b) exploits differential privacy to replace the data-free prior by a differentially private one, making possible to consider the prior as the learning objective (in their case a Gibbs posterior).

Beyond KL divergence. Several works allowed to extend PAC-Bayes beyond KL divergences. The most investigated route is to focus on the more general class of f-divergence, which include, *e.g.*, KL, χ^2, Rényi divergences among others (Alquier and Guedj, 2018; Ohnishi and Honorio, 2021; Picard-Weibel and Guedj, 2022; Viallard *et al.*, 2023a). However, f-divergences still implies absolute continuity of Q *w.r.t.* P. Another route recently emerged (Amit *et al.*, 2022), replacing f-divergences by integral probability metrics (IPMs), finally allowing Dirac distribution in PAC-Bayes.
These works have sometimes been explicitly driven by optimisation considerations (Dziugaite and Roy, 2018b involved differential privacy to numerically tighten their bound without sacrificing data). However, in many cases, the information-theoretic vision of PAC-Bayes remained majoritary. In what follows, the contributions of this manuscript are designed *w.r.t.* the optimisation view of PAC-Bayes detailed above.

Contributions of this thesis

The contributions of this manuscript are motivated by optimisation considerations and are structured as follows:

- In Chapter 2, we propose novel PAC-Bayes bounds for martingales, batch learning, with an application to multi-armed bandits. Those bounds are anytime-valid (*i.e.* for any dataset size simultaneously) and holds at the sole assumption of finite order 2 moments on both the posterior and the data distribution. Such weak statistical assumptions make these results applicable, for instance, for heavy-tailed SGD or many learning problems where optimisation procedure are performed regardless of the training set noise.

- Chapter 3 introduces *Online PAC-Bayes learning*, proposing theoretical bounds and learning algorithms involving a sequence of pairs $(\mathrm{Q}_i, \mathrm{P}_i)$, evolving through the optimisation process. Contrary to PAC-Bayes in a batch setting, the impact of $\mathrm{P} = \mathrm{P}_1$ is attenuated during the learning process, making Online PAC-Bayes useful when there is no prior information available, which is consistent with the vision of P as initialisation of a learning algorithm, while being only applicable, for now, to stochastic predictors as a KL divergence is involved.

- Chapter 4 still consider the prior as initialisation, while focusing on batch learning algorithms. It is shown that the impact of the prior is attenuated by *flat minima*, *i.e.* minima such that their neighbourhood nearly minimise the loss. More generally, this chapter exhibits theoretical links between flat minima and generalisation and thus draw links between the benefits of a successful optimisation process (small gradients) and generalisation.

- Considering P as the learning objective allows to draw more explicit links between optimisation and generalisation. In Chapter 5, it is shown that the convergence guarantees of *Bures-Wasserstein SGD*, a SGD-like algorithm on Gaussian measure spaces, can be directly incorporated within PAC-Bayes bounds, yielding interpretable results. This is possible by exploiting *Wasserstein PAC-Bayes learning*, which uses as $\textsc{Complexity}$ term a 1-Wasserstein distance, allowing to trade statistical assumptions to geometric ones such as lipschtz or gradient-lipschitz losses.

- Wasserstein PAC-Bayes learning can also be exploited when P is seen as an initialisation point. In Chapter 6, we propose Wasserstein PAC-Bayes algorithms with associated theoretical bound for both batch and online learning. A notable strength of these methods is that they hold for deterministic predictors (Dirac distributions), making PAC-Bayes in line with a large part of optimisation algorithms.

We finally recap in Figures 1.1 and 1.2 the classical information-theoretic vision of PAC-Bayes alongside the original optimisation view proposed above.

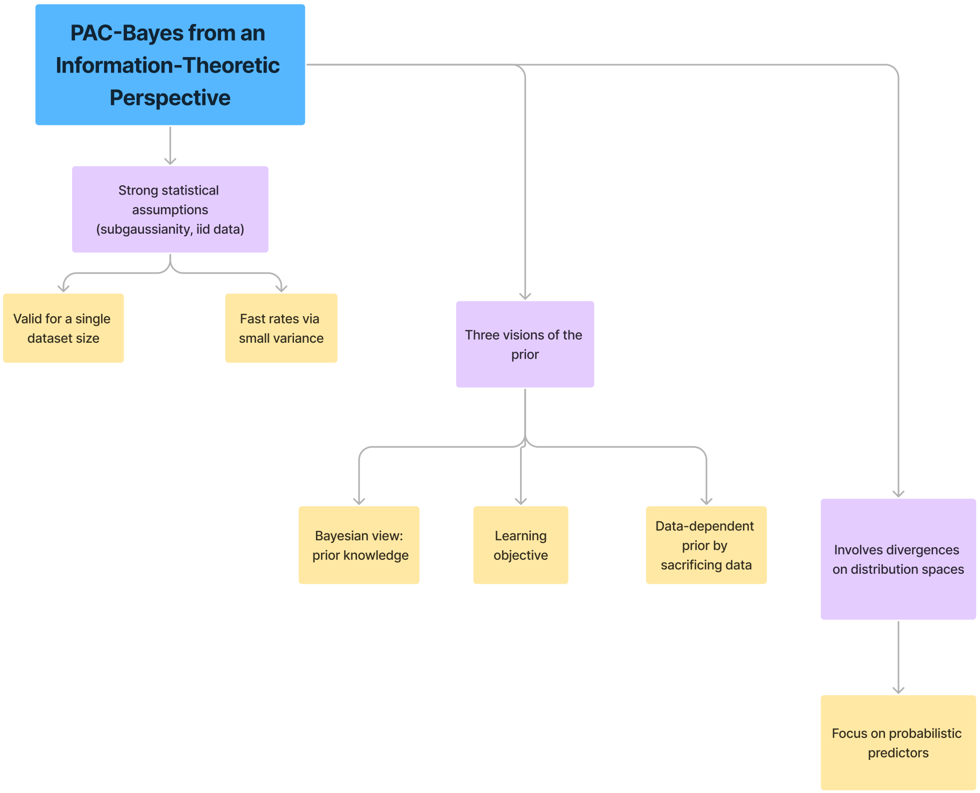
PAC-Bayes from an Information-Theoretic Perspective
Strong statistical assumptions (subgaussianity, iid data)
Valid for a single dataset size
Fast rates via small variance
Three visions of the prior
Bayesian view: prior knowledge
Learning objective
Data-dependent prior by sacrificing data
Involves divergences on distribution spaces
Focus on probabilistic predictors

Figure 1.1. *Recap of the information-theoretic vision of PAC-Bayes.*

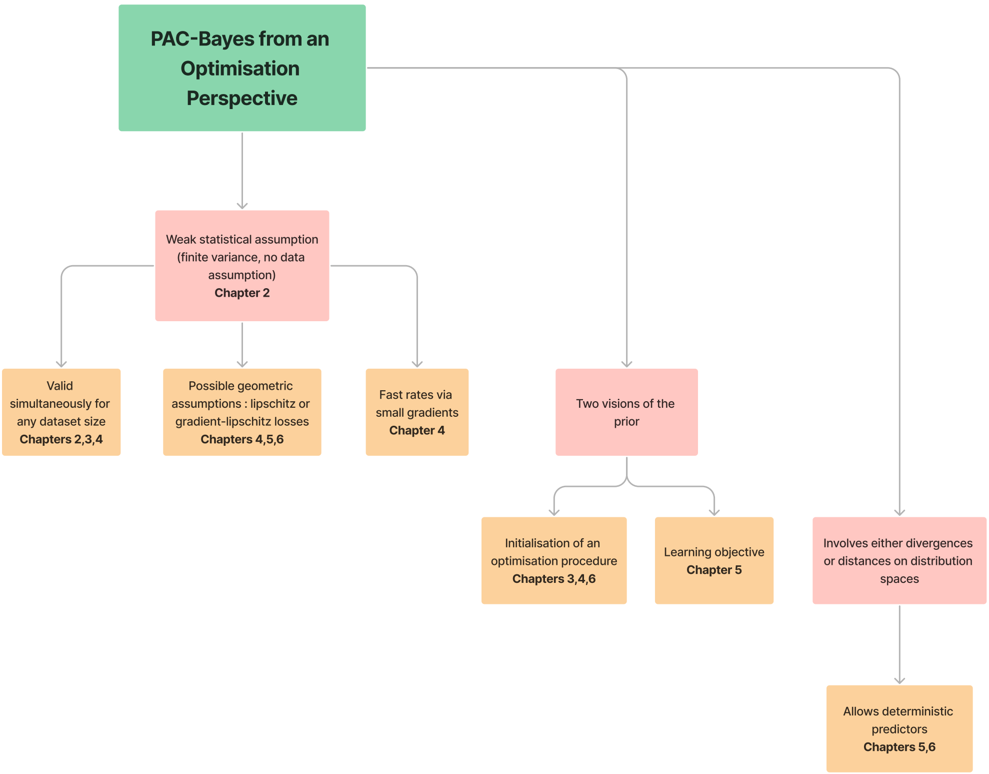
PAC-Bayes from an Optimisation Perspective
Weak statistical assumption (finite variance, no data assumption) Chapter 2
Valid simultaneously for any dataset size Chapters 2,3,4
Possible geometric assumptions : lipschitz or gradient-lipschitz losses Chapters 4,5,6
Fast rates via small gradients Chapter 4
Two visions of the prior
Initialisation of an optimisation procedure Chapters 3,4,6
Learning objective Chapter 5
Involves either divergences or distances on distribution spaces
Allows deterministic predictors Chapters 5,6

Figure 1.2. *Recap of the optimisation vision of PAC-Bayes and where those views are exploited in the manuscript.*

2

PAC-Bayes with Weak Statistical Assumptions: Generalisation Bounds for Martingales and Heavy-Tailed losses

This chapter is based on the following paper

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

Contents

Abstract

Chapter 2 provide PAC-Bayes bounds holding with weak statistical assumptions (finite variance), this is promising to encompass various learning situations involving optimisation algorithms such as heavy-tailed SGD (Gürbüzbalaban *et al.*, 2021) where assumptions such as bounded or subgaussian losses do not hold. Furthermore those results go beyond *i.i.d.* assumption on $\mathcal{S}$ and holds for all datasets $(\mathcal{S}_m)_{m\geq 1}$ simultaneously. Such a flexible setting is in line with various optimisation frameworks, where new data can be available after the beginning of the learning process and be incorporated on-the-fly to the ongoing

training, regardless of their potential correlation with previous data. Then, the theoretical results proposed in this chapter are a promising step toward practical settings where data may exhibit heavy-tailed behaviours and the loss function to be unbounded.

2.1 Introduction

In Chapter 1, McAllester's and Catoni's bound (McAllester, 2003b; Catoni, 2007) have been presented as key theoretical results with practical repercussions through their associated learning algorithm. However, the bounded or subgaussian assumption on the loss makes those results limited to tackle many real-life situations, which are limiting in practice. Indeed, from an optimisation perspective, as stated in Section 1.4 of Chapter 1, generalisation bounds should hold with weak statistical assumptions to make PAC-Bayes general enough to be used for learning settings where data are potentially heavy-tailed. Several works already proposed routes to overcome the boundedness constraint: Catoni (2004, Chapter 5) already proposed PAC-Bayes bounds for classification tasks and regressions ones with quadratic loss under a subexponential assumption. This technique has later been exploited in Alquier and Biau (2013) for the single-index model, and by Guedj and Alquier (2013) for nonparametric sparse additive regression, both under the assumption that the noise is subexponential. However all these works are dealing with light-tailed losses. Alquier and Guedj (2018), Holland (2019), Kuzborskij and Szepesvári (2019), and Haddouche *et al.* (2021) proposed extensions beyond light-tailed losses. This chapter stands in the continuation of this spirit while developing and exploiting a novel technical toolbox. To better highlight the novelty of our approach, we first present the two classical building blocks of PAC-Bayes.

2.1.1 Understanding PAC-Bayes: a celebrated route of proof

In the following subsection, we exploit again, for the sake of pedagagogy, the general pattern of proof for PAC-Bayes bounds described in Equation (1.4) to prove Catoni's bound.

2.1.1.1 Two essential building blocks for a preliminary bound

For the rest of this section, similarly to Chapter 1, we assume access to a non-negative loss function $\ell(h,z)$ taking as argument a predictor $h \in \mathcal{H}$ and data $z \in \mathcal{Z}$ (think of z as a pair input-output (x,y) for supervised learning problems, or as a single datum x in unsupervised learning). We also assume access to a m-sized sample $\mathcal{S}_m = (z_1, ..., z_m) \in \mathcal{Z}^m$. $\mathcal{S}_m$ is then used to learn a posterior distribution Q on $\mathcal{H}$, from a prior P.

PAC-Bayesian proofs are built upon two cornerstones. The first one is the change of measure inequality, recalled in Lemma 1.2.1. This property is applied to a certain function $f_m : \mathcal{Z}^m \times \mathcal{H} \to \mathbb{R}$ of the data and a candidate predictor: for all posteriors Q,

$$\mathbb{E}_{h\sim \mathrm{Q}}[f_m(\mathcal{S}_m, h)] \leq \mathrm{KL}(\mathrm{Q}, \mathrm{P}) + \log\left(\mathbb{E}_{h\sim \mathrm{P}}[\exp(f_m(\mathcal{S}_m, h))]\right). \tag{2.1}$$

To deal with the random variable $X(\mathcal{S}_m) := \mathbb{E}_{h\sim \mathrm{P}}[\exp(f_m(\mathcal{S}_m, h))]$, our second building block is Markov's inequality $\left(\mathbb{P}(X > a) \leq \frac{\mathbb{E}[X]}{a}\right)$ which we apply for a fixed $\delta \in (0,1)$ on $X(\mathcal{S}_m)$ with $a = \mathbb{E}_{\mathcal{S}_m}[X(\mathcal{S}_m)]/\delta$. Taking the complementary event gives that for any m, with probability at least $1-\delta$ over the sample $\mathcal{S}_m$, $X(\mathcal{S}_m) \leq \mathbb{E}_{\mathcal{S}_m}[X(\mathcal{S}_m)]/\delta$, thus:

$$\mathbb{E}_{h\sim \mathrm{Q}}[f_m(\mathcal{S}_m, h)] \leq \mathrm{KL}(\mathrm{Q}, \mathrm{P}) + \log(1/\delta) + \log\left(\mathbb{E}_{h\sim \mathrm{P}}\mathbb{E}_{\mathcal{S}_m}[\exp(f_m(\mathcal{S}_m, h))]\right). \tag{2.2}$$

2.1.1.2 From preliminary to complete bounds

From the preliminary result of Equation (2.2), there exists several ways to obtain PAC-Bayesian generalisation bounds, all being tied to specific choices of f and the assumptions on the dataset $\mathcal{S}_m$. However, they all rely on the control of an exponential moment implied by Markov's inequality: this is a strong constraint which has been at the heart of the classical assumption appearing in PAC-Bayes learning. For instance, McAllester's bound (1.3) and Catoni's bound (1.4), exploits in particular, a data-free prior, an *i.i.d.* assumption on $\mathcal{S}_m$ and a light-tailed loss. Most of the existing results stand with those assumptions (see *e.g.*, CATONI, 2007; GERMAIN *et al.*, 2009; GUEDJ and ALQUIER, 2013; TOLSTIKHIN and SELDIN, 2013; GUEDJ and ROBBIANO, 2018; MHAMMEDI *et al.*, 2019; WU and SELDIN, 2022). Indeed, in many of these works, either a boundedness or a subgaussian assumption on the loss is used. CATONI (2004) extended PAC-Bayes learning to the subexponential case. Many works tried to mitigate at least one of the following three assumptions.

- **Data-free priors.** With an alternative set of techniques, CATONI (2007) obtained bounds with localised (*i.e.*, data-dependent) priors. More recently, LEVER *et al.* (2010), PARRADO-HERNÁNDEZ *et al.* (2012), LEVER *et al.* (2013), ONETO *et al.* (2016), DZIUGAITE and ROY (2017), and MHAMMEDI *et al.* (2019) also obtained PAC-Bayes bound with data-dependent priors.
- **The *i.i.d.* assumption on $\mathcal{S}_m$.** The work of FARD and PINEAU (2010) established links between reinforcement learning and PAC-Bayes theory. This naturally led to the study of PAC-Bayesian bound for martingales instead of iid data (SELDIN *et al.*, 2011, 2012a,b).

- **Light-tailed loss.** PAC-Bayes bounds for heavy-tailed losses (*i.e.*, without subgaussian or subexponential assumptions) have been studied. Audibert and Catoni (2011) provide PAC-Bayes bounds for least square estimators with heavy-tailed random variables. Their results was suboptimal with respect to the intrinsic dimension and was followed by further works from Catoni (2016). More recently, this question has been adressed in the works of Alquier and Guedj (2018), Holland (2019), Kuzborskij and Szepesvári (2019), and Haddouche *et al.* (2021), extending PAC-Bayes to heavy-tailed losses under additional technical assumptions.

Several questions then legitimately arise.

Can we avoid these three assumptions simultaneously? The answer is yes: for instance the work of Rivasplata *et al.* (2020) proposed a preliminary PAC-Bayes bound holding with none of the three assumptions listed above. Building on their theorem, Haddouche and Guedj (2022) only exploited a bounded loss assumption to derive a PAC-Bayesian framework for online learning, requiring no assumption on data and allowing data (history in their context)-dependent priors.

Can we obtain PAC-Bayes bounds without the change of measure inequality? Yes, for instance Alquier and Guedj (2018) proposed PAC-Bayes bounds involving f-divergences and exploiting Holder's inequality instead of Lemma 1.2.1. More recently, Ohnishi and Honorio (2021) and Picard-Weibel and Guedj (2022) developed a broader discussion about generalising the change of measure inequality for a wide range of f-divergences. We note also that Germain *et al.* (2009) proposed a version of the classical route of proof stated above avoiding the use of the change of measure inequality. This comes at the cost of additional technical assumptions (see Haddouche *et al.*, 2021, Theorem 1 for a statement of the theorem in a proper measure-theoretic framework).

Can we avoid Markov's inequality? We mentioned above that several works avoided the change of measure inequality to obtain PAC-Bayesian bounds, but can we do the same with Markov's inequality? This is of interest as avoiding Markov could avoid assumptions such as subgaussiannity to provide PAC-Bayes bound. The answer is yes but this is a rare breed. To the best of our knowledge, only two papers are explicitly not using Markov's inequality: Kakade *et al.* (2008) obtained a PAC-Bayes bound using results on Rademacher complexity based on the McDiarmid concentration inequality, and Kuzborskij and Szepesvári (2019) exploited a concentration inequality from De la Peña *et al.* (2009), up to a technical assumption to obtain results for unbounded losses. Both of those works do not require a bound on an exponential moment to hold.

2.1.2 Originality of our approach

Avoiding Markov's inequality appears challenging in PAC-Bayes but leads to fruitful results as those in Kuzborskij and Szepesvári (2019).
In this work, we exploit a generalisation of Markov's inequality for supermartingales: Ville's inequality (as noticed by Doob 1939). This result has, to our knowledge, never been used in PAC-Bayes before.

Lemma 2.1.1 (Ville's maximal inequality for supermartingales)**.** Let $(\mathcal{F}_t)$ be a filtration adapted to (Z_t), a non-negative super-martingale with $Z_0 = 1$ almost surely, *i.e.* $(Z_t)_{t\geq 1}$ is a discrete process such that for any $t \in \mathbb{N}$, $\mathbb{E}\left[Z_t \mid \mathcal{F}_{t-1}\right] \leq Z_{t-1}$ a.s., $t \geq 1$, then, for any $0 < \delta < 1$, it holds

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

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

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

■

A major interest of Ville's result is that it holds for a countable sequence of random variables simultaneously. This point is new in PAC-Bayes and will allow us to obtain bounds holding for a countable (not necessarily finite) dataset $\mathcal{S}$.

On which supermartingale do we apply Ville's bound ? To fully exploit Lemma 2.1.1, we now take a countable dataset $\mathcal{S} = (\mathbf{z}_i)_{i\geq 1} \in \mathcal{Z}^{\mathbb{N}}$. Recall that, because we use the change of measure inequality, we have to deal with the following exponential random variable appearing in Eq. (2.1) for any $m \geq 1$:

$$Z_m := \mathbb{E}_{h\sim \mathrm{P}}[\exp(f_m(\mathcal{S}, h))].$$

Our goal is to choose a sequence of functions $f_m : \mathcal{Z}^{\mathbb{N}} \times \mathcal{H} \to \mathbb{R}$ such that $(Z_m)_{m\geq 1}$ is a supermartingale. A way to do so comes from Bercu and Touati (2008).

Lemma 2.1.2 (Towards the design of a supermartingale)**.** Let (M_m) be a locally square-integrable martingale with respect to the filtration $(\mathcal{F}_m)$. For all $\eta \in \mathbb{R}$ and $m \geq 0$, one has:

$$\mathbb{E}\left[\exp\left(\eta \Delta M_m - \frac{\eta^2}{2}\left(\Delta[M]_m + \Delta\langle M\rangle_m\right)\right) \mid \mathcal{F}_{m-1}\right] \leq 1,$$

> where $\Delta M_m = M_m - M_{m-1}, \Delta[M]_m = \Delta M_m^2$ and $\Delta\langle M\rangle_m = \mathbb{E}\left[\Delta M_m^2 \mid \mathcal{F}_{m-1}\right]$. We define $V_m(\eta) = \exp\left(\eta M_m - \frac{\eta^2}{2}\left([M]_m + \langle M\rangle_m\right)\right)$. Then, for all $\eta \in \mathbb{R}, (V_m(\eta))$ is a positive supermartingale with $\mathbb{E}\left[V_m(\eta)\right] \leq 1$ where $[M]_m(h) := \sum_{i=1}^m \Delta[M]_m, \langle M\rangle_m(h) := \sum_{i=1}^m \Delta\langle M\rangle_m$.

In the sequel, this lemma will be helpful to design a supermartingale (*i.e.*, to choose a relevant f_m for any m) without further assumption.

2.1.3 Contributions and outline

By avoiding Markov, a key message of (Kuzborskij and Szepesvári, 2019) is that, for learning problems with independent data, PAC-Bayes learning only requires the control of order 2 moment on losses to be used with convergence guarantees. This is strictly less restrictive than the classical subgaussian/subgamma assumptions appearing in the major part of the literature.

We successfully prove this fact remains even for non-independent data: we only need to control order 2 (conditional) moments to perform PAC-Bayes learning. We focus in this chapter on the PAC-Bayesian framework for martingales (Seldin *et al.*, 2011, 2012a,b). We then provide a novel PAC-Bayesian bound holding for data-free priors and unbounded martingales. From this, we recover in PAC-Bayes bounds for unbounded losses and iid data as a significant particular case. We also propose an extension of Seldin *et al.* (2012a)'s result for multi-armed bandits.

More precisely, Section 2.2.1 contains our novel PAC-Bayes bound for unbounded martingales and Section 2.2.3 contains an immediate corollary for learning theory with iid data. We eventually apply our main result for martingales in Section 2.3 to the setting of multi-armed bandit. Doing so, we provably extend a result of Seldin *et al.* (2012a) to the case of unbounded rewards.

Appendix A.1 gathers more details on PAC-Bayes, we draw in Appendix A.2 a detailed comparison between our new results and a few classical ones. We show that adapting our bounds to the assumptions made in those papers allows to recover similar or improved bounds. We defer to Appendix A.3 the proofs of Sections 2.2.3 and 2.3.

2.2 A PAC-Bayesian bound for unbounded martingales

2.2.1 Main result

A line of work led by Seldin *et al.* (2011, 2012a,b) provided PAC-Bayes bounds for almost surely bounded martingales. We provably extend the remits of their result to the case of unbounded martingales.

Framework Our framework is close to the one of SELDIN *et al.*, 2012a: we assume having access to a countable dataset $\mathcal{S} = (\mathbf{z}_i)_{i\geq 1} \in$ with no restriction on the distribution of $\mathcal{S}$ (in particular the $\mathbf{z}_i$ can depend on each others). We denote for any m, $\mathcal{S}_m := (\mathbf{z}_i)_{i=1..m}$ the restriction of $\mathcal{S}$ to its m first points. $(\mathcal{F}_i)_{i\geq 0}$ is a filtration adapted to $\mathcal{S}$. We denote for any $i \in \mathbb{N}$ $\mathbb{E}_{i-1}[.] := \mathbb{E}[. \mid \mathcal{F}_{i-1}]$. We also precise the space $\mathcal{H}$ to be an index (or a hypothesis) space, possibly uncountably infinite. Let $\{X_1(\mathcal{S}_1, h), X_2(\mathcal{S}_2, h), \cdots : h \in \mathcal{H}\}$ be martingale difference sequences, meaning that for any $m \geq 1, h \in \mathcal{H}$, $\mathbb{E}_{m-1}\left[X_m(\mathcal{S}_m, h)\right] = 0$.
For any $h \in \mathcal{H}$, let $M_m(h) = \sum_{i=1}^m X_i(S_i, h)$ be martingales corresponding to the martingale difference sequences and we define, as in BERCU and TOUATI (2008), the following

$$[M]_m(h) := \sum_{i=1}^m X_i(\mathcal{S}_i, h)^2,$$

$$\langle M\rangle_m(h) = \sum_{i=1}^m \mathbb{E}_{i-1}\left[X_i(\mathcal{S}_i, h)^2\right].$$

For a distribution Q over $\mathcal{H}$ define weighted averages of the martingales with respect to Q as $M_m(\mathrm{Q}) = \mathbb{E}_{h\sim\mathrm{Q}}\left[M_m(h)\right]$ (similar definitions hold for $[M]_m(\mathrm{Q}), \langle M\rangle_m(\mathrm{Q})$).
Main result. We now present the main result of this section where we succesfully avoid the boundedness assumption on martingales. This relaxation comes at the cost of additional variance terms $[M]_m, \langle M\rangle_m$.

> **Theorem 2.2.1** (A PAC-Bayesian bound for unbounded martingales)**.** For any data-free prior $\mathrm{P} \in \mathcal{M}(\mathcal{H})$, any $\lambda > 0$, any collection of martingales $(M_m(h))_{m\geq 1}$ indexed by $h \in \mathcal{H}$, the following holds with probability $1-\delta$ over the sample $\mathcal{S} = (\mathbf{z}_i)_{i\in\mathbb{N}}$, for all $m \in \mathbb{N}/\{0\}$, $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:
>
> $$|M_m(\mathrm{Q})| \leq \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{\lambda} + \frac{\lambda}{2}\left([M]_m(\mathrm{Q}) + \langle M\rangle_m(\mathrm{Q})\right).$$

Proof lies in Section 2.2.2.
Analysis of the bound. This theorem involves several terms. The change of measure inequality introduces the KL divergence term, the approximation term $\log(2/\delta)$ comes from Ville's inequality (instead of Markov in classical PAC-Bayes). Finally, the terms $[M]_m(\mathrm{Q}), \langle M\rangle_m(\mathrm{Q})$ come from our choice of supermartingale as suggested by BERCU and TOUATI (2008). The term $[M]_m(\mathrm{Q})$ can be interpreted as an empirical variance term while $\langle M\rangle_m(\mathrm{Q})$ is its theoretical counterpart. Note that $\langle M\rangle_m(\mathrm{Q})$ also appears in SELDIN *et al.* (2012a, Theorem 1).

We recall that this general result stands with no assumption on the martingale difference sequence $(X_i)_{i\geq 1}$ and holds uniformly on all $m \geq 1$. Those two points are, to the best of our knowledge, new within the PAC-Bayes literature. We discuss in Section 2.2.3 and appendix A.2 more concrete instantiations.

Comparison with literature. The closest result from Th. 2.2.1 is the PAC-Bayes Bernstein inequality of SELDIN *et al.*, 2012a. Our bound is a natural extension of theirs as their result only involves the variance term (not the empirical one), but requires two additional assumptions:

1. Bounded variations of the martingale difference sequence: $\forall m, \exists C_m \in \mathbb{R}^2$ such that a.s. for all h $|X_m(\mathcal{S}_m, h)| \leq C_m$.

2. Restriction on the range of the λ: $\forall m, \lambda_m \leq 1/C_m$.

SELDIN *et al.* (2012a) need those assumptions to ensure the *Bernstein assumption* which states that for any h, $\mathbb{E}[\exp(\lambda M_m(h) - \frac{\lambda^2}{2}\langle M\rangle_m(h))] \leq 1$. Our proof technique do not require the Bernstein assumption (and so none of the two conditions described above, which allow us to deal with unbounded martingales) as we exploit the supermartingale structure to obtain our results. More precisely, the price to pay to avoid the Bernstein assumption is to consider the empirical variance term $[M]_m(h)$ and to prove that $\left(\exp\left(\lambda M_m - \frac{\lambda^2}{2}\left([M]_m + \langle M\rangle_m\right)\right)\right)_{m\geq 1}$ is a supermartingale using Lemma 2.1.1 and Lemma 2.1.2 (see Section 2.2.2 for the complete proof). A broader discussion is detailed in appendix A.2.

2.2.2 Proof of Theorem 2.2.1

Proof of Theorem 2.2.1. We fix $\eta \in \mathbb{R}$ and we consider the function f_m to be for all $(\mathcal{S}, h)$:

$$\begin{aligned} f_m(\mathcal{S}, h) &:= \eta M_m(h) - \frac{\eta^2}{2}\left([M]_m(h) + \langle M\rangle_m(h)\right) \\ &= \sum_{i=1}^{m} \eta \Delta M_i(h) - \frac{\eta^2}{2}(\Delta[M]_i(h) + \Delta\langle M\rangle_i(h)), \end{aligned}$$

where $\Delta M_i(h) = X_i(\mathcal{S}_i, h), \quad \Delta[M]_i(h) = X_i(\mathcal{S}_i, h)^2, \quad \Delta\langle M\rangle_i(h) = \mathbb{E}_{i-1}\left[X_i(\mathcal{S}_i, h)^2\right]$.
For the sake of clarity, we dropped the dependency in $\mathcal{S}$ of M_m. Note that, given the definition of M_m, $M_m(h)$ is $\mathcal{F}_m$ measurable for any fixed h.
Let P a fixed data-free prior, we first apply the change of measure inequality to

obtain $\forall m \in \mathbb{N}, \forall \mathrm{Q} \in \mathcal{M}(\mathcal{H})$:

$$\mathbb{E}_{h\sim \mathrm{Q}}[f_m(\mathcal{S},h)] \leq \mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log\left(\underbrace{\mathbb{E}_{h\sim \mathrm{P}}\left[\exp(f_m(\mathcal{S},h))\right]}_{:=Z_m}\right),$$

with the convention $f_0 = 0$. We now have to show that $(Z_m)_m$ is a supermartingale with $Z_0 = 1$. To do so remark that for any m, because P is data free one has the following result.

Lemma 2.2.1. For any data-free prior P, any σ-algebra $\mathcal{F}$ belonging to the filtration $(\mathcal{F}_i)_{i\geq 0}$, any nonnegative function f taking as argument the sample $\mathcal{S}$ and a predictor h, one has almost surely:

$$\mathbb{E}\left[\mathbb{E}_{h\sim \mathrm{P}}[f(\mathcal{S},h)] \mid \mathcal{F}\right] = \mathbb{E}_{h\sim \mathrm{P}}\left[\mathbb{E}[f(\mathcal{S},h) \mid \mathcal{F}]\right].$$

Proof of Lemma 2.2.1. Let A be a $\mathcal{F}$-measurable event. We want to show that

$$\mathbb{E}\left[\mathbb{E}_{h\sim \mathrm{P}}[f(\mathcal{S},h)]\mathbb{1}_A\right] = \mathbb{E}\left[\mathbb{E}_{h\sim \mathrm{P}}\left[\mathbb{E}[f(\mathcal{S},h) \mid \mathcal{F}]\right]\mathbb{1}_A\right],$$

where the first expectation in each term is taken over $\mathcal{S}$. Note that it is possible to take this expectation thanks to the Kolomogorov's extension theorem (see *e.g.* TAO, 2011, Thm 2.4.4) which ensure the existence of a probability space for the discrete-time stochastic process $\mathcal{S} = (\mathbf{z}_i)_{i\geq 1}$.
Thus, this is enough to conclude that

$$\mathbb{E}\left[\mathbb{E}_{h\sim \mathrm{P}}[f(\mathcal{S},h)] \mid \mathcal{F}\right] = \mathbb{E}_{h\sim \mathrm{P}}\left[\mathbb{E}[f(\mathcal{S},h) \mid \mathcal{F}]\right],$$

by definition of the conditional expectation. To do so, notice that because $f(\mathcal{S},h)\mathbb{1}_A$ is a nonnegative function, and that P is data-free, we can apply the classical Fubini-Tonelli theorem.

$$\mathbb{E}\left[\mathbb{E}_{h\sim \mathrm{P}}[f(\mathcal{S},h)]\mathbb{1}_A\right] = \mathbb{E}_{h\sim \mathrm{P}}\left[\mathbb{E}\left[f(\mathcal{S},h)\mathbb{1}_A\right]\right].$$

One now conditions by $\mathcal{F}$ and use the fact that $\mathbb{1}_A$ is $\mathcal{F}$-measurable:

$$= \mathbb{E}_{h\sim \mathrm{P}}\left[\mathbb{E}\left[\mathbb{E}\left[f(\mathcal{S},h) \mid \mathcal{F}\right]\mathbb{1}_A\right]\right].$$

We finally re-apply Fubini-Tonelli to re-intervert the expectations:

$$= \mathbb{E}\left[\mathbb{E}_{h\sim \mathrm{P}}\left[\mathbb{E}\left[f(\mathcal{S},h) \mid \mathcal{F}\right]\mathbb{1}_A\right]\right].$$

This concludes the proof of Lemma 2.2.1. ■

We then use Lemma 2.2.1 with $f = \exp(f_m)$ and $\mathcal{F} = \mathcal{F}_{m-1}$ to obtain:

$$\begin{aligned}\mathbb{E}_{m-1}[Z_m] &= \mathbb{E}_{h\sim \mathrm{P}}\left[\mathbb{E}_{m-1}[(\exp(f_m(\mathcal{S},h))]\right] \\ &= \mathbb{E}_{h\sim \mathrm{P}}\left[\exp(f_{m-1}(\mathcal{S},h))\mathbb{E}_{m-1}\left[\exp(\eta\Delta M_m(h) - \frac{\eta^2}{2}(\Delta[M]_m(h) + \Delta\langle M\rangle_m(h))\right]\right],\end{aligned}$$

with $f_{m-1}(\mathcal{S},h) = \sum_{i=1}^{m-1}\eta(\Delta M_i(h)) - \frac{\eta^2}{2}(\Delta[M]_i(h) + \Delta\langle M\rangle_i(h))$. Using Lemma 2.1.2 ensures that for any h,

$$\mathbb{E}_{m-1}[\exp(\eta\Delta M_m(h) - \frac{\eta^2}{2}(\Delta[M]_m(h) + \Delta\langle M\rangle_m(h))] \leq 1,$$

thus we have

$$\mathbb{E}_{m-1}[Z_m] \leq \mathbb{E}_{h\sim \mathrm{P}}\left[\exp(f_{m-1}(\mathcal{S},h))\right] = Z_{m-1}.$$

Thus $(Z_m)_m$ is a nonnegative supermartingale with $Z_0 = 1$. We can use Ville's inequality (Lemma 2.1.1) which states that

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

Thus, with probability $1-\delta$ over $\mathcal{S}$, for all $m \in \mathbb{N}, Z_m \leq 1/\delta$. We then have the following intermediary result. For all P a data-free prior, $\eta \in \mathbb{R}$, with probability $1-\delta$ over $\mathcal{S}$, for all $m > 0, \mathrm{Q} \in \mathcal{M}(\mathcal{H})$

$$\eta M_m(\mathrm{Q}) \leq \mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta) + \frac{\eta^2}{2}\left([M]_m(\mathrm{Q}) + \langle M\rangle_m(\mathrm{Q})\right)], \tag{2.3}$$

recalling that $M_m(\mathrm{Q}) = \mathbb{E}_{h\sim\mathrm{Q}}[M_m(h)]$, and that similar definitons hold for $[M]_m(\mathrm{Q}), \langle M\rangle_m(\mathrm{Q})$. Thus, applying the bound with $\eta = \pm\lambda$ ($\lambda > 0$) and taking an union bound gives, with probability $1-\delta$ over $\mathcal{S}$, for any $m \in \mathbb{N}$, $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$

$$\lambda\left|M_m(\mathrm{Q})\right| \leq \mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta) + \frac{\lambda^2}{2}\left([M]_m(\mathrm{Q}) + \langle M\rangle_m(\mathrm{Q})\right)].$$

Dividing by λ concludes the proof. ■

2.2.3 A corollary: Batch learning with iid data and unbounded losses

In this section, we instantiate Theorem 2.2.1 onto a learning theory framework with iid data. We show that our bound encompasses several results of literature as particular cases.

Framework We consider a *learning problem* specified by a tuple $(\mathcal{H},\mathcal{Z},\ell)$ consisting of a set $\mathcal{H}$ of predictors, the data space $\mathcal{Z}$, and a loss function $\ell : \mathcal{H}\times\mathcal{Z} \to \mathbb{R}^+$. We consider a countable dataset $\mathcal{S} = (\mathbf{z}_i)_{i\geq 1} \in \mathcal{Z}^{\mathbb{N}}$ and assume that sequence is *i.i.d.* following the distribution $\mathcal{D}$. We also denote by $\mathcal{M}(\mathcal{H})$ is the set of probabilities on $\mathcal{H}$.

Definitions Similarly to Chapter 1, the *population risk* R of a predictor $h \in \mathcal{H}$ is $\forall h, \mathsf{R}(h) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})]$, the *empirical error* of h is $\forall h, \hat{\mathsf{R}}_{\mathcal{S}_m}(h) = \frac{1}{m}\sum_{i=1}^m \ell(h,z_i)$ and finally the *quadratic generalisation error* V of h is $\forall h, Quad(h) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,z)^2]$. We also denote by *generalisation gap* for any h the quantity $\mathsf{R}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)$.

Main result. We now state the main result of this section. This bound is a corollary of Theorem 2.2.1 and fills the gap with learning theory.

> **Theorem 2.2.2** (A PAC-Bayes bound for batch learning with heavy-tailed losses). For any data-free prior $P \in \mathcal{M}(\mathcal{H})$, any $\lambda > 0$ the following holds with probability $1-\delta$ over the sample $\mathcal{S} = (\mathbf{z}_i)_{i\in\mathbb{N}}$, for all $m \in \mathbb{N}/\{0\}$, $Q \in \mathcal{M}(\mathcal{H})$
>
> $$\mathbb{E}_{h\sim \mathrm{Q}}[\mathsf{R}(h)] \leq \mathbb{E}_{h\sim \mathrm{Q}}\left[\hat{\mathsf{R}}_{\mathcal{S}_m}(h) + \frac{\lambda}{2m}\sum_{i=1}^{m}\ell(h,z_i)^2\right] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{\lambda m} + \frac{\lambda}{2}\mathbb{E}_{h\sim\mathrm{Q}}[\mathrm{Quad}(h)].$$

Proof is furnished in Appendix A.3.

About the choice of λ. A novelty in this theorem is that the bound holds *simultaneously on all* $m > 0$ – this is due to the use of Ville's inequality. This sheds a new light on the choice of λ. Indeed, taking a localised λ depending on a given sample size (e.g. $\lambda_m = 1/\sqrt{m}$) ensures convergence guarantees for the expected generalisation gap. Doing so, our bound matches the usual PAC-Bayes literature (i.e. a bound holding with high probability for a single m). However the novelty brought by Theorem 2.2.2 is that our bound holds for unbounded losses for all times simultaneously. This suggests that taking a sample size-dependent λ may not be the best answer. We detail an instance of this fact below when one thinks of λ as a parameter of an optimisation objective. Indeed, our bound suggests a new optimisation objective for unbounded losses which is for any $m > 0$:

$$\mathrm{argmin}_{\mathrm{Q}}\, \mathbb{E}_{h\sim\mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\ell(h,z_i) + \frac{\lambda}{2}\ell(h,z_i)^2\right)\right] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P})}{\lambda m}. \tag{2.4}$$

Equation (2.4) differs from the classical objective of Catoni (2007, Thm 1.2.6) (described in (1.8)) on the additional quadratic term $\frac{\lambda}{2}\ell(h,z_i)^2$. Note that this objective implies a bound on the theoretical order 2 moment to be meaningful as we do not include it in our objective. Note that this constraint is less restrictive than Catoni's objective which requires a bounded loss. This objective stresses the role of the parameter λ as being involved in a new explicit tradeoff between the KL term and the efficiency on training data.

Also, this optimisation objective is valid for any sample size m, this means that our λ should not depend on certain dataset size but should be fixed in order to ensure a learning algorithm with generalisation guarantees at all time. This draws a parallel with Stochastic Gradient Descent with fixed learning step.

About the underlying assumptions in this bound. Our result is empirical (all terms can be computer or approximated) at the exception of the term $\mathbb{E}_{h\sim Q}[\mathrm{Quad}(h)]$. This invites to choose carefully the class of posteriors, in order to bound this second-order moment with minimal assumptions. For instance, if we consider the particular case of the quadratic loss $\ell(h,z)=(h-z)^2$, then we only need to assume that our data have a finite variance if we restrict our posteriors to have both bounded means and variance. This assumption is striclty less restrictive than the classical subgaussian/subgamma assumption classically appearing in the literature.

Comparison with literature. Back to the bounded case, we note that instantiating the boundedness assumption in Th. 2.2.2 make us recover the result of ALQUIER *et al.* (2016, Theorem 4.1) for the subgaussian case. We also remark that instantiating the HYPE condition conditioning HADDOUCHE *et al.* (2021, Theorem 3) allow us to improve their result as we transformed the control of an exponential moment into one on a second-order moment. More details are gathered in Appendix A.2. We also compare Theorem 2.2.2 to KUZBORSKIJ and SZEPESVÁRI (2019, Theorem 3) which is a PAC-Bayes bound for unbounded losses obtained through a concentration inequality from DE LA PEÑA *et al.* (2009). They arrived to what they denote as semi-empirical inequalities which also involve empirical and theoretical variance terms (and not an exponential moment).Their bound holds for independent data and a single posterior. First of all, note that Theorem 2.2.2 holds for any posterior, which is strictly more general. Note also that our bound is a straightforward corollary of Theorem 2.2.1 which holds for any martingale (thus for any data distribution in a learning theory framework) and so, exploits a different toolbox than KUZBORSKIJ and SZEPESVÁRI (2019) (control of a supermartingale vs. concentration bounds for independent data). We insist that a fundamental novelty in our work is to extend the conclusion of KUZBORSKIJ and SZEPESVÁRI, 2019 to the case of non-independent data: it is possible to perform PAC-Bayes learning for unbounded losses at the expense of the control of second-order moments. Note also that their bound is slightly tighter than ours as their result is Theorem 2.2.2 being optimised in λ (which is something we cannot do as the resulting λ would be data-dependent).

2.3 Application to the multi-armed bandit problem

We exploit our main result in the context of the multi-armed bandit problem – we adopt the framework of SELDIN *et al.* (2012a).

Framework. Let $\mathcal{A}$ be a set of actions of size $|\mathcal{A}|=K<+\infty$ and $a\in\mathcal{A}$ be an action. At each round i, the environment furnishes a reward function $R_i:\mathcal{A}\to\mathbb{R}$ which associate a reward $R_i(a)$ to the arm a. Assuming the R_is are iid, we denote for any a, the *expected reward for action* a to be $R(a)=\mathbb{E}_{R_1}[R_1(a)]$. At each round i, the player executes an action A_i according to a policy π_i. We then set the filtration $(\mathcal{F}_i)_{i\geq 1}$ to be $\mathcal{F}_i = \sigma\left(\{\pi_j, A_j, R_j \mid 1 \leq j \leq m\}\right)$.

Assumptions. We suppose here that $(R_i)_{i\geq 1}$ is an iid sequence and that at each time i, A_i and R_i are independent and that π_i is $\mathcal{F}_{i-1}$ measurable. This means that the player is not aware of the rewards each round and performs its current move with regards to the past.
We also add two technical assumptions. First, the order two moment of the expected reward is uniformly bounded: $\sup_{a\in\mathcal{A}} \mathbb{E}_{R_1}[R_1(a)^2] \leq C$. This assumption is strictly less restrictive than the boundedness assumption made in Seldin *et al.*, 2012a. Similarly to this work, we also assume that there exists a sequence $(\varepsilon_i)_{i\geq 1}$ such that $\inf_{a\in\mathcal{A}} \pi_i(a) \geq \varepsilon_i$. We say that $(\pi_i)_{i\geq 1}$ is *bounded from below by* $(\varepsilon_i)_{i\geq 1}$.

Definitions. For $i \geq 1$ and $a \in \{1, \ldots, K\}$, define a set of random variables $(R_i^a)_{i\geq 1}$ (*the importance weighted samples*, Sutton and Barto, 2018)

$$R_i^a := \begin{cases} \frac{1}{\pi_i(a)} R_i, & \text{if } A_i = a, \\ 0, & \text{otherwise.} \end{cases}$$

We define for any time m: $\hat{R}_m(a) = \frac{1}{m}\sum_{i=1}^t R_i^a$. Observe that for all i, $\mathbb{E}\left[R_i^a \mid \mathcal{F}_{i-1}\right] = R(a)$ and $\mathbb{E}[\hat{R}_m(a)] = R(a)$. Let a^* be the "best" action (the action with the highest expected reward, if there are multiple "best" actions pick any of them). Define the *expected and empirical per-round regrets* as

$$\Delta(a) = R\left(a^*\right) - R(a), \quad \hat{\Delta}_m(a) = \hat{R}_m\left(a^*\right) - \hat{R}_m(a).$$

Observe that $m\left(\hat{\Delta}_m(a) - \Delta(a)\right)$ forms a martingale. Let

$$V_m(a) = \sum_{i=1}^m \mathbb{E}\left[\left(R_i^{a^*} - R_i^a - \left[R\left(a^*\right) - R(a)\right]\right)^2 \mid \mathcal{F}_{i-1}\right]$$

be the cumulative variance of this martingale and

$$\hat{V}_m(a) = \sum_{i=1}^m \left(R_i^{a^*} - R_i^a - \left[R\left(a^*\right) - R(a)\right]\right)^2$$

its empirical counterpart. We denote for any distribution Q over $\mathcal{A}$, $\Delta(\mathrm{Q}) = \mathbb{E}_{a\sim \mathrm{Q}}[\Delta(a)]$, $V_m(\mathrm{Q}) = \mathbb{E}_{a\sim\mathrm{Q}}[V_m(a)]$, similar definitions hold for $\hat{\Delta}_m(\mathrm{Q}), \hat{V}_m(\mathrm{Q})$. We can now state the main result of this section – its proof is deferred to Appendix A.3.

Theorem 2.3.1 (PAC-Bayes bounds for heavy-tailed rewards)**.** For any $m \geq 1$, any history-dependent policy sequence $(\pi_i)_{i\geq 1}$ bounded from below by $(\varepsilon_i)_{i\geq 1}$, we have with probability $1-\delta$, for all posterior Q

$$\left|\Delta(\mathrm{Q}) - \hat{\Delta}_m(\mathrm{Q})\right| \leq 2\sqrt{\frac{\left(1+\frac{2K}{\delta}\right)(\log(K)+\log(4/\delta))}{m\varepsilon_m}}.$$

To the best of our knowledge, this result is the first PAC-Bayesian guarantees for multi-armed bandits with unbounded rewards. The proposed bound is as tight as Theorem 2.3 of Seldin *et al.* (2012a), up to a factor $(e-2)$ transformed into $\left(1+\frac{2K}{\delta}\right)$ (which is a huge dependency in K) within the square root. Note that our result comes at the price of the localisation: Theorem 2.3 of Seldin *et al.* (2012a) proposes a bound holding uniformly for all time m while our approach only holds for a single time m.
We believe there is room for improvement in Th. 2.3.1. Indeed, the current approach is naive as it consists in bounding crudely with high probability the empirical variance. Such a naive trick impeach us to consider all times simultaneously. Indeed, in its current form, taking an union bound on Theorem 2.3.1 is costful as we have a dependency in $1/\delta$ in our result (instead of $\log(1/\delta)$ in Seldin *et al.*, 2012a): this would destroy the convergence rate. The question of dealing more subtly with the empirical variance term is left as an open question.

2.4 Conclusion

A first step towards an optimisation perspective of PAC-Bayes We showed that it is possible to generalise the PAC-Bayes toolbox to unbounded martingales and heavy-tailed losses (resp. learning problem with unbounded losses for batch/online learning), the solely implicit assumption being the existence of second order moments on the martingale difference sequence (resp. on the loss function) which is reasonable as many PAC-Bayes bound lies on assumptions on exponential moments (*e.g.* the subgaussian assumption) to work.
Current Limitations. Doing so, we made a first step towards concrete optimisation perspective of PAC-Bayes by showing generalisation bounds are attainable with weak statistical assumptions and thus, compatible with many practical settings where optimisation is performed. However, Chapter 2 still presents some strong links with the information-theoretic approach such as: *(i)* the presence of a prior P in Theorem 2.2.2 which does no fit the optimisation views of the prior (see Figure 1.2), and *(ii)* the presence of a KL divergence, suggesting an information-theoretic perspective of learning. Point *(i)* will be later developed in Chapters 3, 4 and 6 when P is seen as an

initialisation point and in Chapter 5 when P is the learning objective. *(ii)* will be later developed in Chapters 5 and 6.

Extensions of this work. The supermartingale framework presented here are extracted from Haddouche and Guedj (2023a) and has inspired many follow-up works. Chugg *et al.* (2023) extended the approach of this chapter to other supermartingales as well as reversed submartingales, allowing to recover a vast majority of existing PAC-Bayes literature, also, Rodriguez-Galvez *et al.* (2023) tightened the theorems presented here by allowing the optimisation in λ. The tools presented in this work (*e.g.* Ville's inequality) are also useful to obtain fast rate PAC-Bayes bounds based on the coin-betting approach Jang *et al.* (2023) and Kuzborskij *et al.* (2024). The coin-betting approach originally in online learning (Orabona and Pál, 2016). In Chapter 3, we take a deeper focus on online learning, showing that an online approach of PAC-Bayes is possible, and allows to consider prior distribution as an initialisation point of a learning algorithm.

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

This chapter is based on the following papers

Maxime Haddouche and Benjamin Guedj. Online PAC-Bayes Learning. *Advances in Neural Information Processing Systems (NeurIPS)*. (2022)
Maxime Haddouche and Benjamin Guedj. PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales. *Transactions on Machine Learning Research*. (2023)

Contents

Abstract

While Chapter 2 showed weak statistical assumptions were reachable in PAC-Bayes, allowing its use in a wide range of concrete optimisation settings, the role of the prior P remains untreated. To tackle this issue, we propose here to consider P as the initialisation point of a learning algorithm. Then, to attenuate its impact in PAC-Bayes procedures, we develop *Online PAC-Bayes learning*, which consider a sequence (
Q_i,
$P_i)_{i=1\cdots m}$ of pairs (posterior,prior) evolving through time. Thus, the impact of initialisation $\mathrm{P} = \mathrm{P}_1$ is attenuated through the evolution of
P_i during the learning phase. We develop the first Online PAC-Bayes bounds and propose experiments showing that online PAC-Bayes outperforms SGD in several cases.

3.1 Introduction

Batch learning is somewhat the dominant learning paradigm in which we aim to design the best predictor by collecting a training dataset which is then used for inference or prediction. Classical algorithms such as SVMs (see Cristianini, Shawe-Taylor, *et al.*, 2000, among many others) or feedforward neural networks (Svozil *et al.*, 1997) are popular examples of efficient batch learning. While the mathematics of batch learning constitute a vivid and well understood research field, in practice this might not be aligned with the way practitionners collect data, which can be sequential when too much information is available at a given time (*e.g.* the number of micro-transactions made in finance on a daily basis). Indeed batch learning is not designed to properly handle dynamic systems.

Online learning (OL) (Zinkevich, 2003; Shalev-Shwartz, 2012; Hazan, 2016) fills this gap by treating data as a continuous stream with a potentially changing learning goal. OL has been studied with convex optimisation tools and the celebrated notion of regret which measures the discrepancy between the cumulative sum of losses for a specific algorithm at each datum and the optimal strategy. It led to many fruitful results comparing the efficiency of prediction for optimisation algorithms such that Online Gradient Descent (OGD), Online Newton Step through static regret (Zinkevich, 2003; Hazan *et al.*, 2007). OL is flexible enough to incorporate external expert advice onto classical algorithms with the optimistic point of view that such advices are useful for training (Rakhlin and Sridharan, 2013a; Rakhlin and Sridharan, 2013b) and then having optimistic regret bounds. Modern extensions also allow to compare to moving strategies through dynamic regret (see e.g. Yang *et al.*, 2016; Zhao *et al.*, 2020; Zhang *et al.*, n.d.). However, this notion of regret has been challenged recently: for instance, Wintenberger (2021) chose to control an expected cumulative loss through PAC inequalities in order to deal with the case of stochastic loss functions.

While OL tackles problems beyond batch learning, it can also be used as a tool to understand stochastic methods in a batch framework, such as SGD, where data are picked sequentially. In the context of PAC-Bayes, it is then natural to ask whether online learning could explain either the in-training evolution of the generalisation ability of batch methods or provide online variants of classical algorithms (*e.g.* (1.7), (1.8)). In both cases, the online paradigm allows focusing less on the prior P and more on its evolution, being consistent with the optimisation view of the prior as an initialisation point (see Figure 1.2).

Our contributions. Our goal is to provide a general online framework for PAC-Bayesian learning. Our main contribution (Theorem 3.2.1 in Section 3.2) is a general bound valid for bounded losses exploiting the generic PAC-Bayes bound of Rivasplata *et al.* (2020), later used to derive several online PAC-Bayesian results (as developed in Sections 3.3 and 3.4). More specifically, we derive two types of bounds, *online PAC-Bayesian training and test bounds*. Training bounds exhibit online procedures while the test bound provide efficiency guarantees. We propose then several algorithms with their associated training and test bounds as well as a short series of experiments to evaluate the consistency of our online PAC-Bayesian approach. Our efficiency criterion is not the classical regret but an expected cumulative loss close to the one of WINTENBERGER (2021). More precisely, Section 3.3 propose a stable yet time-consuming Gibbs-based algorithm, while Section 3.4 proposes time efficient yet volatile algorithms. However, even if OPB requires no assumption on the data distribution, allows priors to be data-dependent and do not require any convexity assumption on the loss (as commonly assumed in the OL framework), it still requires a bounded loss. We circumvent this limitation in Section 3.6 that it is possible to extend OPB results to the case of heavy-tailed losses, exploiting the supermartingale toolbox of Chapter 2.

Outline. Section 3.2 introduces the theoretical framework as well as our main result. Section 3.3 presents an online PAC-Bayesian algorithm and draws links between PAC-Bayes and OL results. Section 3.4 details online PAC-Bayesian disintegrated procedures with reduced computational time, Section 3.5 gathers supporting experiments and Section 3.6 gathers an extension of Section 3.2 for heavy-tailed losses. We include reminders on OL and PAC-Bayes in Appendices B.1.1 and B.3. Appendix B.2 provide discussion about our main result. All proofs are deferred to Appendix B.4.

3.2 An online PAC-Bayesian bound for bounded losses

We establish a novel PAC-Bayesian theorem (which in turn will be particularised in Section 3.3) overcoming the classical limitation of data-independent prior and *i.i.d.* data. We call our main result an *online PAC-Bayesian bound* as it allows to consider a sequence of priors which may depend on the past and a sequence of posteriors that can dynamically evolve as well. Indeed, we follow the online learning paradigm which considers a continous stream of data that the algorithm has to process on the fly, adjusting its outputs at each time step *w.r.t.* the arrival of new data and the past. In the PAC-Bayesian framework, this paradigm translates as follows: from an initial (still data independent) prior $\mathrm{Q}_1 = \mathrm{P}$ and a data sample $\mathcal{S}_m = (z_1, ..., z_m)$, we design a sequence of posterior $(\mathrm{Q}_i)_{1\leq i\leq m}$ where $\mathrm{Q}_i = f(\mathrm{Q}_1, ..., \mathrm{Q}_{i-1}, z_i)$.

Framework. We fix a countable dataset $\mathcal{S} = (\mathbf{z}_i)_{i\geq 1}$, following a distribution $\mathcal{D}_\mathcal{S}$, an integer $m > 0$ and the training set $\mathcal{S}_m \in \mathcal{Z}^m$, being the restriction of $\mathcal{S}$ to its m first data, drawn from an unknown distribution $\mathcal{D}_m$. We do not make any assumption on $\mathcal{D}_\mathcal{S}, \mathcal{D}_m$ and we fix a filtration $(\mathcal{F}_i)_{i\geq 0}$ adapted to $\mathcal{S}$. We set a sequence of priors, starting with $\mathrm{P}_1 = \mathrm{P}$ a data-free distribution and $(\mathrm{P}_i)_{i\geq 2}$ such that for each i, P_i is $\mathcal{F}_{i-1}$ measurable. For $\mathrm{P}, \mathrm{Q} \in \mathcal{M}(\mathcal{H})$, the notation $\mathrm{Q} \ll \mathrm{P}$ indicates that Q is

absolutely continuous wrt P (i.e. $\mathrm{Q}(A) = 0$ if $\mathrm{P}(A) = 0$ for measurable $A \subset \mathcal{H}$). We also denote by Q_i our sequence of candidate posteriors. There is no restriction on what Q_i could be. In what follows we denote by KL the Kullback-Leibler divergence between two distributions.
We consider a predictor space $\mathcal{H}$ and a loss funtion $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}^+$ bounded by a real constant $K > 0$. We denote by $\mathcal{M}(\mathcal{H})$ the set of all probability distributions on $\mathcal{H}$. We now introduce the notion of *stochastic kernel* (Rivasplata *et al.*, 2020) which formalise properly data-dependent measures within the PAC-Bayes framework. First, for a fixed predictor space $\mathcal{H}$, we set $\Sigma_{\mathcal{H}}$ to be the considered σ-algebra on $\mathcal{H}$.

> **Definition 3.2.1** (Stochastic kernels)**.** A *stochastic kernel* from $\mathcal{Z}^m$ to $\mathcal{H}$ is defined as a mapping $\mathrm{Q} : \mathcal{Z}^m \times \Sigma_{\mathcal{H}} \rightarrow [0; 1]$ where
>
> - For any $B \in \Sigma_{\mathcal{H}}$, the function $\mathcal{S}_m = (\mathbf{z}_1, ..., \mathbf{z}_m) \mapsto Q(\mathcal{S}_m, B)$ is measurable,
> - For any $\mathcal{S}_m \in \mathcal{Z}^m$, the function $B \mapsto Q(\mathcal{S}_m, B)$ is a probability measure over $\mathcal{H}$.
>
> We denote by $\mathtt{Stoch}(\mathcal{Z}^m, \mathcal{H})$ the set of all stochastic kernels from $\mathcal{Z}^m$ to $\mathcal{H}$ and for a fixed S, we set $\mathrm{Q}_{\mathcal{S}_m} := \mathrm{Q}(\mathcal{S}_m, .)$ the data-dependent prior associated to the sample $\mathcal{S}_m$ through Q.

From now, to refer to a distribution $\mathrm{Q}_{\mathcal{S}_m}$ depending on a dataset $\mathcal{S}_m$, we introduce a stochastic kernel $\mathrm{Q}(., .)$ such that $\mathrm{Q}_{\mathcal{S}_m} = \mathrm{Q}(\mathcal{S}_m, .)$. Note that this notation is perfectly suited to the case when $\mathrm{Q}_{\mathcal{S}_m}$ is obtained from an algorithmic procedure A. In this case the stochastic kernel Q of interest is the learning algorithm A. We use this notion to characterise our sequence of priors.

> **Definition 3.2.2** (Online Predictive Sequence)**.** We say that a sequence of stochastic kernels $(\mathrm{P}_i)_{i=1..m}$ is an ***online predictive sequence*** if *(i)* for all $i \geq 1, \mathcal{S}_m \in \mathcal{Z}^m, \mathrm{P}_i(\mathcal{S}_m, .)$ is $\mathcal{F}_{i-1}$ measurable and *(ii)* for all $i \geq 2$, $\mathrm{P}_i(\mathcal{S}_m, .) \ll P_{i-1}(\mathcal{S}_m, .)$.

Note that *(ii)* implies that for all $i, \mathrm{P}_i(\mathcal{S}_m, .) \ll \mathrm{P}_1(\mathcal{S}_m, .)$ with $P_1(\mathcal{S}_m, .)$ a data-free measure (yet a classical prior in the PAC-Bayesian theory).
We can now state our main result.

> **Theorem 3.2.1** (An OPB bound for bounded losses)**.** For any distribution $\mathcal{D}_m$ over $\mathcal{Z}^m$, any $\lambda > 0$ and any online predictive sequence (used as priors) $(\mathrm{P}_i)_{i=1\cdots m}$, for any sequence of stochastic kernels $(\mathrm{Q}_i)_{i=1\cdots m}$ we have with probability $1 - \delta$ over the sample $\mathcal{S}_m \sim \mathcal{D}_m$, the following, holding for the data-dependent measures $\mathrm{Q}_{i,\mathcal{S}_m} := \mathrm{Q}_i(\mathcal{S}_m, .), \mathrm{P}_{i,\mathcal{S}_m} := \mathrm{P}_i(\mathcal{S}_m, .)$:

$$\sum_{i=1}^{m} \mathbb{E}_{h_i \sim \mathrm{Q}_{i,\mathcal{S}_m}} \left[\mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] \right] \leq \sum_{i=1}^{m} \mathbb{E}_{h_i \sim \mathrm{Q}_{i,\mathcal{S}_m}} \left[\ell(h_i, \mathbf{z}_i) \right] + \frac{\mathrm{KL}(\mathrm{Q}_{i,\mathcal{S}_m}, \mathrm{P}_{i,\mathcal{S}_m})}{\lambda} + \frac{\lambda m K^2}{2} + \frac{\log(1/\delta)}{\lambda}.$$

Remark 3.2.1. [Lighter notations for stochastic kernels] For the sake of clarity, we assimilate in what follows the stochastic kernels $\mathrm{Q}_i, \mathrm{P}_i$ to the data-dependent distributions $\mathrm{Q}_i(\mathcal{S}_m, .), \mathrm{P}_i(\mathcal{S}_m, .)$. Then, an online predictive sequence is also assimilated to a sequence of data-dependent distributions. Concretely this leads to the switch of notation $\mathrm{Q}_{i,\mathcal{S}_m} \to \mathrm{Q}_i$ in Theorem 3.2.1. The reason of this switch is that, even though stochastic kernel is the right theoretical structure to state our main result, we consider in Sections 3.3 and 3.4 practical algorithmic extensions which focus only on data-dependent distributions, hence the need to alleviate our notations.

The proof is deferred to Appendix B.4.1. See Appendix B.2 for context and discussions.
A batch to online conversion. First, we remark that our bound slightly exceeds the OL framework: indeed, it would require our posterior sequence to be an online predictive sequence as well, which is not the case here (for any i, the distribution $\mathrm{Q}_{i,\mathcal{S}_m}$ can depend on the whole dataset). This is a consequence of our proof method (see Appendix B.4.1), which is classically denoted as a "batch to online" conversion (in opposition to the "online to batch" procedures as in Dekel and Singer, 2005). In other words, we exploited PAC-Bayesian tools designed for a fixed batch of data to obtain a dynamic result. This is why we refer to our bound as online as it allows considering sequences of priors and posteriors that can dynamically evolve.
Analysis of the different terms in the bound. Our PAC-Bayesian bound formally differs in many points from the classical ones. On the left-hand side of the bound, the sum of the averaged expected loss conditioned to the past appears. Having such a sum of expectations instead of a single one is necessary to assess the quality of all our predictions. Indeed, because data may be dependent, one can not consider a single expectation as in the iid case. We also stress that taking an online predictive sequence as priors leads to control losses conditioned to the past, which differs from classical PAC-Bayes results designed to bound the expected loss. This term, while original in the PAC-Bayesian framework (to the best of our knowledge) recently appeared (in a modified form) in Wintenberger (2021, Prop 3). See Appendix B.2.2 for further disucssions.
On the right hand-side of the bound, online counterparts of classical PAC-Bayes terms appear. At time i, the measure Q_i (i.e. $\mathrm{Q}_{i,\mathcal{S}_m}$ according to Remark 3.2.1) has a

tradeoff to achieve between an overfitted prediction of $\mathbf{z}_i$ (the case $\mathrm{Q}_i = \delta_{z_i}$ where δ is a Dirac measure) and a too weak impact of the new data with regards to our prior knowledge (the case $\mathrm{Q}_i = \mathrm{P}_i$). The quantity $\lambda > 0$ can be seen as a regulariser to adjust the relative impact of both terms.

Influence of λ. The quantity λ also plays a crucial role on the bound as it is involved in an explicit tradeoff between the KL terms, the confidence term $\log(1/\delta)$ and the residual term $mK^2/2$. This idea of seeing λ as a trading parameter is not new (GERMAIN *et al.*, 2016; THIEMANN *et al.*, 2017). However, the results from THIEMANN *et al.* (2017) stand w.p. $1-\delta$ for any λ while ours and the ones from GERMAIN *et al.* (2016) hold for any λ w.p. $1-\delta$ which is weaker and implies to discretise $\mathbb{R}^+$ onto a grid to estimate the optimal λ.

We now move on to the design of online PAC-Bayesian algorithms.

3.3 An online PAC-Bayesian procedure

OL algorithms (we refer to HAZAN, 2016 an introduction to the field) are producing sequences of predictors by learning from a dynamic data stream (see Appendix B.1.1 for an example). Recall that, in the OL framework, an algorithm outputs at time i a predictor which is $\mathcal{F}_{i-1}$-measurable. Here, our goal is to design an online procedure derived from Theorem 3.2.1 which outputs an online predictive sequence (which is assimilated, according to Remark 3.2.1, to a sequence of distributions).

Online PAC-Bayesian (OPB) training bound. We state a corollary of our main result which paves the way to an online algorithm. This constructive procedure motivates the name *Online PAC-Bayesian training bound* (OPBTRAIN in short).

Corollary 3.3.1 (OPBTRAIN)**.** For any distribution $\mathcal{D}_m$ over $\mathcal{Z}^m$, any $\lambda > 0$ and any online predictive sequences $\hat{\mathrm{Q}}, P$, the following holds with probability $1-\delta$ over the sample $\mathcal{S}_m \sim \mathcal{D}_m$:

$$\sum_{i=1}^{m} \mathbb{E}_{h_i \sim \hat{\mathrm{Q}}_{i+1}} \left[\mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] \right] \leq \sum_{i=1}^{m} \mathbb{E}_{h_i \sim \hat{\mathrm{Q}}_{i+1}} \left[\ell(h_i, \mathbf{z}_i) \right] + \frac{\mathrm{KL}(\hat{\mathrm{Q}}_{i+1}, \mathrm{P}_i)}{\lambda} + \frac{\lambda m K^2}{2} + \frac{\log(1/\delta)}{\lambda}.$$

Here, λ is seen as a scale parameter as precised below. The proof consists in applying Theorem 3.2.1 with for all i, $\mathrm{Q}_i = \hat{\mathrm{Q}}_{i+1}$ and P_i. Note that in this case, our posterior sequence is an online predictive sequence in order to fit with the OL framework.

Corollary 3.3.1 suggests to design $\hat{\mathrm{Q}}$ as follows, assuming we have drawn a dataset $S = \{z_1, ..., z_m\}$, fixed a scale parameter $\lambda > 0$ and an online predictive sequence P_i:

$$\hat{\mathrm{Q}}_1 = P_1, \quad \forall i \geq 1 \; \hat{\mathrm{Q}}_{i+1} = \underset{Q \in \mathcal{M}(\mathcal{H})}{\operatorname{argmin}} \mathbb{E}_{h_i \sim \mathrm{Q}} \left[\ell(h_i, \mathbf{z}_i) \right] + \frac{\mathrm{KL}(\mathrm{Q}, \mathrm{P}_i)}{\lambda} \tag{3.1}$$

which leads to the explicit formulation

$$\frac{d\hat{\mathrm{Q}}_{i+1}}{d\mathrm{P}_i}(h) = \frac{\exp\left(-\lambda \ell(h, \mathbf{z}_i)\right)}{\mathbb{E}_{h \sim \mathrm{P}_i}\left[\exp\left(-\lambda \ell(h, \mathbf{z}_i)\right)\right]}. \tag{3.2}$$

Thus, the formulation of Equation (3.2), which has been highlighted by Catoni (2003, Sec. 5.1) shows that our online procedure produces Gibbs posteriors. So, PAC-Bayesian theory provides sound justification for the somewhat intuitive online procedure in Equation (3.1): at time i, we adjust our new measure $\hat{\mathrm{Q}}_{i+1}$ by optimising a tradeoff between the impact of the newly arrived data $\mathbf{z}_i$ and the one of prior knowledge $\hat{\mathrm{Q}}_i$. Notice that $\hat{\mathrm{Q}}$ is an online predictive sequence: $\hat{\mathrm{Q}}_i$ is $\mathcal{F}_{i-1}$-measurable for all i as it depends only on $\hat{\mathrm{Q}}_{i-1}$ and $\mathbf{z}_{i-1}$. Furthermore, one has $\hat{\mathrm{Q}}_i \ll \hat{\mathrm{Q}}_{i-1}$ for all i as $\hat{\mathrm{Q}}_i$ is defined as an argmin and the KL term is finite if and only it is absolutely continuous w.r.t. $\hat{\mathrm{Q}}_{i-1}$.

Remark 3.3.1. In Corollary 3.3.1, while the right hand-side is the reason we considered Equation (3.1), the left hand side still needs to be analysed. It expresses how the posterior $\hat{\mathrm{Q}}_{i+1}$ (designed from $\hat{\mathrm{Q}}_i, \mathbf{z}_i$) generalises well on average to any new draw of $\mathbf{z}_i$. More precisely, this term measures how much the training of $\hat{\mathrm{Q}}_{i+1}$ is overfitting on $\mathbf{z}_i$. A low value of it ensures our online predictive sequence, which is obtained from a single dataset, is robust to the randomness of $\mathcal{S}_m$, hence the interest of optimising the right hand side of the bound. This is a supplementary reason we refer to Corollary 3.3.1 as an OPBTrain bound as it provide robustness guarantees for our training.

Online PAC-Bayesian (OPB) test bound. However, Corollary 3.3.1 does not say if $\hat{\mathrm{Q}}_{i+1}$ will produce good predictors to minimise $\ell(., \mathbf{z}_{i+1})$, which is the objective of $\hat{\mathrm{Q}}_{i+1}$ in the OL framework (we only have access to the past to predict the future). We then need to provide an *Online PAC-Bayesian (OPB) test bound* (OPBTest bound) to quantify our prediction's accuracy. We now derive an OPBTest bound from Theorem 3.2.1.

Corollary 3.3.2 (OPBTest)**.** . For any distribution μ over $\mathcal{Z}^m$, any $\lambda > 0$, and any online predictive sequence $(\hat{\mathrm{Q}}_i)$, the following holds with probability $1-\delta$ over the sample $\mathcal{S}_m \sim \mathcal{D}_m$:

$$\sum_{i=1}^{m} \mathbb{E}_{h_i \sim \hat{\mathrm{Q}}_i} \left[\mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] \right] \leq \sum_{i=1}^{m} \mathbb{E}_{h_i \sim \hat{\mathrm{Q}}_i} \left[\ell(h_i, \mathbf{z}_i) \right] + \frac{\lambda m K^2}{2} + \frac{\log(1/\delta)}{\lambda}.$$

> Optimising in λ gives $\lambda = \sqrt{\frac{2\log(1/\delta)}{mK^2}}$ and ensure that:
>
> $$\sum_{i=1}^{m} \mathbb{E}_{h_i \sim \hat{\mathrm{Q}}_i} \left[\mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] \right] \leq \sum_{i=1}^{m} \mathbb{E}_{h_i \sim \hat{\mathrm{Q}}_i} \left[\ell(h_i, \mathbf{z}_i) \right] + O\left(\sqrt{\log(1/\delta) K^2 m} \right).$$

The proof consists in applying Theorem 3.2.1 with for all i, $\mathrm{Q}_i = \hat{\mathrm{Q}}_i = \mathrm{P}_i$.
Corollary 3.3.2 quantifies how efficient will our predictions be. Indeed, the left hand side of this bound relates for all i, how good $\hat{\mathrm{Q}}_i$ is to predict $\mathbf{z}_i$ (on average) which is what $\hat{\mathrm{Q}}_i$ is designed for. Note that here, the involved λ can differ from the scale parameter of Equation (3.1), it is now a way to compensate for the tradeoff between the two last terms of the bound. The strength of this bound is that since $\hat{\mathrm{Q}}$ is an online predictive sequence, the Kullback-Leibler terms vanished, leaving terms depending only on hyperparameters.

Links with previous approaches

We now present a specific case of Corollary 3.3.1 where we choose as priors the online predictive sequence $\hat{\mathrm{Q}}$ (*i.e.* in Theorem 3.2.1, we choose $\mathrm{Q}_i = \hat{\mathrm{Q}}_{i+1}, \mathrm{P}_i = \hat{\mathrm{Q}}_i$). The reason we focus on this specific case is that it enables to build strong links between PAC-Bayes and OL.
We then adapt our OPBTrain bound (Corollary 3.3.1). The online procedure becomes:

$$\hat{\mathrm{Q}}_1 = P, \quad \forall i \geq 1 \ \hat{\mathrm{Q}}_{i+1} = \operatorname{argmin}_{\mathrm{Q}} \mathbb{E}_{h_i \sim \mathrm{Q}} \left[\ell(h_i, \mathbf{z}_i) \right] + \frac{\mathrm{KL}(\mathrm{Q}, \hat{\mathrm{Q}}_i)}{\lambda}, \tag{3.3}$$

which leads to the explicit formulation

$$\frac{d\hat{\mathrm{Q}}_{i+1}}{d\hat{\mathrm{Q}}_i}(h) = \frac{\exp\left(-\lambda \ell(h, \mathbf{z}_i)\right)}{\mathbb{E}_{h \sim \hat{\mathrm{Q}}_i}\left[\exp\left(-\lambda \ell(h, \mathbf{z}_i)\right)\right]}.$$

Links with classical PAC-Bayesian bounds. We denote that the optimal predictor in this case is such that at any time i, $d\hat{\mathrm{Q}}_{i+1}(h) \propto \exp(-\lambda \ell(h, \mathbf{z}_i)) d\hat{\mathrm{Q}}_i(h)$ hence $d\hat{\mathrm{Q}}_{m+1}(h) \propto \exp\left(-\lambda \sum_{i=1}^m \ell(h, \mathbf{z}_i)\right) d\hat{\mathrm{Q}}_1(h)$. One recognises, up to a multiplicative constant, the optimised predictor of Catoni (2007, Th 1.2.6) which solves $\operatorname{argmin}_{\mathrm{Q}} \mathbb{E}_{h \sim \mathrm{Q}} \left[\frac{1}{m} \sum_{i=1}^m \ell(h, \mathbf{z}_i) \right] + \frac{\mathrm{KL}(\mathrm{Q}, \hat{\mathrm{Q}}_1)}{\lambda}$, thus one sees that in this case, the output of our online procedure after m steps coincides with Catoni's output. This shows consistency of our general procedure which recovers classical result within an online framework: when too many data are available, treating data sequentially until time m leads to the same Gibbs posterior than if we were treating the whole dataset as a batch.

Analogy with Online Gradient Descent (OGD). We propose an analogy between the procedure Equation (3.3) and the celebrated OGD algorithm (see Appendix B.1.1 for a recap). First we remark that our minimisation problem is equivalent to $\operatorname{argmin}_{\mathrm{Q}} \lambda \mathbb{E}_{h_i \sim \mathrm{Q}} \; [\ell(h_i, \mathbf{z}_i)] + \mathrm{KL}(Q \| \hat{\mathrm{Q}}_i)$. Then we assume that for any $i, \hat{\mathrm{Q}}_i = \mathcal{N}(\hat{m}_i, I_d)$ with $\hat{m}_i \in \mathbb{R}^d$ and we set $\mathcal{L}_i(\hat{m}_i) = \mathbb{E}_{h_i \sim \hat{\mathrm{Q}}_i} \; [\ell(h_i, \mathbf{z}_i)]$. The minimisation problem becomes: $\operatorname{argmin}_{\hat{m}} \lambda \mathcal{L}_i(\hat{m}) + \frac{1}{2} \|\hat{m} - \hat{m}_i\|^2$. And so using the first order Taylor expansion, we use the approximation $\mathcal{L}_i(\hat{m}) \approx \mathcal{L}_i(\hat{m}_i) + \langle \hat{m} - \hat{m}_i, \nabla \mathcal{L}_i(\hat{m}_i) \rangle$ which finally transform our argmin into the following optimisation process: $\hat{m}_{i+1} = \hat{m}_i - \lambda \nabla \mathcal{L}_i(\hat{m}_i)$ which is exactly OGD on the loss sequence $\mathcal{L}_i$. We draw an analogy between the scale parameter λ and the step size η in OGD. the KL term translates the influence of the previous point and the expected loss gives the gradient. This analogy has been already exploited in Shalev-Shwartz (2012) where they approximated $\mathbb{E}_{h_i \sim q_\mu}[\ell(h_i, \mathbf{z}_i)] := \bar{L}_i(\mu) \approx \mu^T \nabla \bar{L}_i(\mu_i)$ where μ is their considered online predictive sequence.

Finally, we remark that the optimum rate in Corollary 3.3.2 is a $\mathcal{O}(\sqrt{m})$ which is comparable to the best rate of Shalev-Shwartz (2012, Eq (2.5)) (see Proposition B.1.1).

Comparison with previous work. We acknowledge that the procedure of Equation (3.3) already appeared in literature. Li *et al.* (n.d., Alg. 1) propose a Gibbs procedure somewhat similar to ours, the main difference being the addition of a surrogate of the true loss at each time step. Within the OL literature, the idea of updating measures online has been recently studied for instance in Chérief-Abdellatif *et al.* (2019). More precisely, our procedure is similar to their Streaming Variational Bayes (SVB) algorithm. A slight difference is that they approximated the expected loss similarly to Shalev-Shwartz (2012). The guarantees Chérief-Abdellatif *et al.* (2019) provided for SVB hold for Gaussian priors and comes at the cost of additional constraints that do not allow to consider any aggregation strategies contrary to what Corollary 3.3.1 propose. Their bounds are deterministic and are using tools and assumptions from convex optimisation (such that convex expected losses) while ours are probabilistic and are using measure theory tools which allow to relax these assumptions.

Strength of our result. We emphasize two points. First, to the best of our knowledge, Corollary 3.3.1 is the first bound which theoretically suggests Equation (3.3) as a learning algorithm. Second, we stress that Equation (3.3) is a particular case of Corollary 3.3.1 and our result can lead to other fruitful routes. For instance, we consider the idea of adding noise to our measures at each time step to avoid overfitting (this idea has been used *e.g.* in Neelakantan *et al.*, 2015 in the context of deep neural networks): if our online predicitve sequence $(\hat{\mathrm{Q}}_i)$ can be defined through a sequence of parameter vectors $\hat{\mu}$, then we can define P_i by adding a small noise on $\hat{\mu}_i$ and thus giving more freedom through stochasticity.

Thus, we see that our procedure led us to the use of the Gibbs posteriors of Catoni. However, in practice, Gaussian distributions are preferred (*e.g.* Dziugaite and Roy, 2017; Rivasplata *et al.*, 2019; Perez-Ortiz *et al.*, 2021a,b,c)). That is why we focus next on new online PAC-Bayesian algorithms involving Gaussian distributions.

3.4 Disintegrated online algorithms for Gaussian distributions.

We dig deeper in the field of disintegrated PAC-Bayesian bounds, originally explored by Blanchard and Fleuret (2007) and Catoni (2007), further studied by Alquier and Biau (2013) and Guedj and Alquier (2013) and recently developed by Rivasplata *et al.* (2020) and Viallard *et al.* (2023a) (see Appendix B.3 for a short presentation of the bound we adapted and used). The strength of the disintegrated approach is that we have directly guarantees on the random draw of a single predictor, which avoids to consider expectations over the predictor space. This fact is particularly significant in our work as the procedure precised in Equation (3.2), require the estimation of an exponential moment to be efficient, which may be costful. We then show that disintegrated PAC-Bayesian bounds can be adapted to the OL framework, and that they have the potential to generate proper online algorithms with weak computational cost and sound efficiency guarantees.

Online PAC-Bayesian disintegrated (OPBD) training bounds. We present a general form for *online PAC-Bayes disintegrated (OPBD) training bounds*. The terminology comes from the way we craft those bounds: from PAC-Bayesian disintegrated bounds we use the same tools as in Theorem 3.2.1 to create the first online PAC-Bayesian disintegrated bounds. OPBD training bounds have the following form.

For any online predictive sequences $\hat{\mathrm{Q}}, P$, any $\lambda > 0$ w.p. $1-\delta$ over $\mathcal{S}_m \sim \mathcal{D}_m$ and $(h_1, ..., h_m) \sim \hat{\mathrm{Q}}_2 \otimes ... \otimes \hat{\mathrm{Q}}_{m+1}$:

$$\sum_{i=1}^{m} \mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] \leq \sum_{i=1}^{m} \ell(h_i, \mathbf{z}_i) + \Psi(h_i, \hat{\mathrm{Q}}_{i+1}, \mathrm{P}_i) \ + \Phi(m), \tag{3.4}$$

with Ψ, Φ being real-valued functions. Ψ controls the global behaviour of Q_{i+1} w.r.t. the $\mathcal{F}_{i-1}$-measurable prior P_i. If one has no dependency on h_i this behaviour is global, otherwise it is local. Note that those functions may depend on λ, δ. However, since they are fixed parameters, we do not make these dependencies explicit. Similarly to Corollary 3.3.1, this kind of bound allows to derive a learning algorithm (cf Algorithm 1) which outputs an online predicitve sequence $\hat{\mathrm{Q}}$. Finally we draw $(h_1, ..., h_m) \sim \hat{\mathrm{Q}}_2 \otimes ... \otimes \hat{\mathrm{Q}}_{m+1}$ (and not $\hat{\mathrm{Q}}_1 \otimes ... \otimes \hat{\mathrm{Q}}_m$) since an OPBD bound is designed to justify theoretically an OPBD procedure in the same way Corollary 3.3.1 allowed to justify Equation (3.1).

Why focus on Gaussian measures? The reason is that a Gaussian variable $h \sim \mathcal{N}(w, \sigma^2 \mathbf{I}_d)$ can be written as $h = w + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$, and this expression totally defines h ($\mathbf{I}_d$ being the identity matrix).

A general OPBD algorithm for Gaussian measure with fixed variance We use an idea presented in VIALLARD *et al.* (2023a) which restrict the measure set to Gaussian on $\mathbb{R}^d$ with known and fixed covariance matrix $\sigma^2 \mathbf{I}_d$. Then we present in Algorithm 1 a general algorithm (derived from an OPBD training bound) for Gaussian measures with fixed variance which outputs a sequence of gaussian $\hat{\mathrm{Q}}_i = \mathcal{N}(\hat{w}_i, \sigma^2 \mathbf{I}_d)$ from a prior sequence $\mathrm{P}_i = \mathcal{N}(w_i^0, \sigma^2 \mathbf{I}_d)$ where for each i, w_i^0 is $\mathcal{F}_{i-1}$- measurable. Because the variance is fixed, the distribution is uniquely defined by its mean, thus we identify $\hat{\mathrm{Q}}_i$ and $\hat{w}_i$, P_i and w_i^0.

Algorithm 1: A general OPBD algorithm for Gaussian measures with fixed variance.

```
Parameters : Time m, scale parameter $\lambda$
Initialisation: Variance $\sigma^2$, Initial mean $\hat{w}_1 \in \mathbb{R}^d$, epoch $m$
for each iteration $i$ in $1..m$ do
    Observe $z_i, w_i^0$ and draw $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$
    Update:
        $\hat{w}_{i+1} := \operatorname{argmin}_{w \in \mathbb{R}^d} \ell(w + \varepsilon_i, z_i) + \Psi(w + \varepsilon_i, w, w_i^0)$
end
Return $(\hat{w}_i)_{i=1..m+1}$
```

At each time i, Algorithm 1 requires the draw of $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$. Doing so, we generated the randomness for our h_i (because our bound holds for a single draw of $(h_1, .., h_m) \sim \hat{\mathrm{Q}}_2 \otimes ... \otimes \hat{\mathrm{Q}}_{m+1}$), we then write $h_i = w + \varepsilon_i$ and we optimise w.r.t. Ψ to find $\hat{w}_{i+1}$.

Bounds of interest. We present two possible choices of pairs (Ψ, Φ) derived from the disintegrated results presented in Appendix B.3. Doing so, we explicit two ready-to-use declinations of Algorithm 1.

Corollary 3.4.1 (Two OPB disintegrated learning algorithms)**.** For any distribution μ over $\mathcal{Z}^m$, any online predictive sequences of Gaussian measures with fixed variance $\hat{\mathrm{Q}}_i = \mathcal{N}(\hat{w}_i, \sigma^2 \mathbf{I}_d)$ and $\mathrm{P}_i = \mathcal{N}(w_i^0, \sigma^2 \mathbf{I}_d)$, any $\lambda > 0$, w.p. $1 - \delta$ over $\mathcal{S}_m \sim \mathcal{D}_m$ and $(h_i = \hat{w}_{i+1} + \varepsilon_i)_{i=1..m} \sim \hat{\mathrm{Q}}_2 \otimes ... \otimes \hat{\mathrm{Q}}_{m+1}$, the bound of Equation (3.4) holds for the two following pairs Ψ, Φ:

$$\Psi_1(h_i, \hat{w}_{i+1}, w_i^0) = \frac{||\hat{w}_{i+1} + \varepsilon_i - w_i^0||^2 - ||\varepsilon||^2}{2\lambda\sigma^2} \quad \Phi_1(m) = \frac{\lambda m K^2}{2} + \frac{\log(1/\delta)}{\lambda}, \tag{3.5}$$

$$\Psi_2(h_i, \hat{w}_{i+1}, w_i^0)) = \frac{||\hat{w}_{i+1} - w_i^0||^2}{2\lambda\sigma^2} \quad \Psi_2(m) = \lambda m K^2 + \frac{3\log(1/\delta)}{2\lambda}. \tag{3.6}$$

Where the notation $1, 2$ denote whether the functions have been derived from adapted theorems of Rivasplata *et al.*, 2020; Viallard *et al.*, 2023a recalled in Appendix B.3 We then can use Algorithm 1 with Equation (3.5), Equation (3.6).

Proof is deferred to Appendix B.4.2. Note that in Corollary 3.4.1, we identified $\hat{\mathrm{Q}}_i$ to $\hat{w}_i$ and for the last formula, Ψ has no dependency on h_i.

Comparison with Equation (3.1). The main difference with Equation (3.1) provided by the disintegrated framework is that the optimisation route does not include an expected term within the optimisation objective. The main advantage is a weaker computational cost when we restrict to Gaussian distributions. The main weakness is a lack of stability as our algorithm now depends at time i on $\ell(h + \varepsilon_i, z_i)$ so on ε_i directly. We denote that Equation (3.5) is less stable than Equation (3.6) as it involves another dependency on ε_i through Ψ. The reason is that Rivasplata *et al.*, 2020 proposed a bound involving a disintegrated KL divergence while Viallard *et al.*, 2023a proposed a result involving a Rényi divergence avoiding a dependency on ε_i. We refer to Appendix B.3 for a detailed statement of those properties.

Comparison with Hoeven *et al.*, 2018. Theorem 3 of Hoeven *et al.* (2018) recovers OGD from the exponential weights algorithm by taking a sequence of moving distributions being Gaussians with fixed variance which is exactly what we consider here. From these, they retrieve the classical OGD algorithm as well as its classical convergence rate. Let us compare our results with theirs.

First, if we fix a single step η in their bound and assume two traditional assumptions for OGD (a finite diameter D of the convex set and an uniform bound G on the loss gradients), we recover for the OGD (greedy GD in Hoeven *et al.*, 2018) a rate of $\frac{D^2}{2\sigma^2\eta} + \frac{\eta\sigma^2 T G^2}{2}$. This is, up to constants and notation changes, exactly our Ψ_i ($i \in \{1, 2\}$). Also, we notice a difference in the way to use Gaussian distributions: Theorem 3 of Hoeven *et al.* (2018) is based on their Lemma 1 which provides guarantees for the expected regret. This is a clear incentive to consider as predictors the mean of the sucessive Gaussians of interest. On the contrary, Corollary 3.4.1 involves a supplementary level of randomness by considering predictors h_i drawn from our Gaussians. This additional randomness appears in our optimisation process (Algorithm 1). Finally, notice that Hoeven *et al.* (2018) based their whole work on the use of a KL divergence while Corollary 3.4.1 not only exploit a disintegrated KL (Ψ_1) but also a Rényi α-divergence (Ψ_2). Note that we propose a result only for $\alpha = 2$ for the sake of space constraints but any other value of α leads to another optimisation objective to explore.

OPBD test bounds. Similarly to what we did in Section 3.3, we also provide *OPBD test bounds* to provide efficiency guarantees for online predicitve sequences (e.g. the output of Algorithm 1). Our proposed bounds have the following general form.

For any online predictive sequence $\hat{\mathrm{Q}}$, any $\lambda > 0$ w.p. $1-\delta$ over S and $(h_1, ..., h_m) \sim \hat{\mathrm{Q}}_1 \otimes ... \otimes \hat{\mathrm{Q}}_m$:

$$\sum_{i=1}^{m} \mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] \leq \sum_{i=1}^{m} \ell(h_i, \mathbf{z}_i) + \Phi(m), \tag{3.7}$$

with Φ being a real-valued function(possibly dependent on λ, δ though it is not explicited here).

Note that our predictors $(h_1, ..., h_m)$ are now drawn from $\hat{\mathrm{Q}}_1 \otimes ... \otimes \hat{\mathrm{Q}}_m$. Thus, the left-hand side of the bound considers a h_i drawn from an $\mathcal{F}_{i-1}$-measurable distribution evaluated on $\ell(., \mathbf{z}_i)$: this is effectively a measure of the prediction performance.

We now state a corollary which gives disintegrated guarantees for any online predicitve sequence.

Corollary 3.4.2 (OPB disintegrated test bounds)**.** For any distribution μ over $\mathcal{Z}^m$, any $\lambda > 0$, and any online predictive sequence $(\hat{\mathrm{Q}}_i)$, the following holds with probability $1-\delta$ over the sample $\mathcal{S}_m \sim \mathcal{D}_m$ and the predictors $(h_1, ..., h_m) \sim \hat{\mathrm{Q}}_1 \otimes ... \otimes \hat{\mathrm{Q}}_m$, the bound of Equation (3.7) holds with :

$$\Phi_1(m) = \frac{\lambda m K^2}{2} + \frac{\log(1/\delta)}{\lambda}, \quad \Phi_2(m) = 2\lambda m K^2 + \frac{\log(1/\delta)}{\lambda}.$$

Where the notation $1, 2$ denote whether the functions have been derived from adapted theorems of Rivasplata *et al.*, 2020; Viallard *et al.*, 2023a recalled in Appendix B.3. The optimised λ gives in both cases a $\mathcal{O}(\sqrt{m \log(1/\delta)})$.

Proof is deferred to Appendix B.4.2.

3.5 Experiments

We adapt the experimental framework introduced in Chérief-Abdellatif *et al.* (2019, Sec.5) to our algorithms (anonymised code available here). We conduct experiments on several real-life datasets, in classification and linear regression. Our objective is twofold: check the convergence of our learning methods and compare their efficiencies with classical algorithms. We first introduce our experimental setup.

Algorithms. We consider four online methods of interest: the OPB algorithm of Equation (3.3) which update through time a Gibbs posterior. We instantiate it with two different priors $\hat{\mathrm{Q}}_1$: a Gaussian distribution and a Laplace one. We also implement Algorithm 1 with the functions Ψ_1, Ψ_2 from Corollary 3.4.1. To assess efficiency, we implement the classical OGD (as described in Alg. 1 of ZINKEVICH, 2003) and the SVB method of CHÉRIEF-ABDELLATIF *et al.* (2019).

Binary Classification. At each round i the learner receives a data point $x_i \in \mathbb{R}^d$ and predicts its label $y_i \in \{-1, +1\}$ using $\langle x_i, h_i \rangle$, with $h_i = \mathbb{E}_{h\sim\hat{\mathrm{Q}}_i}[h]$ for OPB methods or h_i being drawn under $\hat{\mathrm{Q}}_i$ for OPBD methods. The adversary reveals the true value y_i, then the learner suffers the loss $\ell(h_i, \mathbf{z}_i) = \left(1 - y_i h_i^T x_i\right)_+$ with $z_i = (x_i, y_i)$ and $a_+ = a$ if $a > 0$ and $a_+ = 0$ otherwise. This loss is unbounded but can be thresholded.

Linear Regression. At each round i, the learner receives a set of features $x_i \in \mathbb{R}^d$ and predicts $y_i \in \mathbb{R}$ using $\langle x_i, h_i \rangle$ with $h_i = \mathbb{E}_{h\sim\hat{\mathrm{Q}}_i}[h]$ for SVB and OPB methods or h_i being drawn under $\hat{\mathrm{Q}}_i$ for OPBD methods. Then the adversary reveals the true value y_t and the learner suffers the loss $\ell(h_i, \mathbf{z}_i) = \left(y_i - h_i^T x_i\right)^2$ with $z_i = (x_i, y_i)$. This loss is unbounded but can be thresholded.

Datasets. We consider four real world dataset: two for classification (Breast Cancer and Pima Indians), and two for regression (Boston Housing and California Housing). All datasets except the Pima Indians have been directly extracted from `sklearn` (PEDREGOSA *et al.*, 2011). Breast Cancer dataset (STREET *et al.*, 1993) is available here and comes from the UCI ML repository as well as the Boston Housing dataset (BELSLEY *et al.*, 2005) which can be obtained here. California Housing dataset (PACE and BARRY, 1997) comes from the StatLib repository and is available here. Finally, Pima Indians dataset (SMITH *et al.*, 1988) has been recovered from this Kaggle repository. Note that we randomly permuted the observations to avoid to learn irrelevant human ordering of data (such that date or label).

Parameter settings. We ran our experiments on a 2021 MacBookPro with an M1 chip and 16 Gb RAM. For OGD, the initialisation point is $\mathbf{0}_{\mathbb{R}^d}$ and the values of the learning rates are set to $\eta = 1/\sqrt{m}$. For SVB, mean is initialised to $\mathbf{0}_{\mathbb{R}^d}$ and covariance matrix to Diag(1). Step at time i is $\eta_i = 0.1/\sqrt{i}$. For both of the OPB algorithms with Gibbs posterior, we chose $\lambda = 1/m$. As priors, we took respectively a centered Gaussian vector with the covariance matrix Diag(σ^2) ($\sigma = 1.5$) and an iid vector following the standard Laplace distribution. For the OPBD algorithm with Ψ_1, we chose $\lambda = 10^{-4}/m$, the initial mean is $\mathbf{0}_{\mathbb{R}^d}$ and our fixed covariance matrix is Diag(σ^2) with $\sigma = 3.10^{-3}$. For the OPBD algorithm with Ψ_1, we chose $\lambda = 2.10^{-3}/m$, the

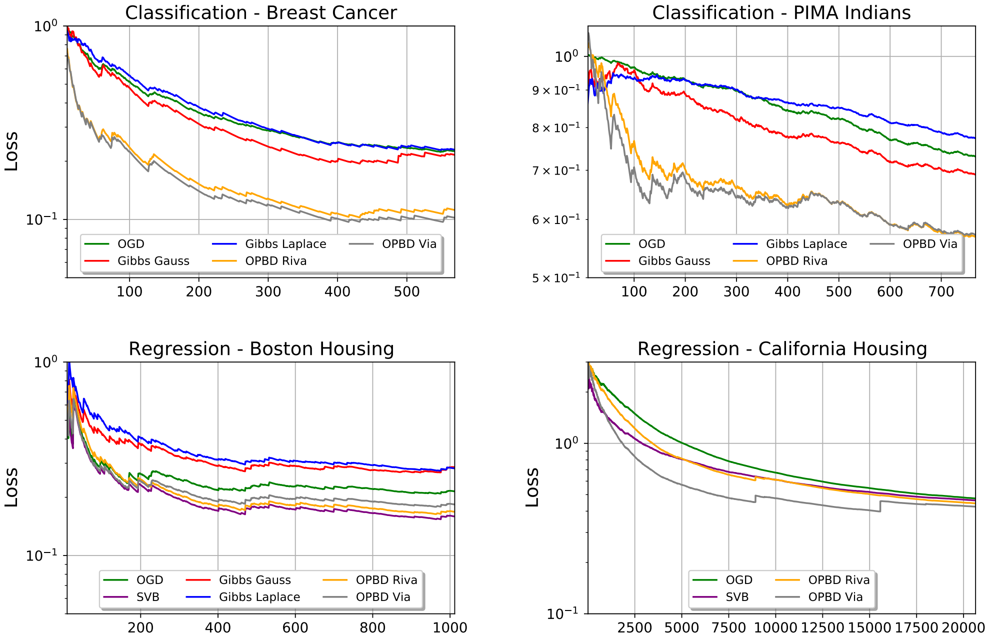

Figure 3.1. *Averaged cumulative losses for all four considered datasets. 'Gibbs Gauss' denotes OPB with Gaussian Prior, 'Gibbs Laplace' denotes OPB with Laplace prior. 'OPBD Riva' denotes OPBD with Ψ_1, 'OPBD Via' denotes OPBD with Ψ_2.*

initial mean is $\mathbf{0}_{\mathbb{R}^d}$ and our covariance matrix is Diag(σ^2) with $\sigma = 10^{-2}$. The reason of those higher scale parameters and variance is that Ψ from Rivasplata *et al.* (2020) is more stochastic (yet unstable) than the one Viallard *et al.* (2023a).

Experimental results. For each dataset, we plot the evolution of the average cumulative loss $\sum_{i=1}^{t} \ell(h_i, \mathbf{z}_i)/t$ as a function of the step $t = 1, \ldots, m$, where m is the dataset size and h_i is the decision made by the learner h_i at step i. The results are gathered in Figure 3.1

Empirical findings. OPB with Gaussian prior ('Gibbs Gauss') outperforms OGD on all datasets except California Housing (on which this method is not implemented) while OPB with Laplace prior ('Gibbs Laplace') always fail w.r.t. OGD. OPB methods fail to compete with SVB on the Boston Housing dataset. OPBD methods compete with SVB on regression problems and clearly outperforms OGD on classification tasks. OPBD with Ψ_2 (labeled as 'OPBD Via' in Figure 3.1) performs better on the California Housing dataset while OPBD with Ψ_1 (labeled as 'OPBD Riva') is more efficient on the Boston Housing dataset. Both methods performs roughly equivalently on classification tasks. This brief experimental validation shows the consistency of all our online procedures as we observe a visible decrease of the cumulative losses through time. It particularly shows that OPBD procedures improve on OGD on these dataset. We refer to Appendix B.5 for additional table gathering the error bars of our OPBD methods.

Why do we perform better than OGD? As stated in Section 3.4, OGD can be recovered as a Gaussian approximation of the exponential weights algorithm (EWA). Thus, a legitimate question is why do we perform better than OGD as our OPBD methods are also based on a Gaussian surrogate of EWA? HOEVEN *et al.*, 2018 only used Gaussians distributions with fixed variance as a technical tool when the considered predictors are the Gaussian means. In our work, we exploited a richer characteristic of our distributions in the sense our predictors are points sampled from our Gaussians and not only the means. This also has consequences in our learning algorithm as at time i of our Algorithm 1, our optimisation step involves a noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. Thus, we believe that OPBD methods should perform at least as well as OGD. We write 'at least' as we think that the higher flexibility due to this additional level of randomness might result in slightly better empirical performances, as seen on the few datasets in Figure 3.1.

3.6 Online PAC-Bayes for heavy-tailed losses.

Results of Section 3.2 exploited a PAC-Bayesian theorem of RIVASPLATA *et al.* (2020) to perform, however, we note that the OL framework, by considering non-*i.i.d.* data is compatible with the supermartingale toolbox of Chapter 2. We then show that it is possible to obtain anytime-valid OPB bounds for heavy-tailed losses, extending our results. Note however that such an extension can have consequences in terms of algorithmic procedures.
We now state the main theorem of this section.

Theorem 3.6.1 (An OPB bound for heavy-tailed losses)**.** For any distribution over the dataset $\mathcal{S}$, any $\lambda > 0$ and any online predictive sequence (used as priors) $(\mathrm{P}_i)_{i\geq 1}$, we have with probability at least $1-\delta$ over the sample $\mathcal{S} \sim \mathcal{D}_\mathcal{S}$, the following, holding for the data-dependent measures $\mathrm{P}_{i,\mathcal{S}} := \mathrm{P}_i(\mathcal{S}, .)$ any posterior sequence $(\mathrm{Q}_i)_{i\geq 1}$ and any $m \geq 1$:

$$\sum_{i=1}^m \mathbb{E}_{h_i\sim \mathrm{Q}_i}\left[\mathbb{E}[\ell(h_i,\mathbf{z}_i) \mid \mathcal{F}_{i-1}]\right] \leq \sum_{i=1}^m \mathbb{E}_{h_i\sim \mathrm{Q}_i}\left[\ell(h_i,\mathbf{z}_i)\right] + \frac{\lambda}{2}\sum_{i=1}^m \mathbb{E}_{h_i\sim \mathrm{Q}_i}\left[\hat{V}_i(h_i,\mathbf{z}_i) + V_i(h_i)\right] + \sum_{i=1}^m \frac{\mathrm{KL}(\mathrm{Q}_i, \mathrm{P}_{i,\mathcal{S}})}{\lambda} + \frac{\log(1/\delta)}{\lambda}.$$

> With for all i, $\hat{V}_i(h_i, \mathbf{z}_i) = (\ell(h_i, \mathbf{z}_i) - \mathbb{E}_{i-1}[\ell(h_i, \mathbf{z}_i)])^2$ is the empirical variance at time i and $V_i(h_i) = \mathbb{E}_{i-1}[\hat{V}(h_i, \mathbf{z}_i)]$ is the true conditional variance.

Proof lies in Appendix B.4.3.

Analysis of the bound. This bound is, to our knowledge, the first Online PAC-Bayes bound in literature holding for heavy-tailed losses. It is semi-empirical as the variance and empirical variance terms have theoretical components. However, these terms can be controlled with assumptions on conditional second-order moments and not on exponential ones (as made in Section 3.2 where the bounded loss assumption was used to obtain conditional subgaussianity). To emphasise our point, we consider as in Section 2.2.3 the case of the quadratic loss $\ell(h, z) = (h - z)^2$. Here, we only need to assume that our data have a finite variance if we restrict our posteriors to have both bounded means and variance. Also the meaning of the online predictive sequence P_i is that we must be able to design properly a sequence of priors before drawing our data, this can be for instance an online algorithm whihc generate a prior distribution from past data at each time step.

Finally, we note that if we assume being able to bound simultaneaously all condtional means and variance (which is strictly less restrictive than bounding the loss),then Theorem 3.6.1 suggests a new online learning objective which is an online counterpart to Equation (2.4).

$$\forall i \geq 1 \ \hat{\mathrm{Q}}_{i+1} = \underset{\mathrm{Q} \in \mathcal{M}(\mathcal{H})}{\operatorname{argmin}} \mathbb{E}_{h_i \sim \mathrm{Q}} \left[\ell(h_i, \mathbf{z}_i) + \frac{\lambda}{2} \ell(h_i, \mathbf{z}_i)^2 \right] + \frac{\mathrm{KL}(\mathrm{Q}, \mathrm{P}_{i,\mathcal{S}})}{\lambda} \tag{3.8}$$

While the algorithm differs from the one derived Theorem 3.2.1, we can still draws many links with this theorem.

- If we assume our loss to be bounded, then we can upper bound our empirical/theoretical variance terms to recover exactly Theorem 3.2.1. Theorem 3.6.1 then shows that finite order two moments are sufficient to perform online PAC-Bayes.
- Another crucial point lies on the range of our result which holds with high probability for any countable posterior sequence $(\mathrm{Q}_i)_{i\geq 1}$, any time m and the priors $(\mathrm{P}_{i,\mathcal{S}_m})_{i\geq 1}$. This is far much general than Theorem 3.2.1 which holds only for a single m and a single posterior sequence $(\mathrm{Q}_{i,\mathcal{S}_m})_{i=1..m}$. This happens because a preliminary theorem from Rivasplata *et al.* (2020) has been used instead of the change of measure inequality (Lemma 1.2.1). This preliminary theorem has imposed conditionnal subgaussianity to deal with the exponential moment. On the contrary, the use of the change of measure inequality alongside the supermartingale toolbox of Chapter 2 allowed a result holding for any posterior sequence, and any time simultaneously.

3.7 Conclusion

Chapter 3 builds a bridge between online learning and generalisation. As seen in Section 3.5, considering online PAC-Bayes procedures mitigates the impact of the prior in the learning process and thus, fit the optimisation view of the prior as in initialisation point (Figure 1.2), yielding performances comparable to online gradient descent. However, while Online PAC-Bayes is a promising step forward optimisation, with time-efficient procedures (Appendix B.3), some questions remains: *(i)* Is it possible to propagate the view of prior as initialisation directly for batch algorithms? *(ii)* Is it possible to obtain PAC-Bayes learning algorithms directly for deterministic predictors instead of using disintegrated results in order to be consistent with practitioners, often avoiding stochastic predictors?

Elements of answer to *(i)* lie in Chapter 4, showing that flat minimum, often attained in the context of deep neural network with much more parameters than training data, allows to attenuate the impact of the prior through a fast convergence rate. *(ii)* is tackled in Chapters 5 and 6 where the KL divergence is traded for a Wasserstein distance.

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

This chapter is based on the following paper

Maxime Haddouche, Paul Viallard, Umut Simsekli, and Benjamin Guedj. A PAC-Bayesian Link Between Generalisation and Flat Minima. (2024)

Contents

Abstract

In Chapter 3 we saw that a way to attenuate the impact of the prior, seen as an initialisation, in PAC-Bayes training is online learning, allowing the prior to evolve alongside the posterior through time. However, a legitimate question is to wonder whether the prior could be attenuated, even in the batch learning setting which is widely used in practice. Maintaining the vision of the prior as initialisation, we propose in this chapter to attenuate the impact of the prior in the batch setting through faster convergence rate. The proposed results hold when a flat minimum has been reached, *i.e.* a minimum whose its neighbourhood nearly minimises the loss as well. Then, a sharper understanding of generalisation can be reached when exploiting the benefits of a successful optimisation process. Indeed, this study is particularly meaningful in the context of deep learning, where it has been shown that flat minimum (also known as sharpness) correlates to a good generalisation ability.

4.1 Introduction

Can we make the impact of the prior vanish at a faster rate than $1/\sqrt{m}$ in the context of batch learning? While this is desirable from an optimisation perspective, this is not what is proposed by classical PAC-Bayes bounds, considering all elements of $\mathcal{M}(\mathcal{H})$ simultaneously. The challenge of this study is to obtain faster rates for a smaller class of posteriors. Doing so, we aim to attenuate the impact of the initialisation (seen as prior) for nonnegative heavy-tailed losses, potentially satisfying geometric assumptions such as gradient-lipschitz, making a promising step towards concrete optimisation settings. The practical way to do so is to obtain results holding only for posteriors distributions focusing on *flat minima*, which can be seen, *e.g.* in deep learning, as a benefit of a successful optimisation process.

Indeed, dating back to Hochreiter and Schmidhuber (1997), it has been hypothesised that the notion of 'flatness' (or sometimes equivalently referred to as 'sharpness') has tight links with the generalisation error: among the minima (belonging to $\hat{\mathsf{R}}_{\mathcal{S}_m}$) that is found by the learning algorithm, the 'flatter' the minimum is, the lower is the generalisation error. While the initial flatness notion was (vaguely) defined through low Kolmogorov complexity, there is no single formal definition of 'flatness'. Hence, several flatness notions have been considered, which typically are based on the second-order derivatives of the empirical risk around the local minimum found by the algorithm, such as $\mathrm{trace}(\nabla^2 \hat{\mathsf{R}}_{\mathcal{S}_m}(h))$, see *e.g.*, Jastrzebski *et al.* (2017) and Wen *et al.* (2023).

While there have been several attempts to link some form of flatness to generalisation in a mathematically rigorous way (Neyshabur *et al.*, 2017; Petzka *et al.*, 2021; Andriushchenko *et al.*, 2023; Yue *et al.*, 2023), mainly in the framework of 'sharpness aware minimisation' (Foret *et al.*, 2020), it has been recently shown that flat minima do not always imply good generalisation. In fact, there exist scenarios such that the flattest minima achieve the worst generalisation performance compared to non-flat ones (Wen *et al.*, 2023).

In this study, we aim at developing novel links between flatness and the generalisation error from a PAC-Bayesian perspective (see *e.g.*, Guedj, 2019; Hellström *et al.*, 2023; Alquier, 2024). Denoting by Q, the probability distribution of the algorithm output h (or the output of a learning algorithm), we identify sufficient conditions on Q such that flatness always implies good generalisation. More precisely, we make the following contributions:

- We show that, when Q satisfies the Poincaré inequality and a technical condition that we identify, we can obtain a 'fast-rate' generalisation bound that diminishes with rate $\frac{1}{m}$ (rather than $\frac{1}{\sqrt{m}}$) and mainly contains two terms:

  (i) The flatness term: $\mathbb{E}_{h\sim\mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^{m}\|\nabla_h \ell(h,\mathbf{z}_i)\|^2\right]$. This term is directly linked to the Hessian of the loss ℓ, due to the connection between the Fisher information and the Hessian of the loss Bickel and Doksum, 2015. For instance, under certain conditions, it can be shown that $\mathrm{trace}(\nabla^2 \hat{\mathsf{R}}_{\mathcal{S}_m}(h)) = \frac{2}{m}\sum_{i=1}^m \|\nabla_h \ell(h,\mathbf{z}_i)\|^2$ (Wen *et al.*, 2023, Lemma 4.1).

(ii) The classical PAC-Bayesian complexity term $\mathrm{KL}(\mathrm{Q},\mathrm{P})$, where KL denotes the Kullback-Leibler divergence and P is data-independent 'prior' distribution.

- We then further analyse the term $\mathrm{KL}(\mathrm{Q},\mathrm{P})$. We show that, when Q is a Gibbs distribution, *i.e.*, $\mathrm{Q}(h) \propto \exp(-\gamma \hat{\mathsf{R}}_{\mathcal{S}_m}(h))\mathrm{P}(h)$ for some $\gamma > 0$ and P satisfies a log-Sobolev inequality, the generalisation error can be controlled *solely* by the term: $\gamma^2 c_{LS}(\mathrm{P})\, \mathbb{E}_{h\sim \mathrm{Q}}[\|\nabla_h \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\|^2]$, where $c_{LS}(\mathrm{P})$ denotes the log-Sobolev constant of the prior P.

- We finally go beyond the KL divergence to link flat minima to deterministic predictors (*i.e.*, when Q is a Dirac distribution) through a novel Wasserstein-based generalisation bound for gradient Lipschitz loss functions.

We provide a numerical assessment of the technical condition underlying our main result, suggesting that it is suitable in the case of neural networks on classification tasks, confirming the relevance of our bounds to better understand the generalisation ability of neural networks. Our results shed further light on the impact of the flatness of the minima over the generalisation error: when the learning algorithm ensures a sufficiently regular distribution over the parameters, the generalisation error can be directly controlled by the flatness of the region found by the algorithm.

4.2 Preliminaries

Framework. We consider a predictor set $\mathcal{H} \subseteq \mathbb{R}^d$ equipped with a norm $\|.\|$, a data space $\mathcal{Z}$ and the space of distributions over $\mathcal{H}, \mathcal{M}(\mathcal{H})$. We also consider a loss function $\ell : \mathcal{H}\times\mathcal{Z} \to \mathbb{R}$. We assume that we have access to a *i.i.d.* dataset $\mathcal{S} = (\mathbf{z}_i)_{i\geq 1} \in \mathcal{Z}^{\mathbb{N}}$ with associated distribution $\mathcal{D}$. For each $m \geq 1$, we define $\mathcal{S}_m := \{\mathbf{z}_1,\cdots,\mathbf{z}_m\}$. In PAC-Bayes learning, we construct a data-driven posterior distribution $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ with respect to a prior distribution P. To assess the generalisation ability of a predictor $h \in \mathcal{H}$, we define the *population risk* to be $\mathsf{R}_{\mathcal{D}}(h) := \mathbb{E}_{\mathbf{z}\sim\mu}[\ell(h,\mathbf{z})]$ and for each m, its empirical counterpart $\hat{\mathsf{R}}_{\mathcal{S}_m}(h) := \frac{1}{m}\sum_{i=1}^m \ell(h,\mathbf{z}_i)$. As PAC-Bayes focuses on elements of $\mathcal{M}(\mathcal{H})$, we also define the expected risk and empirical risks for $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ as $\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) := \mathbb{E}_{h\sim Q}[\mathsf{R}_{\mathcal{D}}(h)]$ and $\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) := \mathbb{E}_{h\sim Q}[\hat{\mathsf{R}}_{\mathcal{S}_m}(h)]$. PAC-Bayes bounds usually aim at controlling the *expected generalisation error (or gap)* for each dataset size m, i.e., $\Delta_{\mathcal{S}_m}(\mathrm{Q}) := \mathsf{R}_{\mathcal{D}}(\mathrm{Q}) - \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q})$.

Background on Poincaré and log-Sobolev inequalities. In this work, we exploit Poincaré and log-Sobolev inequalities in the PAC-Bayes framework. We first recall the definition of Poincaré and log-Sobolev inequalities. To do so, for a fixed distribution Q, we define the *Sobolev space of order* 1 on $\mathbb{R}^d$ as follows:

$$\mathrm{H}^1(\mathrm{Q}) := \left\{ f \in \mathrm{L}^2(\mathrm{Q}) \cap \mathrm{D}_1(\mathbb{R}^d) \mid \|\nabla f\| \in \mathrm{L}^2(\mathrm{Q}) \right\},$$

where $\mathrm{D}_1(\mathbb{R}^d)$ denotes the set of derivable functions $f : \mathbb{R}^d \to \mathbb{R}$.

Definition 4.2.1 (Poincaré and Logarithmic Sobolev inequalites)**.** A measure Q satisfies a *Poincaré inequality* with constant $c_P(\mathrm{Q})$ if for all function $f \in \mathrm{H}^1(\mathrm{Q})$ we have

$$\mathrm{Var}_{\mathrm{Q}}(f) \leq c_P(\mathrm{Q}) \mathop{\mathbb{E}}_{h\sim\mathrm{Q}} \left[\|\nabla f(h)\|^2 \right],$$

where $\mathrm{Var}_{\mathrm{Q}}(f) = \mathbb{E}_{h\sim\mathrm{Q}} \left[f(h) - \mathbb{E}_{h\sim\mathrm{Q}}[f(h)] \right]^2$ is the *variance* of f *w.r.t.* Q. We then say that Q is Poincaré with constant $c_P(\mathrm{Q})$, or that Q is $\texttt{Poinc}(c_P)$. Also, Q satisfies a *log-Sobolev inequality* with constant $c_{LS}(\mathrm{Q})$ if for all function $f \in \mathrm{H}^1(\mathrm{Q})$ we have

$$\mathop{\mathbb{E}}_{h\sim\mathrm{Q}} \left[f^2(h) \log\left(\frac{f^2(h)}{\mathbb{E}_{h\sim\mathrm{Q}}\left[f^2(h)\right]} \right) \right] \leq c_{LS}(\mathrm{Q}) \mathop{\mathbb{E}}_{h\sim\mathrm{Q}} \left[\|\nabla f(h)\|^2 \right],$$

where the term on the left hand side is the *entropy* of f^2, denoted as $\mathrm{Ent}_Q(f^2)$. We then say that Q is log-Sobolev with constant $c_{LS}(\mathrm{Q})$, or that Q is $\texttt{L-Sob}(c_{LS})$.

The class of Gaussian distributions is an important particular case of distributions satisfying both Poincaré and log-Sobolev inequalities, this is the subject of Proposition 4.2.1.

Proposition 4.2.1. For a given pair (μ, Σ) of mean and covariance matrix in $\mathbb{R}^d$, define $\mathrm{Q} = \mathcal{N}(\mu, \Sigma)$. Then we have, for any $f \in \mathrm{H}^1(\mathrm{Q})$:

$$\mathrm{Ent}_{\mathrm{Q}}(f^2) \leq 2\mathbb{E}_{\mathrm{Q}}\left[\langle \Sigma\nabla f, \nabla f\rangle\right], \text{ and } \mathrm{Var}_{\mathrm{Q}}(f^2) \leq \mathbb{E}_{\mathrm{Q}}\left[\langle \Sigma\nabla f, \nabla f\rangle\right].$$

Thus, Q is $\texttt{L-Sob}(c_{LS})$ with constant $c_{LS}(\mathrm{Q}) = 2\|\Sigma\|_{op}$ and also $\texttt{Poinc}(c_{LS})$ with constant $c_{LS}(\mathrm{Q}) = \|\Sigma\|_{op}$, where $\|.\|_{op}$ denotes the operator norm.

In Proposition 4.2.1, the first inequality can be derived from the classical log-Sobolev inequality for $\mathcal{N}(\mathbf{0}, \mathrm{Id})$ stated in Gross (1975), with a change of variable. Similarly, the Poincaré inequality can be obtained through a change of variable from the Poincaré

inequality for $\mathcal{N}(\mathbf{0}, \mathrm{Id})$ which is a particular case of the Brascamp-Lieb inequality for log-concave probability measures (Brascamp and Lieb, 1976) and is stated explicitly in Beckner (1989, Theorem 1).
We now focus on specific posterior distributions called *Gibbs posteriors, or Gibbs distributions*. For a fixed loss ℓ and dataset $\mathcal{S}_m$, the Gibbs posterior, *w.r.t.* prior $\mathrm{P} \in \mathcal{M}(\mathcal{H})$, risk $\hat{\mathsf{R}}_{\mathcal{S}_m}$ and *inverse temperature* $\gamma > 0$.is defined as $\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}$ such that $d\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}(h) \propto \exp(-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}(h))d\mathrm{P}(h)$. Gibbs posteriors are a class of closed-form solutions for relaxation of Catoni (2007, Theorem 1.2.6) stated, for instance, in Alquier *et al.* (2016, Theorem 4.1). Proposition 4.2.2 shows that when the prior and the loss satisfies a few properties, then the associated Gibbs posterior is $\texttt{L-Sob}(c_{LS})$.

Proposition 4.2.2. Assume that P is a probability measure on $\mathbb{R}^d$ such that $d\mathrm{P}(h) \propto \exp(-V(x))$ with V a smooth function such that $Hess(V) \succeq \frac{2}{c_{LS}(\mathrm{P})}\mathrm{Id}$. Assume that $\ell = \ell_1 + \ell_2$ with ℓ_1 convex, twice differentiable and ℓ_2 bounded. Then for any $\gamma > 0$, the Gibbs posterior $\mathrm{Q} = \mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}$ is $\texttt{L-Sob}(c_{LS})$ with constant $c_{LS}(\mathrm{Q}) = c_{LS}(\mathrm{P}) \exp\left(4\|\ell_2\|_\infty\right)$.

Proposition 4.2.2 applies, *e.g.*, when P is a Gaussian prior $\mathrm{P} = \mathcal{N}(\mu_\mathrm{P}, \Sigma_\mathrm{P})$. Notice that in this case $c_{LS}(\mathrm{P}) = 2\|\Sigma_\mathrm{P}\|_{op}$. This property is a straightforward application of Chafai (2004, Corollary 2.1) with Guionnet and Zegarlinksi (2003, Property 2.6) and is stated in Appendix C.1 for completeness. Finally, notice that satisfying a log-Sobolev inequality is stronger than satisfying a Poincaré one. This is stated for instance in Ledoux (2006, Proposition 2.1) and properly recalled in Appendix C.1.

4.3 Reaching a flat minimum allows Poincaré posteriors to generalise well

4.3.1 Fast rate PAC-Bayes bounds for heavy-tailed losses

In order to obtain fast rates, *i.e.*, bounds converging to zero faster than $\frac{1}{\sqrt{m}}$, we exploit the notion of flat minimum (where the loss takes a small value in the neighbourhood of the minimum). Indeed, in an overparametrised setting such as neural networks, it is likely to obtain such a minimum once the optimisation phase has been performed, as there are much more parameters than training data. We exploit this flatness property within PAC-Bayes bounds through the gradient norm $\|\nabla_h \ell(., \mathbf{z})\|$ of the loss *w.r.t.* the predictor h for any $\mathbf{z}$. This is, to the best of our knowledge, the first attempt to do so as Gat *et al.* (2022) focus on gradients with respect to the data $\nabla_\mathbf{z}\ell$ (one does not optimise on those, as the dataset is fixed in practice).

In this section, we consider posterior distributions Q being $\texttt{Poinc}(c_P)$. This assumption covers the important case of Gaussian measures (Proposition 4.2.1) as well as all measures satisfying a log-Sobolev inequality (Proposition C.1.1). We focus on PAC-Bayes bound holding for distributions Q satisfying a particular assumption involving the data distribution $\mathcal{D}$ (contrary to many PAC-Bayes bounds holding for all Q). We then define the *error* of $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ for any datum $\mathbf{z} \in \mathcal{Z}$ as $\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z}) := \mathbb{E}_{h\sim \mathrm{Q}}[\ell(h, \mathbf{z})]$ and identify Assumption 4.3.1 to later involve flat minima.

Assumption 4.3.1. *We say that $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ is* quadratically self-bounded *with respect to ℓ and constant $C > 0$ (namely QSB(ℓ, C)) if*

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z})^2\right] \leq C R_{\mathcal{D}}(\mathrm{Q}) \left(= C\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z})\right]\right).$$

Assumption 4.3.1 is a relaxation of boundedness, as if $\ell \in [0, C]$ then it is $\texttt{QSB}(\ell, C)$. It is an alternative to the bounded expected variance assumption in anytime-valid PAC-Bayes bounds as in Chapter 2 and (Chugg *et al.*, 2023). An issue with such boundedness assumption is that it has to hold for all posteriors, including those providing poor generalisation performances. This is avoided by the $\texttt{QSB}$ assumption which intricate the properties of $\mathcal{D}, \ell$ and Q. Such a design is in line with the conclusions of the recent work of Gastpar *et al.* (2023), inviting to derive generalisation bounds valid for specific pairs $(\mathrm{Q}, \mathcal{D})$ (and not uniformly valid for all such pairs). Finally, we interpret C as a contraction constant attenuating, on average, the local expansion (governed by variances of Q, and $\mathcal{D}$) of the loss around the mean of Q. Exploiting the PAC-Bayes supermartingales bounds of Chapter 2 and Chugg *et al.* (2023) alongside Poincaré inequality leads to the following.

Theorem 4.3.2. For any $C > 0$, any $\frac{2}{C} > \lambda > 0$, any data-free prior P, any $\ell \geq 0$ and any $\delta \in [0, 1]$, we have, with probability at least $1 - \delta$ over the sample $\mathcal{S}$, for any $m > 0$, any Q being $\texttt{Poinc}(c_P)$, $\texttt{QSB}(\ell, C)$ and $\ell(., \mathbf{z}) \in \mathrm{H}^1(\mathrm{Q})$ for all $\mathbf{z}$,

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq \frac{1}{1 - \frac{\lambda C}{2}} \left(\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{\mathrm{KL}(\mathrm{Q}, \mathrm{P}) + \log(1/\delta)}{\lambda m} \right) + \frac{\lambda}{2 - \lambda C} c_P(\mathrm{Q}) \underset{\mathbf{z}\sim\mathcal{D}}{\mathbb{E}} \left[\underset{h\sim\mathrm{Q}}{\mathbb{E}} \left(\|\nabla_h \ell(h, \mathbf{z})\|^2 \right) \right].$$

This theorem shows that, for any posterior being $\texttt{QSB}$ *w.r.t.* the distribution $\mathcal{D}$, fast rates are achievable as long as $\hat{\mathsf{R}}_{\mathcal{S}_m} \approx 0$, and expected gradients are vanishing. While the first condition is often involved for deep neural networks in the overparametrised setting, the second holds if a flat minimum has been reached through the optimisation process. Then, taking $\lambda = \frac{1}{C}$ ensures an anytime-valid PAC-Bayesian bound with a fast rate of $\frac{1}{m}$. Otherwise, for a fixed m, taking $\lambda = \frac{m^{-\alpha}}{C}$, $\alpha \in \left[0; \frac{1}{2}\right]$ allows to adapt the convergence speed *w.r.t.* the behaviour of the gradients. In the case of constant gradients, we recover a convergence rate of $\frac{1}{\sqrt{m}}$, matching ALQUIER *et al.* (2016, Theorem 4.1).

On the role of flat minima in PAC-Bayes learning. Theorem 4.3.2 suggests that, in order to attain good generalisation ability, the mean of Q has to be close from two minima: *(i)* on $\hat{\mathsf{R}}_{\mathcal{S}_m}$ in order to make $\hat{\mathsf{R}}_{\mathcal{S}_m}$ small, and *(ii)* on $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\|\nabla_h \ell(h,\mathbf{z})\|^2]$ to make the gradients small. The variance of Q has to fit the flatness of those minima, the flatter they are, the larger the variance in order to shrink the expected terms on the right-hand-side of Theorem 4.3.2. Finally, the KL term invites, *e.g.* for Gaussian distributions, to consider high variances, hence flat minima to maintain a small value of the bound.

A focus on C. Taking $\lambda = \frac{1}{C}$ in Theorem 4.3.2 attenuates the impact of the prior distribution and amplifies the gradient term. Then, a small C is desirable when working with flat minima to attenuate an ill-designed prior. Having a small C is reachable in practice: we show in Section 4.6, for a classification task on MNIST, that the QSB assumption is verified with C strictly smaller than 1 when considering neural networks.

High probability bounds with fast rates, a paradox? GRUNWALD *et al.* (2021, page 7) showed that, for a trivial $\mathcal{H} = \{h\} \subset \mathbb{R}^d$, for any loss, any *i.i.d.* dataset $\mathcal{S}_m$ with variance σ^2, we have asymptotically, with probability at least α, for a constant C_α depending on α and $\mathcal{N}(\mathbf{0}, \mathrm{Id})$, we have $\mathsf{R}_{\mathcal{D}}(h) \geq \hat{\mathsf{R}}_{\mathcal{S}_m}(h) + C_\alpha \frac{\sigma^2}{\sqrt{m}}$. Is it paradoxical with Theorem 4.3.2? The answer is no: the bound in GRUNWALD *et al.* (2021) gives an asymptotic lower bound on the convergence of $\hat{\mathsf{R}}_{\mathcal{S}_m}(h)$ to $\mathsf{R}_{\mathcal{D}}(h)$. Theorem 4.3.2 informs us on how $\mathsf{R}_{\mathcal{D}}$ is getting closer from $\frac{1}{1-\lambda/2}\hat{\mathsf{R}}_{\mathcal{S}_m}$ which converges to $\frac{1}{1-\lambda/2}\mathsf{R}_{\mathcal{D}} > \mathsf{R}_{\mathcal{D}}$ as the loss is non-negative. Theorem 4.3.2 then show the existence of a 'transition regime' involving a fast rate. Once $\frac{1}{1-\lambda/2}\hat{\mathsf{R}}_{\mathcal{S}_m}$ is reached, the clower bound of GRUNWALD *et al.* (2021) ensures an asymptotic regime with slow convergence rate. Note that such transition regimes already appeared in the literature in TOLSTIKHIN and SELDIN (2013) and MHAMMEDI *et al.* (2019) at the cost of additional variance terms compared to Theorem 4.3.2. However, such fast rates have never been linked before to flat minima (and optimisation in general), highlighting the potential of our bound to explain the ability of deep neural networks to generalise well in the overparametrised setting (m far smaller than the dimension of $\mathcal{H}$), where flat minima are likely to be reached, as studied, *e.g.*, in DZIUGAITE *et al.* (2020), showing correlations between flat minima and generalisation for various learning problems.

Proof of Theorem 4.3.2. We start from CHUGG *et al.* (2023, Corollary 17) instantiated with a single λ, *i.i.d.* data and a prior P. With probability at least $1-\delta$, for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ and $m > 0$:

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda m} + \frac{\lambda}{2}\left(\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})^2]\right]\right),$$

where $\mathbf{z} \sim \mathcal{D}$ is independent of $\mathcal{S}$. We study the last term on the right-hand side. First, applying Fubini's theorem gives:

$$\begin{aligned}\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})^2]\right] &= \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[\ell(h,\mathbf{z})^2]\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathrm{Var}_{h\sim\mathrm{Q}}\left(\ell(h,\mathbf{z})\right) + \left(\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[\ell(h,\mathbf{z})]\right)^2\right].\end{aligned}$$

As for any $\mathbf{z}$, $\ell(.,\mathbf{z}) \in \mathrm{H}^1$, we apply Poincaré's inequality to obtain:

$$\leq \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[c_P(\mathrm{Q})\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right) + \left(\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[\ell(h,\mathbf{z})]\right)^2\right].$$

Using that Q is $\texttt{QSB}(\ell, C)$ and re-organising the terms gives:

$$\begin{aligned}\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq \frac{1}{1-\frac{\lambda C}{2}}\left(\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda m}\right. \\ \left. + \frac{\lambda}{2-\lambda C} c_P(\mathrm{Q})\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right]\right..\end{aligned}$$

■

It is possible to go beyond the $\texttt{QSB}$ assumption. This comes at the cost of an upper bound on $\mathsf{R}_{\mathcal{D}}$ as well as a supplementary Poincaré assumption on $\mathcal{D}$.

Corollary 4.3.1. For any $C > 0$, any $\delta \in (0,1)$ any $\frac{2}{C} > \lambda > 0$, any data-free prior P, any $\ell \geq 0$ such that, for any $\mathbf{z} \in \mathcal{Z}$, we have $\ell(.,\mathbf{z}) \in \mathrm{H}^1$ and for any h, the loss function $\ell(h,.)$ is $\mathcal{C}^1$ almost everywhere on $\mathcal{Z}$. If the data distribution $\mathcal{D}$ is $\texttt{Poinc}(c_P)$, then with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, any posterior Q being $\texttt{Poinc}(c_P)$ with $\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq C$:

$$\begin{aligned}\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq \frac{1}{1-\frac{\lambda C}{2}}\left(\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda m}\right) \\ + \frac{\lambda}{2-\lambda C}\left(c_P(\mathrm{Q})\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right] + c_P(\mathcal{D})\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left(\left\|\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[\nabla_z \ell(h,\mathbf{z})]\right\|^2\right)\right).\end{aligned}$$

Proof is deferred to Section C.3.1. Corollary 4.3.1 states that, if Q reached a flat minimum (meaning $\|\nabla_h \ell\|$ is small), and this minimum is robust to the training dataset (meaning $\|\nabla_{\mathbf{z}} \ell\|$ is small), then a fast rate is attainable while only requiring an upper bound on $\mathrm{R}_{\mathcal{D}}(\mathrm{Q})$. This conclusion holds when $\mathcal{D}$ $\texttt{Poinc}$, encompassing the case of Gaussian mixtures (Schlichting, 2019), which can approximate any smooth density (as recalled in Gat *et al.*, 2022). However, the Poincaré constant of a general mixture is not known, and the upper bound of Schlichting (2019) scales with the number of components, involving potentially high χ^2 divergences.

Comparison with Gat *et al.* (2022). We compare Corollary 4.3.1 with Gat *et al.* (2022, Theorems 3.5, 3.6). First, our result holds with the assumption that $\mathcal{D}$ follows a Poincaré inequality, which is strictly less restrictive than assuming a log-Sobolev inequality (Proposition C.1.1). Second, they assume a bounded loss and their result holds only for classification problem satisfying a technical assumption on the label repartition (see their Lemma 3.3) while ours holds for any learning problem at the sole assumption of a bounded $\mathrm{R}_{\mathcal{D}}(\mathrm{Q})$, allowing ℓ to be non-negative. Moreover, note that to conclude their proof, Gat *et al.* (2022) had to use a uniform bound on $\mathbb{E}\mathbf{z}[\|\nabla_{\mathbf{z}} \ell\|]$ in their Theorem 3.5 to have a tractable bound, thus the benefits of gradient norm is unclear. While they overcome this limitation in Gat *et al.* (2022, Theorem 3.6), the explicit influence of the gradient norm appears within an exponential moment on the losses (attenuated by a logarithm). However, a major limitation is that this exponential moment is averaged *w.r.t.* P, being data-free. Thus, the associated gradients have no apparent reason to be small, and their result cannot be linked to flat minima, contrary to Corollary 4.3.1 involving expected gradients *w.r.t.* Q, being the output of an optimisation process.

4.3.2 Towards fully empirical bound for gradient-Lipschitz functions.

In this section, we assume the loss ℓ is such that, for any $\mathbf{z} \in \mathcal{Z}$, the gradient $\nabla_h \ell(.,\mathbf{z})$ is G-Lipschitz, which is often considered for convergence bounds in optimisation. A large part of high-probability PAC-Bayes bounds are fully empirical: this has numerous advantages including in-training numerical evaluation of generalisation as well as novel PAC-Bayesian algorithms, minimising such empirical bounds; see (Dziugaite and Roy, 2017; Perez-Ortiz *et al.*, 2021b; Viallard *et al.*, 2023b) among others. However, Theorem 4.3.2 and Corollary 4.3.1 are not fully empirical and thus, do not have such desirable properties. We circumvent this issue in Theorem 4.3.3.

Theorem 4.3.3. For any $C_1, C_2, c > 0$, any data-free prior P, any $\ell \geq 0$ being $\mathcal{C}^2$ and any $\delta \in [0,1]$, we have, with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, any Q being $\mathtt{Poinc}(c_P)$ with constant c, $\mathtt{QSB}(\ell, C_1)$, $\mathtt{QSB}\left(\|\nabla_h \ell\|^2, C_2\right)$ and $\ell(.,\mathbf{z}), \|\nabla_h \ell\|^2(.,\mathbf{z}) \in \mathrm{H}^1(\mathrm{Q})$ for all $\mathbf{z}$,

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{2c}{C_1} \mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^m \|\nabla_h \ell(h,\mathbf{z}_i)\|^2\right] + 2\left(C_1 + c\frac{4cG^2 + C_2}{C_1}\right)\frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{m}.$$

Proof is deferred to Section C.3.2. Here, we showed that to attain fast rates, the $\mathtt{QSB}$ assumption has to be reached for both the loss and its gradient. This suggests several things on the flat minimum that has to be reached by Q (designed from $\hat{\mathsf{R}}_{\mathcal{S}}$): first, it needs to be close from a flat minimum of $\mathsf{R}_{\mathcal{D}}$ to satisfy the $\mathtt{QSB}$ assumption. Second, this minimum also ensures the contraction of the gradients. We then are able to derive an empirical generalisation bound, involving both empirical loss and gradients. Not only Theorem 4.3.3 yields, to our knowledge, the first PAC-Bayesian algorithm involving gradient terms, but also can be translated to a generalisation metric in order to understand generalisation. Such an idea has been exploited recently (NEYSHABUR *et al.*, 2017; DZIUGAITE *et al.*, 2020; JIANG *et al.*, 2020). In particular, from $\hat{\mathsf{R}}_{\mathcal{S}}(\mathrm{Q})$, NEYSHABUR *et al.* (2017) derived a notion of *sharpness*, stated in Equation (4.1), aiming to be informative on the flatness of the reached minima for any $\mathrm{Q} = \mathcal{N}(\mu_Q, \sigma^2\mathrm{Id})$. This notion is defined by

$$\mathop{\mathbb{E}}_{\nu\sim\mathcal{N}(\mathbf{0},\sigma^2\mathrm{Id})}\left[\hat{\mathsf{R}}_{\mathcal{S}_m}(\mu_{\mathrm{Q}}+\nu) - \hat{\mathsf{R}}_{\mathcal{S}_m}(\mu_{\mathrm{Q}})\right]. \tag{4.1}$$

Theorem 4.3.3 enhance this notion of sharpness by involving the empirical gradients when Q is $\mathtt{QSB}(\ell, C_1)$:

$$\mathrm{Sharp}_{\frac{\sigma^2}{C_1}}(\mathrm{Q}) := \mathop{\mathbb{E}}_{\nu\sim\mathcal{N}(\mathbf{0},\sigma^2\mathrm{Id})}\left[\left(2\hat{\mathsf{R}}_{\mathcal{S}_m} + \frac{\sigma^2}{C_1}\text{G-}\hat{\mathsf{R}}_{\mathcal{S}_m}\right)(\mu_{\mathrm{Q}}+\nu) - \left(2\hat{\mathsf{R}}_{\mathcal{S}_m} + \frac{\sigma^2}{C_1}\text{G-}\hat{\mathsf{R}}_{\mathcal{S}_m}\right)(\mu_{\mathrm{Q}})\right], \tag{4.2}$$

where $\text{G-}\hat{\mathsf{R}}_{\mathcal{S}_m}(h) = \frac{1}{m}\sum_{i=1}^m \|\nabla_h \ell(h,\mathbf{z}_i)\|^2$. This gradient term can be seen as an empirical Fisher information, linked to the second-order moment derivative. Thus, (4.2) involves a notion of flatness on both the loss and its gradient, contrary to (4.1). For the sake of clarity, we particularise Theorem 4.3.3 in Corollary 4.3.2 with Gaussian distributions and this novel notion of sharpness.

Corollary 4.3.2. For any $C_1, C_2 > 0$, any fixed variance $\sigma^2 > 0$, any data-free prior $\mathrm{P} = \mathcal{N}(\mu_{\mathrm{P}}, \sigma^2 \mathrm{Id})$, any nonnegative loss ℓ being $\mathcal{C}^2$ and any $\delta \in [0,1]$, we have, with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, any $\mathrm{Q} = \mathcal{N}(\mu_{\mathrm{Q}}, \sigma^2 \mathrm{Id})$ being $\mathtt{QSB}(\ell, C_1)$, $\mathtt{QSB}\left(\|\nabla_h \ell\|^2, C_2\right)$ and $\ell(., \mathbf{z}), \|\nabla_h \ell\|^2(., \mathbf{z}) \in \mathrm{H}^1(\mathrm{Q})$ for all $\mathbf{z}$,

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\hat{\mathsf{R}}_{\mathcal{S}_m}(\mu_{\mathrm{Q}}) + \text{G-}\hat{\mathsf{R}}_{\mathcal{S}_m}(\mu_{\mathrm{Q}}) + \mathrm{Sharp}_{\frac{\sigma^2}{C_1}}(\mathrm{Q}) + \mathcal{O}\left(\frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{m}\right).$$

4.4 Generalisation ability of Gibbs distributions with a log-Sobolev prior

One limitation of the results given in Section 4.3 is that the KL divergence term remains uncontrolled in general as its formulation depends on the nature of P and Q. A close form exists for Gaussian distributions for instance, but this class of distribution is limiting. Perpetrating the spirit of Catoni (2007), we go beyond the Gaussian distributions to focus on the Gibbs posteriors which have naturally appeared in PAC-Bayes through the use of tools from statistical physics. We show that log-Sobolev inequalities allow us to control the KL divergence of such distributions *w.r.t.* their priors.

Controlling the KL divergence when Q is a Gibbs posterior. Lemma 4.4.1 exploits the fact that the KL divergence can be formulated as an entropy *w.r.t.* the prior distribution P. It then shows that the KL divergence of the Gibbs posterior $\mathrm{P}_{-\gamma \hat{\mathsf{R}}_{\mathcal{S}_m}}$ *w.r.t.* P is upper bounded by gradient terms as long as P satisfies a log-Sobolev inequality.

Lemma 4.4.1. For any m, P being $\mathtt{L\text{-}Sob}(c_{LS})$, any $\ell \geq 0$ such that for any $\mathbf{z}$, $\ell(., \mathbf{z}) \in \mathrm{H}^1(\mathrm{P})$, we have, for any $\gamma > 0$:

$$\mathrm{KL}\left(\mathrm{P}_{-\gamma \hat{\mathsf{R}}_{\mathcal{S}_m}}, \mathrm{P}\right) \leq \frac{\gamma^2 c_{LS}(\mathrm{P})}{4} \mathop{\mathbb{E}}_{h \sim \mathrm{P}_{-\gamma \hat{\mathsf{R}}_{\mathcal{S}_m}}} \left[\|\nabla_h \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\|^2\right].$$

Proof is deferred to Appendix C.3.3. The crucial message of this lemma is that, a flat minimum of $\hat{\mathsf{R}}_{\mathcal{S}}$ allows controlling the KL divergence. This message is new and independent of Section 4.3 which focus on flat minima reached for $\mathsf{R}_{\mathcal{D}}$. Note that in this case, the KL divergence has an explicit formulation. However it involves to calculate the exponential moment $\mathbb{E}_{h \sim \mathrm{P}}[\exp(-\gamma \hat{\mathsf{R}}_{\mathcal{S}_m})]$ which is costly in practice. On the contrary, we only need to estimate a second-order moment over $\mathrm{P}_{-\gamma \hat{\mathsf{R}}_{\mathcal{S}_m}}$.

Generalisation ability of Gibbs posteriors. When Gibbs posteriors are involved, KL divergence is controllable by a gradient term. An ideal way to conclude would be, as in Section 4.3 to involve Poincaré inequality. However, Gibbs posterior are not necessarily satisfying a Poincaré inequality as in Section 4.3, we then need to make supplementary assumptions on the loss.

Theorem 4.4.1. For any $C > 0$, any $\gamma > 0$, any prior P being $\texttt{L-Sob}(c_{LS})$, any $\ell \geq 0$ and any $\delta \in [0,1]$, we have the following inequalities. If $\ell \in [0,1]$, then with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, and any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:

$$\begin{aligned}&\mathsf{R}_{\mathcal{D}}(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}})\\ &\quad\leq 2\left(\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}) + \frac{\gamma^2 c_{LS}(\mathrm{P})}{4m}\underset{h\sim\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}}{\mathbb{E}}\left[\|\nabla_h \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\|^2\right] + \frac{\log(1/\delta)}{m}\right).\end{aligned}$$

If $\ell = \ell_1 + \ell_2$ with ℓ_1 convex, twice differentiable and ℓ_2 bounded, assume that P satisfies the conditions of Proposition 4.2.2. Then for any $\frac{2}{C} > \lambda > 0$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, such that Q is $\texttt{QSB}(\ell, C)$ and $\ell(.,\mathbf{z}) \in \mathrm{H}^1(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}})$:

$$\begin{aligned}&\mathsf{R}_{\mathcal{D}}(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}})\\ &\leq \frac{1}{1-\frac{\lambda C}{2}}\left(\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}) + \frac{\gamma^2 c_{LS}(\mathrm{P})}{4\lambda m}\underset{h\sim\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}}{\mathbb{E}}\left[\|\nabla_h \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\|^2\right] + \frac{\log(1/\delta)}{\lambda m}\right)\\ &\qquad\qquad + \frac{\lambda e^{4\|\ell_2\|_\infty} c_{LS}(\mathrm{P})}{4-2\lambda C}\underset{\mathbf{z}\sim\mathcal{D}}{\mathbb{E}}\left[\underset{h\sim\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}}{\mathbb{E}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right].\end{aligned}$$

Proof is deferred to Appendix C.3.4. Note that we could have derived analogous to Corollary 4.3.1 at the cost of a supplementary Poincaré assumption on $\mathcal{D}$. The influence of the inverse temperature γ is quadratic: this is the price to pay to fit the dataset and reduce the influence of the prior. This dependency is therefore attenuated by a gradient term, small if a flat minimum on the empirical risk has been reached. This suggests that in the case of Gibbs posteriors with log-Sobolev prior, reaching a flat minima on $\hat{\mathsf{R}}_{\mathcal{S}_m}$ controls not only $\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q})$, but also the KL divergence and this last point is not reachable when considering Poincaré distributions. The other gradient term comes from Section 4.3 and requires to be close from a flat minimum on $\mathsf{R}_{\mathcal{D}}$ to attain fast rates.

4.5 On the benefits of the gradient norm in Wasserstein PAC-Bayes learning

In Sections 4.3 and 4.4, we provided various generalisation bounds, benefiting from flat minima. However, our results involve a KL divergence, implying absolute continuity of Q *w.r.t.* P, incompatible with the case of deterministic predictors (Dirac distributions). To circumvent this issue, a recent line of work emerged, involving integral probability metrics, with a particular focus on the 1-Wasserstein distance as in Chapters 5 and 6 and AMIT *et al.* (2022). The idea behind these works is to replace the change of measure inequality (CSISZÁR, 1975; DONSKER and VARADHAN, 1976) by the Kantorovich-Rubinstein duality (VILLANI, 2009) to trade a KL for a Wasserstein. We go even further here by obtaining the first PAC-Bayesian bound involving directly a 2-Wasserstein distance (see definition C.1.1), trading Lipschitz assumption for gradient-Lipschitz one (well-suited for optimisation). To do so, we first derive a novel change of measure inequality.

> **Theorem 4.5.1.** Assume $\mathcal{H}$ to have a finite diameter $D > 0$. Then for any function $f : \mathcal{H} \to \mathbb{R}$ with G-Lipschitz gradients, the following holds: for all distributions $\mathrm{P}, \mathrm{Q} \in \mathcal{M}(\mathcal{H})^2$,
>
> $$\mathbb{E}_{h\sim \mathrm{Q}}[f(h)] \leq \frac{G}{2}\mathrm{W}_2^2(\mathrm{Q}, P) + \mathbb{E}_{h\sim \mathrm{P}}[f(h)] + D\mathbb{E}_{h\sim \mathrm{Q}}[\|\nabla f(h)\|].$$

Proof is deferred to Appendix C.3.5 Theorem 4.5.1 shows it is possible when gradients are Lipschitz, to obtain a duality formula involving the gradient of the considered function at the price of a linear dependency on the diameter of $\mathcal{H}$. Theorem 4.5.1 is also linked to the change of measure inequality when the prior distribution satisfies a log-Sobolev inequality.

> **Corollary 4.5.1.** Assume that P is such that $d\mathrm{P} \propto \exp(-V)dx$ with V being $\mathcal{C}^2$ and P is $\texttt{L-Sob}(c_{LS})$. Then, for any $R > 0$, any f with gradients G-Lipschitz on $\mathcal{B}(\mathbf{0}, R)$, and any distributions P, Q,
>
> $$\begin{aligned}&\mathbb{E}_{h\sim \mathrm{Q}}\left[f\left(\mathcal{P}_R(h)\right)\right] \\ &\qquad \leq \frac{Gc_{LS}(\mathrm{P})}{4}\mathrm{KL}(\mathrm{Q}, \mathrm{P}) + \mathbb{E}_{h\sim \mathrm{P}}\left[f\left(\mathcal{P}_R(h)\right)\right] + 2R\mathbb{E}_{h\sim \mathrm{Q}}\left[\|\nabla f\left(\mathcal{P}_R(h)\right)\|\right],\end{aligned}$$
>
> where $\mathcal{P}_R$ denotes the Euclidean projection on $\mathcal{B}(\mathbf{0}, R)$.

Proof is deferred to Appendix C.3.6. Corollary 4.5.1 involves a KL divergence and an Euclidean predictor space $\mathcal{H} = \mathbb{R}^d$. This comes at the cost of approximating Q, P by $\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}$. Thus, R is now an hyperparameter which arbitrates a tradeoff between the quality of our approximations and the looseness of the bound (if the gradient norm is large). A notable strength is that the smoothness assumption is relaxed on smoothness over $\mathcal{B}(\mathbf{0}, R)$.

From Theorem 4.5.1, we now derive a novel generalisation bound allowing deterministic predictors.

Theorem 4.5.2. Let $\delta \in (0,1)$ and $\mathrm{P} \in \mathcal{M}(\mathcal{H})$ a data-free prior. Assume $\mathcal{H}$ has a finite diameter $D > 0$, $\ell \geq 0$ and that for any m, the generalisation gap $\Delta_{\mathcal{S}_m}$ is G gradient-Lipschitz. Assume that $\mathbb{E}_{h\sim\mathrm{P}}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,z)^2] \leq \sigma^2$, then the following holds with probability at least $1-\delta$, for any $m > 0$ and any Q:

$$\mathsf{R}_D(\mathrm{Q}) \leq \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{G}{2}\mathrm{W}_2^2(\mathrm{Q},\mathrm{P}) + \sqrt{\frac{2\sigma^2 \log\left(\frac{1}{\delta}\right)}{m}} \\ + D\mathbb{E}_{h\sim\mathrm{Q}}\left(\left\|\nabla_h \mathsf{R}_{\mathcal{D}}(h) - \nabla_h \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right\|\right).$$

Proof is deferred to Appendix C.3.7. Theorem 4.5.2 is not the first generalisation bound to involve a 2-Wasserstein distance (Lugosi and Neu, 2022; Lugosi and Neu, 2023). However, those results involve infinitely smooth loss functions. Also, results from Amit *et al.* (2022), Chapters 5 and 6 using 1-Wasserstein can be directly relaxed on bounds involving the 2-Wasserstein, while still requiring a Lipschitz loss. On the contrary, our result holds for any nonnegative gradient-Lipschitz $\Delta_{\mathcal{S}_m}$, which is well-suited for optimisation. Theorem 4.5.2 involves a slow rate of $\frac{1}{\sqrt{m}}$ as we have to control the generalisation gap *w.r.t.* to P. It is possible to make appear the gradients expected over P using the QSB assumption, but we have no reason to expect those gradients to be small, we then controlled this term uniformly by σ^2. Another restriction of our result compared to previous ones is that it holds for $\mathcal{H}$ having a finite diameter, however, having a small expected $\|\nabla_h \mathsf{R}_{\mathcal{D}} - \nabla_h \hat{\mathsf{R}}_{\mathcal{S}_m}\|$ over Q (which is the case when flat minima on both empirical and true risks are reached) allows taking D large, and thus, having good approximations of measures on a Euclidean space through orthogonal projections as in Corollary 4.5.1.

4.6 An empirical study of Assumption 4.3.1 for neural networks

In this section, we check empirically whether the QSB assumption is verified for neural networks. This allows us to verify if Theorem 4.3.2 is useful to understand the generalisation ability of neural nets.

Experimental protocol. We consider classification tasks on two datasets: MNIST (LeCun, 1998) and FashionMNIST (Xiao *et al.*, 2017). We kept the original training set $\mathcal{S}_m$ and the original test set denoted by $\mathcal{T}_n$ (of size n). We consider the convolutional neural network of Springenberg *et al.* (2015) adapted for MNIST and FashionMNIST. The model is composed of 4 layers containing 10 channels with a $5{\times}5$-kernel; we set the stride and the padding to 1, except for the second layer, where it is fixed to 2. Each of these (convolutional) layers is followed by a Leaky ReLU activation function. Moreover, an average pooling with a $8{\times}8$-kernel is performed before the Softmax activation function. To initialise the weights of the network, we use Glorot and Bengio (2010) uniform initializer, while the biases are initialised in $[-\frac{1}{\sqrt{250}}, +\frac{1}{1/\sqrt{250}}]$ uniformly (except the first layer, the interval is $[-\frac{1}{5}, +\frac{1}{5}]$). Hence, in this case, $\mathcal{H}$ is the set of neural networks with a fixed architecture, and parametrised with a vector $\mathbf{w}$. while the posterior distribution Q is a Gaussian measure $\mathcal{N}(\mathbf{w}, \sigma^2\mathrm{Id})$ centered on the parameters $\mathbf{w}$ associated with the model; σ is set to 10^{-4}. Note that this distribution respects the $\texttt{Poinc}(c_P)$ assumption; see Section 4.3.1. We train the neural network with the (vanilla) stochastic gradient descent algorithm, where the batch size is equal to 512, and the learning rate is fixed to 10^{-2}. We train for at least 10^4 gradient steps and finish the current epoch when this number of iterations is reached. Our loss ℓ is the bounded cross-entropy loss of Dziugaite and Roy (2017, Section D).
In Figure 4.1, we report the evolution of three quantities: *(i)* the estimated value of C, *(ii)* the test risk $\hat{\mathrm{R}}_{\mathcal{T}_n}(\mathrm{Q})$ and *(iii)* the test risk with the 01-loss. More precisely, for computational reasons, the risks and C are estimated by sampling one hypothesis $h \sim \mathrm{Q}$ and by computing the values on a mini-batch of $\mathcal{T}_n$ (with 512 examples) at each iteration. Then, Figure 4.1 represents averaged values on 5 runs, each point of the curve representing the average on 100 iterations of the training process (for 10^4 iterations we only plot 10^2 averaged points for clarity).

Empirical findings. Figure 4.1 illustrates that, when neural networks are involved for two classification tasks, Q evolves during the optimisation process while maintaining the QSB property with constant $C < 1$. For both MNIST and FashionMNIST, the constant C decreases from approximately 0.55 to 0.45. We deduce two things from this: *(i)* the learning phase, while optimising $\hat{\mathrm{R}}_{\mathcal{S}_m}$ also gain in generalisation ability, shrinking the averaged loss on new data which is translated by a smaller C; and (ii),

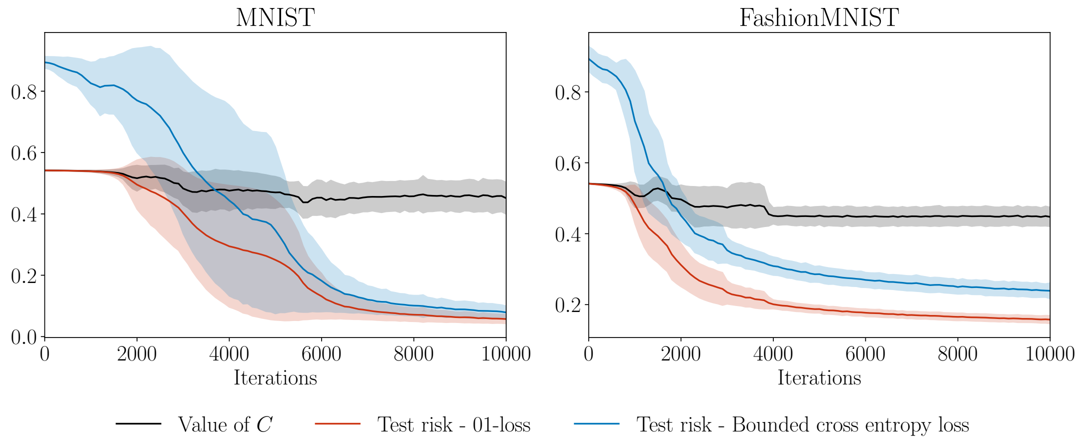

Figure 4.1. *Evolution of the test risks (with the* 01*-loss and the bounded cross-entropy loss) and the value of* C *during the training phase.*

having a data-free P (0 iteration) being QSB with $C < 1$ suggests that the architecture of our neural network also has an influence on the QSB assumption. As precised in Section 4.3, having $C < 1$ attenuates the impact of the KL term, thus P. This is desirable as it allows the optimiser to deeply explore the predictor space when P yields poor performances. We also note that the generalisation ability of Q on the training loss nearly matches the performance on the 0-1 loss for MNIST but is deteriorated for FashionMNIST, this invites to study more deeply the design of such surrogates in future work.

Finally, the take-home message of this study is that the QSB assumption is verified for neural networks on MNIST. Such an empirical confirmation is crucial as it is required for our main result (Theorem 4.3.2) and thus confirms that, for neural networks, reaching flat minima during the optimisation phase translates in increased generalisation ability.

4.7 Conclusion

This chapter showed that it is possible to exploit the benefits of a successful optimisation process to obtain faster rates, making a promising step forward a better understanding of deep neural networks, whose generalisation ability correlates well to flat minima. However, while we exploited potential benefits of optimisation process to make PAC-Bayes in line with optimisaiton, we still do not know whether an optimisation algorithm will reach a flat minima. This is somewhat unconsistent as optimisation processes are often supported by deterministic convergence guarantees. To fill this gap we show in Chapter 5 that it is possible to incorporate directly optimisation guarantees onto PAC-Bayes bounds, making a supplementary step towards optimisation.

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

This chapter is based on the following papers

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

Contents

Abstract

To make PAC-Bayes consistent with practical optimisation which often considers deterministic predictors, we need to consider PAC-Bayes learning beyond the KL divergence term which has been a cornerstone of PAC-Bayes since its emergence. In this chapter, we develop PAC-Bayes learning with Wasserstein distances, allowing to trade statistical assumptions for geometric ones. We also develop an explicit bridge with optimisation by incorporating the convergence guarantees of the *Bures-Wasserstein SGD* into a generalisation bound. This is possible when considering the prior distribution as the learning objective.

5.1 Introduction and state-of-the-art results

In Chapters 2 to 4, we challenged many information-theoretic visions of PAC-Bayes (Figure 1.1) by limiting statistical assumptions, attenuating the impact of the prior seen as initialisation in both batch and online settings. However, we always involved a KL divergence term as a complexity measure of the predictor space, and this term is strongly linked to information theory. Indeed, a KL divergence focuses on posteriors being absolutely continuous *w.r.t.* the prior, meaning that it is possible to transfer information to posteriors with similar shape to the prior (and thus a finite KL divergence). While this vision makes perfectly sense from an information theoretic perspective, it is harder to justify such a condition from an optimisation stance. Indeed, Dirac prior and posterior (*i.e.* deterministic predictors) are often considered in practice and this makes the KL infinite. Furthermore, KL divergence suffers from limitations as it does not satisfy classical properties such as the triangle inequality or even symmetry: it is challenging to exploit geometric properties of the measure space and the loss function through it. Then we might ask whether it is possible to maintain the flexibility of PAC-Bayes by involving another complexity measure, more compatible with optimisation. In this chapter we will develop PAC-Bayes learning theory based on Wasserstein distances, issued from optimal transport and compatible with such considerations.

PAC-Bayes learning with Wasserstein distances. A recent line of work led by Amit *et al.* (2022) investigates PAC-Bayes generalisation bounds with a Wasserstein distance rather than the KL. This idea has been simultaneously developed by Ohana *et al.* (2023) for sliced adaptive Wasserstein distances. Also the recent work of Mbacke *et al.* (2023) provides PAC-Bayesian bounds for adversarial generative models where the quantity of interest is a Wasserstein distance (although the complexity measure remains a KL divergence).

In the present chapter, we propose a major development of the emerging *Wasserstein PAC-Bayes* (WPB) theory. Amit *et al.* (2022) provided the first high-probability WPB bounds with explicit convergence rates (for bounded losses) only for finite predictor classes or for linear regression problems. We extend those results to a broader framework including uncountable predictor classes and unbounded losses. We first propose a novel WPB bound valid on any compact for bounded lipschitz losses. From this, we demonstrate that the WPB framework allows to bypass both the compactness assumption on the predictor class and the bounded loss assumption: Wasserstein PAC-Bayes only requires Lipschitz or smooth functions to be used. We obtain explicit bounds for the case of prior and posterior distributions taken within a compact space of Gaussian measures. We also extend those results to the case of data-dependent priors, which is of interest when one compares the output of an algorithmic procedure to its minimisation objective.

As Wasserstein distance recently appeared as complexity measure in expected generalisation bounds (see *e.g.* RODRIGUEZ-GALVEZ *et al.*, 2021), the high-probability Wasserstein PAC-Bayes bounds presented here investigate deeper this lead. We also go a step further by showing that Wasserstein PAC-Bayes allows to reap the benefits of optimisation guarantees within generalisation. To the best of our knowledge, no previous PAC-Bayes bound has achieved this goal. More precisely, we focus on the Bures-Wasserstein SGD (ALTSCHULER *et al.*, 2021; LAMBERT *et al.*, 2022) and show that the output of this algorithm, with enough data, after enough optimisation steps, is able to generalise well, independently of the quality of the initialisation point. The take-home message is that if an optimisation method has convergence guarantees with respect to a Wasserstein distance, then WPB theory allow us to determine, before any training, whether the algorithmic output will generalise well.

Outline. The reminder of this chapter is structured as follows: we state in Section 5.1.1 the framework and notation. In Section 5.1.2, we describe how current PAC-Bayes procedures are designed and how their efficiency is evaluated, and we discuss current limitations. In Section 5.1.3, we describe our main contributions, showing how we establish a WPB theory (using techniques which differ from those in AMIT *et al.*, 2022) in order to exploit the optimisation results of LAMBERT *et al.* (2022). Section 5.2 gathers results for compact predictor spaces, Section 5.3 gives WPB bounds for Gaussian prior and posterior, Section 5.4 contains a WPB bound with a data-dependent prior for unbounded Lipschitz losses. Section 5.5 establishes a link between optimisation and generalisation by exploiting the results of LAMBERT *et al.* (2022) to establish new generalisation guarantees for the Bures-Wasserstein SGD. We defer to Appendix D.1 additional background notes and to Appendix D.2 proofs which are not essential to the understanding of our contributions.

5.1.1 Framework

Learning theory framework. We consider a *learning problem* specified by a tuple $(\mathcal{H}, \mathcal{Z}, \ell)$ of a set $\mathcal{H}$ of predictors, a data space $\mathcal{Z}$, and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$. We consider a finite dataset $\mathcal{S}_m = (\mathbf{z}_i)_{i\in\{1..m\}} \in \mathcal{Z}^m$ and assume that sequence is *i.i.d.* following the distribution $\mathcal{D}$. We always assume that $\mathcal{H} \subseteq \mathbb{R}^d$, we denote by $\Sigma_{\mathcal{H}}$ the associated Borel σ-algebra and we denote by $||.||$ the classical Euclidean norm. We denote by $\mathcal{M}(\mathcal{H})$ the set of probability measures on $\mathcal{H}$. We denote by $\mathcal{P}_1(\mathcal{H})$ (resp. $\mathcal{P}_2(\mathcal{H})$) the subspace of $\mathcal{M}(\mathcal{H})$ of with finite order 1 (resp. order 2) moments *w.r.t.* $||.||$.

Definitions. The *generalisation error* $\mathsf{R}_{\mathcal{D}}$ of any predictor $h \in \mathcal{H}$ is $\mathsf{R}_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h, \mathbf{z})]$, the *empirical error* of h is $\hat{\mathsf{R}}_{\mathcal{S}_m}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, \mathbf{z}_i)$. The *generalisation* *gap* of any h is the quantity $\Delta_{\mathcal{S}_m}(h) = \mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)$ and, for any $\mathrm{Q} \in \mathcal{M}(H)$, $\Delta_{\mathcal{S}_m}(\mathrm{Q}) = \mathbb{E}_{h\sim\mathrm{Q}}[\Delta_{\mathcal{S}_m}(h)]$. In what follows, we let $\mathcal{B}(\mathbf{x}, r)$ (resp. $\bar{\mathcal{B}}(\mathbf{x}, r)$) denote the ball (resp. closed ball) centered in $\mathbf{x} \in \mathbb{R}^d$ of radius r. We define the *Gibbs posterior* associated to the prior $\mathrm{P} \in \mathcal{M}(\mathcal{H})$ as the measure $\mathrm{P}_{-\lambda\hat{\mathsf{R}}_{\mathcal{S}_m}}$ such that $\mathrm{dP}_{-\lambda\hat{\mathsf{R}}_{\mathcal{S}_m}} \propto \exp(-\lambda\hat{\mathsf{R}}_{\mathcal{S}_m}(.))\mathrm{dP}(.)$.

We denote by $\mathsf{BW}(\mathbb{R}^d) \subset \mathcal{P}_2(\mathbb{R}^d)$ the set of non-degenerate Gaussian distributions, also known as the *Bures-Wasserstein space*. For a measurable function $T : \mathbb{R}^d \to \mathbb{R}^d$, and a measure $\mathrm{P} \in \mathcal{P}_1(\mathbb{R}^d)$ we let $T\#\mathrm{P}$ denote the measure such that for any $B \in \Sigma_{\mathbb{R}^d}, T\#\mathrm{P}(B) = \mathrm{P}(T^{-1}(B))$. For any $R > 0$, we denote by $\mathcal{P}_R$ the projection over $\bar{\mathcal{B}}(0_{\mathbb{R}^d}, R)$. Finally, as we consider compact sets of $\mathsf{BW}(\mathbb{R}^d)$, we define for any $0 \leq \alpha \leq \beta, M \geq 0$ the set

$$C_{\alpha,\beta,M} := \left\{\mathcal{N}(m,\Sigma) \in \mathsf{BW}(\mathbb{R}^d) \mid ||m|| \leq M,\ \alpha\mathrm{Id} \preceq \Sigma \preceq \beta\mathrm{Id}\right\}.$$

5.1.2 PAC-Bayes and optimisation: limits and caveats

Optimisation in PAC-Bayes. PAC-Bayesian generalisation bounds are meant to control how well measures derived from a learning algorithm perform on novel data. Those bounds involves a complexity term which is typically a Kullback Leibler (KL) divergence. A prototypic bound is as follows: with probability $1-\delta$, for all measure Q,

$$\Delta_{\mathcal{S}_m}(\mathrm{Q}) \leq \sqrt{\frac{\mathrm{COMP}(\mathrm{Q})}{m}},$$

where COMP is a complexity term involving a data-free prior P and an approximation term $1-\delta$. From an optimisation perspective, this upper bound can be seen as a learning objective, where COMP acts as a regulariser to avoid overfitting on the empirical risk:

$$\mathrm{Q}^* := \underset{\mathrm{Q}\in\mathcal{M}(\mathcal{H})}{\mathrm{argmin}} \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \sqrt{\frac{\mathrm{COMP}(\mathrm{Q})}{m}}.$$

Such algorithms are build to ensure a candidate measure with a good generalisation ability. However the convergence of the optimisation process remains unclear: as $\sqrt{\mathrm{COMP}}$ is not necessarily convex in Q, it is unclear whether an optimisation procedure on the previous learning objective will lead to $\hat{\mathrm{Q}}$ (or a good approximation of it). A good introductory example is to optimise the PAC-Bayesian learning objective for the following complexity term, holding for a loss ℓ being in $[0,1]$:

$$\sqrt{\frac{\mathrm{COMP}(\mathrm{Q})}{m}} := \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P})}{\lambda} + \frac{\lambda}{2m},$$

with λ being usually fine-tuned over a countable grid. This objective, linear in the KL divergence term is optimised by the Gibbs posterior:

$$\mathrm{dQ}^*(h) \propto \exp(-\lambda \hat{\mathsf{R}}_{\mathcal{S}_m}(h))\mathrm{dP}(h).$$

This distribution, while being known analytically, may be hard to compute in practice. A class of methods dedicated to compute or approximate this posterior distribution are the Markov Chain Monte Carlo (MCMC) methods that rely on carefully constructed Markov chains which (approximately) converge to Q^*. However, MCMC methods can be computationally costly and other methods were studied to obtain quickly surrogates of Q^*. In particular, *Variational Inference* (VI) has been developed as a time-efficient solution. VI algorithms aims to estimate a surrogate $\hat{\mathrm{Q}}$ of Q^*, often chosen within a parametric class of measures such as Gaussian measures. For instance, in order to approximate Q^* it is natural to consider the following surrogate:

$$\hat{\mathrm{Q}} = \operatorname*{argmin}_{Q \in \mathcal{C}} \mathrm{KL}(\mathrm{Q}, \mathrm{Q}^*),$$

where $\mathcal{C}$ is a subset of $\mathcal{M}(\mathcal{H})$. When $\mathcal{C}$ is the set of Gaussian measures (also known as the *Bures-Wasserstein* manifold), the convergence of the associated VI algorithm has been studied (Altschuler *et al.*, 2021; Lambert *et al.*, 2022). This candidate $\hat{\mathrm{Q}}$ is approximated after N optimisation steps by a measure $\hat{\mathrm{Q}}_N$ and is then used in McAllester's bound to assess its efficiency:

$$\Delta_{\mathcal{S}_m}(\hat{\mathrm{Q}}_N) \leq \sqrt{\frac{\mathrm{KL}(\hat{\mathrm{Q}}_N, \mathrm{P}) + \log(m/\delta)}{2m}}. \tag{5.1}$$

Role of the prior P. From an optimisation perspective, the conclusion of (5.1) is that if $\hat{\mathrm{Q}}_N$ is a good approximation of $\hat{\mathrm{Q}}$ and if the initialisation P is well-chosen, then the generalisation ability $\hat{\mathrm{Q}}$ is guaranteed to be high. Assuming such a condition on P may be unrealistic. Furthermore the term $\mathrm{KL}(\hat{\mathrm{Q}}_N, \mathrm{P})$ acts as a blackbox as we do not have a theoretical control on how far $\hat{\mathrm{Q}}$ and $\hat{\mathrm{Q}}_N$ diverge from the prior. In particular if the prior is ill-chosen, then we could have $\mathrm{KL}(\hat{\mathrm{Q}}_N, \mathrm{P}) = \mathcal{O}(m)$, making (5.1) vacuous.

Data-dependent priors are not enough to explain the generalisation gain through optimisation. As shown above, in order to have a sound theoretical control on the generalisation ability of the algorithmic output $\hat{\mathrm{Q}}_N$, it is irrelevant to compare it to the initialisation P. Thus, it is legitimate to wonder if the existing PAC-Bayesian techniques using data-dependent priors are enough to fill this gap. To do so, we identify two strategies.

1. Taking Q^* as a 'prior' distribution (as advised by Dziugaite and Roy, 2017) is, at first sight, a convincing answer. However, the use of KL divergence is problematic. Indeed, we cannot make $\hat{\mathrm{Q}}$ appear easily in Equation (5.1) which is the relevant point of interest. Furthermore, to our knowledge, there is no VI algorithm which guarantees that $\mathrm{KL}(\hat{\mathrm{Q}}_N, \mathrm{Q}^*)$ is decreasing.

2. The prior is obtained from an algorithmic method on a fraction of training data. Then, such a bound does not inform us whether the considered optimisation method has been able to reach an optimum during the training phase: similarly to a test bound, it mainly assesses the post-training efficiency of the output of the learning algorithm. A relevant example is Table 3 of Perez-Ortiz *et al.* (2021a) which considers data-dependent priors obtained through SGD. Then as the performance of the prior and the posterior is roughly similar, it is hard to determine whether the associated theoretical guarantee is more meaningful than a test bound as the prior measure could have already converged near a local optimum.

A strategy to replace (5.1). In order to assess whether the output of a learning algorithm enjoys high generalisation, a PAC-Bayes bound should satisfy the following generic form:

$$\Delta_{\mathcal{S}_m}(\hat{\mathrm{Q}}_N) \leq \sqrt{\frac{f(N)\,\mathrm{D}(\mathrm{P},\hat{\mathrm{Q}}) + \varepsilon + \log(m/\delta)}{2m}}, \tag{5.2}$$

where f is a function decreasing to 0 as N goes to infinity, which comes from the optimisation procedure, D is the way to measure the discrepancy between $\mathrm{P}, \hat{\mathrm{Q}}$ (classically it would be the KL divergence) and ε is a residual term which could contain for instance the discrepancy $\mathrm{KL}(\mathrm{Q}^*, \hat{\mathrm{Q}})$ between the approximation and the true minimiser. Such a guarantee would give theoretical evidence that the generalisation ability of $\hat{\mathrm{Q}}_N$ is independent of the choice of the initialisation point P and tends to $\mathcal{O}\left(\sqrt{\frac{\varepsilon+\log(m/\delta)}{m}}\right)$. To the best of our knowledge, there is no work proposing an optimisation procedure such that $\mathrm{KL}(\hat{\mathrm{Q}}_N, \hat{\mathrm{Q}}) \leq f(N)\,\mathrm{KL}(\mathrm{P}, \hat{\mathrm{Q}})$. This lack is unfortunate but not surprising as the KL divergence is not a distance: it is not easy to incorporate optimisation guarantees, often based on geometric properties of the loss, into the KL divergence.

Our aims in this chapter. A legitimate question is then: is it possible to extend the PAC-Bayes theory beyond the KL divergence in order to explain before training, with a bound of the form of (5.2), whether the output of optimisation procedure have high generalisation ability? We structure the present chapter to provide a positive answer to this question. More precisely we develop a WPB bound of the form of (5.2) for the output of the Bures-Wasserstein SGD (Lambert *et al.*, 2022).

5.1.3 Summary of our contributions

To make PAC-Bayes learning useful to explain the generalisation ability of minimisers reached by optimisation algorithms, we develop theoretical results built around Wasserstein distances whose definitions are recalled below.

> **Definition 5.1.1.** The 1-Wasserstein distance between $\mathrm{P}, \mathrm{Q} \in \mathcal{P}_1(\mathcal{H})$ is defined as
>
> $$\mathrm{W}_1(\mathrm{Q}, \mathrm{P}) = \inf_{\pi \in \Pi(\mathrm{Q},\mathrm{P})} \int_{\mathcal{H}^2} ||x-y|| \mathrm{d}\pi(x,y).$$
>
> where $\Pi(\mathrm{Q}, \mathrm{P})$ denote the set of probability measures on $\mathcal{H}^2$ whose marginals are Q and P. We define the 2-Wasserstein distance on $\mathcal{P}_2(\mathcal{H})$ as
>
> $$\mathrm{W}_2(\mathrm{Q}, \mathrm{P}) = \sqrt{\inf_{\pi \in \Pi(\mathrm{Q},\mathrm{P})} \int_{\mathcal{H}^2} ||x-y||^2 \mathrm{d}\pi(x,y)}.$$

Amit *et al.* (2022) provided a preliminary WPB bound, being explicit for the case of finite predictor classes and linear regression problems. To do so, they exploited the Kantorovich-Rubinstein duality (see, *e.g.*, Remark 6.5 in Villani, 2009) of the 1-Wasserstein distance. We exploit another duality formula (Theorem 5.10 in Villani, 2009) valid for any cost function (in the framework of optimal transport). This leads to a WPB bound valid for *uniformly Lipschitz* loss functions.

> **Definition 5.1.2.** We say that a function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ is *uniformly K-Lipschitz* if for any $\mathbf{z} \in \mathcal{Z}$, $\ell(., \mathbf{z})$ is K-Lipschtiz. We also say that a function is *uniformly L-smooth* (or simply smooth) if for any $\mathbf{z} \in \mathcal{Z}$, its gradient $\nabla \ell(., \mathbf{z})$ is L-Lipschitz.

A WPB bound for compact predictor classes. We first extend the PAC-Bayes framework to the case where the discrepancy between measures is expressed through the 1-Wasserstein distance. It is stated as follows: for uniformly K-lipschitz functions bounded in $[0,1]$ with $\mathcal{H} \subseteq \mathcal{B}_R := \bar{\mathcal{B}}(0_{\mathbb{R}^d}, R)$, we have for any prior $\mathrm{P} \in \mathcal{M}(\mathcal{H})$, with probability at least $1-\delta$, for any posterior distribution $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \mathcal{O}\left(\sqrt{2K(2K+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}(1+\mathrm{W}_1(\mathrm{Q},\mathrm{P})) + \frac{\log\left(\frac{m}{\delta}\right)}{m}} \right).$$

This bound extends the WPB bound of Amit *et al.* (2022) to the case of a compact space of predictors. The proof technique exploits covering number arguments to prove the Lipschitzness (with high probability) of a relevant functional. The duality theorem of Villani (2009, Theorem 5.10) allows us to generate a local change of measure inequality (see, *e.g.*, Donsker and Varadhan, 1976) required to use PAC-Bayes learning. This bound is stated in Theorem 5.2.2 and further discussed in Section 5.2. However, this result does not cover the celebrated case of PAC-Bayes with Gaussian priors and posteriors. We then develop the next result to address this important case.

WPB bounds with Gaussians measures for unbounded losses. Through the calculus of the residuals of Euler's Gamma function we obtain in Theorem 5.3.1, stated in Section 5.3, the following result when $\mathcal{H} = \mathbb{R}^d$, for loss functions lying in $[0,1]$ being uniformly K-lipschitz: for any gaussian prior P in a compact $C_{\alpha,\beta,M} \subseteq \mathrm{BW}(\mathbb{R}^d)$, with probability at least $1-\delta$, for any posterior distribution $\mathrm{Q} \in \mathcal{C}$,

$$\begin{aligned} &|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \\ &\quad \leq \mathcal{O}\left(\sqrt{2K(2K+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(1+\sqrt{\frac{d}{m}}+\mathrm{W}_1(\mathrm{Q},\mathrm{P})\right)+\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right), \end{aligned}$$

where $R = \mathcal{O}(\max\sqrt{d\log(d)}, \sqrt{\log(m)})$. This shows that, using R as an hyperparameter, we are able to maintain nearly the same convergence rate than Theorem 5.2.2 at the cost of an extra factor of $\sqrt{\log(dm)}$. Interestingly, we are able to remove in Corollary 5.3.1 the boundedness assumption to obtain a WPB bound, valid for unbounded uniformly K-lipschitz function with an additional boundedness assumption on $\sup_z \ell(0,\mathbf{z})$. This bound is more sensitive to the dimension of the problem when few data points are available. However, the asymptotic dependency remains (nearly) unchanged, at the cost of an extra polynomial factor in $\log(dm)$:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1+\mathrm{W}_1(\mathrm{Q},\mathrm{P})\right)+\left(1+K^2\log(m)\right)\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right). \tag{5.3}$$

$\tilde{\mathcal{O}}$ hides a polynomial dependency in $(\log(d), \log(m))$. This result is further discussed ion Section 5.3. The underlying proof technique is general enough to deal with (possibly unbounded) convex smooth loss functions. More details are gathered in Theorem 5.3.2 and Cor. 5.3.2.

A WPB bound with data-dependent prior. As we aim to intricate optimisation guarantees with generalisation bounds, we have to overcome the Bayesian paradigm of data-free priors which sets the prior distribution as a comparison point. Here, it is necessary to compare the candidate posterior with the optimisation goal. To do so, we elaborate in Section 5.4 on the idea of Dziugaite and Roy (2018b) who exploit differential privacy to obtain PAC-Bayesian bounds allowing to take data-dependent priors. We show that it is possible to maintain the asymptotic convergence rate of Corollary 5.3.1 when taking as 'prior' a Gibbs posterior. We introduce the following theorem holding again when $\mathcal{H} = \mathbb{R}^d$. For any gaussian prior P living in $C_{\alpha,\beta,M}$, with probability at least $1-\delta$, for any posterior distribution $\mathrm{Q} \in C_{\alpha,\beta,M}$, we have the following asymptotic convergence rate

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1+\mathrm{W}_1(\mathrm{Q},\mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}})\right) + (1+K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

We also study non-asymptotic regimes in Theorem 5.4.1. While Dziugaite and Roy (2018b) exploited differential privacy results for the Gibbs posterior when the loss is bounded, we successfully extended these results to (possibly unbounded) uniformly Lipschitz losses. This is not specific to the WPB framework and may be of independent interest.

PAC-Bayes provides generalisation guarantees for the Bures-Wasserstein SGD. While working on WPB theory, we notice a shift from classical assumptions due to the KL divergence. Indeed, statistical assumptions (such as subgaussiannity, bounded variances) are transformed into geometric assumptions such as Lipschitzness and convex smoothness when Wasserstein distances are involved. We exploit in Section 5.5 WPB theory to provide generalisation guarantees for the Bures-Wasserstein SGD (recalled in Algorithm 2) which approximates the best Gaussian surrogate $\hat{\mathrm{Q}}$ of $\mathrm{Q}^* := \mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}}$ (in the sense of the KL divergence, see Section 5.5 for more details). More precisely, we show that the KL divergence and Wasserstein distances are linked within the WPB framework: the (KL-based) PAC-Bayesian learning objective of Catoni (2007), which outputs the Gibbs posterior Q^*, can be approximated by $\hat{\mathrm{Q}}_N$, the output of the Bures-Wasserstein SGD after N optimisation steps, which is provably close from $\hat{\mathrm{Q}}$ with respect to the 2-Wasserstein distance (see Theorem 5.5.1). Within the WPB framework, this link is translated in Theorem 5.5.2 as a generalisation bound ensuring that asymptotically, the minima reached by the Bures-Wasserstein SGD has a strong generalisation ability.

Concretely, for N large enough, for uniformly K-lipschitz, convex, smooth loss functions we have the following asymptotic guarantee with probability $1-\delta$:

$$|\Delta_{\mathcal{S}_m}(\hat{\mathrm{Q}}_N)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1+\mathrm{W}_1(\hat{\mathrm{Q}},\mathrm{Q}^*)\right) + (1+K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

Thus, the WPB framework is enough to provide an explicit convergence rate for the generalisation gap avoiding the comparison to an arbitrary prior. Instead, this bound shows that a (long enough) run of the Bures-Wasserstein SGD with enough data (or a Lipschitz constant small enough) leads to a minimiser with a high generalisation ability. Furthermore, Theorem 5.5.2 is a reformulation of (5.12) which is, to our knowledge, the first PAC-Bayesian bound of the form (5.2) with $D = \sqrt{d\mathrm{W}_2}$ and

$$\varepsilon = \mathcal{O}(\sqrt{d\mathrm{W}_1(\hat{\mathrm{Q}}, \mathrm{Q}^*)}).$$

This provides elements of answer to the question listed in Section 5.1.2 and concludes this work.

Discussion about the assumptions For the sake of clarity, we provide in Figure 5.1.3 the topography of our main results. We focus on the assumptions required to state each of the results and doing so, we aim to give to the reader a broader vision of when can these bounds be applied. We stress that the Lipschitzness assumption is at the core of all results, except Theorem 5.3.2 and Cor. 5.3.2. Convexity is required to use differential privacy and to obtain Theorem 5.4.1. Finally, we note that while the results of LAMBERT *et al.* (2022) are usable with only smoothness and convexity, we must add the uniform Lipschitz assumption to obtain Theorem 5.5.2. The question of whether all these assumptions are minimal to perform WPB remains open.

5.2 PAC-Bayesian bounds for compact predictor spaces

Here we establish WPB bounds for bounded losses when the predictor space is a compact of $\mathbb{R}^d$. To intricate the 1-Wasserstein distance within the PAC-Bayes proof, we design a surrogate of the change of measure inequality (DONSKER and VARADHAN, 1976) by exploiting the uniform Lipschitz assumption on the loss. To do so we need to exploit the notion of *covering number* recalled below as well as Kantorovich duality (VILLANI, 2009, Theorem 5.10). This notion of duality holds for any cost function (in an optimal transport framework) contrary to the Kantorovich-Rubinstein duality exploited by AMIT *et al.* (2022) which only holds when the cost function is a distance. This result is recalled in Appendix D.1.1.

Definition 5.2.1 (Covering number). Let $\mathcal{H} \subseteq \mathbb{R}^d$. An ε-covering of $\mathcal{H}$ is a subset C of $\mathcal{H}$ such that $\mathcal{H} \subseteq \cup_{x\in C}\bar{\mathcal{B}}(x,\varepsilon)$. The ε-covering number of $\mathcal{H}$ is defined as

$$N(\mathcal{H},\varepsilon) := \min\{n \geq 1 \mid \exists \text{ an } \varepsilon\text{-covering of } \mathcal{H} \text{ of size } n\}.$$

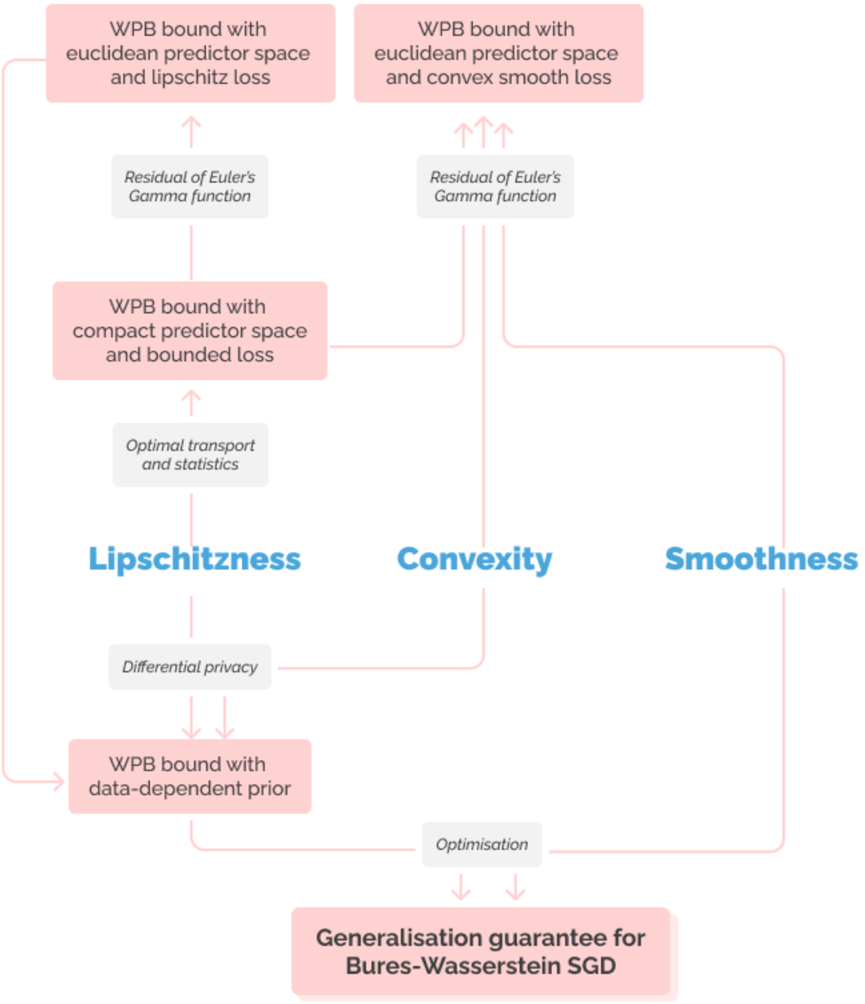
WPB bound with euclidean predictor space and lipschitz loss
WPB bound with euclidean predictor space and convex smooth loss
Residual of Euler's Gamma function
Residual of Euler's Gamma function
WPB bound with compact predictor space and bounded loss
Optimal transport and statistics
Lipschitzness
Convexity
Smoothness
Differential privacy
WPB bound with data-dependent prior
Optimisation
Generalisation guarantee for Bures-Wasserstein SGD

Figure 5.1. *An overwiew of the assumptions required to obtain the main results. Assumptions are stated in blue, main results are in pink boxes and the proof technique exploited to obtain such results are within grey boxes.*

We also define the $\varepsilon, 1$-Wasserstein to be $\mathrm{W}_{\varepsilon}(\mathrm{Q}, \mathrm{P}) = \varepsilon + \mathrm{W}_1(\mathrm{Q}, \mathrm{P})$. This cost function is essential to the analysis. We now state the main results of this section. Additional background is gathered in Appendix D.1.1.

5.2.1 A Catoni-type bound

We propose here a WPB bound analogous to a relaxation of Catoni (2007, Theorem 1.2.6) stated for instance in Alquier *et al.* (2016, Theorem 4.1).

Theorem 5.2.1. For any $\varepsilon, \delta > 0$, assume that $\ell \in [0,1]$ is uniformly K-Lipschitz and that $\mathcal{H}$ is a compact of $\mathbb{R}^d$ bounded by $R > 0$. Let $\mathrm{P} \in \mathcal{P}_1(\mathcal{H})$ be a (data-free) prior distribution and assume we choose a parameter λ such that

$$0 < \lambda \leq \frac{1}{K}\sqrt{\frac{2m}{2d\log(1+\frac{2R}{\varepsilon}) + \log(\frac{2}{\delta})}} := \lambda_{max}.$$

Then, with probability $1-\delta$, for any posterior distribution $\mathrm{Q} \in \mathcal{P}_1(K)$,

$$\Delta_{\mathcal{S}_m}(\mathrm{Q}) \leq 4K\varepsilon + \frac{\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + 2\varepsilon + \log(2/\delta)}{\lambda} + \frac{\lambda}{2m}.$$

Note that we assumed the loss to be bounded, although this can be relaxed to subgaussiannity at no cost. In Theorem 5.2.1, the range of λ is restricted and the loss required to be uniformly Lipschitz. Such restrictions do not exist in Alquier *et al.* (2016, Theorem 4.1) which recovers a similar result with a KL divergence coming from the change of measure inequality (Donsker and Varadhan, 1976). In WPB this is required to have a control on Δ_S which is exploited in Kantorovich duality (Theorem D.1.1). Furthermore, assuming Lipschitzness on a compact space is not restrictive as it covers, *e.g.*, all $\mathcal{C}^1$ functions. Note that the smaller the Lipschitz constant K is, the larger λ_{max}. This is not surprising as, from an optimisation point of view, λ acts as a learning rate which determines the influence of data with respect to the regulariser $\mathrm{W}_1(\mathrm{Q},\mathrm{P})$. A small K says that huge variations between data have a small influence on the loss value, then we can give more influence to the training set without deteriorating much the generalisation ability of the posterior. This bound also says that it is legitimate to consider a WPB learning objective analogous to the one derived from Alquier *et al.* (2016, Theorem 4.1) (which yields Gibbs posteriors):

$$\mathrm{argmin}_{Q \in \mathcal{P}_1(\mathcal{H})} \frac{\mathrm{W}_1(\mathrm{Q},\mathrm{P})}{\lambda} + \frac{\lambda}{2m}.$$

Theorem 5.2.1's proof is stated below and mixes up several arguments from optimal transport with PAC-Bayes learning through covering numbers.

Proof of Theorem 5.2.1 *Step 1: define a good data-dependent function.* We define, for any sample $\mathcal{S}_m$ and predictor $h \in \mathcal{H}$,

$$f_{\mathcal{S}_m}(h) = \lambda\Delta_{\mathcal{S}_m}(h).$$

This function satisfies the following lemma:

Lemma 5.2.1. Let $\varepsilon > 0$ assume that $0 < \lambda \leq \frac{1}{K}\sqrt{\frac{2m}{\log\left(\frac{N(\mathcal{H},\varepsilon)^2}{\delta}\right)}}$. We have, with probability $1-\delta$ for all $h, h' \in \mathcal{H}$, for any P:

$$f_{\mathcal{S}_m}(h) - f_{\mathcal{S}_m}(h') \leq 2(1+2\lambda K)\varepsilon + ||h-h'||.$$

Proof of Lemma 5.2.1. We rename here $N := N(\mathcal{H},\varepsilon)$. There exists an ε-covering $C := \{h_1, ..., h_N\}$ of $\mathcal{H}$ of size N. Then for any $h, h' \in C^2$, we have:

$$f_{\mathcal{S}_m}(h) - f_{\mathcal{S}_m}(h') = \frac{\lambda}{m}\sum_{i=1}^{m}\mathbb{E}[\ell(h,\mathbf{z}) - \ell(h'\mathbf{z})] - (\ell(h,\mathbf{z}_i) - \ell(h',\mathbf{z}_i)).$$

We know that for any h, h', z, $|\ell(h,\mathbf{z}) - \ell(h',\mathbf{z})| \leq \lambda K ||h-h'||$. Then, applying Hoeffding's inequality for all pairs $h, h' \in C^2$ and performing an union bound gives that with probability at least $1-\delta$, for all pairs $(h,h') \in C^2$:

$$f_{\mathcal{S}_m}(h) - f_{\mathcal{S}_m}(h') \leq \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\lambda K||h-h'||.$$

So for any $h, h' \in \mathcal{H}^2$ there exists $h_0, h_0' \in C^2$ such that $||h-h_0|| \leq \varepsilon$ and $||h-h_0|| \leq \varepsilon$. Thus, we have

$$\begin{aligned} f_{\mathcal{S}_m}(h) - f_{\mathcal{S}_m}(h') &= f_{\mathcal{S}_m}(h) - f_{\mathcal{S}_m}(h_0) + f_{\mathcal{S}_m}(h_0) - f(h_0') + f_{\mathcal{S}_m}(h_0') - f_{\mathcal{S}_m}(h') \\ &\leq 2\lambda K\left(||h-h_0|| + ||h'-h_0'||\right) + \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\lambda K||h_0-h_0'|| \\ &\leq 4\lambda K\varepsilon + \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\lambda K||h_0-h_0'||. \end{aligned}$$

By the triangle inequality, $||h_0-h_0'|| \leq ||h-h'|| + 2\varepsilon$ so we finally have with probability at least $1-\delta$, for any $h, h' \in K^2$:

$$f_{\mathcal{S}_m}(h) - f_{\mathcal{S}_m}(h') \leq 4\lambda K\varepsilon + \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\lambda K\left(2\varepsilon + ||h-h'||\right).$$

Using $\lambda \leq \frac{1}{K}\sqrt{\frac{2m}{\log\left(\frac{N^2}{\delta}\right)}}$ and upper bounding concludes the proof. ■

Step 2: A probabilistic change of measure inequality for $f_{\mathcal{S}_m}$. We do not have for the Wasserstein distance such a powerful tool than the change of measure inequality. However, we can generate a probabilistic surrogate on $\mathcal{P}_1(\mathcal{H})$ valid for the function $f_{\mathcal{S}_m}$.

Lemma 5.2.2. For any $\epsilon > 0$, any $\delta > 0$, any

$$0 < \lambda \leq \frac{1}{K}\sqrt{\frac{2m}{\log\left(\frac{N(\mathcal{H},\varepsilon)^2}{\delta}\right)}},$$

we have with probability $1-\delta$ over the sample $\mathcal{S}_m$, for any $\mathrm{P} \in \mathcal{P}_1(K)$

$$\left(\sup_{Q\in\mathcal{P}_1(K)} \mathbb{E}_{h\sim\mathrm{Q}}[f_{\mathcal{S}_m}(h)] - 2(1+\lambda K)\varepsilon - \mathrm{W}_1(\mathrm{Q},\mathrm{P})\right) \leq \mathbb{E}_{h\sim\mathrm{P}}[f_{\mathcal{S}_m}(h)].$$

Proof of Lemma 5.2.2. Firstly, we introduce the cost function $c_\varepsilon(x,y) = \varepsilon + ||x-y||$. From this we notice that we can rewrite the $\varepsilon, 1$- Wasserstein distance:

$$\mathrm{W}_\varepsilon(\mathrm{Q},\mathrm{P}) = \inf_{\pi\in\Pi(\mathrm{Q},\mathrm{P})}\int_{\mathcal{H}^2} c_\varepsilon(x,y)\mathrm{d}\pi(x,y).$$

Remark that because W_1 is a distance, then W_ε is symmetric. Furthermore, if we fix $\mathcal{X} = \mathcal{Y} = \mathcal{H}$ and we notice that $c_\varepsilon \geq 0$, then the condition for Kantorovich duality is satisfied. Thus, we apply Theorem D.1.1 as follows: for all $\mathrm{Q},\mathrm{P} \in \mathcal{P}_1(\mathcal{H})$:

$$\begin{aligned} W_\varepsilon(\mathrm{Q},\mathrm{P}) = W_\varepsilon(\mathrm{P},\mathrm{Q}) &= \min_{\pi\in\Pi(\mathrm{P},\mathrm{Q})}\int_{K^2} c_\varepsilon(h_1,h_2)d\pi(h_1,h_2) \\ &= \sup_{\substack{(\psi,\phi)\in L^1(\mathrm{Q})\times L^1(\mathrm{P}) \\ \psi-\phi\leq c_\varepsilon}} \left[\int_K \psi(h)\mathrm{dQ}(h) - \int_K \phi(h)\mathrm{dP}(h)\right] \\ &= \sup_{\substack{(\psi,\phi)\in L^1(\mathrm{Q})\times L^1(\mathrm{P}) \\ \psi-\phi\leq c_\varepsilon}} \left[\mathbb{E}_{h\sim\mathrm{Q}}[\psi(h)] - \mathbb{E}_{h\sim\mathrm{P}}[\phi(h)]\right]. \end{aligned}$$

A crucial point is that for a well-chosen λ with high probability, the pair $(f_{\mathcal{S}_m}, f_{\mathcal{S}_m})$ satisfies the condition stated under the last supremum. It is formalised in the following lemma.

Lemma 5.2.3. For any $\varepsilon > 0$ any $\delta > 0$, any $0 < \lambda \leq \frac{1}{K}\sqrt{\frac{2m}{\log\left(\frac{N(\mathcal{H},\varepsilon)^2}{\delta}\right)}}$, we have with probability at least $1-\delta$ over the sample $\mathcal{S}_m$ that, for all measures $\mathrm{Q}, \mathrm{P} \in \mathcal{P}_1(\mathcal{H})^2$:

- $f_{\mathcal{S}_m} \in L_1(\mathrm{Q}), L_1(\mathrm{P})$,
- for all $h, h' \in \mathcal{H}^2$, $f_{\mathcal{S}_m}(h) - f_{\mathcal{S}_m}(h') \leq c_{\varepsilon'}(h,h')$ with $\varepsilon' := 2(1+2\lambda K)\varepsilon$.

Thus, Kantorovich duality (Theorem D.1.1) gives:

$$\left(\sup_{Q\in\mathcal{P}_1(\mathcal{H})} \mathbb{E}_{h\sim\mathrm{Q}}[f_{\mathcal{S}_m}(h)] - \mathrm{W}_{\varepsilon'}(\mathrm{Q},\mathrm{P})\right) \leq \mathbb{E}_{h\sim\mathrm{P}}[f_{\mathcal{S}_m}(h)],$$

and using $\mathrm{W}_{\varepsilon'} = \varepsilon' + \mathrm{W}_1$ and the definition of ε' concludes the proof.

Proof Proof of Lemma 5.2.3. Because the space of predictors $\mathcal{H}$ is compact and that for any $\mathbf{z} \in \mathcal{Z}$, the loss function $\ell(., \mathbf{z})$ is K-Lipschitz on $\mathcal{H}$, then both the generalisation and empirical risk are continuous on $\mathcal{H}$. Thus $|f_{\mathcal{S}_m}|$ is also continuous and, by compacity, reaches its maximum M_S on $\mathcal{H}$. Thus for any probability P on $\mathcal{H}$, $\mathbb{E}_{h\sim\mathrm{P}}[|f_{\mathcal{S}_m}(h)|] \leq M_S < +\infty$ almost surely. This proves the first statement. We notice that the second statement, given the choice of λ, is the exact conclusion of Lemma 5.2.1 with probability at least $1-\delta$. So with probability at least $1-\delta$, Kantorovich duality gives us that for any P, Q with $\varepsilon' = 2(1+\lambda K)\varepsilon$,

$$\mathbb{E}_{h\sim\mathrm{Q}}[f_{\mathcal{S}_m}(h)] - \mathbb{E}_{h\sim\mathrm{P}}[f_{\mathcal{S}_m}(h)] \leq \mathrm{W}_{\varepsilon'}(\mathrm{Q},\mathrm{P}).$$

Re-organising the terms and taking the supremum over Q concludes the proof. ■

This concludes the proof of Lemma 5.2.2. ■

Step 3: The PAC-Bayes route of proof for the 1-Wasserstein distance.
We start by exploiting Lemma 5.2.2: for any prior $\mathrm{P} \in \mathcal{P}_1(K)$, for

$$0 < \lambda \leq \frac{1}{K}\sqrt{\frac{2m}{\log\left(\frac{2N(K,\varepsilon)^2}{\delta}\right)}},$$

with probability at least $1-\delta/2$ we have

$$\left(\sup_{Q\in\mathcal{P}_1(K)} \mathbb{E}_{h\sim\mathrm{Q}}[f_{\mathcal{S}_m}(h)] - (2(1+2\lambda K)\varepsilon - \mathrm{W}_1(\mathrm{Q},\mathrm{P})\right) \leq \mathbb{E}_{h\sim\mathrm{P}}[f_{\mathcal{S}_m}(h)].$$

We then notice that by Jensen's inequality,

$$\mathbb{E}_{h\sim \mathrm{P}}[f_{\mathcal{S}_m}(h)] \leq \log\left(\mathbb{E}_{h\sim \mathrm{P}}[\exp(f_{\mathcal{S}_m}(h))]\right).$$

Then, by Markov's inequality we have with probability $1-\delta/2$

$$\mathbb{E}_{h\sim \mathrm{P}}[f_{\mathcal{S}_m}(h)] \leq \log\left(\frac{2}{\delta}\right) + \log\left(\mathbb{E}_S\mathbb{E}_{h\sim \mathrm{P}}\left[\exp(f_{\mathcal{S}_m}(h))\right]\right).$$

By Fubini and Hoeffding lemma applied m times on the iid sample $\mathcal{S}_m$, we have

$$\mathbb{E}_S\mathbb{E}_{h\sim \mathrm{P}}\left[\exp(f_{\mathcal{S}_m}(h))\right] = \mathbb{E}_{h\sim \mathrm{P}}\mathbb{E}_S\left[\exp(f_{\mathcal{S}_m}(h))\right] \leq \frac{\lambda^2}{2m}.$$

Taking an union bound gives us with probability $1-\delta$, for any posterior Q:

$$\mathbb{E}_{h\sim \mathrm{Q}}[\mathsf{R}_{\mathcal{D}}(h)] \leq \mathbb{E}_{h\sim \mathrm{Q}}[\mathsf{R}_{\mathcal{S}_m}(h)] + 4K\varepsilon + \frac{\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + 2\varepsilon + \log(2/\delta)}{\lambda} + \frac{\lambda}{2m}.$$

Finally, we know that $\mathcal{H}$ is bounded by R so by Proposition D.1.1 we have

$$N^2 = N(\bar{\mathcal{B}}(0,R),\varepsilon)^2 \leq (1+2mR)^{2d}.$$

Thus, we can take λ equal to

$$\frac{1}{K}\sqrt{\frac{2m}{2d\log(1+\frac{2R}{\varepsilon}) + \log(\frac{2}{\delta})}}.$$

This concludes the proof.

5.2.2 A McAllester-type bound

We now move on to a McAllester-type bound, which can be tighter than Theorem 5.2.1 for large values of the 1-Wasserstein.

Theorem 5.2.2. For any $\delta > 0$, assume that $\ell \in [0,1]$ is uniformly K-Lipschitz and that $\mathcal{H}$ is a compact of $\mathbb{R}^d$. Let $\mathrm{P} \in \mathcal{P}_1(\mathcal{H})$ a (data-free) prior distribution. Then, with probability $1-\delta$, for any posterior distribution $\mathrm{Q} \in \mathcal{P}_1(\mathcal{H})$:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \sqrt{2K(2K+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \varepsilon_m\right) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}},$$

with $\varepsilon_m = \frac{4}{\log(\frac{3}{\delta})}\left(2 + \sqrt{\frac{\log\left(\frac{3}{\delta}\right)+2d\log(1+2Rm)}{2m}}\right) = \mathcal{O}\left(1 + \sqrt{\frac{d\log(Rm)}{m}}\right)$.

We deteriorate the bound of Amit *et al.*, 2022 by transforming a convergence rate of

$$\sqrt{\frac{\mathrm{W}_1(\mathrm{Q},\mathrm{P})}{m}}$$

for finite predictor classes onto a $\sqrt{(Kd\mathrm{W}_1(\mathrm{Q},\mathrm{P})+1)\,\frac{\log(m)}{m}}$ for compact classes. This deteriorated rate is the price to pay to consider a general WPB bound for an uncountable number of predictors. However, notice that the dimension dependency can be attenuated through the Lipschitz constant, with the limit rate of

$$\mathcal{O}\left(\sqrt{\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right)$$

which is dimension-free and is a consequence of the statistical component of PAC-Bayes learning. Furthermore, note that this proof technique allows us to recover the rate of Amit *et al.* (2022) rate when considering finite classes. The proof of Theorem 5.2.2 involves similar arguments to the one of Theorem 5.2.1, therefore we defer it to Appendix D.2.1.

5.3 PAC-Bayesian bounds for Gaussian distributions

In this section we develop McAllester-type WPB bounds on an Euclidean predictor space. Indeed, in PAC-Bayes learning, considering this predictor space is common as PAC-Bayesian objective often focuses on Gaussian priors and posteriors (see, *e.g.*, Dziugaite and Roy, 2017; Amit and Meir, 2018). Those bounds build up on Theorem 5.2.2 and the overall conclusion is the following: when considering functions with interesting geometric properties (*i.e.*, Lipschitzness or smoothness) on $\mathbb{R}^d$, WPB bounds hold for Gaussian priors and posteriors over $\mathcal{H}=\mathbb{R}^d$ at the cost of negligible extra terms (Theorems 5.3.1 and 5.3.2). More importantly, we show that in this setup, the assumption of a bounded loss is not required anymore to perform WPB: only boundedness on a compact is needed. Thus, we propose WPB bounds for unbounded losses (Corollaries 5.3.1 and 5.3.2).

Two sets of assumption. Previously, we assumed two assumptions on the losses: uniform Lipschitzness (Definition 5.1.2) and boundedness (in $[0,1]$) on a compact of $\mathbb{R}^d$. We provide below to novel sets of hypotheses which encapsulates previous assumptions while allowing the loss to be unbounded on all $\mathbb{R}^d$.

- **(A1)** ℓ is uniformly K-Lipschitz over $\mathcal{H}$, and $\sup_{z\in\mathcal{Z}}||\ell(0,\mathbf{z})||=D<+\infty$.

- **(A2)** For any $\mathbf{z} \in \mathcal{Z}$, $\ell(.,\mathbf{z})$ is continuously differentiable over $\mathcal{H}$, $\ell(.,\mathbf{z})$ is also a convex L- smooth (*i.e*, its gradient is L-Lipschitz) and $\sup_{z\in\mathcal{Z}} ||\nabla_h \ell(0,\mathbf{z})|| = D < +\infty$.

Example 5.3.1. Recall that $\mathcal{H} = \mathbb{R}^d$ and let $\phi : \mathcal{H} \to \mathbb{R}^d$. Also, let $\psi : \mathcal{Z} \to \mathbb{R}^d$ such that $\psi(\mathcal{Z})$ is bounded by $C_\phi > 0$. We assume that both ϕ, ψ are continuously differentiable and that $\nabla\phi$ is G-Lipschitz. Note that the $||\phi||$ is possibly unbounded on $\mathcal{H}$. Then **(A2)** holds for the loss function $\ell(h,\mathbf{z}) = ||\phi(h) - \psi(z)||^2$ Indeed, $\nabla_h \ell(h,\mathbf{z}) = 2(\nabla\phi(h) - \psi(z))$ so on any compact $\mathcal{K}$ bounded by R, $\nabla_h \ell$ is uniformly 2-Lipschitz. Also $\sup_{z\in\mathcal{Z}} ||\nabla_h \ell(0,\mathbf{z})|| \leq 2C$. Note that on $\mathbb{R}^d$, $\ell(.,\mathbf{z})$ is not necessarily Lipschitz for any $\mathbf{z}$ (take the case $\phi = Id_{\mathbb{R}^d}$) so **(A1)** is not satisfied.

A brief summary of the proof technique. To extend Theorem 5.2.2 to the case $\mathcal{H} = \mathbb{R}^d$, we use the push-forward distribution $\mathcal{P}_R\#\mathrm{P}$ where $\mathrm{P} \in C_{\alpha,\beta,M}$ for fixed α, β, M (notation defined in Section 5.1.1). The interest of this is to use Theorem 5.2.2 by considering projections of the Gaussian prior and posterior. When considering Gaussian distributions, the gap between projected distributions and original ones is explicitly controlled. More precisely, for any $R > 0$ large enough, for any $\mathrm{P} \in C_{\alpha,\beta,M}$, $\mathrm{W}_1(\mathrm{P}, \mathcal{P}_R\#\mathrm{P})$ is upper bounded. This is the conclusion of an important technical lemma (Lemma D.1.2), stated with additional background in Appendix D.1.2. We state below new WPB results with Gaussian distributions for Lipschitz functions in Section 5.3.1 and for smooth functions in Section 5.3.2.

5.3.1 PAC-Bayesian bounds for Lipschitz losses

This section focuses on the case of Lipschitz losses. We show that when the loss is uniformly Lipschitz, it is possible to maintain the tightness of Theorem 5.2.2 on all $\mathbb{R}^d$ when the loss remains bounded. We also show that it is also possible to obtain a WPB bound when the loss function satisfies **(A1)** (*i.e.* with an additional boundedness assumption on $\sup_z \ell(0,\mathbf{z})$), while remaining unbounded (Corollary 5.3.1).

Theorem 5.3.1. Assume that $d \geq 3$, $\mathcal{H} = \mathbb{R}^d$ and that the loss is uniformly K-Lipschitz and lies in $[0,1]$ over $\mathcal{H}$. For any $\delta > 0, 0 \leq \alpha \leq \beta, M \geq 0$, let $\mathrm{P} \in C_{\alpha,\beta,M}$ a (data-free) prior distribution. Then, with probability $1-\delta$, for any posterior distribution $\mathrm{Q} \in C_{\alpha,\beta,M}$:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \\ \leq 2\frac{\beta\sqrt{\beta}}{m} + \sqrt{2K(2K+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}(\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \alpha_m) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}},$$

with $R = \mathcal{O}(\max \sqrt{d\log(d)}, \sqrt{\log(m)})$ and $\alpha_m = 2(M+1)\frac{\beta\sqrt{\beta}}{m} + \varepsilon_m = \mathcal{O}\left(1 + \sqrt{\frac{d\log(Rm)}{m}}\right)$ with ε_m defined in Theorem 5.2.2.

Theorem 5.3.1 shows that, at the cost of additional residual terms, it is possible to maintain the convergence rate of Theorem 5.2.2 when considering Gaussian prior and posterior within the compact $C_{\alpha,\beta,M}$. The influence of α, β, γ appear in the explicit value of R described as it is always taken in this work as the smallest value satisfying the assumption Rad described in Appendix D.1.2. As in Theorem 5.2.2, the idea that a small Lipschitz constant tightens the bound is still conveyed here and is of great importance for Corollary 5.3.1 which provides a WPB bound for unbounded losses with higher dimension dependency when few data is available.

Proof of Theorem 5.3.1. We take a specific radius R which is the smallest value satisfying Rad. The proof starts with a straightforward application of Theorem 5.2.2 on the compact $\mathcal{B}(0,R)$, with the prior $\mathcal{P}_R\#\mathrm{P}$, and with high probability, for any posterior $\mathcal{P}_R\#\mathrm{Q}$ with $\mathrm{Q} \in C_{\alpha,\beta,M}$:

$$
\begin{aligned}
&|\Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| \\
&\leq \sqrt{2K(2K+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(\mathrm{W}_1(\mathcal{P}_R\#\mathrm{P},\mathcal{P}_R\#\mathrm{P}) + \varepsilon_m\right) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}}.
\end{aligned}
$$

From this we control the left hand-side term as follows:

$$
|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq |\Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| + |\Delta_{\mathcal{S}_m}(\mathrm{Q}) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})|.
$$

And we also have

$$
\begin{aligned}
|\Delta_{\mathcal{S}_m}(\mathrm{Q}) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| &\leq \mathbb{E}_{h\sim\mathrm{Q}}\left[|\Delta_{\mathcal{S}_m}(h) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R(h))|\right] \\
&= \mathbb{E}_{h\sim\mathrm{Q}}\left[|\Delta_{\mathcal{S}_m}(h) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R(h))|\mathbb{1}(||h|| > R)\right] \\
&\leq 2\mathrm{Q}(||h|| > R) \leq 2\frac{\beta\sqrt{2\beta}}{m},
\end{aligned}
$$

the last line holding thanks to Lemma D.1.2 and because $\Delta_S \in [-1,1]$. Also we have by the triangle inequality:

$$
\mathrm{W}_1(\mathcal{P}_R\#\mathrm{P},\mathcal{P}_R\#\mathrm{P}) \leq \mathrm{W}_1(\mathrm{Q},(\mathcal{P}_R\#\mathrm{Q})) + \mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \mathrm{W}_1(\mathrm{P},\mathcal{P}_R\#\mathrm{P}).
$$

Because both $\mathrm{Q}, \mathrm{P} \in C_{\alpha,\beta,M}$, using again Lemma D.1.2 gives:

$$\mathrm{W}_1(\mathcal{P}_R\#\mathrm{P}, \mathcal{P}_R\#\mathrm{P}) \leq \mathrm{W}_1(\mathrm{Q},\mathrm{P}) + 2(M+1)\frac{\beta\sqrt{2\beta}}{m}.$$

We then have:

$$\begin{aligned}|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq\ & 2\frac{\beta\sqrt{2\beta}}{m} \\ & + \sqrt{2K(2K+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}(\mathrm{W}_1(\mathrm{Q},\mathrm{P})+\alpha_m) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}},\end{aligned}$$

with $\alpha_m = 2(M+1)\frac{\beta\sqrt{\beta}}{m} + \varepsilon_m = \mathcal{O}(1)$. This concludes the proof. ■

A corollary for unbounded losses. We provably extend Theorem 5.3.1 to the case of unbounded Lipschitz losses.

Corollary 5.3.1. Assume that $d \geq 3$, $\mathcal{H} = \mathbb{R}^d$ and that the (unbounded) loss satisfies **(A1)**. For any $\delta > 0, 0 \leq \alpha \leq \beta, M \geq 0$, let $\mathrm{P} \in C_{\alpha,\beta,M}$ a (data-free) prior distribution. Then, with probability $1-\delta$, for any posterior distribution $\mathrm{Q} \in C_{\alpha,\beta,M}$, the three following bounds holds.

Low-data regime $(d \geq m)$

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d^{\frac{3}{2}}}{m}\left(\sqrt{\frac{d}{m}} + \mathrm{W}_1(\mathrm{Q},\mathrm{P})\right) + (1+K^2 d)\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

Transitory regime $(m > d,\ d\log(d) \geq \log(m))$

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d^{\frac{3}{2}}}{m}(1+\mathrm{W}_1(\mathrm{Q},\mathrm{P})) + (1+K^2 d)\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

Asymptotic regime $(d\log(d) < \log(m))$

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}(1+\mathrm{W}_1(\mathrm{Q},\mathrm{P})) + (1+K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

In all these formulas, $\tilde{\mathcal{O}}$ hides a polynomial dependency in $(\log(d), \log(m))$. For an explicit formulation of the bounds, we refer to (5.5).

The message here is that in Wasserstein PAC-Bayes, the bounded loss assumption is not as important as in classical PAC-Bayes using KL divergence. Indeed, the geometric constraints of WPB forced us to consider compact classes of Gaussian distribution and Lipschitz losses. Having such geometric assumptions on the distribution space and the loss is enough to exploit the properties of the 1-Wasserstein distance and to circumvent the boundedness assumption. To avoid boundedness, we transformed the limit rate

$$\mathcal{O}\left(\sqrt{\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right)$$

of Theorem 5.2.2 into

$$\mathcal{O}\left(\sqrt{(1+K^2 d)\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right)$$

for non-asymptotic regimes and

$$\mathcal{O}\left(\sqrt{(1+K^2 \log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right)$$

for the asymptotic one. Thus, even when few data is available, a well constrained (unbounded) Lipschitz loss is able to control the impact of the dimension. Note that, in the small data regime, we have the highest dimension dependency. Note also that the dimensionality of the learning problem is controlled by the Lipschitz constant with the limit rate of

$$\mathcal{O}\left(\sqrt{\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right)$$

which is dimension-free and is a consequence of the statistical component of PAC-Bayes learning. To the best of our knowledge, our work is the first to exploit geometric properties of the loss to propose PAC-Bayes bounds for unbounded and heavy-tailed losses with explicit convergence rates. Indeed, the existing literature on unbounded losses exploits either general divergence properties (ALQUIER and GUEDJ, 2018; PICARD-WEIBEL and GUEDJ, 2022), functional properties for heavy-tailed distribution (HOLLAND, 2019), uniform boundedness assumption on the loss over the data space (HADDOUCHE *et al.*, 2021) or concentration inequalities as in Chapter 2 or in KUZBORSKIJ and SZEPESVÁRI (2019), RIVASPLATA *et al.* (2020), and JANG *et al.* (2023).

Proof of Corollary 5.3.1. First, we start from Theorem 5.2.2 which gives, with probability at least $1-\delta$:

$$
\begin{aligned}
&|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \\
&\quad\leq \sqrt{2K(2K+1)\frac{\log(\frac{3}{\delta}) + 2d\log\left(1+2Rm\right)}{m}\left(\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \varepsilon_m\right) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}}.
\end{aligned} \tag{5.4}
$$

This last bound holds for any uniformly Lipschitz function taking value on $[0,1]$ on a compact predictor space bounded by a certain R. Let $\mathrm{P} \in C_{\alpha,\beta,M}$. We now assume **(A1)** and consider R to be the smallest value satisfying $\mathtt{Rad}$. Let $\ell' = \ell/(D+2KR)$. We note $D_R = D + 2KR$, then on the ball $\mathcal{B}(0,R)$, ℓ' takes value in $[0,1]$ (because the compact is bounded by R and the loss is K-Lipschitz) and is K/D_R-Lipschitz. Applying Equation (5.4) with ℓ' on $\mathcal{B}(0,R)$ and multiplying by D_R gives, with high probability, for any $\mathrm{Q} \in C_{\alpha,\beta,M}$:

$$
\begin{aligned}
&|\Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| \\
&\leq D_R\sqrt{2\frac{K}{D_R}(2\frac{K}{D_R}+1)\frac{\log(\frac{1}{\delta}) + 2d\log\left(1+2Rm\right)}{m}\left(\mathrm{W}_1(\mathcal{P}_R\#\mathrm{P},\mathcal{P}_R\#\mathrm{P}) + \varepsilon_m\right) + \frac{\log\left(\frac{m}{\delta}\right)}{m}} \\
&= \sqrt{2K(2K+D_R)\frac{\log(\frac{1}{\delta}) + 2d\log\left(1+2Rm\right)}{m}\left(\mathrm{W}_1(\mathcal{P}_R\#\mathrm{P},\mathcal{P}_R\#\mathrm{P}) + \varepsilon_m\right) + D_R^2\frac{\log\left(\frac{m}{\delta}\right)}{m}},
\end{aligned}
$$

where $\varepsilon_m = \mathcal{O}\left(1\right)$ defined in Theorem 5.2.2. As in Theorem 5.3.1, we have:

$$
\mathrm{W}_1(\mathcal{P}_R\#\mathrm{P},\mathcal{P}_R\#\mathrm{P}) \leq \mathrm{W}_1(\mathrm{Q},\mathrm{P}) + 2(M+1)\frac{\beta\sqrt{2\beta}}{m}.
$$

We have

$$
|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq |\Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| + |\Delta_{\mathcal{S}_m}(\mathrm{Q}) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})|,
$$

And we have

$$
\begin{aligned}
|\Delta_{\mathcal{S}_m}(\mathrm{Q}) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| &\leq \mathbb{E}_{h\sim\mathrm{Q}}\left[|\Delta_{\mathcal{S}_m}(h) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R(h))|\right] \\
&= \mathbb{E}_{h\sim\mathrm{Q}}\left[|\Delta_{\mathcal{S}_m}(h) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R(h))|\mathbb{1}(||h||>R)\right].
\end{aligned}
$$

And because ℓ is K-Lipschitz, Δ_S is $2K$-Lipschitz and we have:

$$\begin{aligned}|\Delta_{\mathcal{S}_m}(\mathrm{Q}) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| &\leq 2K\mathbb{E}[||h - \mathcal{P}_R(h)||\mathbb{1}(||h|| > R)] \\ &\leq 2K\mathbb{E}[||h||\mathbb{1}(||h|| > R)].\end{aligned}$$

Finally, applying Lemma D.1.2 gives:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q}) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| \leq 2K(M+1)\frac{\beta\sqrt{2\beta}}{m} = \mathcal{O}\left(\frac{1}{m}\right).$$

Then we have:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq 2K(M+1)\frac{\beta\sqrt{2\beta}}{m} + \\ \sqrt{2K(2K+D_R)\frac{\log(\frac{1}{\delta}) + 2d\log\left(1+2Rm\right)}{m}\left(\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \alpha_m\right) + D_R^2\frac{\log\left(\frac{3m}{\delta}\right)}{m}}, \tag{5.5}$$

where $\alpha_m = \mathcal{O}\left(1 + \sqrt{\frac{d\log(Rm)}{m}}\right)$ defined in Theorem 5.3.1. Finally we exploit that $R = \mathcal{O}(\sqrt{d\log(d)}, \sqrt{\log(m)})$ (cf. Remark D.1.1) and $D_R = \mathcal{O}(1 + K^2R)$, to conclude the proof for all the three regimes. ■

5.3.2 PAC-Bayesian bounds for convex smooth functions

This section is focused on convex smooth loss functions, which are well suited for many optimisation objectives. We show that under **(A2)**, it is possible to transform Theorem 5.2.2 into a bound for smooth functions on all $\mathbb{R}^d$ when the loss remain bounded. We also show that it is possible to obtain a PAC-Bayesian bound for smooth unbounded loss functions.

Theorem 5.3.2. Assume that $d \geq 3$, $\mathcal{H} = \mathbb{R}^d$ and that the loss satisfies **(A2)** and lies in $[0, 1]$ over $\mathcal{H}$. For any $\delta > 0, 0 \leq \alpha \leq \beta, M \geq 0$, let $\mathrm{P} \in C_{\alpha,\beta,M}$ a (data-free) prior distribution. Then, with probability $1 - \delta$, for any posterior distribution $\mathrm{Q} \in C_{\alpha,\beta,M}$:

$$\begin{aligned}|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq 2&\frac{\beta\sqrt{2\beta}}{m} \\ &+ \sqrt{2D_R(2D_R+1)\frac{2d\log\left(3^{\frac{1+2Rm}{\delta}}\right)}{m}\left(\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \alpha_m\right) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}},\end{aligned}$$

with $R = \mathcal{O}\left(\max \sqrt{d\log(d)}, \sqrt{\log(m)}\right)$, $D_R = D+LR$ and $\alpha_m = \mathcal{O}(1)$ is defined in Theorem 5.3.1.

The key idea of the proof is to state that on a compact space, a smooth function is also Lipschitz. Therefore, the proof follows the same route as the one of Theorem 5.3.1, with additional technical steps. We then defer it to Appendix D.2.3. We note that, even for bounded losses, the price to pay to consider smooth functions instead of Lipschitz ones is an extra factor $D_R = \mathcal{O}(1+R)$ when $D > 0$. Therefore, in the general case we lose the idea that a tight smooth function will change the convergence rate of the problem as in general the upper bound D of $\sup_z |\ell(0_{\mathbb{R}^d}, z)|$ is greater than zero. However, we are able to obtain results still useful when enough data is available. We also show it is possible to obtain a WPB bound for unbounded convex smooth functions.

Corollary 5.3.2. Assume that $d \geq 3$, $\mathcal{H} = \mathbb{R}^d$ and that the (unbounded) loss satisfies **(A2)**. For any $\delta > 0, 0 \leq \alpha \leq \beta, M \geq 0$, we assume that $R > 0$ is the smallest value satisfying Rad. We assume that $\sup_{z\in\mathcal{Z}} ||\ell(0,\mathbf{z})|| = D_\ell < +\infty$. Let $\mathrm{P} \in C_{\alpha,\beta,M}$ a (data-free) prior distribution. Then, with probability $1-\delta$, for any posterior distribution $\mathrm{Q} \in C_{\alpha,\beta,M}$, the three following bounds holds.
Low-data regime $(d \geq m)$

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{d^{\frac{5}{2}}}{m}\left(\sqrt{\frac{d}{m}} + \mathrm{W}_1(\mathrm{Q},\mathrm{P})\right)}\right).$$

Transitory regime $(d < m, d\log(d) \geq \log(m))$

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{d^{\frac{5}{2}}}{m}\left(1 + \mathrm{W}_1(\mathrm{Q},\mathrm{P})\right)}\right).$$

Asymptotic regime $(d\log(d) < \log(m))$

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{m}}\left(1 + \mathrm{W}_1(\mathrm{Q},\mathrm{P})\right)\right).$$

In all these bounds, $\tilde{\mathcal{O}}$ hides a polynomail factor in $(\log(d), \log(m))$. For a complete formulation of the bounds, we refer to (5.6).

We remark that this theorem is particularly interesting in the transitory and asymptotic regime as, contrary to Corollary 5.3.1, we do not have a Lipschitz constant to attenuate the impact of the dimension (indeed we have $D_R = D + LR$ and in general $D > 0$). However, this bound remains of great interest when many data are available as the smoothness assumption is often used in optimisation.

Proof of Corollary 5.3.2. Firstly, we use Theorem 5.2.2 which state that for any prior on a compact, loss function $\ell \in [0,1]$ being uniformly K-Lipschitz on this compact gives with probability at least $1-\delta$:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \sqrt{2K(2K+1)\frac{\log(\frac{3}{\delta}) + 2d\log\left(1+2Rm\right)}{m}\left(\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \varepsilon_m\right) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}}.$$

Let $\mathrm{P} \in C_{\alpha,\beta,M}$. We fix R to be the smallest value satisfying $\mathtt{Rad}$ and we assume **(A2)**. On $\mathcal{B}(0,R)$, as seen in the proof of Theorem 5.3.2, ℓ is uniformly $D_R := D + LR$-Lipschitz, so ℓ is bounded on this ball by $C_R := D_\ell + RD_R = \mathcal{O}(1+R^2)$. We apply Theorem 5.2.2 on the loss function $\ell' = \ell/C_R$ and we multiply the resulting bound by C_R. Recall that ℓ' takes value in $[0,1]$ and is D_R/C_R-Lipschitz. We then have with high probability, for any $\mathrm{Q} \in C_{\alpha,\beta,M}$:

$$\begin{aligned}&|\Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| \\ &\leq \sqrt{2D_R(2D_R+C_R)\frac{\log(\frac{3}{\delta}) + 2d\log\left(1+2Rm\right)}{m}\left(\mathrm{W}_1(\mathcal{P}_R\#\mathrm{P},\mathcal{P}_R\#\mathrm{P}) + \varepsilon_m\right) + C_R^2\frac{\log\left(\frac{3m}{\delta}\right)}{m}},\end{aligned}$$

where $\varepsilon_m = \mathcal{O}\left(1\right)$ defined in Theorem 5.2.2. As in Theorem 5.3.1, we have:

$$\mathrm{W}_1(\mathcal{P}_R\#\mathrm{P},\mathcal{P}_R\#\mathrm{P}) \leq \mathrm{W}_1(\mathrm{Q},\mathrm{P}) + 2(M+1)\frac{\beta\sqrt{2\beta}}{m}.$$

We have:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq |\Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| + |\Delta_{\mathcal{S}_m}(\mathrm{Q}) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})|,$$

And we have:

$$\begin{aligned}|\Delta_{\mathcal{S}_m}(\mathrm{Q}) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| &\leq \mathbb{E}_{h\sim\mathrm{Q}}\left[|\Delta_{\mathcal{S}_m}(h) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R(h))|\right] \\ &= \mathbb{E}_{h\sim\mathrm{Q}}\left[|\Delta_{\mathcal{S}_m}(h) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R(h))|\mathbb{1}(||h|| > R)\right].\end{aligned}$$

We study the last gap more carefully:

$$\begin{aligned}|\Delta_{\mathcal{S}_m}(h) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R(h))| &= \mathbb{E}_z[|\ell(h,\mathbf{z}) - \ell(\mathcal{P}_R(h),\mathbf{z})|] \\ &+ \frac{1}{m}\sum_{i=1}^{m}|\ell(h,\mathbf{z}_i) - \ell(\mathcal{P}_R(h),z_i)|.\end{aligned}$$

And we know that for any $\mathbf{z}$, because ℓ is convex smooth:

$$\begin{aligned}\ell(h,\mathbf{z}) - \ell(\mathcal{P}_R(h),\mathbf{z}) &\leq \nabla_h \ell(\mathcal{P}_R(h),\mathbf{z})^T(h - \mathcal{P}_R(h)) + \frac{L}{2}||h - \mathcal{P}_R(h)||^2|| \\ &\leq D_R||h - \mathcal{P}_R(h)|| + \frac{L}{2}||h - \mathcal{P}_R(h)||^2||.\end{aligned}$$

We also have by convexity:

$$\begin{aligned}\ell(\mathcal{P}_R(h),\mathbf{z}) - \ell(h,\mathbf{z}) &\leq \nabla_h \ell(\mathcal{P}_R(h),\mathbf{z})^T(\mathcal{P}_R(h) - h) \\ &\leq D_R||h - \mathcal{P}_R(h)||.\end{aligned}$$

In any case, we have for any $h, \mathbf{z}$:

$$|\ell(h,\mathbf{z}) - \ell(\mathcal{P}_R(h),\mathbf{z})| \leq D_R||h - \mathcal{P}_R(h)|| + \frac{L}{2}||h - \mathcal{P}_R(h)||^2.$$

Thus:

$$\begin{aligned}|\Delta_{\mathcal{S}_m}(\mathrm{Q}) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| &\leq D_R\mathbb{E}_{h\sim\mathrm{Q}}\left[||h - \mathcal{P}_R(h)||\mathbb{1}(||h|| > R)\right] \\ &+ \frac{L}{2}\mathbb{E}_{h\sim\mathrm{Q}}\left[||h - \mathcal{P}_R(h)||^2\mathbb{1}(||h|| > R)\right] \\ &\leq D_R\mathbb{E}_{h\sim\mathrm{Q}}\left[||h||\mathbb{1}(||h|| > R)\right] \\ &+ \frac{L}{2}\mathbb{E}_{h\sim\mathrm{Q}}\left[||h||^2\mathbb{1}(||h|| > R)\right].\end{aligned}$$

And thanks to Lemma D.1.2, we finally have:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q}) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| \leq \left(D_R + \frac{L}{2}(M+1)\right)(M+1)\frac{\beta\sqrt{\beta}}{m}.$$

Then we have:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \left(D_R + \frac{L}{2}(M+1)\right)(M+1)\frac{\beta\sqrt{\beta}}{m} + \sqrt{2D_R(2D_R+C_R)\frac{\log(\frac{3}{\delta}) + 2d\log\left(1+2Rm\right)}{m}\left(\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \alpha_m\right) + C_R^2\frac{\log\left(\frac{3m}{\delta}\right)}{m}}, \tag{5.6}$$

where $\alpha_m = \mathcal{O}\left(1 + \sqrt{\frac{d\log(Rm)}{m}}\right)$ defined in Theorem 5.3.1. Finally, we exploit that $R = \mathcal{O}(\sqrt{d\log(d)}, \sqrt{\log(m)})$ (cf Remark D.1.1), that $D_R = \mathcal{O}(1+R)$ and $C_R = \mathcal{O}(1+R^2)$, to conclude the proof for all the three regimes. ■

5.4 Wasserstein PAC-Bayes with data-dependent priors

In PAC-Bayes learning, obtaining results holding with data-dependent priors is a widely studied topic. The reason behind that is that it is more meaningful to compare the posterior distribution, usually obtained via an optimisation procedure to a competitive one (classically the Gibbs posterior) to ensure tight generalisation bounds. A classical way to do so is to use differential privacy as in Dziugaite and Roy (2018b). However, their contribution relies on bounded losses to apply the *exponential mechanism*, a useful tool to determine whether an algorithm is differentially private. We exploit new theorems from Minami *et al.* (2016) and Rogers *et al.* (2016) which allow us to exploit differentially private priors when the loss is unbounded, convex and Lipschitz. We recall in Appendix D.1.3 elements of differential privacy.

A PAC-Bayesian bound for Lipschitz convex losses with data-dependent prior. We now state a PAC-Bayes theorem valid for differentially private probability kernels. The proof elaborates on Dziugaite and Roy (2018b, Theorem 4.2) and is based on the following bound, which is a minor modification of (5.5), making it valid for any prior (and not only Gaussian ones).

Theorem 5.4.1. Assume that $d \geq 3$, $\mathcal{H} = \mathbb{R}^d$ and that the loss is convex and satisfies **(A1)**. Let $\beta_m = \mathcal{O}(\frac{1}{\sqrt{m}})$ and $\lambda \leq \sqrt{m}$. Let $\mathrm{P} \in C_{\alpha,\beta,M}$ a (data-free) prior distribution. Then, for any $\beta_m < \delta < 1$, with probability $1-\delta$, for any posterior distribution $\mathrm{Q} \in C_{\alpha,\beta,M}$ and the Gibbs prior $\mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathrm{R}}_{\mathcal{S}_m}}$, the following bound holds.

Low-data regime $(d \geq m)$

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq$$
$$\tilde{\mathcal{O}}\left(\sqrt{2K\frac{d^{\frac{3}{2}}}{m}\left(\sqrt{\frac{d}{m}} + \mathrm{W}_1(\mathrm{Q}, \mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}}) + f_R\left(\mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}}\right)\right) + (1+K^2 d)\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

Transitory regime $(m > d,\ d\log(d) \geq \log(m))$

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq$$
$$\tilde{\mathcal{O}}\left(\sqrt{2K\frac{d^{\frac{3}{2}}}{m}\left(1 + \mathrm{W}_1(\mathrm{Q}, \mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}}) + f_R\left(\mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}}\right)\right) + (1+K^2 d)\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

Asymptotic regime $(d\log(d) < \log(m))$

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq$$
$$\tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1 + \mathrm{W}_1(\mathrm{Q}, \mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}})\right) + (1+K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right),$$

where $R = \mathcal{O}\left(\max\sqrt{d\log(d)}, \sqrt{\log(m)}\right)$, $f_R(\mathrm{P}) := \mathrm{W}_1(\mathcal{P}_R\#\mathrm{P}, \mathrm{P})$. In the above $\tilde{\mathcal{O}}$ hides a polynomial dependency in $(\log(d), \log(m))$. For an explicit formulation of the bounds, we refer to (5.11).

Note that in the asymptotic bound, the condition to get rid of $f_R(\mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}})$ is that λ is a fixed constant, in particular it does not depend on m. This is essential to apply the law of large numbers: a fixed learning rate in the Gibbs posterior is required for a bound with only explicit terms. Furthermore, an important message is that Lipschitz functions are well suited to the PAC-Bayes framework through Wasserstein distances. Indeed, not only are we able to recover McAllester or Catoni-type WPB bounds, but we also obtain WPB with data-dependent priors using the same techniques than PAC-Bayes learning with KL divergences. Data-dependent WPB bounds have also an additional benefit as they provide guarantees for the Bures-Wasserstein SGD of Lambert *et al.* (2022) as shown in Section 5.5.

Proof of Theorem 5.4.1. Firstly, we start from a slightly modified version of Equation (5.5) which holds for any prior distribution (and not only Gaussian ones). To obtain it we restart from the triangle inequality $\mathrm{W}_1(\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}) \leq \mathrm{W}_1(\mathcal{P}_R\#\mathrm{Q}, \mathrm{Q}) + \mathrm{W}_1(\mathrm{Q}, \mathrm{P}) + f_R(\mathrm{P})$. where $f_R(\mathrm{P}) := \mathrm{W}_1(\mathcal{P}_R\#\mathrm{P}, \mathrm{P})$ and we apply exactly the same route of proof than in Corollary 5.3.1. We then obtain, for any data-free prior P, with probability at least $1-\delta$, for any $\mathrm{Q} \in C_{\alpha,\beta,M}$:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq 2K(M+1)\frac{\beta\sqrt{2\beta}}{m} + \sqrt{C_R \frac{\log(\frac{1}{\delta}) + 2d\log(1+2Rm)}{m}\left(\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \alpha_m + f_R(\mathrm{P})\right) + D_R^2 \frac{\log\left(\frac{m}{\delta}\right)}{m}},$$

where $D_R = D + KR$ and $C_R = 2K(2K + D_R)$ (D, K defined in **(A1)**). We then denote by $\mathtt{Bound}(\mathcal{S}_m, \mathrm{P}, \mathrm{Q}, \delta)$ the bound:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| > 2K(M+1)\frac{\beta\sqrt{2\beta}}{m} + \sqrt{C_R \frac{\log(\frac{1}{\delta}) + 2d\log(1+2Rm)}{m}\left(\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \alpha_m + f_R(\mathrm{P})\right) + D_R^2 \frac{\log\left(\frac{m}{\delta}\right)}{m}}.$$

And for a given δ', let

$$\mathtt{Ev}(\mathrm{P}, \delta') := \{\mathcal{S}_m \in \mathcal{Z}^m \mid \exists \mathrm{Q} \in C_{\alpha,\beta,M} \text{ s.t. } \mathtt{Bound}(\mathcal{S}_m, \mathrm{P}, \mathrm{Q}, \delta') \text{ holds}\}$$

. We know that for a data-free prior P, $\mathbb{P}_{\mathcal{S}_m \in \mathcal{D}^m}(\mathcal{S}_m \in \mathtt{Ev}(\mathrm{P})) \leq \delta$. To exploit the differential privacy framework, we first assume having a differentially private probability kernel $\mathcal{P}$. We fix $\beta > 0$ and re-exploit the idea of DZIUGAITE and ROY (2018b):

$$\mathbb{P}_{\mathcal{S}_m \sim \mathcal{D}^m}\{\mathcal{S}_m \in \mathtt{Ev}(\mathcal{P}(S), \delta')\} \leq \mathrm{e}^{\mathbf{I}_\infty^\beta(\mathcal{P};m)} \underset{(S,\mathcal{S}'_m)\sim\mathcal{D}^{2m}}{\mathbb{P}} \{\mathcal{S}_m \in \mathtt{Ev}\left(\mathcal{P}\left(\mathcal{S}'_m\right)\right)\} + \beta \tag{5.7}$$

$$\leq \mathrm{e}^{\mathbf{I}_\infty^\beta(\mathcal{P};m)}\delta' + \beta = \delta. \tag{5.8}$$

The last line holds for any $\delta > \beta$ by fixing $\delta' = \mathrm{e}^{-I_\infty^\beta(\mathcal{P};m)}(\delta - \beta)$. Note that $\log\left(\frac{1}{\delta'}\right) = \log\left(\frac{1}{\delta-\beta}\right) + I_\infty^\beta(\mathcal{P};m)$, this suggests to bound the β-approximate max-information. To do so, we need to give specific values for the pair (ε, γ). More concretely, let $\varepsilon = \sqrt{\frac{\log(m)}{m}}, \gamma = \frac{\varepsilon}{m^4}$. Then thanks to Proposition D.1.3, we know that for $\beta_m := \mathcal{O}(\frac{1}{m})$, we have:

$$I_\infty^\beta(\mathcal{P}, m) = O\left(\log(m)\right). \tag{5.9}$$

The last thing to do is to prove that the probability kernel $\mathcal{P}_0(\mathcal{S}_m) := P_{-\lambda' m \hat{\mathsf{R}}_{\mathcal{S}_m}}$ is (ε, γ) differentially private. This is true thanks to Proposition D.1.2 which states that $\mathcal{P}_0$ satisfies differential privacy as long as $\lambda' \leq \lambda_m$ with:

$$\lambda_m := \frac{1}{2K}\sqrt{\frac{\alpha \log(m)}{m\left(1 - 2\log\log(m) + 10\log(m)\right)}} = \mathcal{O}\left(\frac{1}{\sqrt{m}}\right). \tag{5.10}$$

Note that α intervenes because for any prior $\mathrm{P} \in C_{\alpha,\beta,M}$, $-\log P(.)$ is α-strongly convex. From now we consider $\lambda' = \frac{\lambda}{2Km}$ where $\lambda \leq \sqrt{m}$. We then have $\lambda' \leq \lambda_m$. We then know, thanks to Equation (5.7) with $\beta = \beta_m$, that for any $\delta > \beta_m$, $\mathbb{P}_{\mathcal{S}_m \sim \mathcal{D}^m}\{\mathcal{S}_m \in \mathtt{Ev}(\mathcal{P}_0(\mathcal{S}_m), \delta') \leq \delta$ with $\delta' = \mathrm{e}^{-I_\infty^\beta(\mathcal{P};m)}(\delta - \beta)$ Taking the complementary event and recalling that thanks to Equation (5.9), $\log\left(\frac{1}{\delta'}\right) = \log\left(\frac{1}{\delta - \beta_m}\right) + \mathcal{O}(\log(m))$ gives, for any data-free Gaussian prior P, for any $\delta > \beta_m$, with probability at least $1 - \delta$, for any $\mathrm{Q} \in C_{\alpha,\beta,M}$:

$$\begin{aligned} |\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq 2K(M+1)\frac{\beta\sqrt{2\beta}}{m} + & \\ \sqrt{C_R \frac{\log(\frac{1}{\delta - \beta_m}) + \mathcal{O}(\log(m)) + 2d\log\left(1 + 2Rm\right)}{m}} & \\ \times \sqrt{\left(\mathrm{W}_1(\mathrm{Q}, \mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}}) + \alpha'_m + f_R(\mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}})\right)} & \\ + \sqrt{D_R^2 \frac{\log\left(\frac{m}{\delta - \beta_m}\right) + \mathcal{O}(\log(m))}{m}}, & \end{aligned} \tag{5.11}$$

where $\alpha'_m = \mathcal{O}(1 + \frac{d\log(m)}{m})$ has the same analytical expression than α_m (defined in Theorem 5.3.1) but where all the occurences of δ have been replaced by δ'. Note that in the last equation, we used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ $(a, b > 0)$ for the sake of readability but we put everything within the same square root in our theorem as it is tighter. Then, exploiting that $R = \mathcal{O}(\sqrt{d\log(d)}, \sqrt{\log(m)})$, gives us the results for the low-data and transitory regimes.

Also, we are able to prove that asymptotically, because $R\sqrt{\log(m)} \to \infty$ when m goes to infinity:

$$f_R(\mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}}) \leq \mathbb{E}[||X - \mathcal{P}_R(X)||] \underset{m\to\infty}{\to} 0,$$

where X follows the Gibbs distribution $\mathrm{P}_{-\frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}}$. The convergence to zero comes from the dominated convergence theorem. Indeed,

$$\mathbb{E}[||X - \mathcal{P}_R(X)||] = \int_{\mathbb{R}^d} g_m(x)\mathrm{d}\mathrm{P}(x),$$

with $g_m(x) = ||x - \mathcal{P}_R(x)||\frac{\exp\left(-\lambda\hat{\mathsf{R}}_{\mathcal{S}_m}(x)\right)}{\mathbb{E}_P[\exp\left(-\lambda\hat{\mathsf{R}}_{\mathcal{S}_m}(x)\right)]}$. Thus, bounding crudely gives:

$$\mathbb{E}[||X - \mathcal{P}_R(X)||] \leq \frac{1}{\inf_{m\geq 1}\mathbb{E}_P[\exp\left(-\lambda\hat{\mathsf{R}}_{\mathcal{S}_m}(x)\right)]}\int_{\mathbb{R}^d} ||x - \mathcal{P}_R(x)||\mathrm{d}\mathrm{P}(x).$$

We know that $\inf_{m\geq 1}\mathbb{E}_P[\exp\left(-\lambda\hat{\mathsf{R}}_{\mathcal{S}_m}(x)\right)] := \inf_{m\geq 1}\mathbb{E}_P[f_m(x)] > 0$ because f_m is λK - lipschitz ($x \to e^{-\lambda x}$ is λ-lipschitz and the loss ℓ is K-lipschitz) and converges almost surely on $\mathbb{R}^{\mathrm{d}}$ towards $x \to \exp -\lambda R(x)$. Indeed, thanks to the law of large numbers, we know that on $\mathbb{Q}^d$, $f_m \to f$ almost surely and using that all the sequence is λK lipschitz extends the result to all $\mathbb{R}^d$. We also notice that for any m, $f_m \leq 1$ so we can use the dominated convergence theorem to conclude that $\mathbb{E}_P[f_m(x)] \to \mathbb{E}_P[\exp(-\lambda R(X))] > 0$. So $\inf_{m\geq 1}\mathbb{E}_P[\exp\left(-\lambda\hat{\mathsf{R}}_{\mathcal{S}_m}(x)\right)] > 0$. The last thing to do is to use Lemma D.1.2 to ensure that $\int_{\mathbb{R}^d} ||x - \mathcal{P}_R(x)||\mathrm{d}\mathrm{P}(x) \to 0$. This allows us to get rid of f_R for the asymptotic regime and then, conclude the proof. ■

5.5 Generalisation ability of the Bures-Wasserstein SGD

For the sake of completeness, we recall (and precise) several elements already defined in Section 5.1.2. In PAC-Bayes learning, the following learning algorithm can be derived from a relaxation of Catoni (2007, Theorem 1.2.6), for any data-free prior P and inverse PAC-Bayesian temperature $\lambda > 0$:

$$\underset{Q\in\mathcal{M}(\mathcal{H})}{\mathrm{argmin}}\mathbb{E}_{h\sim\mathrm{Q}}[\hat{\mathsf{R}}_{\mathcal{S}_m}(h)] + 2K\frac{\mathrm{KL}(\mathrm{Q},\mathrm{P})}{\lambda}.$$

We considered the parameter $\frac{\lambda}{2K}$ as it was suggested by Theorem 5.4.1. A closed form solution is given by the Gibbs posterior $\mathrm{Q}^* := \mathrm{P}_{-\frac{\lambda}{2K}}$ such that $\mathrm{d}\mathrm{Q}^* \propto \exp(-V_{\mathcal{S}_m}(h))\mathrm{d}h$, with $V_{\mathcal{S}_m}(h) = \frac{\lambda}{2K}\hat{\mathsf{R}}_{\mathcal{S}_m}(h) - \log(\mathrm{dP}(h))$ and $\mathrm{d}h$ being the Lebesgue measure. However, such a measure can be difficult to estimate in practice. Two solutions are available. We can estimate the Gibbs posterior through MCMC methods that rely on Markov chains which (approximately) converge to Q^*. However, there is no clear stopping criterion to obtain a good approximate of the true posterior. Otherwise, we can exploit variational inference (VI) to produce rapidly a basic yet informative summary statistics on a subclass of $\mathcal{M}(\mathcal{H})$. In this section, we focus on the VI approach. As Q^* is the result of an optimal trade-off between the empirical loss $\hat{\mathsf{R}}_{\mathcal{S}_m}$ and the KL divergence (weighed by λ) acting as a regulariser, we consider the closest measure of $\mathrm{BW}(\mathbb{R}^d)$ from Q^* with respect to the KL divergence:

$$\hat{\mathrm{Q}} = \mathcal{N}(\hat{m}, \hat{\Sigma}) := \underset{Q \in \mathrm{BW}(\mathbb{R}^d)}{\operatorname{argmin}} \ \mathrm{KL}(\mathrm{Q}, \mathrm{Q}^*).$$

At the cost of this approximation, can we have an optimisation algorithm with convergence guarantees which goes to $\hat{\mathrm{Q}}$? Furthermore, if enough data is available, does $\hat{\mathrm{Q}}$ possess a good generalisation ability? We first state the assumptions holding throughout the whole section.
(A3): We assume that $\mathcal{H} = \mathbb{R}^d$ and

- There exists $M > 0$ such that $||\hat{m}|| \leq M$ almost surely.
- ℓ is twice differentiable, and **(A1), (A2)** hold. In particular, ℓ is L-smooth, convex and uniformly K-Lipschitz over $\mathcal{H}$. We furthermore assume that $L = 1$.
- The prior P used in the definition of Q^* is a Gaussian with mean 0 and covariance matrix $\Sigma = \text{diag}(\gamma), 1 \geq \gamma > 0$. We assume $\lambda \leq 2K$ in the definition of Q^*.

Note that under **(A3)**, we have $0 \prec \alpha I \preceq \nabla^2 V_{\mathcal{S}_m} \preceq I$. The work of Lambert *et al.* (2022, Theorem 4) provides convergence guarantees for SGD over the Bures-Wasserstein space when **(A3)** holds (in particular, they do not even requires the uniformly Lipschitz assumption). We first state their algorithm in Algorithm 2. Note that Algorithm 2 is a slight adaptation of the work of Lambert *et al.* (2022). Indeed, we added a projection step $\mathcal{P}_M$ within the compact of radius M in $\mathbb{R}^d$. This does not change the convergence guarantees stated in Theorem 5.5.1 as long as we assume **(A3)**.

Theorem 5.5.1. Assume **(A3)**. Also, assume that $\eta \leq \frac{\alpha^2}{60}$ and that we initialize Algorithm 2 at a matrix satisfying $\frac{\alpha}{9} I \preceq \Sigma_0 \preceq \frac{1}{\alpha} I$. Then, for all $k \in \mathbb{N}$,

$$\mathbb{E}\mathrm{W}_2^2\left(\hat{\mathrm{Q}}_k, \hat{\mathrm{Q}}\right) \leq \exp(-\alpha k \eta)\mathrm{W}_2^2\left(\hat{\mathrm{Q}}_0, \hat{\mathrm{Q}}\right) + \frac{36 d\eta}{\alpha^2}.$$

```
Algorithm 2: Bures-Wasserstein SGD.
Parameters : Strong convexity parameter $\alpha > 0$, radius $M > 0$; step size
             $\eta > 0$, initial mean $m_0$, initial covariance $\Sigma_0$
Set up $\hat{\mathrm{Q}}_0 = \mathcal{N}(m_0, \Sigma_0)$.
for $k = 0..N-1$ do
    Draw a sample $X_k \sim \hat{\mathrm{Q}}_k$.
    Set $m_k^+ = m_k - \eta \nabla V_{\mathcal{S}_m}(X_k)$.
    Set $M_k = I - \eta(\nabla V^2(X_k) - \Sigma_k^{-1})$.
    Set $\Sigma_k^+ = M_k \Sigma_k M_k$.
    Set $m_{k+1} = \mathcal{P}_M(m_k^+)$, $\Sigma_{k+1} = \mathrm{clip}^{1/\alpha}\, \Sigma_k^+$.
    Set $\hat{\mathrm{Q}}_{k+1} = \mathcal{N}(m_{k+1}, \Sigma_{k+1})$
end
Return $(\hat{\mathrm{Q}}_k)_{k=1\ldots N}$.
```

In particular, we obtain $\mathbb{E}\mathrm{W}_2^2\left(\hat{\mathrm{Q}}_k, \hat{\mathrm{Q}}\right) \leq \varepsilon^2$ provided we set $\eta \asymp \frac{\alpha^2\varepsilon^2}{d}$ and the number of iterations to be $k \gtrsim \frac{d}{\alpha^3\varepsilon^2} \log\left(\mathrm{W}_2\left(\hat{\mathrm{Q}}_0, \hat{\mathrm{Q}}\right)/\varepsilon\right)$.

We want to incorporate Theorem 5.5.1 within Theorem 5.4.1. To do so, we need to make sure that the outputs of Algorithm 2 and $\hat{\mathrm{Q}}$ lie a compact of $BW(\mathbb{R}^d)$. To do so we exploit the following lemma, which sums up the work of Lambert *et al.* (2022) (namely their Lemma 6 and the discussion in Section 3.3).

Lemma 5.5.1. Assume **(A3)** and the step-size η of Algorithm 2 is lesser than $\frac{\alpha^2}{60}$. Also in Algorithm 2, assume that $\frac{\alpha}{9}I \preceq \Sigma_k$. Then $\frac{\alpha}{9}I \preceq \Sigma_k^+$, and so, $\frac{\alpha}{9}I \preceq \Sigma_{k+1} \preceq \frac{1}{\alpha}I$. Furthermore, $I \preceq \hat{\Sigma} \preceq \frac{1}{\alpha}I$. Thus, if the initialisation of Algorithm 2 is such that $\frac{\alpha}{9}I \preceq \Sigma_0 \preceq \frac{1}{\alpha}I$, then the sequence $(\hat{\mathrm{Q}}_k)_{k\geq 0}$ and $\hat{\mathrm{Q}}$ are in the compact $C_{\frac{\alpha}{9},\frac{1}{\alpha},M}$.

Using Lemma 5.5.1, we now can apply Theorem 5.4.1 and obtain the main result of this section.

Theorem 5.5.2. Assume **(A3)**, also assume that $d \geq 3$. Let $\beta_m = \mathcal{O}(\frac{1}{\sqrt{m}})$ and fix any $\beta_m < \delta < 1$. Assume that we perform Algorithm 2, with step size $\eta \asymp \frac{\alpha^2\delta}{d}$ and the number of iterations to be $N \gtrsim \frac{d}{\alpha^3\delta} \log\left(\mathrm{W}_2\left(\mathrm{Q}_0, \hat{\mathrm{Q}}\right)/\delta\right)$. We also set the initialisation such that $\frac{\alpha}{9}I \preceq \Sigma_0 \preceq \frac{1}{\alpha}I$, then we can upper bound the generalisation ability of $\hat{\mathrm{Q}}_N$, with probability $1 - 2\delta$:

Asymptotic regime $(d\log(d) < \log(m))$

$$|\Delta_{\mathcal{S}_m}(\hat{\mathrm{Q}}_N)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1+\mathrm{W}_1(\hat{\mathrm{Q}},\mathrm{Q}^*)\right) + (1+K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right),$$

where $\tilde{\mathcal{O}}$ hides a polynomial dependency in $(\log(d),\log(m))$. We refer to (5.12) for a bound presenting the explicit influence of the Bures-Wasserstein SGD.

Theorem 5.5.2 is based on Equation (5.12) which answers the question stated in the 'Our aims in this chapter' paragraph of Section 5.1.2. We successfully designed a bound of the form of (5.2) by incorporating the optimisation guarantees of Lambert *et al.* (2022) onto a statistical framework. As such, this bound is a bridge between optimisation and PAC-Bayes learning. To the best of our knowledge, it is the first time that PAC-Bayes is able to explain why the minimiser attained by an optimisation procedure on a measure space is also able to generalise well. Until now PAC-Bayes guarantees were used as a check-in procedure, which means that during the optimisation phase it is possible to see whether the candidate predictor is able to generalise well. On the contrary our bound higlights, before any training, that the output of the Bures-Wasserstein SGD will become better at generalising, with the limit rate of $\sqrt{\frac{Kd}{m}\mathrm{W}_1(\hat{\mathrm{Q}},\mathrm{Q}^*) + \frac{\log(m)}{m}}$.

Let us analyse the bound: the convergence rate depends on the quality of the approximation $\hat{\mathrm{Q}}$ of Q^*, this says that if Gaussian measures are not suited to approximate well the Gibbs posterior, then we sacrifice some generalisation ability. However this term is also controlled by the Lipschitz constant K: if K is small, then the learning problem is easy enough to compensate both the curse of dimensionality and a possibly bad approximation $\hat{\mathrm{Q}}$ of Q^*. Again, the limit convergence rate is the statistical ersatz $\mathcal{O}\left(\sqrt{\frac{\log(m)}{m}}\right)$. This roughly says that we cannot hope to converge better than a Hoeffding test bound in this setting. Finally note also that the step η of Algorithm 2 now depends on δ: this suggests that the Bures-Wasserstein SGD needs to be tuned with a smaller step size to ensure not only convergence, but also a good generalisation ability.

Proof Proof of Theorem 5.5.2. We start from Theorem 5.4.1, considering the asymptotic case. We have with probability $1-\delta$, for the posterior $\hat{\mathrm{Q}}_N$ obtained after N steps of Algorithm 2 distribution $\mathrm{Q} \in C_{\alpha,\beta,M}$ and the prior Q^*:

$$|\Delta_{\mathcal{S}_m}(\hat{\mathrm{Q}}_N)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1+\mathrm{W}_1(\hat{\mathrm{Q}}_N,\mathrm{Q}^*)\right)+(1+K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

Then, the triangle inequality gives that $\mathrm{W}_1(\hat{\mathrm{Q}}_N,\mathrm{Q}^*) \leq \mathrm{W}_1(\hat{\mathrm{Q}}_N,\hat{\mathrm{Q}}) + \mathrm{W}_1(\hat{\mathrm{Q}},\mathrm{Q}^*)$. Finally, we exploit Theorem 5.5.1 as follows:

$$\begin{aligned}
\mathrm{W}_1(\hat{\mathrm{Q}}_N,\hat{\mathrm{Q}}) &\leq \sqrt{\mathrm{W}_2^2(\hat{\mathrm{Q}}_N,\hat{\mathrm{Q}})} && \text{by Jensen} \\
&\leq \sqrt{2\frac{\mathbb{E}[\mathrm{W}_2^2(\hat{\mathrm{Q}}_N,\hat{\mathrm{Q}})]}{\delta}} && \text{by Markov} \\
&\leq \sqrt{2\frac{\exp(-\alpha N\eta)\mathrm{W}_2^2\left(\hat{\mathrm{Q}}_0,\hat{\mathrm{Q}}\right)+\frac{36d\eta}{\alpha^2}}{\delta}} && \text{by Theorem 5.5.1.}
\end{aligned}$$

Note that in the last line, we were able to apply Theorem 5.5.1 thanks to Lemma 5.5.1. This leads to the following bound:

$$\begin{aligned}
&|\Delta_{\mathcal{S}_m}(\hat{\mathrm{Q}}_N)| \\
&\leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(f(N,\eta)\sqrt{\mathrm{W}_2^2(\hat{\mathrm{Q}}_0,\hat{\mathrm{Q}})}+1+\varepsilon\right)+(1+K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right),
\end{aligned} \tag{5.12}$$

where $f(N,\eta) = \sqrt{\frac{\exp(-\alpha N\eta)\mathrm{W}_2^2\left(\hat{\mathrm{Q}}_0,\hat{\mathrm{Q}}\right)}{\delta}}$ and $\varepsilon = \sqrt{\frac{36d\eta}{\alpha^2\delta}} + \mathrm{W}_1(\hat{\mathrm{Q}},\mathrm{Q}^*)$. Finally, using that with step size $\eta \asymp \frac{\alpha^2\delta}{d}$ and the number of iterations to be $N \gtrsim \frac{d}{\alpha^3\delta}\log\left(\mathrm{W}_2\left(\hat{\mathrm{Q}}_0,\hat{\mathrm{Q}}\right)/\delta\right)$ allows us to bound: $\sqrt{2\frac{\exp(-\alpha k\eta)\mathrm{W}_2^2\left(\hat{\mathrm{Q}}_0,\hat{\mathrm{Q}}\right)+\frac{36d\eta}{\alpha^2}}{\delta}} \leq 1$. This concludes the proof. ■

5.6 Conclusion

We extended the Wasserstein PAC-Bayes theory beyond the results of AMIT *et al.* (2022). We exploited optimisation results to explain the generalisation ability of existing algorithms and we instantiated this for the Bures-Wasserstein algorithm of LAMBERT *et al.* (2022). We conclude by discussing avenues for future works.

Can we exploit WPB for neural networks? As shown in Figure 5.1.3, we had to assume, Lipschitzness, smoothness and convexity to reach Theorem 5.5.2. Those assumptions are necessary in the current framework and to obtain the results of LAMBERT *et al.* (2022) and thus, do not cover the important case of neural networks. Therefore, an interesting lead to investigate would be to first, avoid smoothness to reach convex neural networks BENGIO *et al.* (2005) and also avoid the convexity assumption to reach the broader subclass of Lipschitz neural networks (*e.g* GOUK *et al.*, 2021). The case of Lipschitz neural networks is particularly interesting as WPB theory shows that a small Lipschitz constant is enough to attenuate the impact of dimensionality.

Are the classical PAC-Bayesian techniques suited to WPB? In Theorems 5.2.1 and 5.2.2, we exploited a surrogate of the change of measure inequality to then exploit the PAC-Bayesian theory. However, those techniques are developed around the control of an exponential moment which appears naturally through the change of measure inequality. The surrogate is tighter as it directly involves the true moment with respect to the prior: an interesting direction would be to check whether tighter concentration bounds(or other bounds exploiting weaker assumptions than a bounded loss) are accessible. Furthermore, we exploited covering numbers to state that, with high probability, the loss is close to a Lipschitz one. Those covering numbers, while crucial, involve explicitly the dimension of the problem. This is challenging as such a dependency do not appear explicitly in KL-based PAC-Bayes learning (although they play a role in the KL term).

We provide elements of answer to those two questions in Chapter 6, where we obtain tractable bounds for heavy-tailed losses, yielding sound learning algorithms for neural networks. Those benefits comes at the cost of no convergence rate for the Wasserstein term but also does not involve explicitly the dimension of $\mathcal{H}$,this practical tradeoff sacrifices the theoretical understanding of convergence of the generalisation gap to zero.

6 Wasserstein PAC-Bayes in Practice: Genrealisation-Driven Learning Algorithms for Deterministic Predictors

This chapter is based on the following paper

Paul Viallard, Maxime Haddouche, Umut Simsekli, and Benjamin Guedj. Learning via Wasserstein-Based High Probability Generalisation Bounds. (2023)

Contents

Abstract

After Chapter 5 which proposed a theoretical study of PAC-Bayes learning with Wasserstein distances, building bridges with the exploiting of convergence guarantees in generalisation, we now focus on practical expansions of Wasserstein PAC-Bayes. The optimisation view of PAC-Bayes learning is deeply exploited here: we derive theory-driven batch and online algorithms (the online paradigm attenuates the impact of the prior) valid for deterministic predictors (and thus consistent with many practical optimisation algorithms) and are derived from bounds valid for heavy-tailed lipschitz losses (weak statistical assumption and a stronger geometric one to be in line with the optimisation literature). This chapter shows that the optimisation view of PAC-Bayes leads to efficient procedures, competing with classical methods.

6.1 Introduction

Chapter 5 introduced Wasserstein PAC-Bayes learning from a theoretical perspective. Indeed, the main goal there was to incorporate the convergence guarantees of existing algorithms onto a generalisation bound. On the contrary, we focus here on deriving novel learning algorithms from Wasserstein PAC-Bayes bounds, circumventing many classical limitations of KL-based PAC-Bayes, which is the major part of the literature. Indeed, the practical use of KL divergence comes with 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 Q to be absolutely continuous with respect to the prior P. 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 Q) 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 P that can dominate Q for almost every $\mathcal{S} \sim \mathcal{D}^m$, which will trivially make $\mathrm{KL}(\mathrm{Q}\|\mathrm{P}) = +\infty$ and the generalisation bound vacuous.

Some works have focused on replacing the Kullback-Leibler divergence with more general divergences in PAC-Bayes (ALQUIER and GUEDJ, 2018; OHNISHI and HONORIO, 2021; PICARD-WEIBEL and GUEDJ, 2022), 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 (GOYAL *et al.*, 2017; 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; RODRIGUEZ-GALVEZ *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 $\mathcal{S} \sim \mathcal{D}^m$.

In the context of PAC-Bayesian learning, the recent works CHEE and LOUSTAU, 2021; AMIT *et al.*, 2022 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 the high-probability generalisation bounds based on the Wasserstein distance is still limited. The aim of this chapter 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.

(i) Using the supermartingale toolbox introduced in Chapter 2, we prove in Section 6.3.1, 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).

(ii) We provide in Section 6.3.2 the first generalisation bounds based on Wasserstein distances for the online PAC-Bayes framework of Chapter 3. 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.

(iii) As our bounds are linear with respect to Wasserstein terms (contrary to those of AMIT *et al.*, 2022 and Chapter 5), 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 in Section 6.4 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.

Outline. Section 6.2 describes our framework and background, Section 6.3 contains our new theoretical results and Section 6.4 gathers our experiments. Appendix E.1 gathers supplementary discussion, Appendix E.2 contains all proofs of our claims, and Appendix E.3 provides insights into our practical results as well as additional experiments.

6.2 Our framework

Framework. We consider a Polish 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 $(\mathcal{F}_i)_{i\geq 1}$ adapted to our data $(\mathbf{z}_i)_{i=1,\dots,m}$, and we assume that the dataset $\mathcal{S}$ follows the distribution $\mathcal{D}_{\mathcal{S}}$. In PAC-Bayes learning, we construct a data-driven posterior distribution $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ with respect to a prior distribution P.

Definitions. 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 $\mathrm{P} : \cup_{m=1}^{\infty}\mathcal{Z}^m \times \Sigma_{\mathcal{H}} \to [0,1]$ where *(i)* for any $B \in \Sigma_{\mathcal{H}}$, the function $\mathcal{S} \mapsto \mathrm{P}(\mathcal{S}, B)$ is measurable, *(ii)* for any dataset $\mathcal{S}$, the function $B \mapsto \mathrm{P}(\mathcal{S}, B)$ 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.

Batch setting. For any m, we assume the dataset $\mathcal{S}_m$ to be *i.i.d.* of size m, so there exists a distribution $\mathcal{D}$ over $\mathcal{Z}$ such that $\mathcal{D}_{\mathcal{S}_m} = \mathcal{D}^m$. We then define, for a given $h \in \mathcal{H}$, the *risk* to be $\mathsf{R}_{\mathcal{D}} := \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})]$ and its empirical counterpart $\hat{\mathsf{R}}_{\mathcal{S}_m} := \frac{1}{m}\sum_{i=1}^{m}\ell(h,\mathbf{z}_i)$. Our results aim to bound the *expected generalisation gap* defined by $\mathbb{E}_{h\sim\mathrm{Q}}[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)]$. We assume that for any $m > 0$, the dataset $\mathcal{S}_m$ is split into K disjoint sets $\mathcal{S}_m^1, \dots, \mathcal{S}_m^K$. We consider K stochastic kernels $\mathrm{P}_1, \dots, \mathrm{P}_K$ such that for any $\mathcal{S}_m$, the distribution $\mathrm{P}_i(\mathcal{S}_m, .)$ *does not* depend on $\mathcal{S}_m^i$.

Online setting. We adapt the online PAC-Bayes framework of Chapter 3. We assume that we have access to a stream of data $\mathcal{S} = (\mathbf{z}_i)_{i\geq 1}$, arriving sequentially, with no assumption on $\mathcal{D}_{\mathcal{S}}$. We assume our sequence of stochastic kernels (used as priors) $(\mathrm{P}_i)_{i=1\cdots m}$ to satisfy: *(i)* for all i and dataset $\mathcal{S}$, the distribution $\mathrm{P}_i(S,.)$ is $\mathcal{F}_{i-1}$ measurable and *(ii)* there exists P_0 such that for all $i \geq 1$, we have $\mathrm{P}_i(S,.) \ll \mathrm{P}_0$. Indeed, all those measures are uniformly continuous with respect to any Gaussian distribution. This last condition covers, in particular, the case where $\mathcal{H}$ is an Euclidean space and for any i, the distribution $\mathrm{P}_{i,\mathcal{S}}$ is a Dirac mass. This is weaker than the condition *(ii)* of the stochastic kernels sequence in Chapter 3, but enough to exploit the conditional Fubini lemma (Lemma B.4.2).

Wasserstein distance. We focus on the Wasserstein distance of order 1 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 β on $\mathcal{A}$, the Wasserstein distance is defined by

$$\mathrm{W}_1(\alpha, \beta) := \inf_{\gamma \in \Gamma(\alpha,\beta)} \mathop{\mathbb{E}}_{(a,b)\sim\gamma} d(a, b), \tag{6.1}$$

where $\Gamma(\alpha, \beta)$ is the set of joint probability measures $\gamma \in \mathcal{M}(\mathcal{A}^2)$ such that the marginals are α and β. 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; PEYRÉ and CUTURI, 2019 for an introduction to optimal transport.

6.3 Wasserstein-based PAC-Bayesian generalisation bounds

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 CHUGG *et al.*, 2023; HADDOUCHE and GUEDJ, 2023a; JANG *et al.*, 2023, to derive bounds for both batch learning (Theorems 6.3.1 and 6.3.2) and online learning (Theorems 6.3.3 and 6.3.4).

6.3.1 PAC-Bayes for batch learning with *i.i.d.* data

In this section, we use the batch setting described in Section 6.2. We state our first result, holding for heavy-tailed losses admitting order 2 moments. Such an assumption is in line, for instance, with reinforcement learning with heavy-tailed reward (see, *e.g.*, LIU and ZHAO, 2011; LU *et al.*, 2019; ZHUANG and SUI, 2021).

> **Theorem 6.3.1.** We assume the loss ℓ to be L-Lipschitz. Then, for any $\delta \in (0, 1]$, for any sequence of positive scalar $(\lambda_i)_{i\in\{1,\ldots,K\}}$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, the following holds for the distributions $\mathrm{P}_{i,\mathcal{S}} := \mathrm{P}_i(\mathcal{S}, .)$ and for any

$\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] \\
&\leq \sum_{i=1}^{K} \frac{2|\mathcal{S}_m^i|L}{m}\mathrm{W}_1(\mathrm{Q}, \mathrm{P}_{i,\mathcal{S}}) + \frac{1}{m}\sum_{i=1}^{K}\frac{\ln\left(\frac{K}{\delta}\right)}{\lambda_i} + \frac{\lambda_i}{2}\left(\mathop{\mathbb{E}}_{h\sim \mathrm{P}_{i,\mathcal{S}}}\left[\hat{V}_{|\mathcal{S}_m^i|}(h) + V_{|\mathcal{S}_m^i|}(h)\right]\right),
\end{aligned}
$$

where $\mathrm{P}_{i,\mathcal{S}}$ *does not* depend on $\mathcal{S}_m^i$. Also, for any $i, |\mathcal{S}_m^i|$, we have $\hat{V}_{|\mathcal{S}_m^i|}(h) = \sum_{\mathbf{z}\in\mathcal{S}_m^i}\left(\ell(h,\mathbf{z}) - R_{\mathcal{D}}(h)\right)^2$ and $V_{|\mathcal{S}_m^i|}(h) = \mathbb{E}_{\mathcal{S}_m^i}\left[\hat{V}_{|\mathcal{S}_m^i|}(h)\right]$.

The proof is deferred to Appendix E.2.1. While Theorem 6.3.1 holds for losses taking values in $\mathbb{R}$, many learning problems rely in practice on more constrained losses. This loss can be bounded as in the case of, *e.g.*, supervised learning or the multi-armed bandit problem (SLIVKINS, 2019), or simply non-negative as in regression problems involving the quadratic loss (studied, for instance, in CATONI, 2016; CATONI and GIULINI, 2017). Using again the supermartingale toolbox, we prove in Theorem 6.3.2 a tighter bound holding for heavy-tailed non-negative losses.

Theorem 6.3.2. We assume our loss ℓ to be non-negative and L-Lipschitz. We also assume that, for any $1 \leq i \leq K$, for any dataset $\mathcal{S}$, we have $\mathbb{E}_{h\sim \mathrm{P}_i(.,\mathcal{S}), z\sim\mathcal{D}}\left[\ell(h,z)^2\right] \leq 1$ (*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 $\mathrm{P}_{i,\mathcal{S}} := \mathrm{P}_i(\mathcal{S}, .)$ and for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:

$$
\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] \leq \sum_{i=1}^{K}\frac{2|\mathcal{S}_m^i|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_{i,\mathcal{S}}) + \sum_{i=1}^{K}\sqrt{\frac{2|\mathcal{S}_m^i|\ln\frac{K}{\delta}}{m^2}},
$$

where $\mathrm{P}_{i,\mathcal{S}}$ *does not* depend on $\mathcal{S}_m^i$.

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 $\frac{1}{2}$. This result is rigorously stated in Theorem E.2.1 of Appendix E.2.3.

High-level ideas of the proofs. Theorems 6.3.1 and 6.3.2 are structured around two tools. First, we exploit the Kantorovich-Rubinstein duality VILLANI, 2009, Remark 6.5 to replace the change of measure inequality CSISZÁR, 1975; DONSKER and VARADHAN, 1976; this allows us to consider a Wasserstein distance instead of a KL term. Then, we exploit the supermartingales used in Chapter 2 and CHUGG *et al.*, 2023 alongside Ville's inequality (instead of Markov's one) to obtain a high probability bound holding for heavy-tailed losses. Combining those techniques provides our PAC-Bayesian bounds.

Analysis of our bounds. Our results hold for Lipschitz losses and allow us to consider heavy-tailed losses with bounded order 2 moments. While such an assumption on the loss is more restrictive than in classical PAC-Bayes, allowing heavy-tailed losses is strictly less restrictive. While Theorem 6.3.1 is our most general statement, Theorem 6.3.2 allows recovering a tighter result (without empirical variance terms) for non-negative heavy-tailed losses. An important point is that the variance terms are considered with respect to the prior distributions $\mathrm{P}_{i,\mathcal{S}}$ and not Q as in Chapter 2. This is crucial as these chapters rely on the implicit assumption of order 2 moments, holding uniformly for all $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$, while we only require this assumption for the prior distributions $(\mathrm{P}_{i,\mathcal{S}})_{i=1,\dots,K}$. Such an assumption is in line with the PAC-Bayesian literature, which often relies on bounding an averaged quantity with respect to the prior. This strength is a consequence of the Kantorovich-Rubinstein duality. To illustrate this, consider *i.i.d.* data with distribution $\mathcal{D}$ admitting a finite variance bounded by V and the loss $\ell(h,z) = |h-z|$ where both h and z lie in the real axis. Notice that in this particular case, we can imagine that z is a data point and h is a hypothesis outputting the same scalar for all data. To satisfy the assumption of Theorem 6.3.2, it is enough, by Cauchy Schwarz, to satisfy $\mathbb{E}_{h\sim \mathrm{P}_{i,\mathcal{S}}, z\sim\mathcal{D}}[\ell(h,z)^2] \leq \mathbb{E}[h^2] + 2V\,\mathbb{E}[|h|] + V^2 \leq 1$ for all $\mathrm{P}_{i,\mathcal{S}}$. On the contrary, Chapter 2 would require this condition to hold for all Q, which is more restrictive. Finally, an important point is that our bound allows us to consider Dirac distributions with disjoint support as priors and posteriors. On the contrary, KL divergence forces us to consider a non-Dirac prior for our bound to be non-vacuous. This allows us to retrieve a uniform-convergence bound described in Corollary E.2.1.

Role of data-dependent priors. Theorems 6.3.1 and 6.3.2 allow the use of prior distributions depending possibly on a fraction of data. Such a dependency is crucial to control our sum of Wasserstein terms as we do not have an explicit convergence rate. For instance, for a fixed K, consider a compact predictor space $\mathcal{H}$, a bounded loss and the *Gibbs posterior* defined as $d\mathrm{Q}(h) \propto \exp\left(-\lambda \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right) dh$ where $\lambda > 0$. Also define for any i and $\mathcal{S}$, the distribution $d\mathrm{P}_{i,\mathcal{S}}(h) \propto \exp\left(-\lambda \mathsf{R}_{\mathcal{S}/\mathcal{S}_m^i}(h)\right) dh$. Then, by the law of large numbers, when m goes to infinity, for any h, both $\mathsf{R}_{\mathcal{S}}(h)$ and $(\mathsf{R}_{\mathcal{S}/\mathcal{S}_m^i}(h))_{i=1,\dots,m}$ converge to $\mathsf{R}_{\mathcal{D}}(h)$. This ensures, alongside with the dominated convergence theorem, that for any i, the Wasserstein distance $\mathrm{W}_1(\mathrm{Q}, \mathrm{P}_{i,\mathcal{S}})$ goes to zero as m goes to infinity.

Comparison with the literature. Amit *et al.*, 2022, Theorem 11 establishes a PAC-Bayes bound with Wasserstein distance valid for bounded losses being Lipschitz with high probability. While we circumvent the first assumption, the second one is less restrictive than actual Lipschitzness and can also be used in our setting. Also Amit *et al.*, 2022, Theorem 12 proposes an explicit convergence for finite predictor classes. We show in Appendix E.1 that we are also able to recover such a convergence.

Towards new PAC-Bayesian algorithms. From Theorem 6.3.2, we derive a new

PAC-Bayesian algorithm for Lipschitz non-negative losses:

$$\operatorname*{argmin}_{\mathrm{Q}\in\mathcal{M}(\mathcal{H})} \mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] + \sum_{i=1}^{K} \frac{2|\mathcal{S}_m^i|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_{i,\mathcal{S}}). \tag{6.2}$$

Equation (6.2) uses Wasserstein distances as regularisers and allows the use of multiple priors. We compare ourselves to the classical PAC-Bayes algorithm derived from Catoni, 2007, Theorem 1.2.6 (which leads to Gibbs posteriors):

$$\operatorname*{argmin}_{\mathrm{Q}\in\mathcal{M}(\mathcal{H})} \mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P})}{\lambda}. \tag{6.3}$$

Considering a Wasserstein distance in Equation (6.2) makes our algorithm more flexible than in Equation (6.3), the KL divergence implies absolute continuity *w.r.t.* the prior P. Such an assumption is not required to use Equation (6.2) and covers the case of prior Dirac distributions. Finally, Equation (6.2) relies on a fixed value K whose value is discussed below.

Role of K. We study the cases $K = 1$, $\sqrt{m}$, and m in Theorem 6.3.2. We refer to Appendix E.1 for a detailed treatment. First of all, when $K = 1$, we recover a classical batch learning setting where all data are collected at once. In this case, we have a single Wasserstein with no convergence rate coupled with a statistical ersatz of $\sqrt{\frac{\ln(1/\delta)}{m}}$. However, similarly to Amit *et al.*, 2022, Theorem 12, in the case of a finite predictor class, we are able to recover an explicit convergence rate. The case $K = \sqrt{m}$ provides a tradeoff between the number of points required to have good data-dependent priors (which may lead to a small $\sum_{i=1}^{\sqrt{m}} \mathrm{W}_1(\mathrm{Q},\mathrm{P}_i)$) and the number of sets required to have an explicit convergence rate. Finally, the case $K = m$ leads to a vacuous bound as we have the incompressible term $\sqrt{\ln\left(\frac{m}{\delta}\right)}$, which makes the bound vacuous for large values of m. This means that the batch setting is not fitted to deal with a data stream arriving sequentially. To mitigate that weakness, we propose in Section 6.3.2 the first online PAC-Bayes bounds with Wasserstein distances.

6.3.2 Wasserstein-based generalisation bounds for online learning

Here, we use the online setting described in Section 6.2 and derive the first online PAC-Bayes bounds involving Wasserstein distances in Theorems 6.3.3 and 6.3.4. Online PAC-Bayes bounds are meant to derive online counterparts of classical PAC-Bayesian algorithms as in Chapter 3, where the KL-divergence acts as a regulariser. We show in Theorems 6.3.3 and 6.3.4 that it is possible to consider online PAC-Bayesian algorithms where the regulariser is a Wasserstein distance, which allows us to optimise on measure spaces without a restriction of absolute continuity.

Theorem 6.3.3. We assume our loss ℓ to be L-Lipschitz. Then, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, the following holds for the distributions $\mathrm{P}_{i,\mathcal{S}} := \mathrm{P}_i(\mathcal{S}, .)$ and for any sequence $(\mathrm{Q}_i)_{i=1\cdots m} \in \mathcal{M}(\mathcal{H})^m$:

$$\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h_i \sim \mathrm{Q}_i} \Big[\mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] - \ell(h_i, \mathbf{z}_i) \Big] \leq 2L \sum_{i=1}^{m} \mathrm{W}_1(\mathrm{Q}_i, \mathrm{P}_{i,\mathcal{S}}) \\ + \frac{\lambda}{2} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{h_i \sim \mathrm{P}_{i,\mathcal{S}}} \Big[\hat{V}_i(h_i, \mathbf{z}_i) + V_i(h_i) \Big] + \frac{\ln(1/\delta)}{\lambda},$$

where for all i, $\hat{V}_i(h_i, \mathbf{z}_i) = (\ell(h_i, \mathbf{z}_i) - \mathbb{E}_{i-1}[\ell(h_i, \mathbf{z}_i)])^2$ is the conditional empirical variance at time i and $V_i(h_i) = \mathbb{E}_{i-1}[\hat{V}(h_i, \mathbf{z}_i)]$ is the true conditional variance.

The proof is deferred to Appendix E.2.4. We also provide the following bound, being an online analogous of Theorem 6.3.2, valid for non-negative heavy-tailed losses.

Theorem 6.3.4. We assume our loss ℓ to be non-negative and L-Lipschitz. We also assume that, for any $i, \mathcal{S}$, $\mathbb{E}_{h \sim \mathrm{P}_i(.,\mathcal{S})} \left[\mathbb{E}_{i-1}[\ell(h, \mathbf{z}_i)^2] \right] \leq 1$ (*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 stochastic kernels sequence (used as priors) $(\mathrm{P}_i)_{i \geq 1}$, we have with probability at least $1-\delta$ over the sample $S \sim \mathcal{D}$, the following, holding for the data-dependent measures $\mathrm{P}_{i,\mathcal{S}} := \mathrm{P}_i(S, .)$ and any posterior sequence $(\mathrm{Q}_i)_{i \geq 1}$:

$$\frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{h_i \sim \mathrm{Q}_i} \Big[\mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] - \ell(h_i, \mathbf{z}_i) \Big] \leq \frac{2L}{m} \sum_{i=1}^{m} \mathrm{W}_1(\mathrm{Q}_i, \mathrm{P}_{i,\mathcal{S}}) + \sqrt{\frac{2 \ln\left(\frac{1}{\delta}\right)}{m}}.$$

The proof is deferred to Appendix E.2.5.

Analysis of our bounds. Theorems 6.3.3 and 6.3.4 are, to our knowledge, the first results involving Wasserstein distances for online PAC-Bayes learning. They are the online counterpart of Theorems 6.3.1 and 6.3.2, and the discussion of Section 6.3.1 about the involved assumptions also apply here. The sum of Wasserstein distances involved here is a consequence of the online setting and must grow sublinearly for the bound to be tight. For instance, when $(\mathrm{Q}_i = \delta_{h_i})_{i \geq 1}$ is the output of an online algorithm outputting Dirac measures and $\mathrm{P}_{i,\mathcal{S}} = \mathrm{Q}_{i-1}$, the sum of Wasserstein is exactly $\sum_{i=1}^{m} d(h_i, h_{i-1})$. This sum has to be sublinear for the bound to be non-vacuous, and the tightness depends on the considered learning problem. An analogous of this sum can be found in dynamic online learning ZINKEVICH, 2003 where similar sums appear as *path lengths* to evaluate the complexity of the problem.

Comparison with literature. We compare our results to existing PAC-Bayes bounds for martingales of SELDIN *et al.*, 2012b. SELDIN *et al.*, 2012b, Theorem 4 is a PAC-

Bayes bound for martingales, which controls an average of martingales, similar to our Theorem 6.3.1. Under a boundedness assumption, they recover a McAllester-typed bound, while Theorem 6.3.1 is more of a Catoni-typed result. Also, SELDIN *et al.*, 2012b, Theorem 7 is a Catoni-typed bound involving a conditional variance, similar to our Theorem 6.3.4. They require to bound uniformly the variance on all the predictor sets, while we only assume averaged variance with respect to priors, which is what we required to perform Theorem 6.3.4.

A new online algorithm. Chapter 3 derived from their main theorem, an online counterpart of Equation (6.3), proving it comes with guarantees. Similarly, we exploit Theorem 6.3.4 to derive the online counterpart of Equation (6.2), from the data-free initialisation Q_1

$$\forall i \geq 1, \quad \mathrm{Q}_i \in \underset{\mathrm{Q}\in\mathcal{M}(\mathcal{H})}{\operatorname{argmin}} \underset{h\sim\mathrm{Q}}{\mathbb{E}} \left[\ell(h_i, \mathbf{z}_i)\right] + 2L\mathrm{W}_1(\mathrm{Q}, \mathrm{P}_{i,\mathcal{S}}). \tag{6.4}$$

We highlight the merits of the algorithm defined by Equation (6.4), alongside with the one from Equation (6.2), in Section 6.4.

6.4 Learning via Wasserstein regularisation

Theorems 6.3.2 and 6.3.4 are designed to be informative on the generalisation ability of a single hypothesis even when Dirac distributions are considered. In particular, our results involve Wasserstein distances acting as regularisers on $\mathcal{H}$. In this section, we show that a Wasserstein regularisation of the learning objective, which comes from our theoretical bounds, helps to better generalise in practice. Inspired by Equations (6.2) and (6.4), we derive new PAC-Bayesian algorithms for both batch and online learning involving a Wasserstein distance (see Section 6.4.1), we describe our experimental framework in Section 6.4.2 and we present some of the results in Section 6.4.3. Additional details, experiments, and discussions are gathered in Appendix E.3 due to space constraints. All the experiments are reproducible with the source code provided on GitHub at `https://github.com/paulviallard/NeurIPS23-PB-Wasserstein`.

6.4.1 Learning algorithms

Classification. In the classification setting, we assume that the data space $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$ is composed of a d-dimensional *input space* $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 \leq 1\}$ and a finite *label space* $\mathcal{Y} = \{1, \ldots, |\mathcal{Y}|\}$ with $|\mathcal{Y}|$ labels. We aim to learn models $h_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$ parameterised by a weight vector $\mathbf{w}$ that outputs, given an input $\mathbf{x} \in \mathcal{X}$, a score $h_{\mathbf{w}}(\mathbf{x})[y'] \in \mathbb{R}$ for each label y'. This score allows us to assign a label to $\mathbf{x} \in \mathcal{X}$; to check if $h_{\mathbf{w}}$ classifies correctly the example $(\mathbf{x}, y)$, we use the *classification loss* defined by $\ell^c(h_{\mathbf{w}}, (\mathbf{x}, y)) := \mathbb{1}\left[h_{\mathbf{w}}(\mathbf{x})[y] - \max_{y' \neq y} h_{\mathbf{w}}(\mathbf{x})[y'] \leq 0\right]$, where $\mathbb{1}$ denotes the indicator function.

Batch algorithm. In the batch setting, we aim to learn a parametrised hypothesis $h_{\mathbf{w}} \in \mathcal{H}$ that minimises the population classification risk, namely, $\mathfrak{R}_{\mathcal{D}}(h_{\mathbf{w}}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\, \ell^c(h_{\mathbf{w}}, (\mathbf{x}, y))$ that we can only estimate through the empirical classification risk $\mathfrak{R}_{\mathcal{S}}(h_{\mathbf{w}}) = \frac{1}{m}\sum_{i=1}^m \ell^c(h_{\mathbf{w}}, (\mathbf{x}_i, y_i))$. To learn the hypothesis, we start from Equation (6.2), when the distributions Q and $\mathrm{P}_1, \ldots, \mathrm{P}_K$ are Dirac masses, localised at $h_{\mathbf{w}}, h_{\mathbf{w}_1}, \ldots h_{\mathbf{w}_K} \in \mathcal{H}$ respectively. Indeed, in this case, $\mathrm{W}_1(\mathrm{Q}, \mathrm{P}_{i,\mathcal{S}}) = d(h_{\mathbf{w}}, h_{\mathbf{w}_i})$ for any i. However, the loss $\ell^c(., \mathbf{z})$ is not Lipschitz and the derivatives are zero for all examples $\mathbf{z} \in \mathcal{X} \times \mathcal{Y}$, which prevents its use in practice to obtain such a hypothesis $h_{\mathbf{w}}$. Instead, for the population risk $\mathsf{R}_{\mathcal{D}}(h)$ and the empirical risk $\hat{\mathsf{R}}_{\mathcal{S}_m}(h)$ (in Theorem 6.3.2 and Equation (6.2)), we consider the loss defined as $\ell(h, (\mathbf{x}, y)) = \frac{1}{|\mathcal{Y}|}\sum_{y' \neq y} \max(0, 1 - \eta(h[y] - h[y']))$, which is η-Lipschitz *w.r.t.* $h[1], \ldots, h[|\mathcal{Y}|]$. This loss has subgradients everywhere, which is convenient in practice. We go a step further by *(a)* setting $L = \frac{1}{2}$ and *(b)* adding a parameter $\varepsilon > 0$ to obtain the objective

$$\underset{h_{\mathbf{w}} \in \mathcal{H}}{\operatorname{argmin}} \left\{ \hat{\mathsf{R}}_{\mathcal{S}_m}(h_{\mathbf{w}}) + \varepsilon \left[\sum_{i=1}^{K} \frac{|\mathcal{S}_m^i|}{m} d\left(h_{\mathbf{w}}, h_{\mathbf{w}_i}\right) \right] \right\}. \tag{6.5}$$

To (approximately) solve Equation (6.5), we propose a two-step algorithm. First, PRIORS LEARNING learns K hypotheses $h_{\mathbf{w}_1}, \ldots, h_{\mathbf{w}_K} \in \mathcal{H}$ by minimising the empirical risk via stochastic gradient descent. Second, POSTERIOR LEARNING learns the hypothesis $h_{\mathbf{w}} \in \mathcal{H}$ by minimising the objective associated with Equation (6.5). More precisely, PRIORS LEARNING outputs the hypotheses $h_{\mathbf{w}_1}, \cdots, h_{\mathbf{w}_K}$, obtained by minimising the empirical risk through mini-batches. Those batches are designed such that for any i, the hypothesis $h_{\mathbf{w}_i}$ does not depend on $\mathcal{S}_m^i$. Then, given $h_{\mathbf{w}_1}, \ldots, h_{\mathbf{w}_K} \in \mathcal{H}$, POSTERIOR LEARNING minimises the objective in Equation (6.5) with mini-batches. Those algorithms are presented in Algorithm 4 of Appendix E.3. While ε is not suggested by Equation (6.2), it helps to control the impact of the regularisation in practice. Equation (6.5) then optimises a tradeoff between the empirical risk and the regularisation term $\varepsilon \sum_{i=1}^K \frac{|\mathcal{S}_m^i|}{m} d(h_{\mathbf{w}}, h_{\mathbf{w}_i})$.

Online algorithm. Online algorithms output, at each time step $i \in \{1, \ldots, m\}$, a new hypothesis $h_{\mathbf{w}_i}$. From Equation (6.4), particularised to a sequence of Dirac distributions (localised in $h_{\mathbf{w}_1}, \cdots, h_{\mathbf{w}_K}$), we design a novel online PAC-Bayesian algorithm with a Wasserstein regulariser:

$$\forall i \geq 1, \quad h_i \in \underset{h_{\mathbf{w}} \in \mathcal{H}}{\operatorname{argmin}}\, \ell(h_{\mathbf{w}}, \mathbf{z}_i) + d\left(h_{\mathbf{w}}, h_{\mathbf{w}_{i-1}}\right) \quad \textit{s.t.} \quad d\left(h_{\mathbf{w}}, h_{\mathbf{w}_{i-1}}\right) \leq 1. \tag{6.6}$$

According to Theorem 6.3.4, such an algorithm aims to bound the *population cumulative classification loss* $\mathfrak{C}_{\mathcal{D}} = \sum_{i=1}^m \mathbb{E}[\ell^c(h_{\mathbf{w}_i}, \mathbf{z}_i) \mid \mathcal{F}_{i-1}]$. Note that we added the constraint $d\left(h_{\mathbf{w}}, h_{\mathbf{w}_{i-1}}\right) \leq 1$ compared to Equation (6.4). This constraint ensures that the new hypothesis $h_{\mathbf{w}_i}$ is not too far from $h_{\mathbf{w}_{i-1}}$ (in the sense of the distance $\|\cdot\|_2$). Note that the constrained optimisation problem in Equation (6.6) can be rewritten in an unconstrained form (see BOYD and VANDENBERGHE, 2004) thanks to a barrier $B(\cdot)$ defined by $B(a) = 0$ if $a \leq 0$ and $B(a) = +\infty$ otherwise; we have

$$\forall i \geq 1, \quad h_i \in \operatorname*{argmin}_{h_{\mathbf{w}} \in \mathcal{H}} \ell(h_{\mathbf{w}}, \mathbf{z}_i) + d\left(h_{\mathbf{w}}, h_{\mathbf{w}_{i-1}}\right) + B(d\left(h_{\mathbf{w}}, h_{\mathbf{w}_{i-1}}\right) - 1). \tag{6.7}$$

When solving the problem in Equation (6.7) is not feasible, we approximate it with a log barrier of KERVADEC *et al.*, 2022 (suitable in a stochastic gradient setting); given a parameter $t > 0$, the log barrier extension is defined by $\hat{B}(a) = -\frac{1}{t}\ln(-a)$ if $a \leq -\frac{1}{t^2}$ and $\hat{B}(a) = ta - \frac{1}{t}\ln(\frac{1}{t^2}) + \frac{1}{t}$ otherwise. We present in Appendix E.3 Algorithm 5 that aims to (approximately) solve Equation (6.7). To do so, for each new example $(\mathbf{x}_i, y_i)$, the algorithm runs several gradient descent steps to optimise Equation (6.7).

6.4.2 Experimental framework

In this part, we assimilate the predictor space $\mathcal{H}$ to the parameter space $\mathbb{R}^d$. Thus, the distance d is the Euclidean distance between two parameters: $d\left(h_{\mathbf{w}}, h_{\mathbf{w}'}\right) = \|\mathbf{w}-\mathbf{w}'\|_2$. This implies that the Lipschitzness of ℓ has to be taken *w.r.t.* $\mathbf{w}$ instead of $h_{\mathbf{w}}$.

Models. We consider that the models are either linear or neural networks (NN). Linear models are defined by $h_{\mathbf{w}}(\mathbf{x}) = W\mathbf{x} + b$, where $W \in \mathbb{R}^{|\mathcal{Y}|\times d}$ is the weight matrix, $b \in \mathbb{R}^{|\mathcal{Y}|}$ is the bias, and $\mathbf{w} = \operatorname{vec}(\{W, b\})$ its vectorisation; the vector $\mathbf{w}$ with the zero vector. Thanks to the definition of $\mathcal{X}$, we know from Lemma E.3.1 (and the composition of Lipschitz functions) that the loss is $\sqrt{2}\eta$-Lipschitz *w.r.t.* $\mathbf{w}$. For neural networks, we consider fully connected ReLU neural networks with L hidden layers and D nodes, where the leaky ReLU activation function $\mathrm{ReLU} : \mathbb{R}^D \to \mathbb{R}^D$ applies elementwise $x \mapsto \max(x, 0.01x)$. More precisely, the network is defined by $h_{\mathbf{w}}(\mathbf{x}) = Wh^L(\cdots h^1(\mathbf{x})) + b$ where $W \in \mathbb{R}^{|\mathcal{Y}|\times D}$, $b \in \mathbb{R}^{|\mathcal{Y}|}$. Each layer $h^i(\mathbf{x}) = \mathrm{ReLU}(W_i\mathbf{x} + b_i)$ has a weight matrix $W_i \in \mathbb{R}^{D\times D}$ and bias $b_i \in \mathbb{R}^D$ except for $i = 1$ where we have $W_1 \in \mathbb{R}^{D\times d}$. The weights $\mathbf{w}$ are also the vectorisation $\mathbf{w} = \operatorname{vec}(\{W, W_L, \ldots, W_1, b, b_L, \ldots, b_1\})$. We have precised in Lemma E.3.2 that our loss is Lipschitz *w.r.t.* the weights $\mathbf{w}$. We initialise the network similarly to DZIUGAITE and ROY, 2017 by sampling the weights from a Gaussian distribution with zero mean and a standard deviation of $\sigma = 0.04$; the weights are further clipped between -2σ and $+2\sigma$. Moreover, the values in the biases $b_1, \ldots, b_L$ are set to 0.1, while the values for b are set to 0. In the following, we consider $D = 600$ and $L = 2$; more experiments are considered in the appendix.

Optimisation. To perform the gradient steps, we use the COCOB-Backprop optimiser Orabona and Tommasi, 2017 (with parameter $\alpha = 10000$).[1] This optimiser is flexible as the learning rate is adaptative and, thus, does not require hyperparameter tuning. For Algorithm 4, which solves Equation (6.5), we fix a batch size of 100, *i.e.*, $|\mathcal{U}| = 100$, and the number of epochs T and T' are fixed to perform at least 20000 iterations. Regarding Algorithm 5, which solves Equation (6.7), we set $t = 100$ for the log barrier, which is enough to constrain the weights and the number of iterations to $T = 10$.

Datasets. We study the performance of Algorithms 4 and 5 on UCI datasets (Dua and Graff, 2017) along with MNIST (LeCun, 1998) and FashionMNIST (Xiao *et al.*, 2017). We also split all the data (from the original training/test set) in two halves; the first part of the data serves in the algorithm (and is considered as a training set), while the second part is used to approximate the population risks $\mathfrak{R}_{\mathcal{D}}(h)$ and $\mathfrak{C}_{\mathcal{D}}$ (and considered as a testing set).

6.4.3 Results

We present in Tables 6.1 and 6.2 the performance of Algorithms 4 and 5 compared to the Empirical Risk Minimisation (ERM) and the Online Gradient Descent (OGD) with the COCOB-Backprop optimiser. Tables 6.1a and 6.2a present the results of Algorithm 4 for the *i.i.d.* setting on linear and neural networks respectively, while Tables 6.1b and 6.2b present the results of Algorithm 5 for the online case.

Analysis of the results. In batch learning, we note that the regularisation term brings generalisation improvements compared to the empirical risk minimisation. Indeed, our batch algorithm (Algorithm 4) has a lower population risk $\mathfrak{R}_{\mathcal{D}}(h)$ on 11 datasets for the linear models and 9 datasets for the neural networks. In particular, notice that NNs obtained from Algorithm 4 are more efficient than the ones obtained from ERM on MNIST and FashionMNIST, which are the more challenging datasets. This suggests that the regularisation term helps to generalise well. For the online case, the performance of the linear models obtained from our algorithm (Algorithm 5) and by OGD are comparable: we have a tighter population classification risk $\mathfrak{R}_{\mathcal{D}}(h)$ on 5 datasets over 13. However, notice that the risk difference is less than 0.05 on 6 datasets. The advantage of Algorithm 5 is more pronounced for neural networks: we improve the performance in all datasets except adult and sensorless. Hence, this confirms that optimising the regularised loss $\ell(h_{\mathbf{w}}, \mathbf{z}_i) + \|\mathbf{w} - \mathbf{w}_{i-1}\|$ brings a good advantage compared to the loss $\ell(h_{\mathbf{w}}, \mathbf{z}_i)$ only. A possible explanation would be that OGD suffers from underfitting (with a high empirical risk $\mathfrak{C}_{\mathcal{D}}$) while we are able to control overfitting through a regularisation term. Indeed, only one gradient descent step is done for each new datum $(\mathbf{x}_i, y_i)$, which might not be sufficient to decrease

[1] The parameter α in COCOB-Backprop can be seen as an initial learning rate; see Orabona and Tommasi, 2017.

Table 6.1. *Performance of Algorithms 4 and 5 compared respectively to ERM and OGD on different datasets on linear models. For the i.i.d. setting, we consider $\varepsilon = \frac{1}{m}$ and $\varepsilon = \frac{1}{\sqrt{m}}$ and with $K = 0.2\sqrt{m}$. For each method, we plot the empirical risk $\mathfrak{R}_{\mathcal{S}}(h)$ or $\mathfrak{C}_{\mathcal{S}}$ with its associated test risk $\mathfrak{R}_{\mathcal{D}}(h)$ or $\mathfrak{C}_{\mathcal{D}}$. The risk in* **bold** *corresponds to the lowest one among the ones considered. For the online case, the two population risks are underlined when the absolute difference is lower than 0.05.*

(a) *Linear model – batch learning*

Dataset	Algo. 4 ($\frac{1}{m}$)		Algo. 4 ($\frac{1}{\sqrt{m}}$)		ERM	
	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$
ADULT	.165	**.166**	.165	.167	.166	.167
FASHIONMNIST	.128	.151	.126	**.148**	.139	.153
LETTER	.285	.297	.287	**.296**	.287	.297
MNIST	.200	.216	.066	.092	.065	**.091**
MUSHROOMS	.001	**.001**	.001	**.001**	.001	**.001**
NURSERY	.766	**.773**	.760	**.773**	.794	.807
PENDIGITS	.049	**.059**	.050	.061	.052	.064
PHISHING	.063	**.067**	.065	.069	.064	**.067**
SATIMAGE	.144	**.200**	.138	.201	.148	.209
SEGMENTATION	.057	**.216**	.164	.386	.087	.232
SENSORLESS	.129	**.129**	.131	.131	.134	.136
TICTACTOE	.388	.299	.013	**.021**	.228	.238
YEAST	.527	.497	.524	.504	.470	**.427**

(b) *Linear model – online learning*

Dataset	Algo. 5		OGD	
	$\mathfrak{C}_{\mathcal{S}}$	$\mathfrak{C}_{\mathcal{D}}$	$\mathfrak{C}_{\mathcal{S}}$	$\mathfrak{C}_{\mathcal{D}}$
ADULT	.230	**.236**	.248	.248
FASHIONMNIST	.223	**.282**	.540	.548
LETTER	.919	.935	.916	**.926**
MNIST	.284	**.310**	.378	.397
MUSHROOMS	.218	.222	.082	**.087**
NURSERY	.794	.807	.789	**.805**
PENDIGITS	.342	**.484**	.589	.600
PHISHING	.226	.242	.226	**.220**
SATIMAGE	.669	.938	.635	**.888**
SEGMENTATION	.749	**.803**	.738	.893
SENSORLESS	.906	.910	.825	**.830**
TICTACTOE	.443	.468	.390	**.303**
YEAST	.699	.713	.667	**.708**

the loss. Instead, our method solves the problem associated with Equation (6.7) and constrains the descent with the norm $\|\mathbf{w} - \mathbf{w}_{i-1}\|$.

Table 6.2. *Performance of Algorithms 4 and 5 compared respectively to ERM and OGD on different datasets on neural network models. For the i.i.d. setting, we consider* $\varepsilon = \frac{1}{m}$ *and* $\varepsilon = \frac{1}{\sqrt{m}}$ *and with* $K = 0.2\sqrt{m}$*. For each method, we plot the empirical risk* $\mathfrak{R}_{\mathcal{S}}(h)$ *or* $\mathfrak{C}_{\mathcal{S}}$ *with its associated test risk* $\mathfrak{R}_{\mathcal{D}}(h)$ *or* $\mathfrak{C}_{\mathcal{D}}$*. The risk in* **bold** *corresponds to the lowest one among the ones considered. For the online case, the two population risks are underlined when the absolute difference is lower than 0.05.*

(a) *NN model – batch learning*

(b) *NN model – online learning*

Dataset	Algo. 4 ($\frac{1}{m}$)		Algo. 4 ($\frac{1}{\sqrt{m}}$)		ERM		Algo. 5		OGD	
	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$	$\mathfrak{C}_{\mathcal{S}}$	$\mathfrak{C}_{\mathcal{D}}$	$\mathfrak{C}_{\mathcal{S}}$	$\mathfrak{C}_{\mathcal{D}}$
ADULT	.164	.164	.166	.165	.165	**.163**	.241	.254	.248	**.248**
FASHIONMNIST	.159	.163	.156	**.160**	.163	.167	.096	**.327**	.397	.446
LETTER	.259	.272	.250	**.260**	.258	.270	.829	**.945**	.958	.963
MNIST	.112	.120	.084	**.094**	.119	.127	.092	**.265**	.470	.521
MUSHROOMS	.000	**.000**	.000	**.000**	.000	**.000**	.082	**.122**	.202	.217
NURSERY	.706	**.719**	.706	**.719**	.706	**.719**	.800	**.805**	.793	.806
PENDIGITS	.009	.023	.021	.032	.009	**.022**	.323	**.537**	.871	.879
PHISHING	.042	**.050**	.039	.054	.046	.055	.164	**.222**	.331	.318
SATIMAGE	.132	.184	.149	**.172**	.141	.189	.401	**.763**	.626	.857
SEGMENTATION	.145	**.250**	.189	.373	.174	.389	.619	**.857**	.739	.913
SENSORLESS	.076	.079	.077	.079	.075	**.078**	.899	.910	.622	**.633**
TICTACTOE	.392	.301	.000	.038	.000	**.023**	.388	**.309**	.397	**.309**
YEAST	.679	.666	.487	**.478**	.644	.682	.662	**.720**	.702	**.720**

6.5 Conclusion

We derived new learning algorithms based on Wasserstein PAC-Bayes bounds. Such remarkable empirical results shows the strengths of the optimisation perspective on PAC-Bayes learning. Indeed, we exploited here various paradigms gathered in Figure 1.2: weak statistical assumptions with stronger geometric ones (heavy-tailed lipschitz losses), the use of deterministic predictors, allowed through the introduction of Wasserstein distance. For the online method, we also involve the vision of prior as initialisation point, which is attenuated theoretically as the learning process goes. Finally, note that for our batch methods to work, we involved data-dependent priors which is an ersatz from the information-theoretic view of PAC-Bayes (Figure 1.1) circumventing this limitation, possibly by exploiting the results of Chapter 4 in the batch learning case is a primising future leads to reach not only strong performances, but also non-vacuous theoretical guarantees, which is not the case here.

Conclusion and Perspectives

Conclusion

In this thesis we studied various interplays between PAC-Bayes learning and optimisation. Doing so, we challenged various prerequisites of PAC-Bayes bounds:

- *Strong statistical assumptions.* The main conclusion of Chapter 2 is that, in order to perform PAC-Bayes, no assumption stronger than finite variance is required. In particular, classical bounded or subgaussian assumptions on the loss can be replaced by bounded variance at no additional cost. Note also that in Chapter 5, it is even possible to relax this assumption to boundedness over a compact alongisde lipschitz or gradient lipschitz assumption. This is consistent with optimisation which often involves such geometric assumptions. Furthermore, the supermartingale toolbox allows bounds holding for all dataset size simultaneously, which is consistent with, *e.g.*, online optimisation.

- *The information-theoretic perspective of the prior.* A major contribution has been to formalise perspectives on the prior differing from the Bayesian view paradigm. Indeed, by considering the prior either as an initialisation point or a learning objective, we derived novel PAC-Bayesian bounds aiming to either reduce the impact of the prior (Chapters 3 and 4) when seen as initialisation or highlight it (Chapter 5) when seen as a learning objective. This drove the emergence of Online PAC-Bayes learning and the introduction of gradient norm or convergence guarantees in PAC-Bayes.

- *PAC-Bayes is useful for stochastic predictors only.*. Following the spirit of Amit *et al.* (2022), we developed Wasserstein PAC-Bayes to incorporate deterministic predictors within PAC-Bayes bounds, as such predictors are often involved in optimisation algorithms. To obtain explicit convergence rates with such bounds we exploited duality results from optimal transport in Chapter 5. We also incorporated directly convergence guarantees of the Bures-Wasserstein SGD in a generalisation bound, at the price of an explicit impact of the dimension of the predictor space. It is then hard to tackle the case of deep neural networks, this is why we developed in Chapter 6, another kind of Wasserstein PAC-Bayes bounds, with no explicit convergence rate, but yielding learning algorithm exploitable for neural nets.

Perspectives

This thesis left unanswered many important questions on the interplays of optimisation and generalisation.

- *Can we relax the finite variance assumption to obtain generalisation bounds?* As proven Chapter 2 and Chugg *et al.* (2023) it is possible to extend a large body of generalisation bounds to the case of finite variance. An interesting question is whether such an assumption is relaxed, this would be of interest, for instance, to understand the case of heavy-tailed SGD (Gürbüzbalaban *et al.*, 2021) which may be modelled by Lévy processes, often having infinite variance.
- *Can we further exploit flat minima to understand generalisation?* Chapter 4 proposed the first PAC-Bayes generalisation bounds exploiting flat minima. However the QSB assumption is required to exploit those results. While we saw that such a condition is verified by small networks, whether this condition is verified for deep neural nets remains an open question. Furthermore, the only empirical bound we have implies gradient lipschitz loss, a condition possibly hard to reach for deep nets. Empirical evaluation of those results is then an interesting future lead.
- *Can we reach Wasserstein PAC-Bayes bound as simple and efficient than a KL one?* As shown in Chapter 5 and Chapter 6 we did not obtain Wasserstein PAC-Bayes bounds with explicit convergence rate and without the explicit impact of the dimension as in KL-based PAC-Bayes bounds. An open question is whether it is possible to obtain a Wasserstein PAC-Bayes with both these desirable properties simultaneously.
- *Investigating the links between Online Learning and PAC-Bayes*. Chapter 3 draws a link from PAC-Bayes toward online learning by deriving novel learning algorithms from Online PAC-Bayes bounds. Recently, the elegant work of Lugosi and Neu (2023) has taken the opposite perspective: starting from an online game they retrieve various generalisation bound, including KL-based and Wasserstein-based ones. Given the direct connection between online learning and the supermartingale framework (Wintenberger, 2021), obtaining a unifying framework encompassing Wasserstein and KL-based PAC-Bayes from online learning for heavy-tailed losses is a promising future lead. A first step in this direction has recently been made Viallard *et al.* (2024) but does not involve online learning and holds only for bounded losses.

Investigating those leads, and then reaching a better understanding of the impact of optimisation on generalisation are exciting questions for future work.

Appendix of Chapter 2 A

A.1 Some PAC-Bayesian background

We present below an immediate corollary of Seldin *et al.* (2012a, Thm 2.1) where we upper bounded the cumulative by an empirical quantity (the sum of squared upper bound of the martingale difference sequence).

Theorem A.1.1 (Seldin *et al.*, 2012a, Theorem 2.1)**.** Let $\{C_1, C_2, \dots\}$ be an increasing sequence set in advance, such that $|X_i(S_i, h)| \leq C_i$ for all S_i, h with probability 1. Let $\{P_1, P_2, \dots\}$ be a sequence of data-free prior distributions over $\mathcal{H}$. Let $(\lambda_i)_{i\geq 1}$ be a sequence of positive numbers such that

$$\lambda_m \leq \frac{1}{C_m}.$$

Then with probability $1-\delta$ over $\mathcal{S} = (\mathbf{z}_i)_{i\geq 1}$, for all $m \geq 1$, any posterior Q over $\mathcal{H}$,

$$|M_m\left(Q\right)| \leq \frac{\mathrm{KL}\left(\mathrm{Q}, \mathrm{P}_m\right) + 2\log(m+1) + \log\frac{2}{\delta}}{\lambda_m} + (e-2)\lambda_m V_m(\mathrm{Q}),$$

where $V_m(\mathrm{Q})$ is defined in appendix A.2.1.
Furthermore, if we bound the variance term, we would have:

$$|M_m\left(\mathrm{Q}\right)| \leq \frac{\mathrm{KL}\left(\mathrm{Q}, \mathrm{P}_m\right) + 2\log(m+1) + \log\frac{2}{\delta}}{\lambda_m} + (e-2)\lambda_m \sum_{i=1}^{m} C_i^2.$$

Below, we use the definitions introduced in Section 2.2.3. We study here a particular case of Alquier *et al.*, 2016 for bounded losses which are especially subgaussian thanks to Hoeffding's lemma.

Theorem A.1.2 (Adapted from Alquier *et al.*, 2016, Theorem 4.1)**.** Let $m > 0, \mathcal{S}_m = (\mathbf{z}_1, ..., \mathbf{z}_m)$ be an *i.i.d.* sample from the same law μ. For any data-free prior P, for any loss function ℓ bounded by K, any $\lambda > 0, \delta \in]0;1[$, one has with probability $1-\delta$ for any posterior $Q \in \mathcal{M}_1(\mathcal{H})$

$$\mathbb{E}_{h\sim\mathrm{Q}}[\mathsf{R}(h)] \leq \mathbb{E}_{h\sim\mathrm{Q}}[\hat{\mathsf{R}}_{\mathcal{S}_m}(h)] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda} + \frac{\lambda K^2}{2m}.$$

Theorem A.1.3 (Haddouche *et al.*, 2021, Theorem 3)**.** Let the loss ℓ be $\mathrm{HYPE}(K)$ compliant. For any $\mathrm{P} \in \mathcal{M}(\mathcal{H})$ with no data dependency, for any $\alpha \in \mathbb{R}$ and for any $\delta \in [0,1]$, we have with probability at least $1-\delta$ over size-m samples S, for any Q

$$\mathbb{E}_{h\sim \mathrm{Q}}\left[\mathsf{R}(h)\right] \leq \mathbb{E}_{h\sim \mathrm{Q}}\left[\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log\left(\frac{1}{\delta}\right)}{m^\alpha} + \frac{1}{m^\alpha}\log\left(\mathbb{E}_{h\sim \mathrm{P}}\left[\exp\left(\frac{K(h)^2}{2m^{1-2\alpha}}\right)\right]\right).$$

A.2 Extensions of previous results

Here we gather several corollaries of our main result in order to show how our Theorem 2.2.1 extends the validity of some classical results in the literature. More precisely we show that our result extends (up to numerical factors) the PAC-Bayes Bernstein inequality of Seldin *et al.* (2012a). Then, going back to the bounded case, we generalise a result from Catoni (2007) reformulated in Alquier *et al.* (2016) and we also show how our work strictly improves on the bound of Haddouche *et al.* (2021).

A.2.1 Extension of the PAC-Bayes Bernstein inequality

Here we rename two terms for consistency with Theorem 2.1 of Seldin *et al.* (2012a) (see Theorem A.1.1). For a martingale $M_m(h) = \sum_{i=1}^m X_i(\mathcal{S}_i,h)$, we define, at time m, *empirical cumulative variance* to be $\hat{V}_m(h) = [M]_m(h) = \sum_{i=1}^m X_i(\mathcal{S}_i,h)^2$ and the *cumulative variance* as $V_m(h) = \langle M\rangle_m(h) = \sum_{i=1}^m \mathbb{E}_{i-1}[X_i(\mathcal{S}_i,h)^2]$.
We provide below a corollary containing two bounds: the first one being a straightforward corollary of Th. 2.2.1, the second being valid for bounded martingales and formally close to Theorem 2.1 of Seldin *et al.* (2012a).

Corollary A.2.1. Let $\{\mathrm{P}_1,\mathrm{P}_2,\ldots\}$ be a sequence of data-free prior distributions over $\mathcal{H}$. Let $(\lambda_i)_{i\geq 1}$ be a sequence of positive numbers. Then the following holds with probability $1-\delta$ over $\mathcal{S} = (\mathbf{z}_i)_{i\geq 1}$: for any tuple $(m,\lambda_k,\mathrm{P}_k)$ with $m,k\geq 1$, any posterior Q over $\mathcal{H}$,

$$|M_m\left(\mathrm{Q}\right)| \leq \frac{\mathrm{KL}\left(\mathrm{Q},\mathrm{P}_k\right) + 2\log(k+1) + \log(2/\delta)}{\lambda_k} + \frac{\lambda_k}{2}\left(\hat{V}_m(\mathrm{Q}) + V_m(\mathrm{Q})\right), \tag{A.1}$$

with $\hat{V}_m(\mathrm{Q}) = \mathbb{E}_{h\sim\mathrm{Q}}[\hat{V}_m(h)], V_m(\mathrm{Q}) = \mathbb{E}_{h\sim\mathrm{Q}}[V_m(h)]$. Furthermore, if we assume that for any i, there exists $C_i > 0$ such that $|X_i(\mathcal{S}_i,h)| \leq C_i$ for all $\mathcal{S}_i, h$ then we have the following corollary: with probability $1-\delta$ over S, for any tuple $(m, \lambda_m, \mathrm{P}_m)$ $m \geq 1$, any posterior Q,

$$|M_m(Q)| \leq \frac{\mathrm{KL}(\mathrm{Q}, \mathrm{P}_m) + 2\log(m+1) + \log(2/\delta)}{\lambda_m} + \lambda_m \sum_{i=1}^{m} C_i^2. \tag{A.2}$$

The proof is deferred to appendix A.3. Note that Eq. (A.1) holds uniformly on all tuples $\{(\lambda_k, \mathrm{P}_k, m) \mid k \geq 1, m \geq 1\}$ while Eq. (A.2), as well as Theorem 2.1 of SELDIN *et al.* (2012a) holds uniformly on the tuples $\{(\lambda_m, \mathrm{P}_m, m) \mid m \geq 1\}$ which is a strictly smaller collection. Hence our approach gives guarantees for a larger event with the same confidence level.

Furthermore, Theorem 2.1 of SELDIN *et al.* (2012a) involves the cumulative variance $V_m(\mathrm{Q})$ (and not its empirical counterpart). Because this term is theoretical, we bound it in Th. A.1.1 by $\sum_{i=1}^m C_i^2$ which is supposedly empirical. In this context, Eq. (A.2), recovers nearly exactly the bound of SELDIN *et al.*, 2012a with the transformation of a factor $(e-2)$ into 1. Notice also that Eq. (A.2) stands with no assumption on the range of the λ_i, which is not the case in Th. A.1.1.

Finally, we stress two fundamental differences between our work and the one of SELDIN *et al.* (2012a). First, we replace Markov's inequality by Ville's inequality; second, we exploited the exponential inequality of Lemma 2.1.2 instead of the Bernstein inequality. These allow for results for unbounded martingales for all m simultaneously.

A.2.2 Extensions of learning theory results

A.2.2.1 A general result for bounded losses

We use definitions from Section 2.2.3 and provide a corollary of our main result when the loss is bounded by a positive constant $K > 0$. We assume our data are iid.

Corollary A.2.2. For any data-free prior $P \in \mathcal{M}(\mathcal{H})$, any $\lambda > 0$ the following holds with probability $1-\delta$ over the sample $S = (z_i)_{i\in\mathbb{N}}$, for all $m \in \mathbb{N}/\{0\}$, $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$

$$\left|\mathbb{E}_{h\sim\mathrm{Q}}[\mathsf{R}(h)] - \mathbb{E}_{h\sim\mathrm{Q}}\left[\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right]\right| \leq \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{\lambda m} + \lambda K^2.$$

We also have the local bound: for any $m \geq 1$, with probability $1-\delta$ over S, for all $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$

$$\mathbb{E}_{h\sim\mathrm{Q}}[\mathsf{R}(h)] \leq \mathbb{E}_{h\sim\mathrm{Q}}\left[\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{\lambda} + \frac{\lambda K^2}{m}.$$

The proof is deferred to appendix A.3. Remark that the second bound of Corollary A.2.2 is exactly the Catoni bound stated in Alquier *et al.* (2016) (see Theorem A.1.2 in Appendix A.1) up to a numerical factor of 2.
The first bound is, to our knowledge, the first PAC-Bayesian bound for bounded losses holding uniformly (for a given parameter λ) on the choice of Q, m and thus extends the scope of Catoni's bound which holds for a single m with high probability. Indeed, if we want for instance Theorem A.1.2 to hold for any $i \in \{1..m\}$, we then have to take an union bound on m events which turns the term $\log(1/\delta)$ into $\log(m/\delta)$ (but with the benefit of holding for m parameters $\lambda_1, ..., \lambda_m$). This point is common to the most classical PAC-Bayesian bounds (including McAllester and Catoni's ones (1.3), (1.4)) and impeach us to have a bound uniformly on all $m \in \mathbb{N}/\{0\}$ as $\log(m)$ goes to infinity asymptotically.

A.2.2.2 An extension of Haddouche *et al.* (2021)

We now focus on the work of Haddouche *et al.* (2021) which provides general PAC-Bayesian bounds for unbounded losses. Their theorems hold for iid data and under the so-called *HYPE* (for HYPothesis-dependent rangE) condition. It states that a loss function ℓ is *HYPE*(K) compliant if there exists a function $K : \mathcal{H} \to \mathbb{R}^+$ (supposedly accessible) such that $\forall z \in \mathcal{Z}, \ell(h, \mathbf{z}) \leq K(h)$. We provide Corollary A.2.3 to compare ourselves with their main result (stated in Theorem A.1.3 for convenience).

Corollary A.2.3. For any data-free prior $\mathrm{P} \in \mathcal{M}(\mathcal{H})$, any loss function ℓ being *HYPE*(K) compliant, any $\alpha \in [0,1], m \geq 1$, the following holds with probability $1-\delta$ over the sample $\mathcal{S} = (\mathbf{z}_i)_{i \in \mathbb{N}}$, for all $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$

$$\mathbb{E}_{h\sim \mathrm{Q}}[\mathsf{R}(h)] \leq \mathbb{E}_{h\sim \mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\ell(h,\mathbf{z}_i) + \frac{1}{2m^{1-\alpha}}\ell(h,\mathbf{z}_i)^2\right)\right] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{m^{\alpha}} + \frac{1}{2m^{1-\alpha}}\mathbb{E}_{h\sim \mathrm{Q}}[K^2(h)].$$

Proof. The proof is a straightforward application of Th. 2.2.2 by fixing $m \geq 1$ choosing $\lambda = m^{\alpha-1}$ (thus we localise Theorem 2.2.2 to a single m), and bounding $\mathrm{Quad}(h)$ by $K^2(h)$. ■

The main improvement of our bound over Theorem A.1.3 is that we do not have to assume the convergence of an exponential moment to obtain a non-trivial bound. Indeed, we transformed the (implicit) assumption $\mathbb{E}_{h\sim \mathrm{P}}\left[\exp\left(\frac{K(h)^2}{2m^{1-2\alpha}}\right)\right] < +\infty$ onto

$\mathbb{E}_{h\sim \mathrm{Q}}[K(h)^2] < +\infty$, which is significantly less restrictive. Furthermore, Theorem A.1.3 holds for a single choice of m while ours still holds uniformly over all integers $m > 0$. Cor. A.2.3 also sheds new light on the *HYPE* condition. Indeed, in HADDOUCHE *et al.* (2021), K only intervenes in an exponential moment involving the prior P, while ours considers a second-order moment on K implying the posterior Q. The difference is major as $\mathbb{E}_{h\sim \mathrm{Q}}[K(h)^2]$ can be controlled by a wise choice of posterior. Thus it can be incorporated in our optimisation route, acting now as an optimisation constraint instead of an environment constraint.

A.3 Proofs

A.3.1 Proof of Th. 2.2.2

Proof. Let P a fixed data-free prior, set $(\mathcal{F}_i)_{i\geq 0}$ such that for all i, $\mathbf{z}_i$ is $\mathcal{F}_i$ measurable. We also set for any fixed $h \in \mathcal{H}, M_m(h) := \sum_{i=1}^m \ell(h,\mathbf{z}_i) - \mathsf{R}(h)$. Note that because data are *i.i.d.*, for any fixed h, the sequence $(M_m(h))_m$ is indeed a martingale. We set for any $m \geq 1, h \in \mathcal{H}$

$$[M]_m(h) = \sum_{i=1}^m \left(\ell(h,\mathbf{z}_i) - \mathsf{R}(h)\right)^2$$

and

$$\langle M\rangle_m(h) = \sum_{i=1}^m \mathbb{E}_{i-1}[(\ell(h,\mathbf{z}_i) - \mathsf{R}(h))^2] = \sum_{i=1}^m \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[(\ell(h,\mathbf{z}) - \mathsf{R}(h))^2].$$

The last equality holds because data is assumed iid. Thus, we can apply Th. 2.2.1 to obtain with probability $1-\delta$

$$|M_m(\mathrm{Q})| \leq \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{\lambda} + \frac{\lambda}{2}\left([M]_m(Q)^2 + \langle M\rangle_m(Q)^2\right).$$

Now, we notice that $|M_m(\mathrm{Q})| = m|\mathbb{E}_{h\sim\mathrm{Q}}[\mathsf{R}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)]|$ and that for any m, h, because ℓ is nonnegative

$$\begin{aligned}[M]_m(h) + \langle M\rangle_m(h) &= \sum_{i=1}^m (\ell(h,\mathbf{z}_i) - \mathsf{R}(h))^2 + \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[(\ell(h,\mathbf{z}) - \mathsf{R}(h))^2] \\ &\leq \sum_{i=1}^m \ell(h,\mathbf{z}_i)^2 + \mathsf{R}(h)^2 + \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})^2] - \mathsf{R}(h)^2.\end{aligned}$$

Thus integrating over h gives:

$$[M]_m(Q) + \langle M \rangle_m(Q) \leq \sum_{i=1}^{m} \mathbb{E}_{h\sim \mathrm{Q}}[\ell(h,\mathbf{z}_i)^2] + m\mathbb{E}_{h\sim \mathrm{Q}}[\mathrm{Quad}(h)].$$

Then dividing by m and applying the last inequality gives

$$\mathbb{E}_{h\sim \mathrm{Q}}[\mathsf{R}(h)] \leq \mathbb{E}_{h\sim \mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\ell(h,\mathbf{z}_i) + \frac{\lambda}{2}\ell(h,\mathbf{z}_i)^2\right)\right] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{\lambda m} + \frac{\lambda}{2}\mathbb{E}_{h\sim \mathrm{Q}}[\mathrm{Quad}(h)].$$

This concludes the proof. ■

A.3.2 Proof of Th. 2.3.1

Proof. Let $(\lambda_m)_{i\geq 1}$ be a countable sequence of positive scalars. As precised earlier $M_m(a) := m\left(\hat{\Delta}_m(a) - \Delta(a)\right)$ is a martingale. We then apply Theorem 2.2.1 with the uniform prior ($\forall a, P(a) = \frac{1}{K}$) and $\lambda = \lambda_m$ (depending possibly on m): with probability $1-\delta/2$, for any tuple (m, λ_m) with $m \geq 1$, any posterior Q,

$$|M_m\left(\mathrm{Q}\right)| \leq \frac{\mathrm{KL}\left(\mathrm{Q},\mathrm{P}\right) + 2 + \log(4/\delta)}{\lambda_m} + \frac{\lambda_m}{2}\left(\hat{V}_m(\mathrm{Q}) + V_m(\mathrm{Q})\right).$$

Notice that for any Q, $\mathrm{KL}(\mathrm{Q},\mathrm{P}) \leq \log(K)$ by concavity of the log. We now fix an horizon $M > 0$, we then have in particular, with probability $1-\delta/2$: for any posterior Q,

$$|M_m\left(Q\right)| \leq \frac{\log(K) + 2\log(k+1) + \log(4/\delta)}{\lambda_k} + \frac{\lambda_m}{2}\left(\hat{V}_m(\mathrm{Q}) + V_m(\mathrm{Q})\right).$$

We now have to deal with $V_k(\mathrm{Q}), \hat{V}_k(\mathrm{Q})$ for all $k \leq m$. To do so, we propose the two following lemmas.

> **Lemma A.3.1.** For all $m \geq 1$, $a \in \mathcal{A}$, $V_m(a) \leq \frac{2Cm}{\varepsilon_m}$. Then, we have for any m, Q, $V_m(\mathrm{Q}) \leq \frac{2Cm}{\varepsilon_m}$.

Proof. We have

$$\begin{aligned}
V_t(a) &= \sum_{i=1}^{m} \mathbb{E}\left[\left(\left[R_i^{a^*} - R_i^{a}\right] - \Delta(a)\right)^2 \mid \mathcal{F}_{i-1}\right] \\
&= \sum_{i=1}^{m} \mathbb{E}\left[\left(R_i^{a^*} - R_i^{a}\right)^2 \mid \mathcal{F}_{i-1}\right] - m\Delta(a)^2 \\
&\leq \sum_{i=1}^{m} \mathbb{E}\left[\left(R_i^{a^*} - R_i^{a}\right)^2 \mid \mathcal{F}_{i-1}\right] \\
&= \sum_{i=1}^{m} \mathbb{E}\left[\mathbb{E}_{A_i\sim\pi_i}\mathbb{E}_{R_i}\left[\frac{1}{\pi_i(a^*)^2}R_i(a^*)^2\mathbb{1}(A_i = a^*) + \frac{1}{\pi_i(a)^2}R_i(a)^2\mathbb{1}(A_i = a)\right] \mid \mathcal{F}_{i-1}\right].
\end{aligned}$$

The last line holding because R_i is independent of $\mathcal{F}_{i-1}$, A_i is independent of R_i and π is $\mathcal{F}_{i-1}$ measurable. We now use that for all i, a, $\mathbb{E}_{R_i}[R_i(a)^2] \leq C$

$$\begin{aligned}
&= \sum_{i=1}^{m} \mathbb{E}\left[\mathbb{E}_{A_i\sim\pi_i}\left[\frac{1}{\pi_i(a^*)^2}C\mathbb{1}(A_i = a^*) + \frac{1}{\pi_i(a)^2}C\mathbb{1}(A_i = a)\right] \mid \mathcal{F}_{i-1}\right] \\
&= \sum_{i=1}^{m} C\left(\frac{\pi_i(a)}{\pi_i(a)^2} + \frac{\pi_i\left(a^*\right)}{\pi_i\left(a^*\right)^2}\right) \\
&= \sum_{i=1}^{m} C\left(\frac{1}{\pi_i(a)} + \frac{1}{\pi_i\left(a^*\right)}\right) \\
&\leq \frac{2Cm}{\varepsilon_m}.
\end{aligned}$$

■

Lemma A.3.2. Let $m \geq 1$, with probability $1 - \delta/2$, for any posterior Q, we have

$$\hat{V}_m(\mathrm{Q}) \leq \frac{4CKm}{\varepsilon_m \delta}.$$

Proof. Let Q a distribution over $\mathcal{A}$. Recall that

$$\begin{aligned}\hat{V}_m(\mathrm{Q}) &= \sum_{i=1}^{m} \left(R_i^{a^*} - R_i^a - \left[R\left(a^*\right) - R(a)\right]\right)^2 \\ &= \sum_{a \in \mathcal{A}} Q(a)\hat{V}_m(a).\end{aligned}$$

Notice that for any a, $(S\hat{M}_m^a)_m$ is a nonnegative random variable. We then apply Markov's inequality for any a, with probability $1 - \delta/2K$

$$\hat{V}_m(a) \leq \frac{2K\mathbb{E}[\hat{V}_m(a)]}{\delta}.$$

Noticing that $\mathbb{E}[\hat{V}_m(a)] = \mathbb{E}[V_m(a)]$, we can apply lemma A.3.1 to conclude that

$$\mathbb{E}[\hat{V}_m(a)] \leq \frac{2Cm}{\varepsilon_m}.$$

Finally, taking an union bound on thoser events for all $a \in \mathcal{A}$ gives us, with probability $1 - \delta/2$, for any posterior Q

$$\begin{aligned}V_m(\mathrm{Q}) &\leq \sum_{a \in \mathcal{A}} Q(a)\hat{V}_m(a) \\ &\leq \sum_{a \in \mathcal{A}} Q(a)\frac{4CKm}{\varepsilon_m \delta} \\ &= \frac{4CKm}{\varepsilon_m \delta}.\end{aligned}$$

This concludes the proof. ■

To conclude, we apply lemmas A.3.1 and A.3.2 to get that with probability $1 - \delta$, for any posterior Q

$$|M_m\left(\mathrm{Q}\right)| \leq \frac{\mathrm{KL}\left(\mathrm{Q}, \mathrm{P}\right) + \log(4/\delta)}{\lambda_m} + \frac{Cm\lambda_m}{\varepsilon_m}\left(1 + \frac{2K}{\delta}\right).$$

Dividing by m and taking

$$\lambda_m = \sqrt{\frac{\left(\log(K) + \log(4/\delta)\right)\varepsilon_m}{Cm\left(1 + \frac{2K}{\delta}\right)}}$$

concludes the proof.

■

A.3.3 Proof of Cor. A.2.1

Proof. Fix $\delta > 0$. For any pair $(\lambda_k, P_k), k \geq 1$, we apply Theorem 2.2.1 with

$$\delta_k := \frac{\delta}{k(k+1)} \geq \frac{\delta}{(k+1)^2}.$$

Notice that we have $\sum_{k=1}^{+\infty} \delta_k = \delta$. We then have with probability $1 - \delta_k$ over S, for any $m \geq 1$, any posterior Q,

$$|M_m(\mathrm{Q})| \leq \frac{\mathrm{KL}(\mathrm{Q}, \mathrm{P}_k) + 2\log(k+1) + \log(2/\delta)}{\lambda_k} + \frac{\lambda_k}{2}\left(\hat{V}_m(\mathrm{Q}) + V_m(\mathrm{Q})\right).$$

Taking an union bound on all those event, gives the final result, valid with probability $1 - \delta$ over the sample S, for any any tuple (m, λ_k, P_k) with $m, k \geq 1$, any posterior Q over $\mathcal{H}$. This gives Equation (A.1).
To obtain Eq. (A.2), we restrict the range of Eq. (A.1) to the tuples $(m, \lambda_m, P_m), m \geq 1$ (the restricted set of tuples where $k = m$) and we bound both $\hat{V}_m(\mathrm{Q}), V_m(\mathrm{Q})$ by $\sum_{i=1}^m C_i^2$ to conclude. ■

A.3.4 Proof of Cor. A.2.2

Proof. For the first bound we start from the intermediary result Eq. (2.3) of Th. 2.2.1. Using the same marrtingale as in Th. 2.2.2 gives, for any $\eta \in \mathbb{R}$, holding with probability $1 - \delta$ for any $m > 0, Q \in \mathcal{M}(\mathcal{H})$

$$\eta\left(\sum_{i=1}^m \mathbb{E}_{h\sim\mathrm{Q}}[\ell(h, \mathbf{z}_i)] - m\mathbb{E}_{h\sim\mathrm{Q}}[\mathsf{R}(h)]\right)$$
$$\leq \mathrm{KL}(\mathrm{Q}, \mathrm{P}) + \log(1/\delta) + \frac{\eta^2}{2}\sum_{i=1}^m \mathbb{E}_{h\sim\mathrm{Q}}[\Delta[M]_i(h) + \Delta\langle M\rangle_i(h)].$$

Taking $\eta = \pm\lambda$ with $\lambda > 0$ gives

$$\lambda m \left|\mathbb{E}_{h\sim\mathrm{Q}}[\mathsf{R}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)]\right| \leq \mathrm{KL}(\mathrm{Q}, \mathrm{P}) + \log(1/\delta) \tag{A.3}$$
$$+ \frac{\lambda^2}{2}\sum_{i=1}^m \mathbb{E}_{h\sim\mathrm{Q}}[\Delta[M]_i(h) + \Delta\langle M\rangle_i(h)]. \tag{A.4}$$

Finally, divide by λm and bound $\Delta[M]_i(h) + \Delta\langle M\rangle_i(h)$ by $2K^2$ to conclude.
For the second bound, we start from Equation (A.3) again and for a fixed m, we now apply our result with $\lambda' = \lambda/m$. We then have for any m, with probability

$1-\delta$, for any Q

$$\lambda\left|\mathbb{E}_{h\sim\mathrm{Q}}[\mathsf{R}(h)-\hat{\mathsf{R}}_{\mathcal{S}_m}(h)]\right| \le \mathrm{KL}(\mathrm{Q},\mathrm{P})+\log(1/\delta)+\frac{\lambda^2}{2m^2}\sum_{i=1}^{m}\mathbb{E}_{h\sim\mathrm{Q}}[\Delta[M]_i(h)+\Delta\langle M\rangle_i(h)].$$

Finally, dividing by λ, bounding $\Delta[M]_i(h)+\Delta\langle M\rangle_i(h)$ by $2K^2$ and rearranging the terms concludes the proof. ■

Appendix of Chapter 3 B

B.1 Background

B.1.1 Reminder on Online Gradient Descent

For the sake of completeness we re-introduce the projected Online Gradient Descent (OGD) on a convex set $\mathcal{K}$. This is a first example of online learning philosophy. It may be the algorithm that applies to the most general setting of online convex optimization. This algorithm, which is based on standard gradient descent from offline optimization, was introduced in its online form by ZINKEVICH, 2003. In each iteration, the algorithm takes a step from the previous point in the direction of the gradient of the previous cost. This step may result in a point outside of the underlying convex set. In such cases, the algorithm projects the point back to the convex set, i.e. finds its closest point in the convex set. We precise this algorithm works with the assumptions of a convex set $\mathcal{K}$ bounded in diameter by D and of bounded gradients (by a certain G). We also assume here to have a dataset $\mathcal{S}_T = (\mathbf{z}_t)_{t=1..T}$ and to be coherent with the online learning philosophy, we assume that for each $t > 0$, we possess a loss function ℓ_t depending on the points $(\mathbf{z}_1, ..., \mathbf{z}_t)$. We present OGD in Algo. 3

Algorithm 3: Projected OGD onto a convex $\mathcal{K}$ with fixed step η.

```
Parameters : Epoch T, step-size ($(\eta)$)
Initialisation: Convex set $\mathcal{K}$, Initial point $\theta_0 \in \mathcal{K}$, T, step sizes $(\eta_t)_t$
for each iteration $t$ in $1..T$ do
    Compute $f'(\theta_n)$
    Play (observe) $\theta_t$ and compute the cost $f_t(\theta_t)$ Update and project
        $\zeta_t = \theta_{t-1} - \eta \nabla \ell_t(\theta_{t-1})$
        $\theta_t = \Pi_{\mathcal{K}}(\zeta_t)$
end
Return $\theta_T$
```

One now defines the notion of regret which is the classical quantity to evaluate the performance of an online algorithm.

Definition B.1.1. One defines the *regret* of a decision sequence (θ_t) at time T w.r.t. a point θ as:

$$Regret_T(\theta) := \sum_{t=1}^{T} \ell_t(\theta_t) - \sum_{t=1}^{T} \ell_t(\theta)$$

Now we state a regret bound which can be found in Shalev-Shwartz, 2012, Eq 2.5 although we slightly modified the result, which uses additional hypotheses from Hazan, 2016.

Proposition B.1.1. Assume that $\mathcal{K}$ has a fixed diameter D and that the gradients of any point is bounded by G. Then for any $\theta \in \mathcal{K}$, the regret of projected OGD with fixed step η satisfies:

$$Regret_T(\theta) \leq \frac{D^2}{2\eta} + \eta T G^2$$

B.2 Discussion about Th. 3.2.1

B.2.1 Comparison with classical PAC-Bayes

The goal of this section is to show how good Th. 3.2.1 compared to a naive approach which consists in applying classical PAC-Bayes results sequentially. The interest of this section is twofold:

- First, presenting a classical PAC-Bayes result extracted and adapted from Alquier *et al.*, 2016 which is formally close to what we propose.
- Second, showing that a naive (yet natural) approach to obtain online PAC-Bayes bound leads to a deteriorated bound.

We first state our PAC-Bayes bound of interest.

Theorem B.2.1 (Adapted from Alquier *et al.*, 2016, Thm 4.1)**.** Let $\mathcal{S}_m = (\mathbf{z}_1, ..., \mathbf{z}_m)$ be an *i.i.d.* sample from the same law $\mathcal{D}$. For any data-free prior P, for any loss function ℓ bounded by K, any $\lambda > 0, \delta \in (0,1)$, one has with probability $1-\delta$ for any posterior $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:

$$\mathbb{E}_{h\sim\mathrm{Q}}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})] \leq \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{h\sim\mathrm{Q}}[\ell(h,\mathbf{z}_i)] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda} + \frac{\lambda K^2}{2m}$$

Remark B.2.1. Two remarks about this result:

- Th. B.2.1 is a particular case of the original theorem from Alquier *et al.*, 2016 as we take the case of a bounded loss which implies the subgaussianity of the random variables $\ell(.,z_i)$ and then allows us to recover the factor $\frac{\lambda K^2}{m}$
- This theorem is derived from Catoni, 2007 and constitutes a good basis to compare ourselves with as it similar formally similar.

Naive approach A naive way to obtain OPB bounds is to apply m times Th. B.2.1 (one per data) on batches of size 1 and then summing up the associated bounds. Thus one has the benefits of classical PAC-Bayes bound without having no more the need of data-free priors nor the iid assumption. The associated result is stated below:

Theorem B.2.2. For any distributions $\mathcal{D}_1,...,\mathcal{D}_m$ over $\mathcal{Z}$ (such that $\mathbf{z}_i \sim \mathcal{D}_i$), any $\lambda > 0$ and any online predictive sequence (used as priors) $(\mathrm{P}_i)_{i=1\cdots m}$, the following holds with probability $1-\delta$ over the sample $\mathcal{S}_m \sim \mathcal{D}^m$ for any posterior sequence (Q_i) :

$$\sum_{i=1}^m \mathbb{E}_{h_i\sim \mathrm{Q}_i}\left[\mathbb{E}_{\mathbf{z}_i\sim\mathcal{D}_i}[\ell(h_i,\mathbf{z}_i)]\right] \leq \sum_{i=1}^m \mathbb{E}_{h_i\sim \mathrm{Q}_i}\left[\ell(h_i,\mathbf{z}_i)\right] + \frac{\mathrm{KL}(\mathrm{Q}_i\|\mathrm{P}_i)}{\lambda} + \frac{\lambda m K^2}{2} + \frac{m\log(m/\delta)}{\lambda}.$$

Recall that here again we assimilate the stochastic kernels $\mathrm{Q}_i, \mathrm{P}_i$ to the data-dependent distributions $\mathrm{Q}_i(\mathcal{S}_m,.), \mathrm{P}_i(\mathcal{S}_m,.)$

Proof. First of all, for any i, we apply Th. B.2.1 m to the batch $\{z_i\}$. This allows us to consider P_i as a prior as it does not depend on the current data. We then have, taking $\delta' = \delta/m$, for any $i \in \{1..m\}$ with probability $1-\delta/m$:

$$\mathbb{E}_{h_i\sim \mathrm{Q}_i}\left[\mathbb{E}_{\mathbf{z}_i\sim\mathcal{D}_i}[\ell(h_i,\mathbf{z}_i)]\right] \leq \mathbb{E}_{h_i\sim \mathrm{Q}_i}\left[\ell(h_i,\mathbf{z}_i)\right] + \frac{\mathrm{KL}(\mathrm{Q}_i\|\mathrm{P}_i)}{\lambda} + \frac{\lambda K^2}{2} + \frac{\log(m/\delta)}{\lambda}.$$

Then, taking an union bound on those m events ensure us that with probability $1-\delta$, for any $i \in \{1..m\}$:

$$\mathbb{E}_{h_i\sim \mathrm{Q}_i}\left[\mathbb{E}_{\mathbf{z}_i\sim\mathcal{D}_i}[\ell(h_i,\mathbf{z}_i)]\right] \leq \mathbb{E}_{h_i\sim \mathrm{Q}_i}\left[\ell(h_i,\mathbf{z}_i)\right] + \frac{\mathrm{KL}(\mathrm{Q}_i\|\mathrm{P}_i)}{\lambda} + \frac{\lambda K^2}{2} + \frac{\log(m/\delta)}{\lambda}.$$

Finally, summing those m inequalities ensure us the final result with probability $1-\delta$.

■

Comparison between Th. 3.2.1 and Th. B.2.2 Three points are noticeable between those two theorems:

- First of all, the main issue with Th. B.2.2 is that has a strongly deteriorated rate of $O\left(\frac{m\log(m/\delta)}{\lambda}\right)$ instead of the rate in $O\left(\frac{\log(1/\delta)}{\lambda}\right)$ proposed in Th. 3.2.1. More precisely, the problem is that we do not have a sublinear bound: one cannot ensure any learning through time. This point justifies the need of the heavy machinery exploited in Th. 3.2.1 proof as it allows a tighter convergence rate.
- The second point point lies in the controlled quantity on the left hand-side of the bound. Th. B.2.2 controls $A := \sum_{i=1}^m \mathbb{E}_{h_i\sim \mathrm{Q}_i}\left[\mathbb{E}_{\mathbf{z}_i\sim\mathcal{D}_i}[\ell(h_i,\mathbf{z}_i)]\right]$ instead of $B := \sum_{i=1}^m \mathbb{E}_{h_i\sim \mathrm{Q}_i}\left[\mathbb{E}[\ell(h_i,\mathbf{z}_i)\mid\mathcal{F}_{i-1}]\right]$.

  A is a less dynamic quantity than B in the sense that it does not imply any evolution through time, it just considers global expectations. Doing so, A does not take into account that at each time step we have acces to all te past data to predict the future, this may explain the deteriorated convergence rate. Thus B, which appears to be a suitable quantity to control to perform online PAC-Bayes (see appendix B.2.2 for additional explanations)
- Finally, an interesting point is that in Th. B.2.2 the bound, while looser, holds unformly for any posterior sequence contrary to Th. 3.2.1 which holds only for a specific posterior sequence. This point will have a consequence for optimisation. We will come back later on this in appendix B.2.3.

B.2.2 A deeper analysis of Th. 3.2.1

This section includes discussion about our proof technique and why all the assumptions made are necessary. We also propose a short discussion about the benefits and limitations of an online PAC-Bayesian framework as well as a deeper reflexion about the new term our bound introduce.

Why do we need an online predictive sequence as priors? This condition is fully exploited when dealing with the exponential moment ξ_m in the proof (see lemma B.4.1 proof). Indeed, the fact of having P_i being $\mathcal{F}_{i-1}$-measurable is essential to apply conditional Fubini (lemma B.4.2). Note that the condition $\forall i, P_{i-1} \gg \mathrm{P}_i$ is not necessary as the weaker condition $\forall i, P_1 \gg \mathrm{P}_i$ would suffice here. However, note that when we particularise our theorem, for instance if we choose in Cor. 3.3.1 $\mathrm{P}_i = \hat{\mathrm{Q}}_i$, one recovers the condition $\hat{\mathrm{Q}}_i \gg \hat{\mathrm{Q}}_{i+1}$ to have finite KL divergences. Hence the interest of taking directly an online predictive sequence.

About the boundedness assumption The only moment where we invoke the boundedness assumption is in lemma B.4.1's proof where we apply the conditionnal Hoeffding lemma. This lemma actually translates that the sequence of r.v. $(\ell(., z_i)_{i=1..m}$ is *conditionally subgaussian* wrt the past i.e for any i, $h_i \in \mathcal{H}; \lambda \in \mathbb{R}$:

$$\mathbb{E}[\exp(\lambda\tilde{\ell}_i(h_i, \mathbf{z}_i)) \mid \mathcal{F}_{i-1}] \leq \exp\left(\frac{\lambda^2 K^2}{2}\right)$$

where $\tilde{\ell}_i(h_i, \mathbf{z}_i) = \mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] - \ell(h_i, \mathbf{z}_i)$.
This condition is the one truly involved in our heavy machinery. However, we chose to restrict ourselves to the stronger assumption of bounded loss function for the sake of clarity. However, an interesting open direction is to find whether there exists concrete classes of unbounded losses which may satisfy either conditional subgaussianity or others conditions (such as conditional Bernstein condition for instance).

Reflections about the left hand side of Th. 3.2.1. We study in this paragraph the following term

$$B := \sum_{i=1}^{m} \mathbb{E}_{h_i \sim \mathrm{Q}_i}\left[\mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}]\right]$$

has naturally arisen in our work as the right term to compare our empirical loss with to perform the conditional Hoeffding lemma. Taking a broader look, we now interpret this term as the right quantity to control if one wants to perform online PAC-Bayes learning. Indeed this term is a 'best of both world' quantity bridging PAC-Bayes and online learning:

- From the PAC-Bayes point of view one keeps the control on average (cf the conditional expectation in B) on a novel data drawn at each time step. This point is crucial in the PAC-Bayes literature as our posteriors are designed to generalise well to unseen data.
- From the Online Learning point of view, one keeps the control of a sequence of points generated from an online algorithm. Because an online learning algorithm generate a prediction for future points while having access to past data, the conditional expectation in B translates this.

Finally this conditional expectation appears to be a good tradeoff between the classical expectation on data appearing in the PAC-Bayes literature (see e.g. Th. B.2.1) and the local control that we have in online learning by only dealing with the performance of a sequence of points generated from a learning algorithm (see e.g. proposition B.1.1)

About the interest of an Online PAC-Bayesian framework The main shift our work does with classical online learning literature is that it does not consider the celebrated regret but instead focuses on B which is a cumulative expected loss conditionned to the past. This shift does not invalidate our work but put some relief to hte guarantees Online PAC-Bayes learning can provide that Online Learning cannot and reversely.

- Online PAC-Bayes ensure a good potential for generalisation as it deals with the control of conditional expectation. This can be useful if one wants to deal with a periodic process for instance.
- Online Learning through the regret compares the studied sequence of predictors (typically generated from an online learning algorithm) and tries to compare it to the best fixed strategy (static regret) or the best dynamic one (dynamic regret). In this way, OL algorithms want to ensure that their predictions are closed from the optimal solution. This point is not guaranteed by our online PAC-Bayesian study.
- However the limitations of online learning can arise if the studied problem has a huge variance (for instance micro-transactions in finance). In this case those algorithms can follow an unpredictable optimisation route while PAC-Bayes still ensure a good performance on average (knowing the past) in this case.
- Finally, we want to emphasize that PAC-Bayesian learning circumvent a problem of *memoryless learning* which appears in classical OL algorithms. For instance, the OGD algorthm (see appendix B.1.1) uses once a data and do not memorise it for further use. This problem does not happen in Online PAC-Bayes learning. Indeed, we take the example of the procedure Eq. (3.3) which generates Gibbs posterior which keep in mind the influence of past data.

B.2.3 Th. 3.2.1 and optimisation

In this section we discuss about the way Thm 2.2 can be thought in the framework of an optimisation process as we did in sections 3.3 and 3.4.

A significant change compared to classical PAC-Bayes Th. 3.2.1 holds 'for any posterior sequence (Q_i) the following holds with probability $1-\delta$ over the sample $\mathcal{S}_m \sim \mathcal{D}^m$ ' while most classical PAC-Bayesian results such that Th. B.2.1 holds 'with probability $1-\delta$ over the sample $\mathcal{S}_m \sim \mathcal{D}^m$ for any posterior Q'. This change is significant as our theorem does not control simultaneaously all possible sequences of posteriors but only holds for one. Thus, Th. 3.2.1 has to be seen as a local or pointwise theorem and not as a global one. In classical PAC-Bayes, this local behavior is a brake on the optimisation process. But as we develop below, it is not the case in our online framework.

Th. 3.2.1 is compatible with online optimisation We first recall that classically, an online algorithm like OGD (see appendix B.1.1) performs one optimisation step per arriving data. Thus, at time m, such algorithm will perform m optimisation steps and generate m predictors. Similarly the OPB algorithm of Eq. (3.1) generates m distribution in m time steps.
We insist on the fact that, Th. 3.2.1 **and all its corollaries throughout our paper are valid for a sequence of m posteriors and not only a single one.** A key point is that whatever the number m of data, our theoretical guarantee wil still be valid for m posterior distributions with the approximation term $\log(1/\delta)$ (and not $\log(m/\delta)$ as an union bound would provide for a classical PAC-Bayes theorem).
For this reason, given an online PAC-Bayes algorithm, Th. 3.2.1 is suited for optimisation. Indeed, having a bound valid for a sequence of posteriors ensures guarantees for a single run of our OPB algorithm. This point is crucial to bridge a link with online learning as regret bounds (e.g. proposition B.1.1) also provide guarantees for a single sequence of predictors. In online learning however, those guarantees are mainly deterministic (because based on convex optimisation properties) but not totally: the recent work of WINTENBERGER, 2021 proposed PAC regret bounds for its general Stochastic Online Convex Optimisation framework.
An interesting open challenge is to overcome the pointwise behavior of our theorem, for that, we need to rethought RIVASPLATA *et al.*, 2020, Thm 2.1 as this basis is pointwise itself. Given we consider a sequence of data-dependent priors one cannot apply the classical change of measure inequality to ensure guarantees holding uniformly on posterior sequences.

A crucial point: having an explicit OPB/OPBD algorithm In our previous paragraph we said that our bound were suitable for optimisation given an OPB/OPBD algorithm. We now provide some precision about this point. All the procedures provided in the paper (i.e. Eq. (3.1), Algo. 1) take into account an update phase implying an argmin. Luckily for our procedures, this argmin is explicit:

- For the OPB algorithm of Eq. (3.1), the argmin is solved thanks to the variational formulation of the Gibbs posterior
- For OPBD algorithms, given the explicit choices of Ψ given in Cor. 3.4.1, argmin becomes explicit when one has a derivable loss function.

In both cases, this explicit argmin ensure our procedure of interest generates explictly a single posterior per time step: we have a well-defined sequence of m posteriors at time m. Doing so the guarantees of Th. 3.2.1 holds for this sequence.

B.3 A reminder on PAC-Bayesian disintegrated bounds

We present two PAC-Bayesian disintegrated bounds valid with data-dependent priors (i.e. any stochastic kernels).

- The first one is Th. 1) i) from Rivasplata *et al.*, 2020 which provides a disintegrated version of Th. 3.2.1.
- The second one is Thm 2. from Viallard *et al.*, 2023a which involves Rényi divergence instead of the classical KL. Note that this bound has originally been stated for data-indepedent prior, which is why we revisit the proof to adapt it to the stochastic kernel framework.

Proposition B.3.1 (Th 1) i) Rivasplata *et al.*, 2020)**.** Let $P \in \mathcal{M}(\mathcal{H})$, $\mathrm{Q}^0 \in \mathtt{Stoch}(\mathcal{Z}^m, \mathcal{F})$. Let $f : \mathcal{S}_m \times \mathcal{H} \to \mathbb{R}$ be any measurable function. Then for any $Q \in \mathtt{Stoch}(\mathcal{Z}^m, \mathcal{F})$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of $S \sim P$ and $h \sim \mathrm{Q}_{\mathcal{S}_m}$, we have:

$$f(\mathcal{S}_m, h) \leq \log\left(\frac{d\mathrm{Q}_{\mathcal{S}_m}}{d\mathrm{Q}^0_{\mathcal{S}_m}}(h)\right) + \log(\xi_m/\delta).$$

where $\xi_m := \int_{\mathcal{S}_m} \int_{\mathcal{H}} e^{f(s,h)} \mathrm{Q}^0_{\mathcal{S}_m}(dh) P(ds)$ and $\frac{d\mathrm{Q}_{\mathcal{S}_m}}{d\mathrm{Q}^0_{\mathcal{S}_m}}$ is the Radon Nykodym derivative of $\mathrm{Q}_{\mathcal{S}_m}$ w.r.t. $\mathrm{Q}^0_{\mathcal{S}_m}$.

Proposition B.3.2 (Adapted from Th. 2 of Viallard *et al.*, 2023a)**.** Let $\mu \in \mathcal{M}(\mathcal{Z}^m)$, $\mathrm{Q}^0 \in \mathtt{Stoch}(\mathcal{Z}^m, \mathcal{F})$. Let $\alpha > 1$ and $f : \mathcal{S}_m \times \mathcal{H} \to \mathbb{R}^+$ be any measurable function.
Then for any $\mathrm{Q} \in \mathtt{Stoch}(\mathcal{Z}^m, \mathcal{F})$ such that for any $\mathcal{S}_m \in \mathcal{Z}^m, \mathrm{Q}_{\mathcal{S}_m} \ll \mathrm{Q}^0_{\mathcal{S}_m}$, $\mathrm{Q}^0_{\mathcal{S}_m} \gg \mathrm{Q}_{\mathcal{S}_m}$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of $\mathcal{S}_m \sim \mathcal{D}_m$ and $h \sim \mathrm{Q}_{\mathcal{S}_m}$, we have:

$$\begin{aligned}\frac{\alpha}{\alpha-1}\log(f(\mathcal{S}_m,h)) \leq \frac{2\alpha-1}{\alpha-1}\log\frac{2}{\delta} + D_\alpha\left(\mathrm{Q}_{\mathcal{S}_m}\|\mathrm{Q}^0_{\mathcal{S}_m}\right)& \\ + \log\left(\mathop{\mathbb{E}}_{\mathcal{S}'_m\sim\mathcal{D}_m}\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}'_m}} f(\mathcal{S}'_m,h')^{\frac{\alpha}{\alpha-1}}\right)&\end{aligned}$$

where $D_\alpha(\mathrm{Q},\mathrm{P}) = \frac{1}{\alpha-1}\log\left(\mathbb{E}\left[\mathbb{E}_{h\sim\mathrm{P}}\left(\frac{d\mathrm{Q}}{d\mathrm{P}}(h)\right)^\alpha\right]\right)$ is the Rényi diverence of order α.

Note that Viallard et al. original bound only stand for data-free priors and i.i.d data. However it appears their proof works with any stochastic kernel as prior and any distribution over the dataset. We propose below an adaptation of their proof below to fit with those more general assumptions.

B.3.1 Proof of proposition B.3.2

Proof. For any sample $\mathcal{S}_m$ and any stochastic kernel Q, note that $f(\mathcal{S}_m, h)$ is a non-negative random variable. Hence, from Markov's inequality we have

$$\begin{aligned}
&\underset{h\sim \mathrm{Q}_{\mathcal{S}_m}}{\mathbb{P}}\left[f(\mathcal{S}_m,h) \leq \frac{2}{\delta}\underset{h'\sim \mathrm{Q}_{\mathcal{S}_m}}{\mathbb{E}} f\left(\mathcal{S}_m,h'\right)\right] \geq 1-\frac{\delta}{2}\\
&\qquad\iff \underset{h\sim \mathrm{Q}_{\mathcal{S}_m}}{\mathbb{E}} \mathbb{1}\left[f(\mathcal{S}_m,h) \leq \frac{2}{\delta}\underset{h'\sim \mathrm{Q}_{\mathcal{S}_m}}{\mathbb{E}} f\left(S,h'\right)\right] \geq 1-\frac{\delta}{2}
\end{aligned}$$

Taking the expectation over $\mathcal{S}_m \sim \mathcal{D}_m$ to both sides of the inequality gives

$$\begin{aligned}
&\underset{\mathcal{S}_m\sim\mathcal{D}_m}{\mathbb{E}}\ \underset{h\sim \mathrm{Q}_{\mathcal{S}_m}}{\mathbb{E}} \mathbb{1}\left[f(\mathcal{S}_m,h) \leq \frac{2}{\delta}\underset{h'\sim \mathrm{Q}_{\mathcal{S}_m}}{\mathbb{E}} f(\mathcal{S}_m,h')\right] \geq 1-\frac{\delta}{2}\\
&\qquad\iff \underset{\mathcal{S}_m\sim\mathcal{D}_m, h\sim \mathrm{Q}_{\mathcal{S}_m}}{\mathbb{P}}\left[f(\mathcal{S}_m,h) \leq \frac{2}{\delta}\underset{h'\sim \mathrm{Q}_{\mathcal{S}_m}}{\mathbb{E}} f(\mathcal{S}_m,h')\right] \geq 1-\frac{\delta}{2}.
\end{aligned}$$

Taking the logarithm to both sides of the equality and multiplying by $\frac{\alpha}{\alpha-1} > 0$, we obtain

$$\underset{\mathcal{S}_m\sim\mathcal{D}_m, h\sim \mathrm{Q}_{\mathcal{S}_m}}{\mathbb{P}}\left[\frac{\alpha}{\alpha-1}\log(f(\mathcal{S}_m,h)) \leq \frac{\alpha}{\alpha-1}\log\left(\frac{2}{\delta}\underset{h'\sim \mathrm{Q}_{\mathcal{S}_m}}{\mathbb{E}} f(\mathcal{S}_m,h')\right)\right] \geq 1-\frac{\delta}{2}.$$

We develop the right side of the inequality in the indicator function and make the expectation of the hypothesis over $\mathrm{Q}^0_{\mathcal{S}_m}$ our "prior" stochadtic kernel appears. Indeed, because for any $S \in \mathcal{S}_m, \mathrm{Q}_{\mathcal{S}_m} \gg \mathrm{Q}^0_{\mathcal{S}_m}$ and $\mathrm{Q}^0_{\mathcal{S}_m} \ll \mathrm{Q}_{\mathcal{S}_m}$ one can write properly $\frac{d\mathrm{Q}_{\mathcal{S}_m}}{d\mathrm{Q}^0_{\mathcal{S}_m}}$ and $\frac{d\mathrm{Q}^0_{\mathcal{S}_m}}{d\mathrm{Q}_{\mathcal{S}_m}} = \left(\frac{d\mathrm{Q}_{\mathcal{S}_m}}{d\mathrm{Q}^0_{\mathcal{S}_m}}\right)^{-1}$ the Radon-Nykodym derivatives. Thus we have

$$\begin{aligned}&\frac{\alpha}{\alpha-1}\log\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}_{\mathcal{S}_m}} f(\mathcal{S}_m,h')\right)\\ &\qquad=\frac{\alpha}{\alpha-1}\log\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}_{\mathcal{S}_m}}\frac{d\mathrm{Q}_{\mathcal{S}_m}}{d\mathrm{Q}^0_{\mathcal{S}_m}}(h')\frac{d\mathrm{Q}^0_{\mathcal{S}_m}}{d\mathrm{Q}_{\mathcal{S}_m}}(h')f(\mathcal{S}_m,h')\right)\\ &\qquad\qquad=\frac{\alpha}{\alpha-1}\log\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}}\frac{d\mathrm{Q}_{\mathcal{S}_m}}{d\mathrm{Q}^0_{\mathcal{S}_m}}(h')f(\mathcal{S}_m,h')\right).\end{aligned}$$

Remark that $\frac{1}{r}+\frac{1}{s}=1$ with $r=\alpha$ and $s=\frac{\alpha}{\alpha-1}$. Hence, we can apply Hölder's inequality:

$$\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}}\frac{d\mathrm{Q}_{\mathcal{S}_m}}{d\mathrm{Q}^0_{\mathcal{S}_m}}(h')f(\mathcal{S}_m,h')\le\left[\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}}\left(\frac{d\mathrm{Q}_{\mathcal{S}_m}}{d\mathrm{Q}^0_{\mathcal{S}_m}}(h')\right)^{\alpha}\right]^{\frac{1}{\alpha}}\left[\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m,h')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}.$$

Then, by taking the logarithm; adding $\log\left(\frac{2}{\delta}\right)$ and multiplying by $\frac{\alpha}{\alpha-1}>0$ to both sides of the inequality, we obtain

$$\begin{aligned}&\frac{\alpha}{\alpha-1}\log\left(\frac{2}{\delta}\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}}\frac{d\mathrm{Q}_{\mathcal{S}_m}}{d\mathrm{Q}^0_{\mathcal{S}_m}}(h')f(\mathcal{S}_m,h')\right)\\ &\quad\le\frac{\alpha}{\alpha-1}\log\left(\frac{2}{\delta}\left[\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}}\left(\frac{d\mathrm{Q}_{\mathcal{S}_m}}{d\mathrm{Q}^0_{\mathcal{S}_m}}(h')\right)^{\alpha}\right]^{\frac{1}{\alpha}}\left[\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m,h')^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\\ &=\frac{1}{\alpha-1}\log\left(\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}}\left[\frac{d\mathrm{Q}_{\mathcal{S}_m}}{d\mathrm{Q}^0_{\mathcal{S}_m}}(h')\right]^{\alpha}\right)+\frac{\alpha}{\alpha-1}\log\frac{2}{\delta}+\log\left(\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m,h')^{\frac{\alpha}{\alpha-1}}\right)\\ &\qquad=D_\alpha\left(\mathrm{Q}_{\mathcal{S}_m}\|\mathrm{Q}^0_{\mathcal{S}_m}\right)+\frac{\alpha}{\alpha-1}\log\frac{2}{\delta}+\log\left(\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m,h')^{\frac{\alpha}{\alpha-1}}\right)\end{aligned}$$

From this inequality, we can deduce that

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathcal{S}_m\sim\mathcal{D}_m,h\sim\mathrm{Q}_{\mathcal{S}_m}}\Bigg[&\frac{\alpha}{\alpha-1}\log(f(\mathcal{S}_m,h))\le D_\alpha\left(\mathrm{Q}_{\mathcal{S}_m}\|\mathrm{Q}^0_{\mathcal{S}_m}\right)\\ &+\frac{\alpha}{\alpha-1}\log\frac{2}{\delta}+\log\left(\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m,h')^{\frac{\alpha}{\alpha-1}}\right)\Bigg]\\ &\ge 1-\frac{\delta}{2}.\end{aligned}\tag{B.1}$$

Note that $\mathbb{E}_{h'\sim \mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m, h')^{\frac{\alpha}{\alpha-1}}$ is a non-negative random variable, hence, we apply Markov's inequality to have

$$\mathop{\mathbb{P}}_{\mathcal{S}_m\sim\mathcal{D}_m}\left[\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m,h')^{\frac{\alpha}{\alpha-1}} \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'_m\sim\mathcal{D}_m}\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}'_m,h')^{\frac{\alpha}{\alpha-1}}\right] \ge 1-\frac{\delta}{2}.$$

Since the inequality does not depend on the random variable $h\sim\mathrm{Q}_{\mathcal{S}_m}$, we have

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathcal{S}_m\sim\mathcal{D}_m}\left[\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m,h')^{\frac{\alpha}{\alpha-1}} \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'_m\sim\mathcal{D}_m}\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}'_m,h')^{\frac{\alpha}{\alpha-1}}\right]\\
&\quad=\mathop{\mathbb{E}}_{\mathcal{S}_m\sim\mathcal{D}_m}\mathbb{1}\left[\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m,h')^{\frac{\alpha}{\alpha-1}} \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'_m\sim\mathcal{D}_m}\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}'_m,h')^{\frac{\alpha}{\alpha-1}}\right]\\
&=\mathop{\mathbb{E}}_{\mathcal{S}_m\sim\mathcal{D}_m}\mathop{\mathbb{E}}_{h\sim\mathrm{Q}_{\mathcal{S}_m}}\mathbb{1}\left[\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m,h')^{\frac{\alpha}{\alpha-1}} \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'_m\sim\mathcal{D}_m}\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}'_m,h')^{\frac{\alpha}{\alpha-1}}\right]\\
&\quad=\mathop{\mathbb{P}}_{\mathcal{S}_m\sim\mathcal{D}_m,h\sim\mathrm{Q}_{\mathcal{S}_m}}\left[\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m,h')^{\frac{\alpha}{\alpha-1}} \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'_m\sim\mathcal{D}_m}\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}'_m,h')^{\frac{\alpha}{\alpha-1}}\right].
\end{aligned}$$

Taking the logarithm to both sides of the inequality and adding $\frac{\alpha}{\alpha-1}\log\frac{2}{\delta}$ give us

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathcal{S}_m\sim\mathcal{D}_m,h\sim\mathrm{Q}_{\mathcal{S}_m}}\left[\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m,h')^{\frac{\alpha}{\alpha-1}} \le \frac{2}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}'_m\sim\mathcal{D}_m}\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}'_m,h')^{\frac{\alpha}{\alpha-1}}\right] \ge 1-\frac{\delta}{2} \iff\\
&\qquad\mathop{\mathbb{P}}_{\mathcal{S}_m\sim\mathcal{D}_m,h\sim\mathrm{Q}_{\mathcal{S}_m}}\left[\frac{\alpha}{\alpha-1}\log\frac{2}{\delta}+\log\left(\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}_m,h')^{\frac{\alpha}{\alpha-1}}\right)\le\right.\\
&\qquad\qquad\left.\frac{2\alpha-1}{\alpha-1}\log\frac{2}{\delta}+\log\left(\mathop{\mathbb{E}}_{\mathcal{S}'_m\sim\mathcal{D}_m}\mathop{\mathbb{E}}_{h'\sim\mathrm{Q}^0_{\mathcal{S}_m}} f(\mathcal{S}'_m,h')^{\frac{\alpha}{\alpha-1}}\right)\right]\ge 1-\frac{\delta}{2}. \qquad \text{(B.2)}
\end{aligned}$$

Combining Equation Eq. (B.1) and Eq. (B.2) with a union bound gives us the desired result. ■

B.4 Proofs

B.4.1 Proof of Th. 3.2.1

Background We first recall Rivasplata *et al.*, 2020, Thm 2.

> **Theorem B.4.1.** Let $\mathcal{D}_m \in \mathcal{M}(\mathcal{Z}^m)$, $\mathrm{Q}^0 \in \mathtt{Stoch}(\mathcal{Z}^m, \mathcal{F})$. Let k be a positive integer, any $A : \mathcal{S}_m \times \mathcal{H} \to \mathbb{R}^k$ a measurable function and $F : \mathbb{R}^k \to \mathbb{R}$ be a convex

function . Then for any $Q \in \texttt{Stoch}(\mathcal{Z}^m, \mathcal{F})$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ over the random draw of $\mathcal{S}_m \sim \mathcal{D}_m$ we have

$$F\left(\mathrm{Q}_{\mathcal{S}_m}[A_S]\right) \leq \mathrm{KL}\left(\mathrm{Q}_{\mathcal{S}_m} \| \mathrm{Q}^0_{\mathcal{S}_m}\right) + \log(\xi_m/\delta).$$

where $\xi_m := \int_{\mathcal{S}_m}\int_{\mathcal{H}} e^{f(s,h)}\mathrm{Q}^0_{\mathcal{S}_m}(dh)P(ds)$ and $\mathrm{Q}_{\mathcal{S}_m}[A_{\mathcal{S}_m}] := \mathrm{Q}_{\mathcal{S}_m}[A(\mathcal{S}_m,.)] = \int_{\mathcal{H}} A(\mathcal{S}_m,h)\mathrm{Q}_{\mathcal{S}_m}(dh)$.

Proof of Th. 3.2.1. To fully exploit the generality of Th. B.4.1, we aim to design a m-tuple of probabilities. Thus, our predictor set of interest is $\mathcal{H}_m := \mathcal{H}^{\otimes m}$ and then, our predictor h is a tuple $(h_1,..,h_m) \in \mathcal{H}$. Throughout our study, our stochastic kernels Q, Q^0 will belong to the specific class $\mathcal{C}$ defined below:

$$\mathcal{C} := \{Q \mid \exists(\mathrm{Q}_i)_{i=1..m}\text{s.t. } \forall S,\ \mathrm{Q}(\mathcal{S}_m,.) = \mathrm{Q}_1(\mathcal{S}_m,.)\otimes ... \otimes \mathrm{Q}_m(\mathcal{S}_m,.)\}\,. \quad \text{(B.3)}$$

Thus our kernels are such that conditionally to a given sample, our predictors $(h_1,...,h_m)$ are drawn independently.
We now apply Th. B.4.1. To do so, we consider the following function A : $\mathcal{S}_m \times \mathcal{H}_m \to \mathbb{R}^2$ such that $\forall \mathcal{S}_m = (\mathbf{z}_i)_{i=1..m}, h = (h_i)_{i=1..m} \in \mathcal{S}_m \times \mathcal{H}_m$:

$$A(\mathcal{S}_m,h) = \left(\sum_{i=1}^m \mathbb{E}[\ell(h_i,\mathbf{z}_i) \mid \mathcal{F}_{i-1}], \sum_{i=1}^m \ell(h_i,\mathbf{z}_i)\right)$$

A is indeed measurable in both of its variables. For a fixed $\lambda > 0$, we set the function F to be $F(x,y) = \lambda(x-y)$.
The only thing left to set up is our stochastic kernels. To do so, let $P = (P_1,...P_m)$ be an online predictive sequence, we then define $Q^0 \in \mathcal{C}$ (defined in Eq. (B.3)) s.t. for any sample $\mathcal{S}_m$,
$Q^0_{\mathcal{S}_m} = \mathrm{P}_1(\mathcal{S}_m,.)\otimes ... \otimes \mathrm{P}_m(\mathcal{S}_m,.)$. We also fix $\mathrm{Q}_1,...,\mathrm{Q}_m$ to be any (posterior) stochastic kernels and similarly we define the stochastic kernel $\mathrm{Q} \in \mathcal{C}$ such that for any sample $\mathcal{S}_m$, $\mathrm{Q}(\mathcal{S}_m,.) = \mathrm{Q}_1(\mathcal{S}_m,.)\otimes ... \otimes \mathrm{Q}_m(\mathcal{S}_m,.)$.
From now, we fix a dataset $\mathcal{S}_m$ and, for the sake of clarity, we assimilate in what follows the stochastic kernels $\mathrm{Q}_i, \mathrm{P}_i$ to the data-dependent distributions $\mathrm{Q}_i(\mathcal{S}_m,.), \mathrm{P}_i(\mathcal{S}_m,.)$ (i.e. we drop the dependency in $\mathcal{S}_m$).
Under those choices, one has:

$$\mathrm{Q}_{\mathcal{S}_m}[A_{\mathcal{S}_m}] = \int_{h\in\mathcal{H}_m} A(\mathcal{S}_m,h)\mathrm{Q}_{\mathcal{S}_m}(dh_1,...,dh_m)$$
$$= \left(\int_{h\in\mathcal{H}_m}\sum_{i=1}^m \mathbb{E}[\ell(h_i,\mathbf{z}_i)\mid\mathcal{F}_{i-1}]\mathrm{Q}_{\mathcal{S}_m}(dh_1,...,dh_m), \int_{h\in\mathcal{H}_m}\sum_{i=1}^m \ell(h_i,\mathbf{z}_i)\mathrm{Q}_{\mathcal{S}_m}(dh_1,...,dh_m)\right).$$

Furthermore, $Q\in\mathcal{C}$, thus $\mathrm{Q}_{\mathcal{S}_m}(dh_1,...,dh_m)=\Pi_{i=1}^m \mathrm{Q}_i(dh_i)$ so:‘

$$\mathrm{Q}_{\mathcal{S}_m}[A_{\mathcal{S}_m}] = \left(\sum_{i=1}^m \mathbb{E}_{h_i\sim\mathrm{Q}_i}[\mathbb{E}\left[\ell(h_i,\mathbf{z}_i)\mid\mathcal{F}_{i-1}\right]], \sum_{i=1}^m \mathbb{E}_{h_i\sim\mathrm{Q}_i}[\ell(h_i,\mathbf{z}_i)]\right).$$

Finally:

$$F(\mathrm{Q}_{\mathcal{S}_m}[A_{\mathcal{S}_m}]) = \lambda\left(\sum_{i=1}^m \mathbb{E}_{h_i\sim\mathrm{Q}_i}[\mathbb{E}\left[\ell(h_i,\mathbf{z}_i)\mid\mathcal{F}_{i-1}\right]] - \sum_{i=1}^m \mathbb{E}_{h_i\sim\mathrm{Q}_i}[\ell(h_i,\mathbf{z}_i)]\right).$$

Applying Th. B.4.1 and re-organising the terms gives us with probability $1-\delta$:

$$\sum_{i=1}^m \mathbb{E}_{h_i\sim\mathrm{Q}_i}\left[\mathbb{E}[\ell(h_i,\mathbf{z}_i)\mid\mathcal{F}_{i-1}]\right] \leq \sum_{i=1}^m \mathbb{E}_{h_i\sim\mathrm{Q}_i}[\ell(h_i,\mathbf{z}_i)] + \frac{KL(\mathrm{Q}_{\mathcal{S}_m}\|\mathrm{Q}^0_{\mathcal{S}_m})}{\lambda} + \frac{\log(\xi_m/\delta)}{\lambda}.$$

Thus:

$$\sum_{i=1}^m \mathbb{E}_{h_i\sim\mathrm{Q}_i}\left[\mathbb{E}[\ell(h_i,\mathbf{z}_i)\mid\mathcal{F}_{i-1}]\right] \leq \sum_{i=1}^m \mathbb{E}_{h_i\sim\mathrm{Q}_i}[\ell(h_i,\mathbf{z}_i)] + \sum_{i=1}^m \frac{KL(\mathrm{Q}_i\|\mathrm{P}_i)}{\lambda} + \frac{\log(\xi_m/\delta)}{\lambda}. \tag{B.4}$$

The last line holding because for a fixed $\mathcal{S}_m$, $\mathrm{Q}_{\mathcal{S}_m}=Q_1\otimes...\otimes Q_m$ and $\mathrm{Q}^0_{\mathcal{S}_m} = P_1\otimes...\otimes P_m$.
The last term to control is

$$\xi_m = \mathbb{E}_S\left[\mathbb{E}_{h_1,...,h_m\sim\mathrm{Q}^0_{\mathcal{S}_m}}\left[\exp\left(\lambda\sum_{i=1}^m \tilde{\ell}_i(h_i,\mathbf{z}_i)\right)\right]\right],$$

with $\tilde{\ell}_i(h_i,\mathbf{z}_i)=\mathbb{E}[\ell(h_i,\mathbf{z}_i)\mid\mathcal{F}_{i-1}]-\ell(h_i,\mathbf{z}_i)$. Hence the following lemma.

Lemma B.4.1. One has for any m, $\xi_m \leq \exp\left(\frac{\lambda^2 m K^2}{2}\right)$ with K bounding ℓ.

The proof of this lemma is deferred to Section B.4.1.1
To conclude the proof, we just bound ξ_m by the result of lemma B.4.1 within Eq. (B.4). ■

B.4.1.1 Proof of lemma B.4.1

Proof of lemma B.4.1. We prove our result by recursion: for $m = 1$, $\mathcal{S}_1 = \mathbf{z}_1$ and one knows that P_1 is $\mathcal{F}_0$ measurable yet it does not depend on $\mathcal{S}_m$. Thus for any $h_1 \in \mathcal{H}$, $\mathbb{E}[\ell(h_1, \mathbf{z}_1) \mid \mathcal{F}_0] = \mathbb{E}[\ell(h_1, \mathbf{z}_1)]$. We then has:

$$\begin{aligned}
\xi_1 &= \mathbb{E}_{\mathcal{S}_1}\mathbb{E}_{h_1\sim \mathrm{P}_1}[\tilde{\ell}_1(h_1, \mathbf{z}_1)] \\
&= \mathbb{E}_{h_1\sim P_1}\mathbb{E}_{\mathcal{S}_1}[\tilde{\ell}_1(h_1, \mathbf{z}_1)] && \text{by Fubini} \\
&\leq \exp\frac{\lambda^2 K^2}{2}
\end{aligned}$$

The last line holding because for any $h_1 \in \mathcal{H}$, $\tilde{\ell}_1(h_1, \mathbf{z}_1)$ is a centered variable belonging in $[-K, K]$ a.s. and so one can apply Hoeffding's lemma to conclude.
Assume the result is true at rank $m - 1 \geq 0$. We then has to prove the result at rank m. Our strategy consists in conditioning by $\mathcal{F}_{m-1}$ within the expectation over $\mathcal{S}_m$:

$$\xi_m = \mathbb{E}_{\mathcal{S}_m}\left[\mathbb{E}_{h_1,\ldots,h_m\sim \mathrm{Q}^0_{\mathcal{S}_m}}\left[\exp\left(\lambda\sum_{i=1}^{m}\tilde{\ell}_i(h_i, \mathbf{z}_i)\right)\right]\right].$$

First, we use that $Q^0 \in \mathcal{C}$, thus $\mathrm{Q}^0_{\mathcal{S}_m} = P_1 \otimes \ldots \otimes P_m$ (i.e. our data are drawn independently for a given $\mathcal{S}_m$):

$$= \mathbb{E}_S\left[\Pi_{i=1}^{m}\mathbb{E}_{h_i\sim \mathrm{P}_i}\left[\exp\left(\lambda\tilde{\ell}_i(h_i, \mathbf{z}_i)\right)\right]\right].$$

We now condition by $\mathcal{F}_{m-1}$ and use that $\Pi_{i=1}^{m-1}\mathbb{E}_{h_i\sim \mathrm{P}_i}\left[\exp\left(\lambda\tilde{\ell}_i(h_i, \mathbf{z}_i)\right)\right]$ is a $\mathcal{F}_{m-1}$-measurable r.v.

$$\xi_m = \mathbb{E}_S\left[\Pi_{i=1}^{m-1}\mathbb{E}_{h_i\sim \mathrm{P}_i}\left[\exp\left(\lambda\tilde{\ell}_i(h_i, \mathbf{z}_i)\right)\right]\mathbb{E}\left[\mathbb{E}_{h_m\sim P_m}[\exp(\lambda\tilde{\ell}_m(h_m, \mathbf{z}_m))] \mid \mathcal{F}_{m-1}\right]\right].$$

Now our next step is to use a variant of Fubini valid for $\mathcal{F}_{m-1}$- measurable measures.

Lemma B.4.2 (Conditional Fubini)**.** Let $f : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$. For a *sigma-*algebra $\mathcal{F}$ over $\mathcal{Z}$ and a measure P over $\mathcal{H}$ such that

- P is a $\mathcal{F}$-measurable r.v.
- There exists a constant measure (a.s.) P_0 such that $\mathrm{P} \ll \mathrm{P}_0$.

Then one has almost surely, for any r.v. z over $\mathcal{Z}$:

$$\mathbb{E}\left[\mathbb{E}_{h\sim\mathrm{P}}[f(h,\mathbf{z})] \mid \mathcal{F}\right] = \mathbb{E}_{h\sim\mathrm{P}}\left[\mathbb{E}[f(h,\mathbf{z}) \mid \mathcal{F}]\right].$$

The proof of this lemma lies at the end of this section.
We then fix $\mathcal{F} = \mathcal{F}_{m-1}$ and $f(h,\mathbf{z}) = \exp(\lambda\tilde{\ell}_i(h,z))$. Furthermore, because we assumed the sequence $(\mathrm{P}_i)_{i=1\cdots m}$ to be an online predictive sequence, P_m is $\mathcal{F}_{m-1}$-measurable and $P_m >> P_1$ with P_1 a data-free prior. One then applies lemma B.4.2:

$$\mathbb{E}\left[\mathbb{E}_{h_m\sim P_m}[\exp(\lambda\tilde{\ell}_m(h_m,\mathbf{z}_m))] \mid \mathcal{F}_{m-1}\right] = \mathbb{E}_{h_m\sim P_m}\left[\mathbb{E}[\exp(\lambda\tilde{\ell}_m(h_m,\mathbf{z}_m)) \mid \mathcal{F}_{m-1}]\right].$$

Yet, injecting this result onto ξ_m provides:

$$\xi_m = \mathbb{E}_S\left[\Pi_{i=1}^{m-1}\mathbb{E}_{h_i\sim\mathrm{P}_i}\left[\exp\left(\lambda\tilde{\ell}_i(h_i,\mathbf{z}_i)\right)\right]\mathbb{E}_{h_m\sim P_m}\left[\mathbb{E}[\exp(\lambda\tilde{\ell}_m(h_m,\mathbf{z}_m)) \mid \mathcal{F}_{m-1}]\right]\right]$$

The final remark is to notice that for any $h_m \in \mathcal{H}$, $\mathbb{E}[\tilde{\ell}_m(h_m,\mathbf{z}_m) \mid \mathcal{F}_{m-1}] = 0$ and $\tilde{\ell}_m(h_m,\mathbf{z}_m) \in [-K,K]$ a.s. then one can apply the conditional Hoeffding's lemma which ensure us that for any $\lambda > 0$:

$$\mathbb{E}[\exp(\lambda\tilde{\ell}_m(h_m,\mathbf{z}_m)) \mid \mathcal{F}_{m-1}] \leq \exp\left(\frac{\lambda^2K^2}{2}\right).$$

One then has $\xi_m \leq \exp\left(\frac{\lambda^2K^2}{2}\right)\xi_{m-1}$. The recursion assumption concludes the proof.

■

Proof Proof of lemma B.4.2. Let A be a $\mathcal{F}$-measurable event. One wants to show that

$$\mathbb{E}\left[\mathbb{E}_{h\sim\mathrm{P}}[f(h,\mathbf{z})]\mathbb{1}_A\right] = \mathbb{E}\left[\mathbb{E}_{h\sim\mathrm{P}}\left[\mathbb{E}[f(h,\mathbf{z}) \mid \mathcal{F}]\right]\mathbb{1}_A\right].$$

Where the first expectation in each term is taken over z. This will be enough to conclude that

$$\mathbb{E}\left[\mathbb{E}_{h\sim\mathrm{P}}[f(h,\mathbf{z})] \mid \mathcal{F}\right] = \mathbb{E}_{h\sim\mathrm{P}}\left[\mathbb{E}[f(h,\mathbf{z}) \mid \mathcal{F}]\right]$$

thanks to the definition of conditional expectation. We first start by using the fact that P is $\mathcal{F}$-measurable and that $P_0 \gg P$ with P_0 a constant measure. This is enough to obtain that the Radon-Nykodym derivative $\frac{d\mathrm{P}}{d\mathrm{P}_0}$ is a $\mathcal{F}$-measurable function, thus:

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{h\sim\mathrm{P}}[f(h,\mathbf{z})]\mathbb{1}_A\right] &= \mathbb{E}\left[\mathbb{E}_{h\sim\mathrm{P}_0}\left[f(h,\mathbf{z})\frac{d\mathrm{P}}{d\mathrm{P}_0}(h)\right]\mathbb{1}_A(\mathbf{z})\right],\\ &= \mathbb{E}\left[\mathbb{E}_{h\sim\mathrm{P}_0}\left[f(h,\mathbf{z})\frac{d\mathrm{P}}{d\mathrm{P}_0}(h)\mathbb{1}_A(\mathbf{z})\right]\right].\end{aligned}$$

Because $f(h,\mathbf{z})\frac{d\mathrm{P}}{d\mathrm{P}_0}(h)\mathbb{1}_A(\mathbf{z})$ is a positive function, and that P_0 is fixed, one can apply the classical Fubini-Tonelli theorem:

$$= \mathbb{E}_{h\sim\mathrm{P}_0}\left[\mathbb{E}\left[f(h,\mathbf{z})\frac{d\mathrm{P}}{d\mathrm{P}_0}(h)\mathbb{1}_A(\mathbf{z})\right]\right].$$

One now conditions by $\mathcal{F}$ and use the fact that $\frac{d\mathrm{P}}{d\mathrm{P}_0}, \mathbb{1}_A$ are $\mathcal{F}$-measurable:

$$= \mathbb{E}_{h\sim\mathrm{P}_0}\left[\mathbb{E}\left[\mathbb{E}\left[f(h,\mathbf{z}) \mid \mathcal{F}\right]\frac{d\mathrm{P}}{d\mathrm{P}_0}(h)\mathbb{1}_A(\mathbf{z})\right]\right].$$

We finally re-apply Fubini-Tonelli to re-intervert the expectations:

$$\begin{aligned}&= \mathbb{E}\left[\mathbb{E}_{h\sim\mathrm{P}_0}\left[\mathbb{E}\left[f(h,\mathbf{z}) \mid \mathcal{F}\right]\frac{d\mathrm{P}}{d\mathrm{P}_0}(h)\mathbb{1}_A(\mathbf{z})\right]\right],\\ &= \mathbb{E}\left[\mathbb{E}_{h\sim\mathrm{P}}\left[\mathbb{E}\left[f(h,\mathbf{z}) \mid \mathcal{F}\right]\mathbb{1}_A(\mathbf{z})\right]\right].\end{aligned}$$

This finally proves the announced results, yet concludes the proof.

■

B.4.2 Proofs of section 3.4

We prove here Cor. 3.4.1 and Cor. 3.4.2.

B.4.2.1 Proof of Cor. 3.4.1

We fix $\hat{\mathrm{Q}}, \mathrm{P}$ to be online predictive sequences (with $\hat{\mathrm{Q}}_1, \mathrm{P}_1$ being data-free priors). Recall that we assimilated the stochastic kernels $\hat{\mathrm{Q}}_i, \mathrm{P}_i$ to the their associated data-dependent sitribution given a sample $\mathcal{S}_m$ $\hat{\mathrm{Q}}_i(\mathcal{S}_m, .), \mathrm{P}_i(\mathcal{S}_m, .)$.
As in Th. 3.2.1, our predictor set of interest is $\mathcal{H}_m := \mathcal{H}^{\otimes m}$ and then, our predictor h is a tuple $(h_1, .., h_m) \in \mathcal{H}$. We consider the stochastic kernel Q belonging to the class $\mathcal{C}$ defined in Eq. (B.3) such that for any $S \in \mathcal{S}_m, \mathrm{Q}(\mathcal{S}_m, .) = \hat{\mathrm{Q}}_2 \otimes ... \otimes \hat{\mathrm{Q}}_{m+1}$. Similarly one defines $Q^0 \in \mathcal{C}$ such that for any $S \in \mathcal{S}_m, Q^0(\mathcal{S}_m, .) = P_1 \otimes ... \otimes P_m$

Proof for (Ψ_1, Φ_1): For $\lambda > 0$, we set our function f to be for any dataset $\mathcal{S}_m$ and predictor tuple $h = (h_1, ..., h_m)$,

$$f(\mathcal{S}_m, h) = \lambda \left(\sum_{i=1}^m \mathbb{E}\left[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}\right] - \sum_{i=1}^m \ell(h_i, \mathbf{z}_i) \right).$$

We then apply proposition B.3.1 with the function f, Q, Q^0 defined above. One then has by dividing by λ with probability $1 - \delta$ over $S \sim \mu$ and $h = (h_1, ..., h_m) \sim \hat{\mathrm{Q}}_2 \otimes ... \otimes \hat{\mathrm{Q}}_{m+1}$:

$$\sum_{i=1}^m \mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] \leq \sum_{i=1}^m \ell(h_i, \mathbf{z}_i) + \frac{1}{\lambda} \log \left(\frac{d\mathrm{Q}_{\mathcal{S}_m}}{d\mathrm{Q}^0_{\mathcal{S}_m}}(h_i) \right) + \frac{1}{\lambda} \log(\xi_m) + \frac{\log(1/\delta)}{\lambda}.$$

And then using the fact that $S \in \mathcal{S}_m, \mathrm{Q}_{\mathcal{S}_m} = \hat{\mathrm{Q}}_2 \otimes ... \otimes \hat{\mathrm{Q}}_{m+1}, \mathrm{Q}^0_{\mathcal{S}_m} = P_1 \otimes ... \otimes P_m$ gives us:

$$\sum_{i=1}^m \mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] \leq \sum_{i=1}^m \ell(h_i, \mathbf{z}_i) + \frac{1}{\lambda} \sum_{i=1}^m \log \left(\frac{d\hat{\mathrm{Q}}_{i+1}}{d\mathrm{P}_i}(h_i) \right) + \frac{1}{\lambda} \log(\xi_m) + \frac{\log(1/\delta)}{\lambda},$$

with $\xi_m = \mathbb{E}_S \left[\mathbb{E}_{h_1, ..., h_m \sim \mathrm{Q}_{\mathcal{S}_m}} \left[\exp\left(\lambda \sum_{i=1}^m \tilde{\ell}_i(h_i, \mathbf{z}_i) \right) \right] \right]$ and for any i, $\tilde{\ell}_i(h_i, \mathbf{z}_i) = \mathbb{E}\left[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}\right] - \ell(h_i, \mathbf{z}_i)$
Notice that, because P is an online predictive sequence, then one can apply directly lemma B.4.1 to conclude that $\xi_m \leq \exp\left(\frac{\lambda^2 K^2 m}{2} \right)$.
We also use VIALLARD *et al.*, 2023a, Lemma 11 which derives the calculation of the disintegrated KL divergence between two Gaussians. One then has for any i, with $h_i = \hat{w}_{i+1} + \varepsilon_i$:

$$\log \left(\frac{d\hat{\mathrm{Q}}_{i+1}}{d\mathrm{P}_i}(h_i) \right) = \frac{||\hat{w}_{i+1} + \varepsilon_i - w_i^0||^2 - ||\varepsilon||^2}{2\sigma^2}.$$

Combining those facts altogether allows us to conclude.

Proof for (Ψ_2, Φ_2)**:** For $\lambda > 0$, we set our function f to be for any dataset $\mathcal{S}_m$ and predictor tuple $(h = h_1, ..., h_m)$,

$$f(\mathcal{S}_m, h) = \exp\left(\lambda\left(\sum_{i=1}^m \mathbb{E}\left[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}\right] - \sum_{i=1}^m \ell(h_i, \mathbf{z}_i)\right)\right).$$

We take $\alpha = 2$ and apply this time proposition B.3.2. One then has by dividing by 2λ with probability $1-\delta$ over $S \sim \mu$ and $h = (h_1, ..., h_m) \sim \hat{\mathrm{Q}}_2 \otimes ... \otimes \hat{\mathrm{Q}}_{m+1}$:

$$\sum_{i=1}^m \mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] \leq \sum_{i=1}^m \ell(h_i, \mathbf{z}_i) + \frac{3}{2\lambda}\log\frac{2}{\delta} + \frac{D_2\left(\mathrm{Q}_{\mathcal{S}_m} \| \mathrm{Q}^0_{\mathcal{S}_m}\right)}{2\lambda} + \frac{1}{2\lambda}\log\left(\underbrace{\mathop{\mathbb{E}}_{\mathcal{S}'_m \sim \mathcal{D}_m} \mathop{\mathbb{E}}_{h' \sim \mathrm{Q}^0_{\mathcal{S}'_m}} f(\mathcal{S}'_m, h')^2}_{:=\xi'_m}\right).$$

We first notice that $D_2\left(\mathrm{Q}_{\mathcal{S}_m} \| \mathrm{Q}^0_{\mathcal{S}_m}\right) = \sum_{i=1}^m D_2(\hat{\mathrm{Q}}_{i+1} \| \mathrm{P}_i)$ as our predictors are drawn independently once $\mathcal{S}_m$ is given.

We also use that for any i, the Rényi divergence with $\alpha = 2$ between $\hat{\mathrm{Q}}_{i+1}$ and P_i (two multivariate Gaussians with same covariance matrix) is $\frac{\|\hat{w}_{i+1} - w_i^0\|^2}{\sigma^2}$ (as recalled in GIL *et al.*, 2013).

We then remark that:

$$\xi'_m = \mathop{\mathbb{E}}_{\mathcal{S}'_m \sim \mathcal{D}_m} \mathop{\mathbb{E}}_{h' \sim \mathrm{Q}^0_{\mathcal{S}'_m}} \exp\left(2\lambda\left(\sum_{i=1}^m \mathbb{E}\left[\ell(h'_i, \mathbf{z}'_i) \mid \mathcal{F}_{i-1}\right] - \sum_{i=1}^m \ell(h'_i, \mathbf{z}'_i)\right)\right).$$

Thus we recover the exponential moment ξ_m from the Rivasplata's case up to a factor 2 within the exponential. We then apply lemma B.4.1 with $\lambda' = 2\lambda$ to obtain that $\xi'_m \leq \exp(2\lambda^2 K^2 m)$.

Combining all those facts allows us to conclude.

B.4.2.2 Proof of Cor. 3.4.2

We apply the exact same proof than Cor. 3.4.1. The only difference is the way to define our stochastic kernels. We now take, for a single online predictive sequence $\hat{\mathrm{Q}}$ the following stochastic kernels:

We consider the stochastic kernel Q belonging to the class $\mathcal{C}$ defined in Eq. (B.3) such that for any $S \in \mathcal{S}_m, \mathrm{Q}(\mathcal{S}_m, .) = \hat{\mathrm{Q}}_1 \otimes ... \otimes \hat{\mathrm{Q}}_m$ and we take $\mathrm{Q}_0 = \mathrm{Q}$.

This fact allows the divergence terms (Rényi or KL depending on which bound we consider) to vanish. The rest of the proof remains unchanged.

B.4.3 Proof of Theorem 3.6.1

Proof. We fix $m \geq 1$, $\mathcal{S}$ a countable dataset and $(\mathrm{P}_i)_{i\geq 1}$ an online predictive sequence. We aim to design a m-tuple of probabilities. Thus, our predictor set of interest is $\mathcal{H}_m := \mathcal{H}^{\otimes m}$ and then, our predictor h is a tuple $(h_1, .., h_m) \in \mathcal{H}$.
Our goal is to apply the change of measure inequality on $\mathcal{H}_m$ to a specific function f_m inspired from Lemma 2.1.2. We define this function below, for any sample $\mathcal{S}$ and any predictor $h^m = (h_1, ..., h_m)$

$$f_m(\mathcal{S}, h^m) := \sum_{i=1}^{m} \lambda X_i(h_i, \mathbf{z}_i) - \frac{\lambda^2}{2} \sum_{i=1}^{m} (\hat{V}_i(h_i, \mathbf{z}_i) + V_i(h_i)),$$

where $X_i(h_i, \mathbf{z}_i) = \mathbb{E}_{i-1}[\ell(h_i, \mathbf{z}_i)] - \ell(h_i, \mathbf{z}_i)$. Notice that for fixed h, the sequence $(f_m(\mathcal{S}, h))_{m\geq 1}$ is a supermartingale according to Lemma 2.1.2.
Now for a given posterior tuple $Q_1, ...Q_m$ we define $Q = Q_1 \otimes ... \otimes Q_m$ and also $P_S^m = P_{1,S} \otimes ... \otimes \mathrm{P}_{m,\mathcal{S}}$. We can now properly apply the change of measure inequality for any m:

$$\sum_{i=1}^{m} \mathbb{E}_{h_i \sim \mathrm{Q}_i}[\lambda X_i(h_i, \mathbf{z}_i) - \frac{\lambda^2}{2}(\hat{V}_i(h_i, \mathbf{z}_i) + V_i(h_i))] = \mathbb{E}_{h^m \sim Q}\left[f_m(\mathcal{S}, h^m)\right]$$
$$\leq \mathrm{KL}(Q, P_S^m) + \log\left(\mathbb{E}_{h^m \sim P_S^m} \exp(f_m(\mathcal{S}, h^m))\right).$$

Noticing that $\mathrm{KL}(Q, P_S^m) = \sum_{i=1}^m \mathrm{KL}(\mathrm{Q}_i, \mathrm{P}_{i,\mathcal{S}_m})$, the only remaining term to deal with is the exponential rv.
To do so we prove the following lemma:

> **Lemma B.4.3.** The sequence $(M_m := \mathbb{E}_{h^m \sim P_S^m} \exp(f_m(\mathcal{S}, h^m)))_{m\geq 1}$ is a non-negative supermartingale.

The proof of this lemma lies at the end of this section.
Now we can apply Ville's inequality which implies that with probability at least $1 - \delta$, for any $m \geq 1$:

$$\mathbb{E}_{h^m \sim P_S^m} \exp(f_m(\mathcal{S}, h^m)) \leq \frac{1}{\delta}.$$

Thus we have with probability at least $1 - \delta$, for any posterior sequence $(\mathrm{Q}_i)_{i\geq 1}$, the data-dependent measures $P_{1,S}, ..., \mathrm{P}_{m,\mathcal{S}}$ and any $m \geq 1$:

$$\sum_{i=1}^{m} \mathbb{E}_{h_i\sim \mathrm{Q}_i}\left[\lambda X_i(h_i,\mathbf{z}_i) - \frac{\lambda^2}{2}(\hat{V}_i(h_i,\mathbf{z}_i) + V_i(h_i))\right] \leq \sum_{i=1}^{m} \mathrm{KL}(\mathrm{Q}_i, \mathrm{P}_{i,\mathcal{S}_m}) + \log\left(\frac{1}{\delta}\right).$$

Re-organising the terms in this bound and dividing by λ concludes the proof. ■

Proof of Lemma B.4.3. We fix $m \geq 1$ and we recall that for any i, $\mathrm{P}_{i,\mathcal{S}_m}$ is $\mathcal{F}_{i-1}$-measurable. We show that $\mathbb{E}_{m-1}[M_m] \leq M_{m-1}$. We first recover M_{m-1} from $\mathbb{E}_{m-1}[M_m]$.

$$\begin{aligned}\mathbb{E}_{m-1}[M_m] &= \mathbb{E}_{m-1}\left[\mathbb{E}_{h^m\sim P_S^m} \exp(f_m(\mathcal{S},h^m))\right]\\ &= \mathbb{E}_{m-1}\left[\mathbb{E}_{h_1,..,h_m\sim P_{1,S}\otimes...\otimes \mathrm{P}_{m,\mathcal{S}}} \exp(f_m(\mathcal{S},h^m))\right]\\ &= \mathbb{E}_{m-1}\left[\mathbb{E}_{h_1,..,h_m\sim P_{1,S}\otimes...\otimes \mathrm{P}_{m,\mathcal{S}}}\left[\Pi_{i=1}^m \exp\left(\lambda X_i(h_i,\mathbf{z}_i) - \frac{\lambda^2}{2}(\hat{V}_i(h_i,\mathbf{z}_i) + V_i(h_i))\right)\right]\right]\\ &= M_{m-1}\mathbb{E}_{m-1}\left[\mathbb{E}_{h_m\sim \mathrm{P}_{m,\mathcal{S}}}\left[\exp\left(\lambda X_m(h_m,\mathbf{z}_m) - \frac{\lambda^2}{2}(\hat{V}_m(h_m,\mathbf{z}_m) + V_m(h_m))\right)\right]\right].\end{aligned}$$

The last line holding because $P_S^{m-1} = P_{1,S} \otimes ... \otimes P_{m-1,S}$ is $\mathcal{F}_{m-1}$ measurable. Now we exploit the fact that $\mathrm{P}_{m,\mathcal{S}}$ is $\mathcal{F}_{m-1}$ measurable to apply Lemma B.4.2. We have:

$$\begin{aligned}&\mathbb{E}_{m-1}\left[\mathbb{E}_{h_m\sim \mathrm{P}_{m,\mathcal{S}}}\left[\exp\left(\lambda X_m(h_m,\mathbf{z}_m) - \frac{\lambda^2}{2}(\hat{V}_m(h_m,\mathbf{z}_m) + V_m(h_m))\right)\right]\right]\\ &\quad= \mathbb{E}_{h_m\sim \mathrm{P}_{m,\mathcal{S}}}\left[\mathbb{E}_{m-1}\left[\exp\left(\lambda X_m(h_m,\mathbf{z}_m) - \frac{\lambda^2}{2}(\hat{V}_m(h_m,\mathbf{z}_m) + V_m(h_m))\right)\right]\right].\end{aligned}$$

Now we can apply Lemma 2.1.2 for any $h_m \in \mathcal{H}$ with $\Delta M_m = X_m(h_m,\mathbf{z}_m)$, $\Delta[M]_m = \hat{V}(h_m,\mathbf{z}_m)$ and $\Delta\langle M\rangle_m = V_m(h_m)$. We then have for all $h_m \in \mathcal{H}$:

$$\mathbb{E}_{m-1}\left[\exp\left(\lambda X_m(h_m,\mathbf{z}_m) - \frac{\lambda^2}{2}(\hat{V}_m(h_m,\mathbf{z}_m) + V_m(h_m))\right)\right] \leq 1.$$

Thus $\mathbb{E}_{m-1}[M_m] \leq M_{m-1}$, this concludes the lemma's proof. ■

B.5 Additional experiment

In this section we perform error bars for our OPBD methods in order to evaluate their volatility. We ran $n = 50$ times our algorithms and then show in the table below for each data set the means and the standard deviation of our averaged cumulative losses at regular time steps. We denote for $i \in \{1, 2\}$ 'OPBD Ψ_i' to indicate that this algorithm is our OPBD method used with thev optimisation objective Ψ_i.

Analysis Those tables shows the robustness of our OPBD methods to their intrinsic randomness: we always have a decreasing mean through time as well as an overall variance reduction. Note that for the most complicated problem (California Housing dataset), the variance is the highest. More precisely, we notice that the standard deviation of OPBD with Ψ_1 is always greater than the one of OPBD with Ψ_2 which is not a surprise as Ψ_1 involves a disintegrated KL divergence while Ψ_2 is a proper Rényi divergence. Hence the additional volatility for OPBD with Ψ_1.
This fact is particurlaly noticeable on the California Housing dataset where both the means and variance of OPBD with Ψ_1 increase drastically between t=16000 and t=20000 while the increase is more attenuated for OPBD with Ψ_2. This fact is also visible on fig. 3.1.

	means OPBD Ψ_1	std OPBD Ψ_1	means OPBD Ψ_2	std OPBD Ψ_2
t=200	0.2014	0.0034	0.1993	0.0007
t=400	0.1888	0.0030	0.1861	0.0004
t=600	0.1867	0.0023	0.1839	0.0003
t=800	0.1714	0.0020	0.1686	0.0003
t=1000	0.1760	0.0016	0.1731	0.0003

Table B.1. *Error bars for the Boston Housing dataset*

	means OPBD Ψ_1	std OPBD Ψ_1	means OPBD Ψ_2	std OPBD Ψ_2
t=100	0.1619	0.0063	0.1601	0.0030
t=200	0.1350	0.0057	0.1361	0.0008
t=300	0.1214	0.0044	0.1241	0.0009
t=400	0.1210	0.0043	0.1238	0.0021
t=500	0.1131	0.0037	0.1159	0.0015

Table B.2. *Error bars for the Breast Cancer dataset*

	means OPBD Ψ_1	std OPBD Ψ_1	means OPBD Ψ_2	std OPBD Ψ_2
t=150	0.7102	0.0061	0.7069	0.0007
t=300	0.6455	0.0056	0.6422	0.0007
t=450	0.6134	0.0042	0.6103	0.0007
t=600	0.5860	0.0035	0.5837	0.0008
t=750	0.5685	0.0031	0.5664	0.0008

Table B.3. *Error bars for the PIMA Indians dataset*

	means OPBD Ψ_1	std OPBD Ψ_1	means OPBD Ψ_2	std OPBD Ψ_2
t=4000	0.9320	0.0572	0.8905	0.0003
t=8000	0.6325	0.0335	0.5947	0.0003
t=12000	0.5314	0.0254	0.4954	0.0002
t=16000	0.4967	0.0299	0.4477	0.0004
t=20000	0.5273	0.1056	0.4355	0.0030

Table B.4. *Error bars for the California Housing dataset*

Appendix of Chapter 4 C

C.1 Supplementary background

C.1.1 Additional details on Poincaré and Log-Sobolev inequalities.

Proof of proposition 4.2.2.

Proof. We define P_1 such that $d\mathrm{P}_1(h) \propto \exp(-V(h) - \gamma \hat{\mathsf{R}}_{\mathcal{S}_m}(h))dh$, then note that, by convexity assumption over ℓ_1, $Hess(V + \gamma \hat{\mathsf{R}}_{\mathcal{S}_m}) \succeq \frac{1}{c_{LS}}\mathrm{Id}$. Then, applying Chafai (2004, Corollary 2.1), we know that P_1 satisfies a Poincaré inequality with constant $c_{LS}(\mathrm{P})$.
Finally, defining P_2 such that $d\mathrm{P}_2(h) \propto \exp(-\frac{\gamma}{m}\sum_{i=1}^m \ell_2(h, \mathbf{z}_i))$, thanks to the boundedness of ℓ_2, we use Guionnet and Zegarlinksi (2003, Property 2.6), which ensure that $P_2 = \mathrm{P}_{-\gamma \hat{\mathsf{R}}_{\mathcal{S}_m}} d\mathrm{P}_1(h)$ satisfies a Log-Sobolev inequality with constant $2c_{LS}(\mathrm{P})\exp(4\|\ell_2\|_\infty)c_P(\mathrm{P})$. Noting that $\mathrm{P}_2 = \mathrm{P}_{-\gamma \hat{\mathsf{R}_{\mathcal{S}}}}$ concludes the proof. ■

Proof of Ledoux (2006, Propsition 2.1) We prove here Proposition C.1.1, stated below, showing that Log-Sobolev implies Poincaré.

> **Proposition C.1.1.** If Q is $\texttt{L-Sob}(c_{LS})$, then it is also $\texttt{Poinc}(c_P)$.
> We then have $c_P(\mathrm{Q}) = \frac{c_{LS(\mathrm{Q})}}{2}$.

We provide the proof for completeness.

Proof. Let $f \in \mathrm{H}^1(\mathrm{Q})$, such that $\mathbb{E}_Q[f] = 0$ and $\mathbb{E}[f^2] = 1$. For any $\varepsilon > 0$, $1 + \varepsilon f \in \mathrm{H}^1$. We then apply the Log-Sobolev inequality on $1 + \varepsilon f$:

$$\mathbb{E}_Q\left[(1+\varepsilon f)^2\left(2\log(1+\varepsilon f) - \log(1+\varepsilon^2)\right)\right] \leq c_{LS}(\mathrm{Q})\varepsilon^2 \mathbb{E}_{\mathrm{Q}}\left[\|\nabla f\|^2\right].$$

Note that, by a Taylor expansion, $\log(1+\varepsilon f) = \varepsilon f - \frac{(\varepsilon f)^2}{2} + o\left(\varepsilon^2\right)$ and also that $\log(1+\varepsilon^2) = \varepsilon^2 + o(\varepsilon^2)$. Then, plugging this into the previous equation gives:

$$\mathbb{E}_Q\left[2\varepsilon f + 3(\varepsilon f)^2 - \varepsilon^2 + o(\varepsilon^2)\right] \leq c_{LS}(\mathrm{Q})\varepsilon^2 \mathbb{E}_{\mathrm{Q}}\left[\|\nabla f\|^2\right].$$

We use that $\mathbb{E}[f] = 0$ and we then divide by ε^2. Taking the limit $\varepsilon \to 0$ gives:

$$\mathbb{E}_Q\left[3f^2 - 1\right] \leq c_{LS}(\mathrm{Q})\mathbb{E}_{\mathrm{Q}}\left[\|\nabla f\|^2\right].$$

Using that $\mathbb{E}[f^2] = 1$ gives:

$$1 \leq \frac{c_{LS}(\mathrm{Q})}{2}\mathbb{E}_{\mathrm{Q}}\left[\|\nabla f\|^2\right]$$

Then, for any $g \in \mathrm{H}^1(\mathrm{Q})$ applying this proof on $f = \frac{g - \mathbb{E}_{\mathrm{Q}}[g]}{\sqrt{\mathrm{Var}_{\mathrm{Q}}(g)}}$ concludes the proof. ■

C.1.2 Wasserstein distances

We recall here the definition of Wasserstein distances, valid for any Polish space $\mathcal{H}$ equipped with a distance d.

Definition C.1.1. The 1-Wasserstein distance between $P, Q \in \mathcal{M}(\mathcal{H})^2$ is defined as

$$\mathrm{W}_1(Q,P) = \inf_{\pi \in \Pi(Q,P)} \int_{\mathcal{H}^2} ||x-y|| \mathrm{d}\pi(x,y).$$

where $\Pi(Q,P)$ denote the set of probability measures on $\mathcal{H}^2$ whose marginals are Q and P. We define the 2-Wasserstein distance on $\mathcal{P}(\mathcal{H})$ as

$$\mathrm{W}_2(Q,P) = \sqrt{\inf_{\pi \in \Pi(Q,P)} \int_{\mathcal{H}^2} ||x-y||^2 \mathrm{d}\pi(x,y)}.$$

C.2 PAC-Bayes bounds for Lipschitz losses through log-Sobolev inequalities

Extending Catoni's bound to Lipschitz losses. A well-known relaxation of Catoni (2007, Theorem 1.2.6) (see *e.g.* Alquier *et al.*, 2016, Theorem 4.1) holding for subgaussian losses has been widely used in practice as a tractable PAC-Bayesian algorithm exhibiting a linear dependency on the KL divergence. We exploit below a consequence of the Herbst argument as stated, *e.g.*, in Ledoux (2006, Section 2.3), stating that a L-Lipschitz function of a random variable following a distribution $\mathcal{D}$ being $\texttt{L-Sob}(c_{LS})$ is $L\sqrt{c_{LS}(\mathcal{D})}$ subgaussian. This yields the following corollary.

Corollary C.2.1. Let $\lambda > 0$, $m \geq 1$ and a data-free prior P. Assume that for any $h \in \mathcal{H}$, $\ell(h,.)$ is L-Lipschitz and that the data distribution $\mathcal{D}$ is $\texttt{L-Sob}(c_{LS})$. Then for with probability at least $1-\delta$ over $\mathcal{S}$, for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$,

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda} + \frac{2\lambda^2 L^2 c_{LS}(\mathcal{D})}{m}.$$

Proof. We take $f(h) = \lambda\Delta_{\mathcal{S}}(h) := \lambda(\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) - \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}))$ first use the change of measure inequality (Csiszár, 1975; Donsker and Varadhan, 1976) to state that, for any Q,

$$\mathbb{E}_{h\sim\mathrm{Q}}[f(h)] \leq \mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log\left(\mathbb{E}_{h\sim\mathrm{P}}\left[\exp\left(f(h)\right)\right]\right).$$

Markov's inequality alongside Fubini's theorem gives, with probability at least $1-\delta$,

$$\mathbb{E}_{h\sim\mathrm{Q}}[f(h)] \leq \mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta) + \log\left(\mathbb{E}_{h\sim\mathrm{P}}\mathbb{E}_{\mathcal{S}}\left[\exp\left(f(h)\right)\right]\right).$$

Now, we use that f is L-Lipschitz for all h, then the function $\mathcal{S} \to \Delta_{\mathcal{S}}(h)$ is $\frac{2}{\sqrt{m}}$-Lipschitz on $\mathcal{S}$ for each h. As $\mathcal{D}$ is $\texttt{L-Sob}(c_{LS})$ inequality, $\mathcal{D}^{\otimes m}$ is also $\texttt{L-Sob}(c_{LS})$ with identical constant (Ane *et al.*, 2000, Corollary 3.2.3). Then, using Herbst argument similarly as in Ledoux (2006, Section 2.3) allow us to conclude that f is $2L\sqrt{c_{LS}(\mathcal{D})}$-subgaussian, thus,

$$\log\left(\mathbb{E}_{h\sim\mathrm{P}}\mathbb{E}_{\mathcal{S}}\left[\exp\left(f(h)\right)\right]\right) \leq \frac{2\lambda^2 L^2}{m}.$$

This concludes the proof. ■

Disintegrated PAC-Bayes bounds Numerical estimation of PAC-Bayes bounds is usually challenging as it often involves Monte-Carlo approximations of the expectation over the posterior Q. A recent line of work developed in Chapter 3 and Rivasplata *et al.* (2020) and Viallard *et al.* (2023a) studies *disintegrated PAC-Bayes bounds e.g.*, bounds holding with high-probability on both the dataset $\mathcal{S}$ and a single predictor h drawn from the posterior Q. Those bounds are relevant for practitioners as they require little computational time. However, a drawback of these bounds is that existing disintegrated bounds do not allow the KL divergence to be used as a complexity measure. Either disintegrated KL (Rivasplata *et al.*, 2020) or Rényi divergences

(Viallard *et al.*, 2023a), which can be seen as a relaxation of the KL one, are considered.
Using again the subgaussianity behavior of Lipschitz losses, it is possible to attain PAC-Bayesian disintegrated bounds as long as the posterior distribution satisfies a log-Sobolev inequality with sharp constant (achievable for instance for Gaussian distribution with small operator norm).

Lemma C.2.1. Assume that for any $\mathbf{z}, \ell(.,\mathbf{z})$ is L-Lipschitz and that Q is $\texttt{Poinc}(c_P)$ with $c_P(\mathrm{Q}) \leq 1/m$. Then, with probability $1-\delta$ over the draw of $h \sim \mathrm{Q}$:

$$\Delta_{\mathcal{S}_m}(h) \leq \Delta_{\mathcal{S}_m}(\mathrm{Q}) + \sqrt{\frac{2L^2\log(1/\delta)}{m}}.$$

This lemma states that, as long as we assume our loss to be Lipschitz *w.r.t.* h, then it is possible to easily derive disintegrated PAC-Bayesian bounds. Also notice that Lemma C.2.1 is easily completed by Corollary C.2.1 which makes appear a KL divergence as complexity. Note also that as the loss is Lipschitz, it is also possible to make appear 1-Wasserstein distance through the bounds of Chapters 5 and 6. Thus Having a Log-Sobolev assumption with sharp constant on the posterior distribution is enough to provide disintegrated PAC-Bayesian bounds involving KL or Wasserstein terms instead of Rényi divegerences or disintegrated KL.

C.3 Proofs

C.3.1 Proof of Corollary 4.3.1

To start this proof, we first state an important intermediary theorem, holding at no assumption on the data-distribution.

Theorem C.3.1. For any $C>0$, any $\frac{2}{C} > \lambda > 0$, any data-free prior P, any nonnegative loss function ℓ such that, for any $\mathbf{z} \in \mathcal{Z}, \ell(.,\mathbf{z}) \in \mathrm{H}^1$, and any $\delta \in [0,1]$, we have, with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m>0$, any posterior Q being $\texttt{Poinc}(c_P)$, such that $\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq C$ and such that for any $\mathbf{z}$, $\ell(.,\mathbf{z}) \in \mathrm{H}^1(\mathrm{Q})$:

$$\begin{aligned}\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq & \frac{1}{1-\frac{\lambda C}{2}}\left(\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{KL(\mathrm{Q},\mathrm{P})+\log(1/\delta)}{\lambda m}\right) \\ & + \frac{\lambda}{2-\lambda C}\left(c_P(\mathrm{Q}) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right] + \mathrm{Var}_{\mathbf{z}\sim\mathcal{D}}\left(\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[\ell(h,\mathbf{z})]\right)\right).\end{aligned}$$

Theorem C.3.1 exhibits the influence of the gradient norm of $\nabla_h \ell$ on the generalisation ability: small gradients makes the bound vanish, the remaining variance term is not treated for now and can be assumed bounded, but we cannot then recover a fast rate. We show next that assuming additional assumption over the data distribution circumvent this issue.

Proof. We re-start from CHUGG *et al.* (2023, Corollary 17), for any $\lambda > 0$, with probability at least $1-\delta$, for any $m > 0$, any posterior Q:

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda m} + \frac{\lambda}{2}\left(\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})^2]\right]\right).$$

Then, the last term is controlled as follows,

$$\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})^2]\right] \leq \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[c_P(\mathrm{Q})\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right) + \left(\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[\ell(h,\mathbf{z})]\right)^2\right].$$

We then make appear a supplementary variance term:

$$\begin{aligned}&= \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[c_P(\mathrm{Q})\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right] + \mathrm{Var}_{\mathbf{z}\sim\mathcal{D}}\left(\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[\ell(h,\mathbf{z})]\right)\\ &+ \left(\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[\ell(h,\mathbf{z})]\right)^2.\end{aligned}$$

Note that by Fubini, the last term on the right-hand side is exactly $\mathsf{R}_{\mathcal{D}}(\mathrm{Q})^2$, then using that the averaged true risk is lesser than C, and re-organising the terms in CHUGG *et al.* (2023, Corollary 17) gives, for $\lambda \in \left(0, \frac{2}{C}\right)$:

$$\begin{aligned}\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq \frac{1}{1-\frac{\lambda C}{2}}\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{KL(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda\left(1-\frac{\lambda C}{2}\right)m}\\ + \frac{\lambda}{2-\lambda C}\left(c_P(\mathrm{Q})\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right] + \mathrm{Var}_{\mathbf{z}\sim\mathcal{D}}\left(\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[\ell(h,\mathbf{z})]\right)\right).\end{aligned}$$

■

Now Theorem C.3.1 is proven, we only need to exploit the Poincaré assumption on the data distribution on the variance term to obtain Corollary 4.3.1.

C.3.2 Proof of Theorem 4.3.3

Proof. We start again from Theorem 4.3.2, with $\lambda = 1/C_1$ then have with probability $1-\delta/2$:

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\left(\hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + 2C_1 \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{m}\right) + \frac{c_P(\mathrm{Q})}{C_1} \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right]. \quad \text{(C.1)}$$

We now remark that $g(h,\mathbf{z}) := \|\nabla_h \ell(h,\mathbf{z})\|^2$ is nonnegative. Then, given our assumptions, we apply the route of proof of Theorem 4.3.2 on g *i.e.* we start again from the (CHUGG *et al.*, 2023, Corollary 17), apply Poincaré's inequality on Q and use the `QSB` assumption on g. We then have for any $\lambda > 0$, with probability at least $1-\delta/2$, any Q being $\texttt{Poinc}(c_P)$, $\texttt{QSB}\,(g, C_2)$ and $g(.,\mathbf{z}) \in \mathrm{H}^1(\mathrm{Q})$ for all $\mathbf{z}$:

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right] \leq \mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^m \|\nabla_h \ell(h,\mathbf{z}_i)\|^2\right] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{\lambda m} + \frac{\lambda c_P(\mathrm{Q})}{2} \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h g(h,\mathbf{z})\|^2\right)\right] + \frac{\lambda C_2}{2} \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right]. \quad \text{(C.2)}$$

Finally, notice that, by definition of g, $\nabla_h g(h,\mathbf{z}) = 2Hess_h(\ell)(h,\mathbf{z})\nabla_h \ell(h,\mathbf{z})$, where $Hess_h(\ell)$ denotes the Hessian of ℓ. Thus, using that $\ell(.,\mathbf{z})$ is G gradient Lipschitz for any $\mathbf{z}$ gives, for any $(h,\mathbf{z})$ that $\|\nabla_h g(h,\mathbf{z})\| \leq 2G\|\nabla_h \ell(h,z)\|$. Plugging this in (C.2) gives:

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right] \leq \mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^m \|\nabla_h \ell(h,\mathbf{z}_i)\|^2\right] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{\lambda m} + \frac{\lambda}{2}\left(4c_P(\mathrm{Q})G^2 + C_2\right) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right]. \quad \text{(C.3)}$$

Finally, using that $c_P(\mathrm{Q}) = c$, taking $\lambda = \frac{1}{4cG^2 + C_2}$ and re-organising the terms in

(C.3) gives:

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h\ell(h,\mathbf{z})\|^2\right)\right]\leq 2\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^{m}\|\nabla_h\ell(h,\mathbf{z}_i)\|^2\right]$$
$$+\,2(4cG^2+C_2)\frac{\mathrm{KL}(\mathrm{Q},\mathrm{P})+\log(2/\delta)}{m}\qquad\text{(C.4)}$$

Finally, taking an union bound and plugging (C.4) in (C.1) concludes the proof. ■

C.3.3 Proof of Lemma 4.4.1

Proof. For conciseness, we rename $\mathrm{Q}:=\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}$. We first notice that, denoting by $\frac{d\mathrm{Q}}{d\mathrm{P}}$ the Radon-Nikodym derivative of Q with respect to P:

$$\mathrm{KL}\left(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}},\mathrm{P}\right)=\mathbb{E}_{h\sim\mathrm{Q}}\left[\log\left(\frac{d\mathrm{Q}}{d\mathrm{P}}(h)\right)\right]$$
$$=\mathrm{Ent}_{\mathrm{P}}\left(\frac{d\mathrm{Q}}{d\mathrm{P}}\right)=\mathrm{Ent}_{P}[g^2],$$

where $g=\sqrt{\frac{d\mathrm{Q}}{d\mathrm{P}}}$.
Recall that $\frac{d\mathrm{Q}}{d\mathrm{P}}(h)=\frac{1}{Z}\exp\left(-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right)$ where $Z=\mathbb{E}_{h\sim\mathrm{P}}\left[\exp\left(-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right)\right]$.
Then, $g(h)=\frac{1}{\sqrt{Z}}\exp(-\frac{\gamma}{2}\hat{\mathsf{R}}_{\mathcal{S}_m}(h))$ belongs in $\mathrm{H}^1(\mathrm{P})$ as long as $\ell\in\mathrm{H}^1$. Indeed, as $\exp$ is infinitely smooth, $g\in D_1(\mathbb{R}^d)$, also as the loss is nonnegative, then $g\leq\frac{1}{\sqrt{Z}}$ thus $g\in L^2(\mathrm{P})$. Finally, $\nabla g=-\frac{\gamma}{2}g(h)\nabla\hat{\mathsf{R}}_{\mathcal{S}_m}(h)$. As $g(h)\leq\frac{1}{\sqrt{K}}$, we only need to bound $\|\nabla\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\|^2$ to ensure that $g\in\mathrm{H}^1(\mathrm{P})$:

$$\|\nabla\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\|^2=\frac{1}{m^2}\sum_{1\leq i,j\leq m}\langle\nabla\ell(h,\mathbf{z}_i),\nabla\ell(h,\mathbf{z}_j)\rangle$$
$$\leq\frac{1}{2m^2}\sum_{1\leq i,j\leq m}\|\nabla\ell(h,\mathbf{z}_i)\|^2+\|\nabla\ell(h,\mathbf{z}_j)\|^2$$

As we assumed $\|\nabla\ell(.,\mathbf{z})\|^2\in L^2(\mathrm{P})$ for all $\mathbf{z}$, we conclude that $g\in\mathrm{H}^1$.
We then can apply the log-Sobolev inequality to conlude that

$$\begin{aligned}
\mathrm{KL}\left(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}, \mathrm{P}\right) &\leq c_{LS}(\mathrm{P}) \mathop{\mathbb{E}}_{h\sim\mathrm{P}}[\|\nabla g(h)\|^2] \\
&= \frac{\gamma^2 c_{LS}(\mathrm{P})}{4} \mathop{\mathbb{E}}_{h\sim\mathrm{P}}\left[\|\nabla_h \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\|^2 g^2(h)\right] \\
&= \frac{\gamma^2 c_{LS}(\mathrm{P})}{4} \mathop{\mathbb{E}}_{h\sim\mathrm{P}}\left[\|\nabla_h \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\|^2 \frac{d\mathrm{Q}}{d\mathrm{P}}(h)\right] \\
&= \frac{\gamma^2 c_{LS}(\mathrm{P})}{4} \mathop{\mathbb{E}}_{h\sim\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}}\left[\|\nabla_h \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\|^2\right]
\end{aligned}$$

■

C.3.4 Proof of Theorem 4.4.1

Proof. We start again from Chugg *et al.* (2023, Corollary 17) instantiated with a single λ, *i.i.d.* data and a prior P. Then with probability at least $1-\delta$, for any posterior Q and $m > 0$:

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda m} + \frac{\lambda}{2}\left(\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})^2]\right]\right),$$

where $\mathbf{z}\sim\mathcal{D}$ is independent of $\mathcal{S}$.
For the first inequality, we just take $\lambda = 1$, we use that $\ell(h,\mathbf{z})^2 \leq \ell(h,\mathbf{z})$ and re-organise the terms. Finally, we upper bound the KL term thanks to Lemma 4.4.1.
For the second inequality, we exploit Proposition 4.2.2 to use the fact that $\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}$ is $\texttt{L-Sob}(c_{LS})$ alongside Proposition C.1.1 which ensures that $\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}$ is $\texttt{Poinc}(c_P)$ with constant equal to $c_{LS}\left(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}\right)/2$.
We then apply a route of proof similar to Theorem 4.3.2. We have :

$$\mathop{\mathbb{E}}_{h\sim\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}}\left[\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})^2]\right] = \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathrm{Var}_{h\sim\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}}(\ell(h,\mathbf{z})) + \left(\mathop{\mathbb{E}}_{h\sim\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}}[\ell(h,\mathbf{z})]\right)^2\right]$$

Applying Poincaré's inequality then gives:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{h\sim \mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}} \left[\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,\mathbf{z})^2]\right] \\
&\quad\leq \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[c_P(\mathrm{P})e^{4\|\ell_2\|_\infty} \mathop{\mathbb{E}}_{h\sim \mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}} \left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right) + \left(\mathop{\mathbb{E}}_{h\sim \mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}} [\ell(h,\mathbf{z})]\right)^2\right].
\end{aligned}
$$

Finally, using that $\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}_m}}$ is $\mathtt{QSB}(\ell, C)$ allow us to re-organise the terms as in Theorem 4.3.2. Combining this with Lemma 4.4.1 to bound the KL divergence concludes the proof.

■

C.3.5 Proof of Theorem 4.5.1

Proof. We assume first $G=1$. We start from the Kantorovich duality formula (Villani, 2009, Theorem 5.10) instantiated with the cost function $c(x,y) = \|x-y\|^2$. We have for any Q,P, because W_2 is a distance:

$$
W^2(\mathrm{Q},\mathrm{P}) = W^2(\mathrm{P},\mathrm{Q}) = \sup_{\phi,\psi} \; \mathbb{E}_{h\sim\mathrm{Q}}[\phi(h)] - \mathbb{E}_{h\sim\mathrm{P}}[\psi(h)], \tag{C.5}
$$

where the supremum is taken over the functions $\phi,\psi \in L^1(\mathrm{Q})\times L^1(\mathrm{P})$ such that for all $h,h' \in \mathcal{H}^2$, $\phi(h)-\psi(h') \leq \|h-h'\|^2$.
We claim that if $\phi(h) = f(h) - D\|\nabla f(h)\|$ and $\psi(h') = f(h')$ then the pair Φ,Ψ satisfies $\phi(h)-\psi(h') \leq \frac{\|h-h'\|^2}{2}$.
Indeed,

$$
\begin{aligned}
\phi(h)-\psi(h') &= f(h) - f(h') - D\|\nabla f(h)\| \\
&= f\circ g(1) - f\circ g(0) - D\|\nabla f(h)\|,
\end{aligned}
$$

where $g(t) = th + (1-t)h'$. Then, by the fundamental theorem of calculus, we have

$$
\begin{aligned}
\phi(h)-\psi(h') &= \int_0^1 (f\circ g)'(t)dt - D\|\nabla f(h)\| \\
&= \int_0^1 \left\langle \nabla f\left(th+(1-t)h'\right), h-h'\right\rangle dt - D\|\nabla f(h)\|.
\end{aligned}
$$

We now control the last term using that $\|h-h'\| \leq D$ and Cauchy-Schwarz:

$$\begin{aligned}\phi(h)-\psi(h') &\leq \int_0^1 \langle \nabla f\left(th+(1-t)h'\right), h-h'\rangle\, dt - \langle \nabla f(h), h-h'\rangle \\ &= \int_0^1 \langle \nabla f\left(th+(1-t)h'\right) - \nabla f(h), h-h'\rangle\, dt.\end{aligned}$$

Then by Cauchy-Schwarz alongside Lipschitz gradient,

$$\begin{aligned}\phi(h)-\psi(h') &\leq \|h-h'\| \int_0^1 \|\nabla f\left(th+(1-t)h'\right) - \nabla f(h)\|\, dt \\ &\leq \|h-h'\| \int_0^1 (1-t)dt\, \|h-h'\|\, dt \\ &= \frac{\|h-h'\|^2}{2}.\end{aligned}$$

We then conclude by applying (C.5) to the pair $(2\phi, 2\psi)$. The general case with $G \neq 1$ is immediate when considering the pair $(\frac{2}{G}\phi, \frac{2}{G}\psi)$. ■

C.3.6 Proof of Corollary 4.5.1

Proof. We fix $R > 0$ and we start from Theorem 4.5.1 with predictor space $\mathcal{H}_0 = \mathcal{B}(\mathbf{0}, R)$, f being gradient-Lipschitz on this ball and prior and posterior $\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}$,

$$\mathbb{E}_{h\sim\mathrm{Q}}\left[f\left(\mathcal{P}_R(h)\right)\right] \leq \frac{G}{2}\mathrm{W}_2^2\left(\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}\right) + \mathbb{E}_{h\sim\mathrm{P}}\left[f\left(\mathcal{P}_R(h)\right)\right] + 2R\mathbb{E}_{h\sim\mathrm{Q}}\left[\|\nabla f\left(\mathcal{P}_R(h)\right)\|\right].$$

We first prove that $\mathrm{W}_2^2\left(\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}\right) \leq \mathrm{W}_2^2(\mathrm{Q},\mathrm{P})$. Let $\pi \in \Gamma(\mathrm{Q},\mathrm{P})$ being the optimal transport coupling from P to Q, *i.e.*

$$\mathrm{W}_2^2(\mathrm{Q},\mathrm{P}) = \mathbb{E}_{(X,Y)\sim\pi}\left[\|X-Y\|^2\right].$$

Then notice that if we denote by $\pi_1 = (\mathcal{P}_R, \mathcal{P}_R)\#\pi$, then $\pi_1 \in \Gamma\left(\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}\right)$ and so:

$$\begin{aligned}\mathrm{W}_2^2\left(\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}\right) &\leq \mathbb{E}_{(X,Y)\sim\pi_1}\left[\|X-Y\|^2\right] \\ &= \mathbb{E}_{(X,Y)\sim\pi_1}\left[\|\mathcal{P}_R(X)-\mathcal{P}_R(Y)\|^2\right].\end{aligned}$$

Using that $\mathcal{P}_R$ is 1-Lipschitz gives,

$$\begin{aligned}\mathrm{W}_2^2\left(\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}\right) &\leq \mathbb{E}_{(X,Y)\sim\pi_1}\left[\|X-Y\|^2\right] \\ &= \mathrm{W}_2^2(\mathrm{Q},\mathrm{P}).\end{aligned}$$

Then we need to control $\mathrm{W}_2^2(\mathrm{Q},\mathrm{P})$. To do so, we use the fact that P is $\mathtt{L-Sob}(c_{LS})$ to affirm, through Otto-Villani's theorem (Otto and Villani, 2000, Theorem 1) that the following holds: $\mathrm{W}_2^2(\mathrm{Q},\mathrm{P}) \leq \frac{c_{LS}(\mathrm{P})}{2}\mathrm{KL}(\mathrm{Q},\mathrm{P})$. This concludes the proof. ■

C.3.7 Proof of Theorem 4.5.2

Proof. We start from Theorem 4.5.1, using that $\Delta_{\mathcal{S}_m}$ is G-gradient-Lipschitz for any m to obtain:

$$\mathbb{E}_{h\sim\mathrm{Q}}[\Delta_{\mathcal{S}_m}(h)] \leq \frac{G}{2}\mathrm{W}_2^2(Q,P) + \mathbb{E}_{h\sim\mathrm{P}}[\Delta_{\mathcal{S}_m}(h)] + D\mathbb{E}_{h\sim\mathrm{Q}}[\|\nabla\Delta_{\mathcal{S}_m}(h)\|]$$

The only thing left to control is $\mathbb{E}_{h\sim\mathrm{P}}[\Delta_{\mathcal{S}_m}(h)]$. For this, we use that P alongisde the supermartingale concentration inequality of Chugg *et al.* (2023, Corollary 17) instantiated with prior equal to posterior, *i.i.d.* data showing that, for any $\lambda > 0$, with probability at least $1-\delta$,

$$\mathbb{E}_{h\sim\mathrm{P}}[\Delta_{\mathcal{S}_m}(h)] \leq \frac{\log(1/\delta)}{\lambda} + \frac{\lambda}{2}\mathbb{E}_{h\sim\mathrm{P}}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,z)^2].$$

The last term on the right-hand side is bounded by σ^2 by assumption, then, taking $\lambda = \sqrt{\frac{2\log(1/\delta)}{\sigma^2}}$ gives finally $\mathbb{E}_{h\sim\mathrm{P}}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,z)^2] \leq \sqrt{2\log(1/\delta)/m}$, concludes the proof. ■

Appendix of Chapter 5 D

D.1 Additional background

D.1.1 Background on optimal transport and covering numbers

We recall a basic property on covering numbers.

Proposition D.1.1. For any R, ε, $N(\bar{\mathcal{B}}(0, R)), \varepsilon) \leq \left(1 + \frac{2R}{\varepsilon}\right)^d$.

The following theorem is initially stated in (Villani, 2009, Theorem 5.10).

Theorem D.1.1 (Kantorovich duality)**.** Let $(\mathcal{X}, Q)$ and $(\mathcal{Y}, P)$ be two Polish probability spaces and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous cost function, such that

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad c(x, y) \geq a(x) + b(y)$$

for some real-valued upper semicontinuous functions $a \in L^1(\mathrm{Q})$ and $b \in L^1(\mathrm{P})$. Then there is duality:

$$\min_{\pi \in \Pi(\mathrm{Q},\mathrm{P})} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) \quad = \sup_{\substack{(\psi,\phi) \in L^1(\mathrm{Q}) \times L^1(\mathrm{P}) \\ \phi - \psi \leq c}} \left[\int_{\mathcal{Y}} \phi(y) dP(y) - \int_{\mathcal{X}} \psi(x) d\mathrm{Q}(x) \right],$$

where $L_1(\mathrm{P})$ refers to the set of all functions integrable with respect to P and the condition $\phi - \psi \leq c$ means that for all $x, y \in \mathcal{X} \times \mathcal{Y}, \phi(y) - \psi(x) \leq c(x, y)$.

D.1.2 Technical background for Section 5.3

The theorems of Section 5.3 all rely on a well-chosen radius R (seen here as an hyperparameter) verifying the following set of (non-restrictive) assumptions.

The set of assumptions Rad*.*

We say that $R > 0$ is satisfying $\mathtt{Rad}(\alpha, \beta, M, m, d)$ (abbreviated as Rad when clear from context) for $0 < \alpha \leq \beta$ and $d \in \mathbb{N}/\{0\}, M > 0$ if:

1. $R \geq M + 1$,

2. $R \geq M + \sqrt{2\beta}\sqrt{d\log\left(d\frac{m^{\frac{2}{d}}\sqrt{\beta}}{\sqrt{\pi\alpha}}\right)} = M + \sqrt{2\beta}\sqrt{d\log\left(d\frac{\sqrt{\beta}}{\sqrt{\pi\alpha}}\right) + 2\log(m)}$,

3. $R \geq M + \sqrt{2\beta}\sqrt{1+\frac{d}{2}}$.

> **Remark D.1.1.** Note that $R = \mathcal{O}\max(\sqrt{d\log(d)}, \sqrt{\log(m)})$ when R is the smallest value satisfying Rad.

We state a lemma from Panaretos and Zemel, 2020 which controls the Wasserstein distance between a measure and its projection on a ball.

> **Lemma D.1.1** (Adapted from Panaretos and Zemel, 2020, Equation 2.3). Let $\mathrm{P} \in \mathcal{M}(\mathbb{R}^d)$ and $R > 0$. The 1-Wasserstein distance between P and $\mathcal{P}_R\#\mathrm{P}$ is controlled as follows:
>
> $$\mathrm{W}_1(\mathrm{P}, \mathcal{P}_R\#\mathrm{P}) \leq \int_{||\mathbf{x}||>R} ||\mathbf{x} - \mathcal{P}_R(\mathbf{x})||dP(\mathbf{x}) \leq \int_{||\mathbf{x}||>R} ||\mathbf{x}||dP(\mathbf{x}).$$

Lemma D.1.1 suggests to consider projected distributions and to control them through the residual moments of the norm of gaussian vectors – which is done in the following result.

> **Lemma D.1.2.** For $d \geq 3$, R satisfying Rad, any $\mathrm{Q} = \mathcal{N}(\mu, \Sigma) \in C_{\alpha,\beta,M}$,
>
> $$\mathrm{Q}(||h|| > R) \leq \frac{\beta\sqrt{2\beta}}{m}.$$
>
> Also, for any $\mathrm{Q} \in C_{\alpha,\beta,M}$:
>
> $$\mathrm{W}_1(\mathrm{Q}, \mathcal{P}_R\#\mathrm{Q}) \leq \mathbb{E}_{h\sim\mathrm{Q}}\left[||h||\ \mathbb{1}(||h|| > R)\right] \leq (M+1)\frac{\beta\sqrt{2\beta}}{m}.$$
>
> Finally:
>
> $$\mathbb{E}_{h\sim\mathrm{Q}}\left[||h||^2\ \mathbb{1}(||h|| > R)\right] \leq (M+1)^2\frac{\beta\sqrt{2\beta}}{m}.$$

The proof of Lemma D.1.2 is gathered in Appendix D.2.2.

D.1.3 Differential privacy background

Definition D.1.1 (Probability kernels)**.** A *probability kernel* $\mathcal{P}$ from $\mathcal{Z}^m$ to $\mathcal{M}(\mathcal{H})$ is defined as a mapping $\mathcal{P} : Z^m \to \mathcal{M}(\mathcal{H})$.

Definition D.1.2. A probability kernel $\mathcal{P} : \mathcal{Z}^m \to T$ is (ε, γ)-differentially private if, for all pairs $\mathcal{S}_m, \mathcal{S}'_m \in Z^m$ that differ at only one coordinate, and all measurable subsets $B \in \Sigma_{\mathcal{H}}$, we have

$$\mathbb{P}\{\mathcal{P}(\mathcal{S}_m) \in B\} \leq \mathrm{e}^{\varepsilon} \mathbb{P}\left\{\mathcal{P}\left(\mathcal{S}'_m\right) \in B\right\} + \gamma.$$

Further, ε-differentially private means $(\varepsilon, 0)$-differentially private.

Remark D.1.2. Note that classically, differential privacy do not consider stochastic kernels but *randomised algorithms*. Note that this is equivalent to consider probability kernels as precised in Dziugaite and Roy (2018b, footnote 3, Appendix A).

For our purposes, max-information is the key quantity controlled by differential privacy.

Definition D.1.3 (Dwork *et al.* (2015), paragraph 3)**.** Let $\beta \geq 0$, let X and Y be random variables in arbitrary measurable spaces, and let X' be independent of Y and equal in distribution to X. The *β-approximate max-information* between X and Y, denoted $I_{\infty}^{\beta}(X; Y)$, is the least value k such that, for all product-measurable events E,

$$\mathbb{P}\{(X, Y) \in E\} \leq e^{k} \mathbb{P}\left\{\left(X', Y\right) \in E\right\} + \beta.$$

The max-information $I_{\infty}(X; Y)$ is defined to be $I_{\infty}^{\beta}(X; Y)$ for $\beta = 0$. For $m \in \mathbb{N}$ and stochastic kernel $\mathcal{P} : \mathcal{Z}^m \to \mathcal{M}(\mathcal{Z})$, the β-approximate max-information of $\mathcal{P}$, denoted $I_{\infty}^{\beta}(\mathcal{P}, m)$, is the least value k such that, for all $\mu \in \mathcal{M}_1(\mathcal{Z}), I_{\infty}^{\beta}(\mathcal{S}_m; \mathcal{P}(\mathcal{S}_m)) \leq k$ when $\mathcal{S}_m \sim \mathcal{D}^m$. The max-information of $\mathcal{P}$ is defined similarly.

Dziugaite and Roy (2018b) exploited a boundedness assumption to control the exponential mechanism of McSherry and Talwar (2007). This ensures that the Gibbs posterior $\mathcal{P}(\mathcal{S}_m) = P_{-\lambda m \hat{\mathrm{R}}_{\mathcal{S}_m}}$ is ε-diffrentially private for ε given in Dziugaite and Roy (2018b, Corollary 5.2). Here, we use a theorem from Minami *et al.* (2016) to ensure that for uniformly Lipschitz losses (possibly unbounded), the Gibbs posterior remain (ε, γ)-differentially private.

Proposition D.1.2 (MINAMI *et al.* (2016), Corollary 8)**.** Assume $\mathcal{H} = \mathbb{R}^d$. Assume the loss function to be convex and satisfying **(A1)**. Finally assume that the (data-free) distribution P is such that $-\log P(.)$ is twice differentiable and m_P-strongly convex. Let $\varepsilon > 0, 0 < \gamma < 1$. Take $\lambda > 0$ such that

$$\lambda \leq \frac{\varepsilon}{2K}\sqrt{\frac{m_P}{1 + 2\log\left(\frac{1}{\gamma}\right)}}.$$

Then the probability kernel $\mathcal{P} : \mathcal{S}_m \to P_{-\lambda m \hat{\mathsf{R}}_{\mathcal{S}_m}}$ is (ε, γ)-differentially private.

Note that, as we mainly focus on Gaussian priors lying on the compact $C_{\alpha,\beta,M}$, the condition on P will always be satisfied with $m_P \geq \alpha$. The last result in this appendix is Theorem 3.1 of ROGERS *et al.* (2016) which upper bounds the β-approximate max-information of any (ε, γ) differentially private probability kernel.

Proposition D.1.3. Let $\mathcal{P} : \mathcal{Z}^n \to \mathcal{M}(\mathcal{H})$ be an (ϵ, γ)-differentially private probability kernel for $\epsilon \in (0, 1/2]$ and $\gamma \in (0, \epsilon)$. For $\beta = e^{-\epsilon^2 m} + O\left(m\sqrt{\frac{\gamma}{\epsilon}}\right)$, we have

$$I_\infty^\beta(\mathcal{P}, m) = O\left(\epsilon^2 m + m\sqrt{\frac{\gamma}{\epsilon}}\right).$$

D.2 Additional proofs

D.2.1 Proof of Theorem 5.2.2

We fix $\lambda > 0$.

Step 1: define a good data-dependent function. We define, for any sample $\mathcal{S}_m$ and predictor $h \in \mathcal{H}$

$$f_{\mathcal{S}_m}(h) = \lambda \Delta {\mathcal{S}_m}^2(h).$$

This function satisfies the following lemma.

Lemma D.2.1. We fix

$$\varepsilon = \frac{1}{m}, \quad \lambda^{-1} = K\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}}\left(\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}} + 2K\varepsilon\right),$$

with $N = N(\mathcal{H}, \varepsilon)$ the ε-covering number of $\mathcal{H}$. We then have with probability $1 - 2\delta$ for all $h, h' \in \mathcal{H}$:

$$f_{\mathcal{S}_m}(h) - f_{\mathcal{S}_m}(h') \leq \varepsilon_m + ||h - h'||,$$

with $\varepsilon_m = \frac{4}{\log(\frac{1}{\delta})}\left(2+\sqrt{\frac{\log\left(\frac{1}{\delta}\right)+2d\log(1+2Rm)}{2m}}\right) = \mathcal{O}\left(1+\sqrt{\frac{d}{m}}\right)$.

Proof of Lemma D.2.1. We rename $N := N(\mathcal{H}, \varepsilon)$. For any $h, h' \in \mathcal{H}^2$, we have:

$$f_{\mathcal{S}_m}(h) - f_{\mathcal{S}_m}(h') = \lambda\left(\Delta_{\mathcal{S}_m}(h) - \Delta_{\mathcal{S}_m}(h')\right).\left(\Delta_{\mathcal{S}_m}(h) + \Delta_{\mathcal{S}_m}(h')\right).$$

The proof of Lemma 5.2.1 gives with probability at least $1-\delta$, for any $h, h' \in \mathcal{H}^2$,

$$\lambda(\Delta_{\mathcal{S}_m}(h) - \Delta_{\mathcal{S}_m}(h') \leq 4\lambda K\varepsilon + \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\lambda K\left(2\varepsilon + ||h-h||\right).$$

Thus with probability $1-\delta$:

$$f_{\mathcal{S}_m}(h) - f_{\mathcal{S}_m}(h') \leq \left(4\lambda K\varepsilon + \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\lambda K\left(2\varepsilon + ||h-h||\right)\right).\left(2\sup_{h\in K}\Delta_{\mathcal{S}_m}(h)\right)$$

$$= \lambda\left(2K\varepsilon\left(2+\sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\right) + K\sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}||h-h'||\right).\left(2\sup_{h\in K}\Delta_{\mathcal{S}_m}(h)\right).$$

Because $\mathcal{H}$ is compact and ℓ is K-lipschitz, $\Delta\mathcal{S}_m$ is continuous so there exists $h_{\mathcal{S}_m}$ such that $\sup_{h\in\mathcal{H}}\Delta_{\mathcal{S}_m}(h) = \Delta_{\mathcal{S}_m}(h_{\mathcal{S}_m})$.

We consider an ε-covering $C := \{h_1, ..., h_N\}$ of $\mathcal{H}$ of size N. Thus, there exists $h_0 \in C$ such that $||h_{\mathcal{S}_m} - h_0|| \leq \varepsilon$. Furthermore, because $\ell \in [0,1]$, by Hoeffding inequality applied for every $h \in C$ and an union bound, we have with probability at least $1-\delta$, for all $h \in C$:

$$\Delta_{\mathcal{S}_m}(h) \leq \sqrt{\frac{\log(\frac{N}{\delta})}{2m}} \leq \sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}}.$$

Finally using that $\Delta\mathcal{S}_m$ is $2K$-Lipschitz gives with probability at least $1-\delta$:

$$\begin{aligned}\sup_{h\in K}\Delta_{\mathcal{S}_m}(h) = \Delta_{\mathcal{S}_m}(h_{\mathcal{S}_m}) &= \Delta_{\mathcal{S}_m}(h_0) + \left(\Delta_{\mathcal{S}_m}(h_{\mathcal{S}_m}) - \Delta_{\mathcal{S}_m}(h_0)\right)\\ &\leq \sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}} + 2K\varepsilon.\end{aligned}$$

So finally, with probability $1-2\delta$, we have, for any $h,h'\in\mathcal{H}^2$:

$$\frac{1}{\lambda}\left(f_{\mathcal{S}_m}(h)-f_{\mathcal{S}_m}(h')\right)$$
$$\leq\left(2K\varepsilon\left(2+\sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\right)+K\sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}||h-h'||\right)\times 2\left(\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}}+2K\varepsilon\right).$$

Taking $\lambda^{-1}=2K\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}}\left(\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}}+2K\varepsilon\right)$ gives:

$$\begin{aligned}
f_{\mathcal{S}_m}(h)-f_{\mathcal{S}_m}(h') &\leq \frac{2\varepsilon\left(2+\sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\right)}{\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}}\left(\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}}+2K\varepsilon\right)}+||h-h'|| \\
&\leq \frac{4m\varepsilon}{\log\left(\frac{N^2}{\delta}\right)}\left(2+\sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\right)+||h-h'|| \\
&\leq \frac{4}{\log(\frac{1}{\delta})}\left(2+\sqrt{\frac{\log\left(\frac{1}{\delta}\right)+2d\log(1+2Rm)}{2m}}\right)+||h-h'||.
\end{aligned}$$

The last line holds as $N\geq 1$ and that $N\leq N(\bar{\mathcal{B}}(0,R)),\varepsilon)\leq\left(1+\frac{2R}{\varepsilon}\right)^d$ thanks to Proposition D.1.1 ($\varepsilon=1/m$). This proves the lemma. ■

Step 2: A probabilistic change of measure inequality for $f_{\mathcal{S}_m}$. We do not have for the Wasserstein distance such a powerful tool than the change of measure inequality. However, we can generate a probabilistic surrogate on $\mathcal{P}_1(\mathcal{H})$ valid for the function $f_{\mathcal{S}_m}$ described below.

Lemma D.2.2. For any λ,ε_m defined as in Lemma D.2.1, any $\delta>0$, we have with probability $1-2\delta$ over the sample $\mathcal{S}_m$, for any $\mathrm{P}\in\mathcal{P}_1(\mathcal{H})$:

$$\left(\sup_{Q\in\mathcal{P}_1(\mathcal{H})}\mathbb{E}_{h\sim\mathrm{Q}}[f_{\mathcal{S}_m}(h)]-\varepsilon_m-\mathrm{W}_1(\mathrm{Q},\mathrm{P})\right)\leq\mathbb{E}_{h\sim\mathrm{P}}[f_{\mathcal{S}_m}(h)].$$

Proof Proof of Lemma D.2.2. For any $\varepsilon > 0$, we introduce the cost function $c_\varepsilon(x,y) = \varepsilon + ||x-y||$. From this we notice that we can rewrite the $\varepsilon, 1$- Wasserstein distance introduced in Definition 5.1.1 the same way we did in Lemma 5.2.2. This leads to

$$W_\varepsilon(\mathrm{Q},\mathrm{P}) = \sup_{\substack{(\psi,\phi)\in L^1(\mathrm{Q})\times L^1(\mathrm{P}) \\ \psi-\phi\leq c_\varepsilon}} \left[\mathbb{E}_{h\sim\mathrm{Q}}[\psi(h)] - \mathbb{E}_{h\sim\mathrm{P}}[\phi(h)]\right].$$

A crucial point is that for a well-chosen λ with high probability, the pair $(f_{\mathcal{S}_m}, f_{\mathcal{S}_m})$ satisfies the condition stated under the last supremum. It is formalised in the lemma below.

> **Lemma D.2.3.** Given our choices of λ, ε_m, we have with probability at least $1-2\delta$ over the sample $\mathcal{S}_m$ that, for all measures $\mathrm{Q},\mathrm{P} \in \mathcal{P}_1(\mathcal{H})^2$:
>
> - $f_{\mathcal{S}_m} \in L_1(\mathrm{Q}), L_1(\mathrm{P})$,
> - for all $h,h' \in \mathcal{H}^2, f_{\mathcal{S}_m}(h) - f_{\mathcal{S}_m}(h') \leq c_{\varepsilon_m}(h,h')$.
>
> Thus, Kantorovich duality gives us:
>
> $$\left(\sup_{Q\in\mathcal{P}_1(\mathcal{H})} \mathbb{E}_{h\sim\mathrm{Q}}[f_{\mathcal{S}_m}(h)] - W_{\varepsilon_m}(\mathrm{Q},\mathrm{P})\right) \leq \mathbb{E}_{h\sim\mathrm{P}}[f_{\mathcal{S}_m}(h)],$$
>
> and using $W_{\varepsilon_m} = \varepsilon_m + \mathrm{W}_1$ concludes the proof.

Proof of Lemma D.2.3. Because our space of predictors is compact and that for any $\mathbf{z} \in \mathcal{Z}$, the loss function $\ell(.,\mathbf{z})$ is K-lipschitz on $\mathcal{H}$, then both the generalisation and empirical risk are continuous on $\mathcal{H}$. Thus $|f_{\mathcal{S}_m}|$ is also continuous and, by compacity, reaches its maximum M_S on $\mathcal{H}$. Thus for any probability P on $K, \mathbb{E}_{h\sim\mathrm{P}}[|f_{\mathcal{S}_m}(h)|] \leq M_S < +\infty$ almost surely. This proves the first statement. We notice that the second bullet, given our choice of λ, is the exact conclusion of Lemma D.2.1 with probability at least $1-2\delta$. So with probability at least $1-2\delta$, Kantorovich duality gives us that for any P,Q

$$\mathbb{E}_{h\sim\mathrm{Q}}[f_{\mathcal{S}_m}(h)] - \mathbb{E}_{h\sim\mathrm{P}}[f_{\mathcal{S}_m}(h)] \leq W_{\varepsilon_m}(\mathrm{Q},\mathrm{P}).$$

Re-organising the terms and taking the supremum over Q concludes the proof. ■

This concludes the proof of Lemma D.2.2. ■

Step 3: The PAC-Bayes proof for the 1-Wasserstein distance.
We start by exploiting lemma D.2.2: for any prior $\mathrm{P} \in \mathcal{P}_1(\mathcal{H})$, for λ, ε_m defined as in Lemma D.2.1, with probability at least $1-2\delta$ we have:

$$\left(\sup_{Q \in \mathcal{P}_1(\mathcal{H})} \mathbb{E}_{h\sim \mathrm{Q}}[f_{\mathcal{S}_m}(h)] - \varepsilon_m - \mathrm{W}_1(\mathrm{Q},\mathrm{P}) \right) \leq \mathbb{E}_{h\sim \mathrm{P}}[f_{\mathcal{S}_m}(h)].$$

We then notice that by Jensen's inequality

$$\mathbb{E}_{h\sim \mathrm{P}}[f_{\mathcal{S}_m}(h)] \leq \frac{\lambda}{2(m-1)} \log \left(\mathbb{E}_{h\sim \mathrm{P}}[\exp(2(m-1)\Delta{\mathcal{S}_m}^2(h))] \right).$$

Then, by Markov's inequality we have with probability $1-\delta$:

$$\mathbb{E}_{h\sim \mathrm{P}}[f_{\mathcal{S}_m}(h)] \leq \frac{\lambda}{2(m-1)} \log \left(\frac{\mathbb{E}_{\mathcal{S}}\mathbb{E}_{h\sim \mathrm{P}}\left[\exp\left(2(m-1)\Delta{\mathcal{S}_m}^2(h)\right)\right]}{\delta} \right).$$

By Fubini and Lemma 5 of McAllester (2003a), we have

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{h\sim \mathrm{P}}\left[\exp(f_{\mathcal{S}_m}(h))\right] \leq m.$$

Taking an union bound and dividing by λ gives with probability $1-3\delta$, for any posterior Q

$$\mathbb{E}_{h\sim \mathrm{Q}}[\Delta{\mathcal{S}_m}^2(h)] \leq \frac{\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \varepsilon_m}{\lambda} + \frac{\log\left(\frac{m}{\delta}\right)}{2(m-1)}.$$

We also remark that we can upper bound λ:

$$\begin{aligned}\lambda^{-1} &= 2K\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}} \left(\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}} + \frac{2K}{m} \right) \\ &\leq 2K(2K+1)\frac{\log(\frac{1}{\delta}) + 2d\log(1+2Rm)}{2m}.\end{aligned}$$

The last line holding because $1/m \leq \sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}}$. Also $N = N(\mathcal{H}, 1/m) \leq (1+2Rm)^d$ thanks to proposition D.1.1. Then, bounding $1/2m, 1/(2m-1)$ by $1/m$ gives, with probability at least $1-3\delta$, for any posterior Q

$$\mathbb{E}_{h\sim \mathrm{Q}}[\Delta{\mathcal{S}_m}^2(h)] \leq 2K(2K+1)\frac{\log(\frac{1}{\delta}) + 2d\log(1+2Rm)}{m}\left(\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \varepsilon_m\right) + \frac{\log\left(\frac{m}{\delta}\right)}{m}.$$

We finally exploit Jensen's inequality once more to remark that for any Q, $\mathbb{E}_{h\sim \mathrm{Q}}[\Delta{\mathcal{S}_m}^2(h)] \geq \left(\mathbb{E}_{h\sim \mathrm{Q}}[\Delta_{\mathcal{S}_m}(h)]\right)^2$. Then, with probability at least $1-3\delta$, for any posterior Q

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq \sqrt{2K(2K+1)\frac{2d\log\left(\frac{1+2Rm}{\delta}\right)}{m}\left(\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \varepsilon_m\right) + \frac{\log\left(\frac{m}{\delta}\right)}{m}}$$

Taking $\delta' = \delta/3$ concludes the proof.

D.2.2 Proof of Lemma D.1.2

Proof of Lemma D.1.2. We denote by $\mathbf{x}$ a vector of $\mathbb{R}^d$, by $d\mathbf{x} = d_{x_1}...dx_d$ the Lebesgue measure on $\mathbb{R}^d$ and $f_{\mu,\Sigma}(\mathbf{x}) = \exp\left(\frac{1}{2}(\mathbf{x}^T - m)\Sigma^{-1}(\mathbf{x} - m)\right)$.
First bound. First we use that $||\mu|| \leq M$ to say that $\bar{\mathcal{B}}(0_{\mathbb{R}^d}, R-M) \subseteq \bar{\mathcal{B}}(-m, R)$ and so:

$$\sqrt{(2\pi)^d|\Sigma|}.\mathrm{Q}(||x|| > R) = \int_{||\mathbf{x}||>R} f_{\mu,\Sigma}(\mathbf{x})d\mathbf{x} \leq \int_{||\mathbf{x}||>R-M} f_{0,\Sigma}(\mathbf{x})d\mathbf{x}$$

where $|\Sigma|$ the determinant of Σ. We now use that because $\mathrm{Q} \in C_{\alpha,\beta,M}, \alpha Id \preceq \Sigma \preceq \beta Id$. We then have: $|\Sigma| \geq \alpha^d$ and for any $\mathbf{x}$, $\mathbf{x}^T\Sigma^{-1}\mathbf{x} \geq ||\mathbf{x}||^2/\beta$. Thus we have:

$$\mathrm{Q}(||h|| > R) = \frac{1}{\sqrt{2\pi\alpha}^d} \int_{||\mathbf{x}||>(R-M)} \exp\left(\frac{1}{2\beta}||\mathbf{x}||^2\right) d\mathbf{x}$$

We use the hyperspherical coordinate (see *e.g.* Blumenson, 1960) to obtain:

$$\begin{aligned}\int_{||\mathbf{x}||>(R-M)} \exp\left(\frac{1}{2\beta}||\mathbf{x}||^2\right) d\mathbf{x} &= \int_{R-M}^{+\infty} r^{d-1} \exp\left(-\frac{r^2}{2\beta}\right) dr \\ &\leq \int_{R-M}^{+\infty} r^{d+1} \exp\left(-\frac{r^2}{2\beta}\right) dr \\ &= \beta\sqrt{2\beta}^{d+1} \int_{\frac{(R-M)^2}{2\beta}}^{+\infty} r^{\frac{d}{2}} \exp^{-r} dr.\end{aligned}$$

The second line holding because we assumed $R-M \geq 1$ thanks to Rad. We define the *residual of Euler's Gamma function* as: $\Gamma\left(1+\frac{d}{2}, \frac{(R-M)^2}{2\beta}\right) := \int_{\frac{(R-M)^2}{2\beta}}^{+\infty} r^{\frac{d}{2}} \exp^{-r} dr$.
Then we use Gabcke (1979, Lemma 4.4.3, p.84) which ensure us that (because point 3 of Rad gives $\frac{(R-M)^2}{2\beta} \geq 1 + \frac{d}{2}$):

$$\Gamma\left(1+\frac{d}{2}, \frac{(R-M)^2}{2\beta}\right) \leq \frac{d+2}{2} \exp\left(-\frac{(R-M)^2}{2\beta}\right) \left(\frac{(R-M)^2}{2\beta}\right)^{\frac{d}{2}}.$$

We now control this quantity through the following lemma.

Lemma D.2.4. Let $d \geq 3$, $f(r) = \frac{d}{2}\log(r) - r$ Then for any $r = \frac{(R-M)^2}{2\beta}$ with R satisfying Rad, we have :

$$f(r) \leq -\frac{d}{2}\log\left(\sqrt{\frac{\beta}{\pi\alpha}}\right) - \log(m) - \log\left(\frac{d+2}{2}\right).$$

The proof of Lemma D.2.4 lies at the end of this section. We then have

$$\exp\left(-\frac{(R-M)^2}{2\beta}\right)\left(\frac{(R-M)^2}{2\beta}\right)^{\frac{d}{2}}$$
$$= \exp\left(f\left(\frac{(R-M)^2}{2\beta}\right)\right) \leq \sqrt{\frac{\pi\alpha}{\beta}}^d \times \frac{2}{d+2} \times \frac{1}{m}.$$

Hence the final bound:

$$\mathrm{Q}(||h|| > R) \leq \frac{\beta\sqrt{2\beta}}{m}.$$

Second bound. We use lemma D.1.1 to have

$$\mathrm{W}_1(\mathrm{Q}, \mathcal{P}_R\#\mathrm{Q}) \leq \int_{||\mathbf{x}||>R} ||\mathbf{x} - \mathcal{P}_R(\mathbf{x})|| dP(\mathbf{x}).$$

By definition of the projection on a closed convex, $||\mathbf{x} - \mathcal{P}_R(\mathbf{x})|| \leq ||\mathbf{x}||$. Thus:

$$\begin{aligned}
&\leq \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \int_{||\mathbf{x}||>R} ||\mathbf{x}|| f_{\mu,\Sigma}(\mathbf{x}) d\mathbf{x} \\
&\leq \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \int_{||\mathbf{x}||>R} ||\mathbf{x} - \mu|| f_{\mu,\Sigma}(\mathbf{x}) d\mathbf{x} + M\mathrm{Q}(||h|| > R) \\
&\leq \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \int_{||\mathbf{x}||>R} ||\mathbf{x} - \mu|| f_{\mu,\Sigma}(\mathbf{x}) d\mathbf{x} + \frac{M\beta\sqrt{2\beta}}{m}.
\end{aligned}$$

The last line holding thanks to the first part of the proof, then using again that $||\mu|| \leq M$ gives:

$$\mathrm{W}_1(\mathrm{Q}, \mathcal{P}_R\#\mathrm{Q}) \leq \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \int_{||\mathbf{x}||>R-M} ||\mathbf{x}|| f_{0,\Sigma}(\mathbf{x}) d\mathbf{x} + \frac{M\beta\sqrt{2\beta}}{m}.$$

Then using the same arguments than in the first part of the proof gives:

$$\mathrm{W}_1(\mathrm{Q}, \mathcal{P}_R\#\mathrm{Q}) \leq \frac{1}{\sqrt{2\pi\alpha}^d}\int_{||\mathbf{x}||>R-M} ||\mathbf{x}|| \exp\left(-\frac{||\mathbf{x}||^2}{2\beta}\right) d\mathbf{x} + \frac{M\beta\sqrt{2\beta}}{m}.$$

We use the hyperspherical coordinate to obtain:

$$\begin{aligned}\int_{||\mathbf{x}||>R-M} ||\mathbf{x}|| \exp\left(-\frac{||\mathbf{x}||^2}{2\beta}\right) d\mathbf{x} &= \int_{R-M}^{+\infty} r^d \exp\left(-\frac{r^2}{2\beta}\right) dr \\ &\leq \int_{R-M}^{+\infty} r^{d+1} \exp\left(-\frac{r^2}{2\beta}\right) dr \\ &= \beta\sqrt{2\beta}^{d+1} \int_{\frac{(R-M)^2}{2\beta}}^{+\infty} r^{\frac{d}{2}} \exp^{-r} dr \\ &= \beta\sqrt{2\beta}^{d+1} \Gamma\left(\frac{d+1}{2}, \frac{(R-M)^2}{2\beta}\right).\end{aligned}$$

The second line holding because $R - M \geq 1$. Then applying again Lemma D.2.4 gives:

$$\begin{aligned}\mathbb{E}_{h\sim\mathrm{Q}}\left[||h||\ \mathbb{1}(||h|| > R)\right] &\leq \beta\sqrt{2\beta}\sqrt{\frac{\beta}{\pi\alpha}}^d \times \frac{d+2}{2}\sqrt{\frac{\pi\alpha}{\beta}}^d \times \frac{2}{d+2} \times \frac{1}{m} + \frac{M\beta\sqrt{2\beta}}{m} \\ &= (M+1)\frac{\beta\sqrt{2\beta}}{m}.\end{aligned}$$

Hence the final bound:

$$\mathrm{W}_1(\mathrm{Q}, \mathcal{P}_R\#\mathrm{Q}) \leq \mathbb{E}_{h\sim\mathrm{Q}}\left[||h||\ \mathbb{1}(||h|| > R)\right] \leq (M+1)\frac{\beta\sqrt{2\beta}}{m}.$$

Third bound. We start again as

$$\begin{aligned}\mathbb{E}_{h\sim\mathrm{Q}}[||h||^2\mathbb{1}(||h|| > R)] &= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{||\mathbf{x}||>R} ||\mathbf{x}||^2 f_{\mu,\Sigma}(\mathbf{x})d\mathbf{x} \\ &= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{||\mathbf{x}||>R} ||\mathbf{x}-\mu||^2 + 2\langle\mu, \mathbf{x}-\mu\rangle + ||\mu||^2 f_{\mu,\Sigma}(\mathbf{x})d\mathbf{x}.\end{aligned}$$

Then, using that μ is the mean of Q and that $||\mu|| \leq M$ gives:

$$\begin{aligned}\mathbb{E}_{h\sim\mathrm{Q}}[||h||^2\mathbb{1}(||h||>R)] \leq \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{||\mathbf{x}||>R}||\mathbf{x}-\mu||^2 f_{\mu,\Sigma}(\mathbf{x})d\mathbf{x}\\ + 2M\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{||\mathbf{x}||>R}||\mathbf{x}-\mu||f_{\mu,\Sigma}(\mathbf{x})d\mathbf{x} + M^2\mathrm{Q}(||h||>R).\end{aligned}$$

Then, the first and second bounds of lemma D.1.2 give

$$\mathbb{E}_{h\sim\mathrm{Q}}[||h||^2\mathbb{1}(||h||>R)] \leq \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{||\mathbf{x}||>R}||\mathbf{x}-\mu||^2 f_{\mu,\Sigma}(\mathbf{x})d\mathbf{x} + (M^2+2M)\frac{\beta\sqrt{2\beta}}{m}.$$

Finally,

$$\begin{aligned}&\mathbb{E}_{h\sim\mathrm{Q}}[||h||^2\mathbb{1}(||h||>R)]\\ &\quad\leq \frac{1}{\sqrt{2\pi\alpha}^d}\int_{||\mathbf{x}||>R-M}||\mathbf{x}||^2\exp\left(-\frac{||\mathbf{x}||^2}{2\beta}\right)d\mathbf{x} + (M^2+2M+2)\frac{\beta\sqrt{2\beta}}{m}.\end{aligned}$$

We use the hyperspherical coordinate to obtain:

$$\begin{aligned}\int_{||\mathbf{x}||>R-M}||\mathbf{x}||^2\exp\left(-\frac{||\mathbf{x}||^2}{2\beta}\right)d\mathbf{x} &= \int_{R-M}^{+\infty} r^{d+1}\exp\left(-\frac{r^2}{2\beta}\right)dr\\ &= \beta\sqrt{2\beta}^{d+1}\int_{\frac{(R-M)^2}{2\beta}}^{+\infty} r^{\frac{d}{2}}\exp^{-r}dr\\ &= \beta\sqrt{2\beta}^{d+1}\Gamma\left(\frac{d+1}{2},\frac{(R-M)^2}{2\beta}\right).\end{aligned}$$

Then applying again Lemma D.2.4 gives:

$$\begin{aligned}\mathbb{E}_{h\sim\mathrm{Q}}\left[||h||^2\ \mathbb{1}(||h||>R)\right] &\leq \frac{\beta\sqrt{2\beta}}{m} + (M^2+2M)\frac{\beta\sqrt{2\beta}}{m}\\ &= (M+1)^2\frac{\beta\sqrt{2\beta}}{m}.\end{aligned}$$

This concludes the proof.

■

Proof of Lemma D.2.4. First of all, f is decreasing on $[\frac{d}{2}, +\infty)$. Notice that if $r_0 = d\log\left(d\frac{m^{\frac{2}{d}}\sqrt{\beta}}{\sqrt{\pi\alpha}}\right)$, then $r_0 \geq \frac{d}{2}$ because $d \geq 3$. Thus, $r = \frac{(R-M)^2}{2\beta}$, with R satisfying Rad. We then know that $r \geq r_0$ so $f(r) \leq f(r_0)$. The only thing left to prove is that

$$f(r_0) \leq -\frac{d}{2}\log\left(\sqrt{\frac{\beta}{\pi\alpha}}\right) - \log(m) - \log\left(\frac{d+2}{2}\right).$$

To do so, notice that:

$$\log(r_0) = \log(d) + \log\left(\log\left(dm^{\frac{2}{d}}\sqrt{\frac{\beta}{\pi\alpha}}\right)\right).$$

So, multiplying by $d/2$ gives:

$$\frac{d}{2}\log(r_0) = -\frac{d}{2}\log\left(m^{\frac{2}{d}}\sqrt{\frac{\beta}{\pi\alpha}}\right) + \frac{r_0}{2} + \frac{d}{2}\log\log\left(dm^{\frac{2}{d}}\sqrt{\frac{\beta}{\pi\alpha}}\right).$$

Finally:

$$f(r_0) = -\frac{d}{2}\log\left(m^{\frac{2}{d}}\sqrt{\frac{\beta}{\pi\alpha}}\right) - \frac{r_0}{2} + \frac{d}{2}\log\log\left(dm^{\frac{2}{d}}\frac{\beta}{\pi\alpha}\right)$$

We conclude the proof by proving

$$-\frac{r_0}{2} + \frac{d}{2}\log\log\left(dm^{\frac{2}{d}}\sqrt{\frac{\beta}{\pi\alpha}}\right) \leq -\log\left(\frac{d+2}{2}\right).$$

Note that this is equivalent to

$$dm^{\frac{2}{d}}\sqrt{\frac{\beta}{\pi\alpha}} - \left(1+\frac{d}{2}\right)^{\frac{2}{d}}\log\left(dm^{\frac{2}{d}}\sqrt{\frac{\beta}{\pi\alpha}}\right) \geq 0.$$

This is true because for $d \geq 3$, $\left(1+\frac{d}{2}\right)^{\frac{2}{d}} \leq 2$ and the function $\mathbb{R}^+$, $x \to x - 2\log(x)$ is positive. This concludes the proof. ■

D.2.3 Proof of Theorem 5.3.2

Proof of Theorem 5.3.2. We take a specific radius R which is the smallest value satisfying Rad. We first notice that because for all z, $\ell(.,\mathbf{z})$ is L-smooth, then on $\mathcal{B}(0,R)$, the gradients of $\ell(.,\mathbf{z})$ are bounded by $D_R = D + LR$. Thus ℓ is uniformly D_R-Lipschitz on the closed ball of radius R. This allow us a straightforward application of Theorem 5.2.2 on the compact $\mathcal{B}(0,R)$, with the prior $\mathcal{P}_R\#\mathrm{P}$, and with high probability, for any posterior $\mathcal{P}_R\#\mathrm{Q}$ with $\mathrm{Q}\in C_{\alpha,\beta,M}$:

$$|\Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| \leq \sqrt{2D_R(2D_R+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(\mathrm{W}_1(\mathcal{P}_R\#\mathrm{Q},\mathcal{P}_R\#\mathrm{P})+\varepsilon_m\right)+\frac{\log\left(\frac{3m}{\delta}\right)}{m}}.$$

From this we control the left hand-side term as follows:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq |\Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| + |\Delta_{\mathcal{S}_m}(\mathrm{Q}) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})|$$

And we also have as in the proof of Theorem 5.3.1:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q}) - \Delta_{\mathcal{S}_m}(\mathcal{P}_R\#\mathrm{Q})| \leq 2\mathrm{Q}(||h|| > R) \leq 2\frac{\beta\sqrt{2\beta}}{m}.$$

Also we have by the triangle inequality:

$$\mathrm{W}_1(\mathcal{P}_R\#\mathrm{Q},\mathcal{P}_R\#\mathrm{P}) \leq \mathrm{W}_1(\mathrm{Q},\mathcal{P}_R\#\mathrm{Q}) + \mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \mathrm{W}_1(\mathrm{P},\mathcal{P}_R\#\mathrm{P}).$$

Because both $\mathrm{Q},\mathrm{P}\in C_{\alpha,\beta,M}$, using again Lemma D.1.2 gives:

$$\mathrm{W}_1(\mathcal{P}_R\#\mathrm{Q},\mathcal{P}_R\#\mathrm{P}) \leq \mathrm{W}_1(\mathrm{Q},\mathrm{P}) + 2(M+1)\frac{\beta\sqrt{2\beta}}{m}.$$

We then have:

$$|\Delta_{\mathcal{S}_m}(\mathrm{Q})| \leq 2\frac{\beta\sqrt{2\beta}}{m} + \sqrt{2D_R(2D_R+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(\mathrm{W}_1(\mathrm{Q},\mathrm{P})+\alpha_m\right)+\frac{\log\left(\frac{3m}{\delta}\right)}{m}}.$$

with $\alpha_m = 2(M+1)\frac{\beta\sqrt{\beta}}{m} + \varepsilon_m = \mathcal{O}\left(1+\sqrt{\frac{d\log(Rm)}{m}}\right)$. This concludes the proof. ■

Appendix of Chapter 6 E

The supplementary material is organized as follows:

1. We provide more discussion about Theorems 6.3.1 and 6.3.2 in Appendix E.1;
2. The proofs of Theorems 6.3.1 to 6.3.4 are presented in Appendix E.2;
3. We present in Appendix E.3 additional information about the experiments.

E.1 Additional insights on Section 6.3.1

In Appendix E.1.1, we provide additional discussion about Theorem 6.3.1 while Appendix E.1.2 discuss about the convergence rates for Theorem 6.3.2.

E.1.1 Supplementary discussion about Theorem 6.3.1

Haddouche and Guedj, 2023b, Corollary 10 proposed PAC-Bayes bounds with Wasserstein distances on a Euclidean predictor space with Gaussian prior and posteriors. The bounds have an explicit convergence rate of $\mathcal{O}(\sqrt{\frac{dW_1(\mathrm{Q},\mathrm{P})}{m}})$ where the predictor space is Euclidean with dimension d. While our bound does not propose such an explicit convergence rate, it allows us to derive learning algorithms as described in Section 6.4. A broader discussion about the role of K is detailed in Theorem 6.3.2. Furthermore, our bound holds for any Polish predictor space and does not require Gaussian distributions. Furthermore, our result exploits data-dependent priors and deals with the dimension only through the Wasserstein distance, which can attenuate the impact of the dimension.

E.1.2 Convergence rates for Theorem 6.3.2

In this section, we discuss more deeply the values of K in Theorem 6.3.2. This implies a tradeoff between the number of sets K and the cardinal of each $\mathcal{S}_m^i$. The tightness of the bound depends highly on the sets $\mathcal{S}_m^1, \dots, \mathcal{S}_m^K$.
Full batch setting K=1. When $\mathcal{S}_m^1 = \mathcal{S}_m$ with $K = 1$, the bound of Theorem 6.3.2 becomes, with probability $1-\delta$, for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$

$$\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] \leq 2L\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + 2\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\ ,$$

where $\mathrm{P} = \mathrm{P}_1$ is data-free. This bound can be seen as the high-probability (PAC-Bayesian) version of the expected bound of WANG *et al.*, 2019. Furthermore, in this setting, we are able, through our proof technique, to recover an explicit convergence rate similar to the one of AMIT *et al.*, 2022, Theorem 12. It is stated below.

Corollary E.1.1. For any distribution $\mathcal{D}$ on $\mathcal{Z}$, for any finite hypothesis space $\mathcal{H}$ equipped with a distance d, for any L-Lipschitz loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$, for any $\delta \in (0,1]$, we have, with probability $1-\delta$ over the sample $\mathcal{S}$, for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:

$$\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] \leq L\sqrt{\frac{2\ln\left(\frac{4|\mathcal{H}|^2}{\delta}\right)}{m}}\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + 2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{m}}$$

where P is a data-free prior.

Proof. We exploit AMIT *et al.*, 2022, Equation 35 to state that with probability at least $1-\frac{\delta}{2}$, for any $(h,h') \in \mathcal{H}^2$:

$$\left|\frac{1}{m}\sum_{i=1}^{m}\left[\ell\left(h',\mathbf{z}_i\right) - \ell\left(h,\mathbf{z}_i\right)\right] - \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\ell\left(h',\mathbf{z}\right) - \ell(h,\mathbf{z})\right]\right| \leq L\sqrt{\frac{2\ln\left(\frac{4|\mathcal{H}|^2}{\delta}\right)}{m}}d\left(h,h'\right).$$

So, with high probability, we can exploit the Kantorovich-Rubinstein duality with this new Lipschitz constant: with probability at least $1-\delta/2$:

$$\begin{aligned}&\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right]\\ &\qquad\qquad \leq L\sqrt{\frac{2\ln\left(\frac{4|\mathcal{H}|^2}{\delta}\right)}{m}}\mathrm{W}_1(\mathrm{Q},\mathrm{P}) + \mathop{\mathbb{E}}_{h\sim \mathrm{P}}\frac{1}{m}\left[\sum_{i=1}^{m}\mathsf{R}_{\mathcal{D}}(h) - \ell(h,\mathbf{z}_i)\right],\end{aligned}$$

To conclude, we control the quantity on the right-hand side the same way as in Theorem 6.3.1 and Theorem 6.3.2. We then have, with probability at least $1-\delta/2$, for a loss function in $[0,1]$:

$$\frac{1}{m}\sum_{i=1}^{m}\mathsf{R}_{\mathcal{D}}(h) - \ell(h,\mathbf{z}_i) \leq 2\sqrt{\frac{\ln\frac{K}{\delta}}{m}}.$$

Taking the union bound concludes the proof. ■

Mini-batch setting $K = \sqrt{m}$. When a tradeoff is desired between the quantity of data we want to infuse in our priors and an explicit convergence rate, a meaningful candidate is when $K = \sqrt{m}$. Theorem 6.3.2's bound becomes, in this particular case:

$$\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] \leq \frac{2L}{\sqrt{m}}\sum_{i=1}^{\sqrt{m}} \mathrm{W}_1(\mathrm{Q},\mathrm{P}_i) + 2\sqrt{\frac{\ln\frac{\sqrt{m}}{\delta}}{\sqrt{m}}}. \tag{E.1}$$

Towards online learning: $K = m$. When $K = m$, the sets $\mathcal{S}_m^i$ contain only one example. More precisely, we have for all $i \in \{1, \dots, m\}$ the set $\mathcal{S}_m^i = \{\mathbf{z}_i\}$. In this case, the bound becomes:

$$\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] \leq \frac{2L}{m}\sum_{i=1}^{m} \mathrm{W}_1(\mathrm{Q},\mathrm{P}_i) + 2\sqrt{\ln\frac{m}{\delta}}.$$

This bound is vacuous since the last term is incompressible, hence the need for a new technique detailed in Section 6.3.2 to deal with it.

E.2 Proofs

The proof of Theorem 6.3.1 is presented in Appendix E.2.1. Appendices E.2.2 and E.2.3 introduce two proofs of Theorem 6.3.2. Theorem 6.3.3's proof is presented in Appendix E.2.4. Appendix E.2.5 provides the proof of Theorem 6.3.3.

E.2.1 Proof of Theorem 6.3.1

> **Theorem 6.3.1.** We assume the loss ℓ to be L-Lipschitz. Then, for any $\delta \in (0, 1]$, for any sequence of positive scalar $(\lambda_i)_{i\in\{1,\dots,K\}}$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, the following holds for the distributions $\mathrm{P}_{i,\mathcal{S}} := \mathrm{P}_i(\mathcal{S}, .)$ and for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:
>
> $$\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right]$$
> $$\leq \sum_{i=1}^{K} \frac{2|\mathcal{S}_m^i|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_{i,\mathcal{S}}) + \frac{1}{m}\sum_{i=1}^{K} \frac{\ln\left(\frac{K}{\delta}\right)}{\lambda_i} + \frac{\lambda_i}{2}\left(\mathop{\mathbb{E}}_{h\sim \mathrm{P}_{i,\mathcal{S}}}\left[\hat{V}_{|\mathcal{S}_m^i|}(h) + V_{|\mathcal{S}_m^i|}(h)\right]\right),$$
>
> where $\mathrm{P}_{i,\mathcal{S}}$ *does not* depend on $\mathcal{S}_m^i$. Also, for any $i, |\mathcal{S}_m^i|$, we have $\hat{V}_{|\mathcal{S}_m^i|}(h) = \sum_{\mathbf{z}\in\mathcal{S}_m^i}\left(\ell(h,\mathbf{z}) - R_{\mathcal{D}}(h)\right)^2$ and $V_{|\mathcal{S}_m^i|}(h) = \mathbb{E}_{\mathcal{S}_m^i}\left[\hat{V}_{|\mathcal{S}_m^i|}(h)\right]$.

Proof. For the sake of readability, we identify, for any i, P_i and $\mathrm{P}_{i,\mathcal{S}}$.

Step 1: Exploit the Kantorovich duality Villani, 2009, Remark 6.5. First of all, note that for a L-Lipschitz loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$, we have

$$\left|\left(|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h_1) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h_1,\mathbf{z})\right) - \left(|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h_2) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h_2,\mathbf{z})\right)\right| \\ \leq 2|\mathcal{S}_m^i|Ld(h_1,h_2). \quad \text{(E.2)}$$

Indeed, we can deduce Equation (E.2) from Jensen inequality, the triangle inequality, and by definition that we have

$$\begin{aligned}
&\left|\left(|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h_1) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h_1,\mathbf{z})\right) - \left(|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h_2) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h_2,\mathbf{z})\right)\right| \\
&= \left|\left(\sum_{\mathbf{z}\in\mathcal{S}_m^i} \mathsf{R}_{\mathcal{D}}(h_1) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h_1,\mathbf{z})\right) - \left(\sum_{\mathbf{z}\in\mathcal{S}_m^i} \mathsf{R}_{\mathcal{D}}(h_2) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h_2,\mathbf{z})\right)\right| \\
&\leq \sum_{\mathbf{z}\in\mathcal{S}_m^i} \mathop{\mathbb{E}}_{\mathbf{z}'\sim\mathcal{D}} \Big[|\ell(h_1,\mathbf{z}') - \ell(h_2,\mathbf{z}')| + |\ell(h_2,\mathbf{z}) - \ell(h_1,\mathbf{z})|\Big] \\
&\leq \mathop{\mathbb{E}}_{\mathbf{z}'\sim\mathcal{D}} \sum_{\mathbf{z}\in\mathcal{S}_m^i} 2Ld(h_1,h_2) \\
&= 2|\mathcal{S}_m^i|Ld(h_1,h_2).
\end{aligned}$$

We are now able to upper-bound $\mathbb{E}_{h\sim\mathrm{Q}}[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)]$. Indeed, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\Big[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\Big] &= \frac{1}{m}\sum_{i=1}^{K} \mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h,\mathbf{z})\right] \\
&\leq \sum_{i=1}^{K} \frac{2|\mathcal{S}_m^i|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_i) + \sum_{i=1}^{K} \mathop{\mathbb{E}}_{h\sim\mathrm{P}_i} \frac{1}{m}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h,\mathbf{z})\right], \quad \text{(E.3)}
\end{aligned}$$

where the inequality comes from the Kantorovich-Rubinstein duality theorem.

Step 2: Define an adapted supermartingale. For any $1 \leq i \leq K$, we fix $\lambda_i > 0$ and we provide an arbitrary order to the elements of $\mathcal{S}_m^i := \{\mathbf{z}_{i,1}, \cdots, \mathbf{z}_{i,|S_i|}\}$.

Then we define for any h:

$$M_{|\mathcal{S}_m^i|}(h) := |\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h,\mathbf{z}) = \sum_{j=1}^{|\mathcal{S}_m^i|} R_{\mathcal{D}}(h) - \ell(h,\mathbf{z}_{i,j}).$$

Remark that, because our data are *i.i.d.*, $(M_{|\mathcal{S}_m^i|})_{|\mathcal{S}_m^i|\geq 1}$ is a martingale. We then exploit the technique Chapter 2 to define a supermartingale. More precisely, we exploit a result from Bercu and Touati, 2008 cited in Lemma 1.3 of Chapter 2 coupled with Lemma 2.2 of Chapter 2 to ensure that the process

$$SM_{|\mathcal{S}_m^i|} := \mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\left[\exp\left(\lambda_i M_{|\mathcal{S}_m^i|}(h) - \frac{\lambda_i^2}{2}\left(\hat{V}_{|\mathcal{S}_m^i|}(h) + V_{|\mathcal{S}_m^i|}(h)\right)\right)\right],$$

is a supermartingale, where $\hat{V}_{|\mathcal{S}_m^i|}(h) = \sum_{j=1}^{|\mathcal{S}_m^i|}\left(\ell(h,\mathbf{z}_{i,j}) - R_{\mathcal{D}}(h)\right)^2$ and $V_{|\mathcal{S}_m^i|}(h) = \mathbb{E}_{\mathcal{S}_m^i}\left[\hat{V}_{|\mathcal{S}_m^i|}(h)\right]$.

Step 3. Combine steps 1 and 2. We restart from Equation (E.3) to exploit again the Kantorovich-Rubinstein duality.

$$\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] &= \frac{1}{m}\sum_{i=1}^{K}\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z})\right] \\
&\leq \sum_{i=1}^{K}\frac{2|\mathcal{S}_m^i|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_i) + \sum_{i=1}^{K}\frac{1}{m\lambda_i}\lambda_i\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z})\right], \\
&= \sum_{i=1}^{K}\frac{2|\mathcal{S}_m^i|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_i) + \sum_{i=1}^{K}\frac{1}{m\lambda_i}\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\left[\lambda_i M_{|\mathcal{S}_m^i|}\right], \\
&\leq \sum_{i=1}^{K}\frac{2|\mathcal{S}_m^i|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_i) + \sum_{i=1}^{K}\frac{1}{m\lambda_i}\ln\left(SM_{|\mathcal{S}_m^i|}\right) \\
&\qquad + \frac{1}{m}\sum_{i=1}^{K}\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\left[\frac{\lambda_i}{2}\left(\hat{V}_{|\mathcal{S}_m^i|}(h) + V_{|\mathcal{S}_m^i|}(h)\right)\right].
\end{aligned}$$

The last line holds thanks to Jensen's inequality. We now apply Ville's inequality (see *e.g.*, Section 1.2 of Chapter 2). We have for any i:

$$\mathop{\mathbb{P}}_{\mathcal{S}_m^i\sim\mathcal{D}^{|\mathcal{S}_m^i|}}\left(\forall|S_i|\geq 1, SM_{|S_i|}\leq\frac{1}{\delta}\right)\geq 1-\delta.$$

Applying an union bound and authorising λ_i to be a function of $|S_i|$ (thus the inequality does not hold for all $|\mathcal{S}_m^i|$ simultaneously) finally gives with probability at least $1-\delta$, for all $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:

$$\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] \leq \sum_{i=1}^{K} \frac{2|\mathcal{S}_m^i|L}{m} \mathrm{W}_1(\mathrm{Q},\mathrm{P}_i) + \sum_{i=1}^{K} \frac{\ln\left(\frac{K}{\delta}\right)}{\lambda_i m} + \frac{\lambda_i}{2m} \mathop{\mathbb{E}}_{h\sim \mathrm{P}_i}\left[\hat{V}_{|\mathcal{S}_m^i|}(h) + V_{|\mathcal{S}_m^i|}(h)\right].$$

■

E.2.2 Proof of Theorem 6.3.2

Theorem 6.3.2. We assume our loss ℓ to be non-negative and L-Lipschitz. We also assume that, for any $1 \leq i \leq K$, for any dataset $\mathcal{S}$, we have $\mathbb{E}_{h\sim \mathrm{P}_i(.,\mathcal{S}), z\sim \mathcal{D}}\left[\ell(h,z)^2\right] \leq 1$ (*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 $\mathrm{P}_{i,\mathcal{S}} := \mathrm{P}_i(\mathcal{S},.)$ and for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:

$$\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] \leq \sum_{i=1}^{K} \frac{2|\mathcal{S}_m^i|L}{m} \mathrm{W}_1(\mathrm{Q},\mathrm{P}_{i,\mathcal{S}}) + \sum_{i=1}^{K} \sqrt{\frac{2|\mathcal{S}_m^i| \ln \frac{K}{\delta}}{m^2}},$$

where $\mathrm{P}_{i,\mathcal{S}}$ *does not* depend on $\mathcal{S}_m^i$.

Proof. For the sake of readability, we identify, for any i, P_i and $\mathrm{P}_{i,\mathcal{S}}$.

Step 1: Exploit the Kantorovich duality Villani, 2009, Remark 6.5. First of all, note that for a L-Lipschitz loss function $\ell : \mathcal{H}\times\mathcal{Z} \to [0,1]$, we have

$$\left|\left(|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h_1) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h_1,\mathbf{z})\right) - \left(|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h_2) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h_2,\mathbf{z})\right)\right| \leq 2|\mathcal{S}_m^i|L d(h_1,h_2). \quad \text{(E.4)}$$

Indeed, we can deduce Equation (E.4) from Jensen inequality, the triangle inequality, and by definition that we have

$$\begin{aligned}
&\left|\left(|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h_1) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h_1,\mathbf{z})\right) - \left(|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h_2) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h_2,\mathbf{z})\right)\right| \\
&= \left|\left(\sum_{\mathbf{z}\in\mathcal{S}_m^i}\mathsf{R}_{\mathcal{D}}(h_1) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h_1,\mathbf{z})\right) - \left(\sum_{\mathbf{z}\in\mathcal{S}_m^i}\mathsf{R}_{\mathcal{D}}(h_2) - \sum_{\mathbf{z}\in\mathcal{S}_m^i} \ell(h_2,\mathbf{z})\right)\right| \\
&\leq \sum_{\mathbf{z}\in\mathcal{S}_m^i} \mathop{\mathbb{E}}_{\mathbf{z}'\sim\mathcal{D}}\Big[|\ell(h_1,\mathbf{z}') - \ell(h_2,\mathbf{z}')| + |\ell(h_2,\mathbf{z}) - \ell(h_1,\mathbf{z})|\Big] \\
&\leq \mathop{\mathbb{E}}_{\mathbf{z}'\sim\mathcal{D}} \sum_{\mathbf{z}\in\mathcal{S}_m^i} 2Ld(h_1,h_2) \\
&= 2|\mathcal{S}_m^i|Ld(h_1,h_2).
\end{aligned}$$

We are now able to upper-bound $\mathbb{E}_{h\sim\mathrm{Q}}[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)]$. Indeed, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\Big[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\Big] &= \frac{1}{m}\sum_{i=1}^{K}\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z})\right] \\
&\leq \sum_{i=1}^{K}\frac{2|\mathcal{S}_m^i|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_i) + \sum_{i=1}^{K}\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\frac{1}{m}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z})\right], \quad \text{(E.5)}
\end{aligned}$$

where the inequality comes from the Kantorovich-Rubinstein duality theorem.

Step 2: Define an adapted supermartingale. For any $1 \leq i \leq K$, we fix $\lambda_i > 0$ and we provide an arbitrary order to the elements of $\mathcal{S}_m^i := \{\mathbf{z}_{i,1}, \cdots, \mathbf{z}_{i,|\mathcal{S}_i|}\}$. Then we define for any h:

$$M_{|\mathcal{S}_m^i|}(h) := |\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z}) = \sum_{j=1}^{|\mathcal{S}_m^i|} R_{\mathcal{D}}(h) - \ell(h,\mathbf{z}_{i,j}).$$

Remark that, because our data are *i.i.d.*, $(M_{|\mathcal{S}_m^i|})_{|\mathcal{S}_m^i|\geq 1}$ is a martingale. We then exploit the technique Chugg *et al.*, 2023 to define a supermartingale. More precisely, we exploit Chugg *et al.*, 2023, Lemma A.2 and Lemma B.1 to ensure that the process

$$SM_{|\mathcal{S}_m^i|} := \mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\left[\exp\left(\lambda_i M_{|\mathcal{S}_m^i|}(h) - \frac{\lambda_i^2}{2}L_{|\mathcal{S}_m^i|}(h)\right)\right],$$

is a supermartingale, where, because $\mathcal{S}$ is *i.i.d.*, $L_{|\mathcal{S}^i_m|}(h) = \mathbb{E}_{\mathcal{S}}\left[\sum_{j=1}^{|S_i|}\ell(h,\mathbf{z}_{i,j})^2\right] = |\mathcal{S}^i_m|\,\mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)^2]$.

Step 3. Combine steps 1 and 2. We restart from Equation (E.5) to exploit the Kantorovich-Rubinstein duality again.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h)-\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] &= \frac{1}{m}\sum_{i=1}^{K}\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[|\mathcal{S}^i_m|\mathsf{R}_{\mathcal{D}}(h)-\sum_{\mathbf{z}\in\mathcal{S}^i_m}\ell(h,\mathbf{z})\right]\\
&\leq \sum_{i=1}^{K}\frac{2|\mathcal{S}^i_m|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_i)+\sum_{i=1}^{K}\frac{1}{m\lambda_i}\lambda_i\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\left[|\mathcal{S}^i_m|\mathsf{R}_{\mathcal{D}}(h)-\sum_{\mathbf{z}\in\mathcal{S}^i_m}\ell(h,\mathbf{z})\right],\\
&= \sum_{i=1}^{K}\frac{2|\mathcal{S}^i_m|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_i)+\sum_{i=1}^{K}\frac{1}{m\lambda_i}\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\left[\lambda_i M_{|\mathcal{S}^i_m|}\right],\\
&\leq \sum_{i=1}^{K}\frac{2|\mathcal{S}^i_m|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_i)+\sum_{i=1}^{K}\frac{1}{m\lambda_i}\ln\left(SM_{|\mathcal{S}^i_m|}\right)+\frac{1}{m}\sum_{i=1}^{K}\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\left[\frac{\lambda_i}{2}L_{|\mathcal{S}^i_m|}(h)\right].
\end{aligned}
$$

The last line holds thanks to Jensen's inequality. We now apply Ville's inequality (see *e.g.*, section 1.2 of Chapter 2). We have for any i:

$$
\mathop{\mathbb{P}}_{\mathcal{S}^i_m\sim\mathcal{D}^{|\mathcal{S}^i_m|}}\left(\forall|S_i|\geq 1, SM_{|S_i|}\leq\frac{1}{\delta}\right)\geq 1-\delta.
$$

Applying an union bound and authorising λ_i to be a function of $|S_i|$ (thus the inequality does not hold for all $|\mathcal{S}^i_m|$ simultaneously) finally gives with probability at least $1-\delta$, for all $\mathrm{Q}\in\mathcal{M}(\mathcal{H})$:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h)-\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] \leq \sum_{i=1}^{K}\frac{2|\mathcal{S}^i_m|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_i)&\\
+\sum_{i=1}^{K}\frac{\ln\left(\frac{K}{\delta}\right)}{\lambda_i m}+\frac{\lambda_i}{2m}\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\left[L|\mathcal{S}^i_m|(h)\right].&
\end{aligned}
$$

Finally, using the assumption $\mathbb{E}_{h\sim\mathrm{P}_i}\mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)^2]\leq 1$ gives, with probability at least $1-\delta$, for all $\mathrm{Q}\in\mathcal{M}(\mathcal{H})$:

$$
\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h)-\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right]\leq\sum_{i=1}^{K}\frac{2|\mathcal{S}^i_m|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_i)+\sum_{i=1}^{K}\frac{\ln\left(\frac{K}{\delta}\right)}{\lambda_i m}+\frac{\lambda_i[\mathcal{S}^i_m|]}{2m}.
$$

Taking for each i, $\lambda_i = \sqrt{\frac{2\ln(K/\delta)}{|\mathcal{S}_m^i|}}$ concludes the proof. ■

E.2.3 Alternative proof of Theorem 6.3.2

We state here a slightly tighter version of Theorem 6.3.2 for bounded losses, which relies on an application of McDiarmid's inequality instead of supermartingale techniques. This is useful for the numerical evaluations of our bound.

Theorem E.2.1. We assume our loss ℓ to be in $[0,1]$ and L-Lipschitz. Then, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, the following holds for the distributions $\mathrm{P}_{i,\mathcal{S}} := \mathrm{P}_i(\mathcal{S},.)$ and for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:

$$\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] \leq \sum_{i=1}^{K} \frac{2|\mathcal{S}_m^i|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_{i,\mathcal{S}}) + \sum_{i=1}^{K}\sqrt{\frac{|\mathcal{S}_m^i|\ln\frac{K}{\delta}}{2m^2}}$$

where P_i *does not* depend on $\mathcal{S}_m^i$.

Proof. For the sake of readability, we identify, for any i, P_i and $\mathrm{P}_{i,\mathcal{S}}$.
First of all, note that for a L-Lipschitz loss function $\ell : \mathcal{H}\times\mathcal{Z} \to [0,1]$, we have

$$\left|\left(|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h_1) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h_1,\mathbf{z})\right) - \left(|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h_2) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h_2,\mathbf{z})\right)\right| \leq 2|\mathcal{S}_m^i|Ld(h_1,h_2). \tag{E.6}$$

Indeed, we can deduce Equation (E.6) from Jensen's inequality, the triangle inequality, and by definition that we have

$$\begin{aligned}
&\left|\left(|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h_1) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h_1,\mathbf{z})\right) - \left(|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h_2) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h_2,\mathbf{z})\right)\right| \\
&= \left|\left(\sum_{\mathbf{z}\in\mathcal{S}_m^i}\mathsf{R}_{\mathcal{D}}(h_1) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h_1,\mathbf{z})\right) - \left(\sum_{\mathbf{z}\in\mathcal{S}_m^i}\mathsf{R}_{\mathcal{D}}(h_2) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h_2,\mathbf{z})\right)\right| \\
&\leq \sum_{\mathbf{z}\in\mathcal{S}_m^i}\mathop{\mathbb{E}}_{\mathbf{z}'\sim\mathcal{D}}\left[|\ell(h_1,\mathbf{z}') - \ell(h_2,\mathbf{z}')| + |\ell(h_2,\mathbf{z}) - \ell(h_1,\mathbf{z})|\right] \\
&\leq \mathop{\mathbb{E}}_{\mathbf{z}'\sim\mathcal{D}}\sum_{\mathbf{z}\in\mathcal{S}_m^i} 2Ld(h_1,h_2) \\
&= 2|\mathcal{S}_m^i|Ld(h_1,h_2).
\end{aligned}$$

We are now able to upper-bound $\mathbb{E}_{h\sim\mathrm{Q}}[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)]$. Indeed, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] &= \frac{1}{m}\sum_{i=1}^{K}\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z})\right]\\ &\le \sum_{i=1}^{K}\frac{2|\mathcal{S}_m^i|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_i) + \sum_{i=1}^{K}\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\frac{1}{m}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z})\right], \qquad \text{(E.7)}\end{aligned}$$

where the inequality comes from the Kantorovich-Rubinstein duality theorem. Let $f(\mathcal{S}_m^i) = \mathbb{E}_{h\sim\mathrm{P}_i}\frac{1}{m}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z}_i)\right]$, the function has the bounded difference inequality, *i.e.*, for two datasets $\mathcal{S}_m^i$ and $\mathcal{S}_i'$ that differs from one example (the k-th example, without loss of generality), we have

$$\begin{aligned}&\left|f(\mathcal{S}_m^i) - f(\mathcal{S}_i')\right|\\ &= \left|\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\frac{1}{m}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z})\right] - \mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\frac{1}{m}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}'\in\mathcal{S}_i'}\ell(h,\mathbf{z}')\right]\right|\\ &= \left|\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\left[\frac{1}{m}|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \frac{1}{m}\sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z}) - \frac{1}{m}|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) + \frac{1}{m}\sum_{\mathbf{z}'\in\mathcal{S}_i'}\ell(h,\mathbf{z}')\right]\right|\\ &= \left|\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\left[\frac{1}{m}\sum_{\mathbf{z}'\in\mathcal{S}_i'}\ell(h,\mathbf{z}') - \frac{1}{m}\sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z})\right]\right|\\ &= \left|\mathop{\mathbb{E}}_{h\sim\mathrm{P}_i}\left[\frac{1}{m}\ell(h,\mathbf{z}_k') - \frac{1}{m}\ell(h,\mathbf{z}_k)\right]\right| \le \frac{1}{m}.\end{aligned}$$

Hence, from Mcdiarmid's inequality, we have with probability at least $1-\frac{\delta}{K}$ over

$\mathcal{S} \sim \mathcal{D}^m$

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{h\sim \mathrm{P}_i} \frac{1}{m}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z})\right]\\
&\le \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim \mathrm{P}_i} \frac{1}{m}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z})\right] + \sqrt{\frac{|\mathcal{S}_m^i|\ln\frac{K}{\delta}}{2m^2}}\\
&= \mathop{\mathbb{E}}_{\mathcal{S}_i^c\sim\mathcal{D}^{m-|\mathcal{S}_m^i|}}\ \mathop{\mathbb{E}}_{\mathcal{S}_m^i\sim\mathcal{D}^{|\mathcal{S}_m^i|}}\ \mathop{\mathbb{E}}_{h\sim \mathrm{P}_i} \frac{1}{m}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z})\right] + \sqrt{\frac{|\mathcal{S}_m^i|\ln\frac{K}{\delta}}{2m^2}}\\
&= \mathop{\mathbb{E}}_{\mathcal{S}_i^c\sim\mathcal{D}^{m-|\mathcal{S}_m^i|}}\ \mathop{\mathbb{E}}_{h\sim \mathrm{P}_i} \frac{1}{m}\left[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - \mathop{\mathbb{E}}_{\mathcal{S}_m^i\sim\mathcal{D}^{|\mathcal{S}_m^i|}}\ \sum_{\mathbf{z}\in\mathcal{S}_m^i}\ell(h,\mathbf{z})\right] + \sqrt{\frac{|\mathcal{S}_m^i|\ln\frac{K}{\delta}}{2m^2}}\\
&= \mathop{\mathbb{E}}_{\mathcal{S}_i^c\sim\mathcal{D}^{m-|\mathcal{S}_m^i|}}\ \mathop{\mathbb{E}}_{h\sim \mathrm{P}_i} \frac{1}{m}\Big[|\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h) - |\mathcal{S}_m^i|\mathsf{R}_{\mathcal{D}}(h)\Big] + \sqrt{\frac{|\mathcal{S}_m^i|\ln\frac{K}{\delta}}{2m^2}}\\
&= \sqrt{\frac{|\mathcal{S}_m^i|\ln\frac{K}{\delta}}{2m^2}}.
\end{aligned}
$$

From the union bound, we have with probability at least $1-\delta$ over $\mathcal{S}\sim\mathcal{D}^m$, for any $\mathrm{Q}\in\mathcal{M}(\mathcal{H})$,

$$
\mathop{\mathbb{E}}_{h\sim \mathrm{Q}}\left[\mathsf{R}_{\mathcal{D}}(h) - \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] \le \sum_{i=1}^{K}\frac{2|\mathcal{S}_m^i|L}{m}\mathrm{W}_1(\mathrm{Q},\mathrm{P}_i) + \sum_{i=1}^{K}\sqrt{\frac{|\mathcal{S}_m^i|\ln\frac{K}{\delta}}{2m^2}},
$$

which is the claimed result. ■

We are now able to give a corollary of Theorem E.2.1.

Corollary E.2.1. We assume our loss ℓ to be in $[0,1]$ and L-Lipschitz. Then, for any $\delta\in(0,1]$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, the following holds for the hypotheses $h_{i,\mathcal{S}}\in\mathcal{H}$ associated with the Dirac distributions $\mathrm{P}_{i,\mathcal{S}}$ and for any $h\in\mathcal{H}$:

$$
\mathsf{R}_{\mathcal{D}}(h) \le \hat{\mathsf{R}}_{\mathcal{S}_m}(h) + \sum_{i=1}^{K}\frac{2|\mathcal{S}_m^i|L}{m}d(h,h_{i,\mathcal{S}}) + \sum_{i=1}^{K}\sqrt{\frac{|\mathcal{S}_m^i|\ln\frac{K}{\delta}}{2m^2}}.
$$

Such a bound was impossible to obtain from the PAC-Bayesian bounds based on a KL divergence. Indeed, the KL divergence is infinite for two distributions with disjoint supports. Hence, the PAC-Bayesian framework based on the Wasserstein distance allows us to provide uniform-convergence bounds from a proof technique different from the ones based on the Rademacher complexity Koltchinskii and Panchenko, 2000; Bartlett and Mendelson, 2001, 2002 or the VC-dimension Vapnik and Chervonenkis, 1968, 1974. In Section 6.4, we provide an algorithm minimising such a bound.

E.2.4 Proof of Theorem 6.3.3

Theorem 6.3.3. We assume our loss ℓ to be L-Lipschitz. Then, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, the following holds for the distributions $\mathrm{P}_{i,\mathcal{S}} := \mathrm{P}_i(\mathcal{S}, .)$ and for any sequence $(\mathrm{Q}_i)_{i=1\cdots m} \in \mathcal{M}(\mathcal{H})^m$:

$$\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h_i \sim \mathrm{Q}_i} \Big[\mathbb{E}[\ell(h_i, \mathbf{z}_i) \mid \mathcal{F}_{i-1}] - \ell(h_i, \mathbf{z}_i) \Big] \leq 2L \sum_{i=1}^{m} \mathrm{W}_1(\mathrm{Q}_i, \mathrm{P}_{i,\mathcal{S}}) \\ + \frac{\lambda}{2} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{h_i \sim \mathrm{P}_{i,\mathcal{S}}} \Big[\hat{V}_i(h_i, \mathbf{z}_i) + V_i(h_i) \Big] + \frac{\ln(1/\delta)}{\lambda},$$

where for all i, $\hat{V}_i(h_i, \mathbf{z}_i) = (\ell(h_i, \mathbf{z}_i) - \mathbb{E}_{i-1}[\ell(h_i, \mathbf{z}_i)])^2$ is the conditional empirical variance at time i and $V_i(h_i) = \mathbb{E}_{i-1}[\hat{V}(h_i, \mathbf{z}_i)]$ is the true conditional variance.

Proof. First of all, note that for a L-Lipschitz loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, we have

$$\left| \left(\mathop{\mathbb{E}}_{i-1}[\ell(h_i, \mathbf{z}_i)] - \ell(h_i, \mathbf{z}_i) \right) - \left(\mathop{\mathbb{E}}_{i-1}[\ell(h_i', \mathbf{z}_i)] - \ell(h_i', \mathbf{z}_i) \right) \right| \leq 2Ld(h_i, h_i'). \tag{E.8}$$

Indeed, we can deduce Equation (E.8) from Jensen inequality, the triangle inequality, and by definition that we have

$$\left| \left(\mathop{\mathbb{E}}_{i-1}[\ell(h_i, \mathbf{z}_i)] - \ell(h_i, \mathbf{z}_i) \right) - \left(\mathop{\mathbb{E}}_{i-1}[\ell(h_i', \mathbf{z}_i)] - \ell(h_i', \mathbf{z}_i) \right) \right| \\ \leq \mathop{\mathbb{E}}_{i-1} \Big[|\ell(h_i, \mathbf{z}_i') - \ell(h_i', \mathbf{z}_i')| + |\ell(h_i, \mathbf{z}_i) - \ell(h_i', \mathbf{z}_i)| \Big] \\ \leq \mathop{\mathbb{E}}_{i-1} 2Ld(h_i, h_i') = 2Ld(h_i, h_i').$$

From the Kantorovich-Rubinstein duality theorem Villani, 2009, Remark 6.5, we have

$$\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h_i\sim \mathrm{Q}_i}\left[\mathop{\mathbb{E}}_{i-1}[\ell(h_i,\mathbf{z}_i)] - \ell(h_i,\mathbf{z}_i)\right]$$
$$\leq 2L\sum_{i=1}^{m} W_1(\mathrm{Q}_i,\mathrm{P}_{i,\mathcal{S}}) + \sum_{i=1}^{m}\mathop{\mathbb{E}}_{h\sim \mathrm{P}_{i,\mathcal{S}}}\left[\mathsf{R}_{\mathcal{D}}(h_i) - \ell(h_i,\mathbf{z}_i)\right].$$

Now, we define $X_i(h_i,\mathbf{z}_i) := \mathbb{E}_{i-1}[\ell(h_i,\mathbf{z}_i)] - \ell(h_i,\mathbf{z}_i)$. We also recall that for any i, we have $\hat{V}_i(h_i,\mathbf{z}_i) = (\ell(h_i,\mathbf{z}_i) - \mathbb{E}_{i-1}[\ell(h_i,\mathbf{z}_i)])^2$ and $V_i(h_i) = \mathbb{E}_{i-1}[\hat{V}(h_i,\mathbf{z}_i)]$. To apply the supermartingales techniques of Chapter 2, we define the following function:

$$f_m(S,h_1,...,h_m) := \sum_{i=1}^{m}\lambda X_i(h_i,\mathbf{z}_i) - \frac{\lambda^2}{2}\sum_{i=1}^{m}(\hat{V}_i(h_i,\mathbf{z}_i) + V_i(h_i)).$$

Now, Lemma 2.2.1 state that the sequence $(SM_m)_{m\geq 1}$ defined for any m as:

$$SM_m := \mathop{\mathbb{E}}_{(h_1,\cdots,h_m)\sim \mathrm{P}_{1,\mathcal{S}}\otimes\cdots\otimes \mathrm{P}_{m,\mathcal{S}}}\left[\exp\Big(f_m(\mathcal{S},h_1,...,h_m)\Big)\right],$$

is a supermartingale. We exploit this fact as follows:

$$\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h\sim \mathrm{Q}_{i-1}}\left[\mathop{\mathbb{E}}_{i-1}[\ell(h_i,\mathbf{z}_i)] - \ell(h_i,\mathbf{z}_i)\right]$$
$$= \mathop{\mathbb{E}}_{(h_1,\cdots,h_m)\sim \mathrm{P}_{1,\mathcal{S}}\otimes\cdots\otimes \mathrm{P}_{m,\mathcal{S}}}\left[\sum_{i=1}^{m} X_i(h_i,\mathbf{z}_i)\right]$$
$$= \frac{1}{\lambda}\mathop{\mathbb{E}}_{(h_1,\cdots,h_m)\sim \mathrm{P}_{1,\mathcal{S}}\otimes\cdots\otimes \mathrm{P}_{m,\mathcal{S}}}[f_m(\mathcal{S},h_1,\cdots,h_m)] + \frac{\lambda}{2}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{h_i\sim \mathrm{P}_{i,\mathcal{S}}}\left[\hat{V}_i(h_i,\mathbf{z}_i) + V_i(h_i)\right]$$
$$\leq \frac{\ln(SM_m)}{\lambda} + \frac{\lambda}{2}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{h_i\sim \mathrm{P}_{i,\mathcal{S}}}\left[\hat{V}_i(h_i,\mathbf{z}_i) + V_i(h_i)\right]$$

The last line holds thanks to Jensen's inequality. Now using Ville's inequality ensures us that:

$$\mathop{\mathbb{P}}_{\mathcal{S}}\left(\forall m, SM_m \leq \frac{1}{\delta}\right) \geq \frac{1}{\delta}.$$

Thus, with probability $1-\delta$, for any m we have $\ln(SM_m) \leq \ln\left(\frac{1}{\delta}\right)$. This concludes the proof. ■

E.2.5 Proof of Theorem 6.3.4

> **Theorem 6.3.4.** We assume our loss ℓ to be non-negative and L-Lipschitz. We also assume that, for any $i, \mathcal{S}$, $\mathbb{E}_{h\sim \mathrm{P}_i(.,\mathcal{S})}\left[\mathbb{E}_{i-1}[\ell(h,\mathbf{z}_i)^2]\right] \leq 1$ (*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 stochastic kernels sequence (used as priors) $(\mathrm{P}_i)_{i\geq 1}$, we have with probability at least $1-\delta$ over the sample $S \sim \mathcal{D}$, the following, holding for the data-dependent measures $\mathrm{P}_{i,\mathcal{S}} := \mathrm{P}_i(S,.)$ and any posterior sequence $(\mathrm{Q}_i)_{i\geq 1}$:
>
> $$\frac{1}{m}\sum_{i=1}^m \mathop{\mathbb{E}}_{h_i\sim \mathrm{Q}_i}\Big[\mathbb{E}[\ell(h_i,\mathbf{z}_i)\mid \mathcal{F}_{i-1}] - \ell(h_i,\mathbf{z}_i)\Big] \leq \frac{2L}{m}\sum_{i=1}^m \mathrm{W}_1(\mathrm{Q}_i,\mathrm{P}_{i,\mathcal{S}}) + \sqrt{\frac{2\ln\left(\frac{1}{\delta}\right)}{m}}.$$

Proof. The proof starts similarly to the one of Theorem 6.3.3. Indeed, note that for a L-Lipschitz loss function $\ell : \mathcal{H}\times\mathcal{Z}\to\mathbb{R}$, we have

$$\left|\left(\mathop{\mathbb{E}}_{i-1}[\ell(h_i,\mathbf{z}_i)] - \ell(h_i,\mathbf{z}_i)\right) - \left(\mathop{\mathbb{E}}_{i-1}[\ell(h_i',\mathbf{z}_i)] - \ell(h_i',\mathbf{z}_i)\right)\right| \leq 2Ld(h_i,h_i'). \qquad \text{(E.9)}$$

Indeed, we can deduce Equation (E.9) from Jensen inequality, the triangle inequality, and by definition that we have

$$\begin{aligned}\Big|\Big(\mathop{\mathbb{E}}_{i-1}[\ell(h_i,\mathbf{z}_i)] - \ell(h_i,\mathbf{z}_i)\Big) &- \Big(\mathop{\mathbb{E}}_{i-1}[\ell(h_i',\mathbf{z}_i)] - \ell(h_i',\mathbf{z}_i)\Big)\Big| \\ &\leq \mathop{\mathbb{E}}_{i-1}\Big[|\ell(h_i,\mathbf{z}_i') - \ell(h_i',\mathbf{z}_i')| + |\ell(h_i,\mathbf{z}_i) - \ell(h_i',\mathbf{z}_i)|\Big] \\ &\leq \mathop{\mathbb{E}}_{i-1} 2Ld(h_i,h_i') = 2Ld(h_i,h_i').\end{aligned}$$

From the Kantorovich-Rubinstein duality theorem Villani, 2009, Remark 6.5, we have

$$\sum_{i=1}^m \mathop{\mathbb{E}}_{h_i\sim \mathrm{Q}_i}\left[\mathop{\mathbb{E}}_{i-1}[\ell(h_i,\mathbf{z}_i)] - \ell(h_i,\mathbf{z}_i)\right] \leq 2L\sum_{i=1}^m W_1(\mathrm{Q}_i,\mathrm{P}_{i,\mathcal{S}}) + \sum_{i=1}^m \mathop{\mathbb{E}}_{h\sim \mathrm{P}_{i,\mathcal{S}}}\left[\mathrm{R}_{\mathcal{D}}(h_i) - \ell(h_i,\mathbf{z}_i)\right].$$

Now, we define $X_i(h_i,\mathbf{z}_i) := \mathbb{E}_{i-1}[\ell(h_i,\mathbf{z}_i)] - \ell(h_i,\mathbf{z}_i)$. To apply the supermartingales techniques of Chugg *et al.*, 2023, we define the following function:

$$f_m(S,h_1,...,h_m) := \sum_{i=1}^m \lambda X_i(h_i,\mathbf{z}_i) - \frac{\lambda^2}{2}\sum_{i=1}^m \mathop{\mathbb{E}}_{i-1}[\ell(h_i,\mathbf{z}_i)^2].$$

Now, because our loss is nonnegative, CHUGG *et al.*, 2023, Lemma A.2 and Lemma B.1 state that the sequence $(SM_m)_{m\geq 1}$ defined for any m as:

$$SM_m := \mathop{\mathbb{E}}_{(h_1,\cdots,h_m)\sim \mathrm{P}_{1,\mathcal{S}}\otimes\cdots\otimes\mathrm{P}_{m,\mathcal{S}}}\left[\exp\left(f_m(\mathcal{S},h_1,...,h_m)\right)\right],$$

is a supermartingale. We exploit this fact as follows:

$$\begin{aligned}\sum_{i=1}^m \mathop{\mathbb{E}}_{h\sim \mathrm{Q}_{i-1}}\left[\mathop{\mathbb{E}}_{i-1}[\ell(h_i,\mathbf{z}_i)]-\ell(h_i,\mathbf{z}_i)\right] &= \mathop{\mathbb{E}}_{(h_1,\cdots,h_m)\sim \mathrm{P}_{1,\mathcal{S}}\otimes\cdots\otimes\mathrm{P}_{m,\mathcal{S}}}\left[\sum_{i=1}^m X_i(h_i,\mathbf{z}_i)\right]\\ &= \frac{1}{\lambda}\mathop{\mathbb{E}}_{(h_1,\cdots,h_m)\sim \mathrm{P}_{1,\mathcal{S}}\otimes\cdots\otimes\mathrm{P}_{m,\mathcal{S}}}[f_m(\mathcal{S},h_1,\cdots,h_m)]\\ &+\frac{\lambda}{2}\sum_{i=1}^m \mathop{\mathbb{E}}_{h_i\sim \mathrm{P}_{i,\mathcal{S}}}\left[\mathop{\mathbb{E}}_{i-1}[\ell(h_i,\mathbf{z}_i)^2]\right]\\ &\leq \frac{\ln(SM_m)}{\lambda}+\frac{\lambda}{2}\sum_{i=1}^m \mathop{\mathbb{E}}_{h_i\sim \mathrm{P}_{i,\mathcal{S}}}\left[\mathop{\mathbb{E}}_{i-1}[\ell(h_i,\mathbf{z}_i)^2]\right]\end{aligned}$$

The last line holds thanks to Jensen's inequality. Now using Ville's inequality ensures us that:

$$\mathop{\mathbb{P}}_{\mathcal{S}}\left(\forall m, SM_m\leq\frac{1}{\delta}\right)\geq\frac{1}{\delta}$$

Thus, with probability $1-\delta$, for any m we have $\ln(SM_m)\leq\ln\frac{1}{\delta}$. We conclude the proof by exploiting the boundedness assumption on conditional order 2 moments and optimising the bound in λ. ■

E.3 Supplementary insights on experiments

In this section, Appendix E.3.1 presents the learning algorithm for the *i.i.d.* setting. We also introduce the online algorithm in Appendix E.3.2. We prove the Lipschitz constant of the loss for the linear models in Appendix E.3.3. Finally, we provide more experiments in Appendix E.3.5.

E.3.1 Batch algorithm for the *i.i.d.* setting

The pseudocode of our batch algorithm is presented in Algorithm 4.

Algorithm 4: (Mini-)Batch Learning Algorithm with Wasserstein distances

```
1: procedure PRIORS LEARNING
2:     $h_1, \ldots, h_K \leftarrow$ initialize the hypotheses
3: for $t \leftarrow 1, \ldots, T$ do
       for each mini-batch $\mathcal{U} \subseteq \mathcal{S}$ do
           for $i \leftarrow 1, \ldots, K$ do
               $\mathcal{U}_i \leftarrow \mathcal{U} \setminus \mathcal{S}^i_m$
           $h_i \leftarrow$ perform a gradient descent step with $\nabla \mathsf{R}_{\mathcal{U}_i}(h_i)$
4:     return hypotheses $h_1, \ldots, h_K$

5: procedure POSTERIOR LEARNING
6:     $h \leftarrow$ initialize the hypothesis
7: for $t \leftarrow 1, \ldots, T'$ do
       for each mini-batch $\mathcal{U} \subseteq \mathcal{S}$ do
           $h \leftarrow$ perform a gradient descent step with
       $\nabla[\mathsf{R}_{\mathcal{U}}(h) + \varepsilon \sum_{i=1}^{K} \frac{|\mathcal{S}^i_m|}{m} d(h, h_i)]$
8:     return hypothesis $h$
```

Priors Learning minimises the empirical risk through mini-batches $\mathcal{U} \subseteq \mathcal{S}$ for T epochs. More precisely, for each epoch, we *(a)* sample a mini-batch $\mathcal{U}$ (line 4) by excluding the set $\mathcal{S}^i_m$ from $\mathcal{U}$ for each $h_i \in \mathcal{H}$ (line 5-6), then *(b)* the hypotheses $h_1, \ldots, h_K \in \mathcal{H}$ are updated (line 7). In Posterior Learning, we perform a gradient descent step (line 14) on the objective function associated with Equation (6.5) for T' epochs in a mini-batch fashion.

E.3.2 Learning algorithm for the online setting

Algorithm 5 presents the pseudocode of our online algorithm.

Algorithm 5: Online Learning Algorithm with Wasserstein distances

```
1: Initialize the hypothesis $h_0 \in \mathcal{H}$
2: for $i \leftarrow 1, \ldots, m$ do
       for $t \leftarrow 1, \ldots, T$ do
           $h_i \leftarrow$ perform a gradient step with
       $\nabla[\ell(h_i, \mathbf{z}_i) + \hat{B}(d(h_i, h_{i-1}) - 1)]$ (Eq. (6.7) with $\hat{B}$)
3: return hypotheses $h_1, \ldots, h_m$
```

For each time step i, we perform T gradient descent steps on the objective associated with Equation (6.6) (line 4). Note that we can retrieve OGD from Algorithm 5 by *(a)* setting $T = 1$ and *(b)* removing the regularisation term $\hat{B}(d(h_i, h_{i-1}) - 1)$.

E.3.3 Lipschitzness for the linear model

Recall that we use, in our experiments, the multi-margin loss function from the Pytorch module defined for any linear model with weights $W \in \mathbb{R}^{|\mathcal{Y}| \times d}$ and biases $b \in \mathbb{R}^{|\mathcal{Y}|}$, any data point $\mathbf{z} \in \mathcal{X} \times \mathcal{Y}$

$$\ell(W, b, \mathbf{z}) = \frac{1}{|\mathcal{Y}| - 1} \sum_{y' \neq y} \max\left(0, f(W, b, \mathbf{z}, y')\right),$$

where $f(W, b, \mathbf{z}, y') = 1 + \langle W[y'] - W[y], \mathbf{x} \rangle + b[y'] - b[y]$, and $W[y] \in \mathbb{R}^d$ and $b[y] \in \mathbb{R}$ are respectively the vector and the scalar for the y-th output.
To apply our theorems, we must ensure that our loss function is Lipschitz with respect to the linear model, hence the following lemma.

Lemma E.3.1. For any $\mathbf{z} = (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ with the norm of $\mathbf{x}$ bounded by 1, the function $W, b \mapsto \ell(W, b, \mathbf{z})$ is 2-Lipschitz.

Proof. Let $(W, b), (W', b')$ both in $\mathbb{R}^{|\mathcal{Y}| \times d} \times \mathbb{R}^{|\mathcal{Y}|}$, we have

$$\begin{aligned} &|\ell(W, b, \mathbf{z}) - \ell(W', b', \mathbf{z})| \\ &\qquad\qquad \leq \frac{1}{|\mathcal{Y}| - 1} \sum_{y' \neq y} |\max\left(0, f(W, b, \mathbf{z}, y')\right) - \max\left(0, f(W', b', \mathbf{z}, y')\right)|. \end{aligned}$$

Note that because $\alpha \mapsto \max(0, \alpha)$ is 1-Lipschitz, we have:

$$|\ell(W, b, \mathbf{z}) - \ell(W', b', \mathbf{z})| \leq \frac{1}{|\mathcal{Y}| - 1} \sum_{y' \neq y} |f(W, b, \mathbf{z}, y') - f(W', b', \mathbf{z}, y')|.$$

Finally, notice that:

$$
\begin{aligned}
\frac{1}{|\mathcal{Y}|-1}&\sum_{y'\neq y}|f(W,b,\mathbf{z},y')-f(W',b',\mathbf{z},y')| \\
&\qquad\le \frac{1}{|\mathcal{Y}|-1}\sum_{y'\neq y}|\langle (W-W')[y']-(W-W')[y],\mathbf{x}\rangle| \\
&\qquad\quad+\frac{1}{|\mathcal{Y}|-1}\sum_{y'\neq y}|(b-b')[y']-(b-b')[y]| \\
&\qquad\le \frac{1}{|\mathcal{Y}|-1}\sum_{y'\neq y}\|(W-W')[y']-(W-W')[y]\|\,\|\mathbf{x}\| \\
&\qquad\qquad\qquad+\frac{1}{|\mathcal{Y}|-1}\sum_{y'\neq y}|(b-b')[y']-(b-b')[y]|.
\end{aligned}
$$

Because we consider the Euclidean norm, we have for any $y'\in\mathcal{Y}$:

$$
\begin{aligned}
\|(W-W')[y']-(W-W')[y]\| &= \sqrt{\|(W-W')[y']-(W-W')[y]\|^2} \\
&\le \sqrt{2\left(\|(W-W')[y']\|^2+\|(W-W')[y]\|^2\right)} \\
&\le \sqrt{2}\|W-W'\|.
\end{aligned}
$$

The second line holding because for any scalars a,b, we have $(a-b)^2\le 2(a^2+b^2)$ and the last line holding because $\|W-W'\|^2=\sum_{y\in\mathcal{Y}}\|(W-W')[y]\|^2$. A similar argument gives

$$
\frac{1}{|\mathcal{Y}|-1}\sum_{y'\neq y}|(b-b')[y']-(b-b')[y]|\le\sqrt{2}||b-b'||.
$$

Then, using that $\|x\|\le 1$ and summing on all y' gives:

$$
|\ell(W,b,\mathbf{z})-\ell(W',b',\mathbf{z})|\le\sqrt{2}\left(\|W-W'\|+\|b-b'\|\right).
$$

Finally, notice that $(\|W-W'\|+\|b-b'\|)^2\le 2(\|W-W'\|^2+\|b-b'\|^2)=2\|(W,b)-(W',b')\|^2$.
Thus $\|W-W'\|+\|b-b'\|\le\sqrt{2}\|(W,b)-(W',b')\|$. This concludes the proof. ■

E.3.4 Lipschitzness for neural networks

Recall that we use, in our experiments, the multi-margin loss function from the Pytorch module defined we consider the loss $\ell(h,(\mathbf{x},y))=\frac{1}{|\mathcal{Y}|}\sum_{y'\neq y}\max(0,1-\eta(h[y]-h[y']))$, which is η-Lipschitz *w.r.t.* the outputs $h[1],\ldots,h[|\mathcal{Y}|]$. For neural networks, h is the

output of the neural network with input $\mathbf{x}$. Note that this loss is η-lipschitz with respect to the outputs. To apply our theorems, we must ensure that our loss function is Lipschitz with respect to the weights of the neural networks, hence the following lemma with associated background.
We define a FCN recursively as follows: for a vector $\mathbf{W}_1 = \mathrm{vec}(\{W_1, b\})$, (*i.e.*, the vectorisation of a weight matrix W_1 and a bias b) and an input datum $\mathbf{x}$, $\mathsf{FCN}_1(\mathbf{W}_1, \mathbf{x}) = \sigma_1(W_1\mathbf{x} + b_1)$, where σ_1 is the activation function. Also, for any $i \geq 2$ we define for a vector $\mathbf{W}_i = (W_i, b_i, \mathbf{W}_{i-1})$ (defined recursively as well), $\mathsf{FCN}_i(\mathbf{W}_i, \mathbf{x}) = \sigma_i(W_i\mathsf{FCN}_{i-1}(\mathbf{W}_{i-1}, \mathbf{x}) + b_i)$. Then, setting $\mathbf{z} = (\mathbf{x}, y)$ a datum and $h_i(\mathbf{x}) := \mathsf{FCN}_i(\mathbf{W}_i, \mathbf{x})$ we can rewrite our loss as a function of $(\mathbf{W}_i, \mathbf{z})$.

Lemma E.3.2. Assume that all the weight matrices of $\mathbf{W}_i$ are bounded and that the activation functions are Lipschitz continuous with constant bounded by K_σ. Then for any datum $\mathbf{z} = (\mathbf{x}, y)$, any i, $\mathbf{W}_i \to \ell(\mathbf{W}_i, \mathbf{z})$ is Lipschitz continuous.

Proof. We consider the Frobenius norm on matrices as $\mathbf{W}_2$ is a vector as we consider the L2-norm on the vector. We prove the result for $i = 2$, assuming it is true for $i = 1$. We then explain how this proof generalises the case $i = 1$ and works recursively. Let $\mathbf{z}, \mathbf{W}_2, \mathbf{W}_2'$, for clarity we write $\mathsf{FCN}_2(\mathbf{x}) := \mathsf{FCN}(\mathbf{W}_2, \mathbf{x})$ and $\mathsf{FCN}_2'(\mathbf{x}) := \mathsf{FCN}(\mathbf{W}_2', \mathbf{x})$. As ℓ is Lipschitz on the outputs $\mathsf{FCN}_2(\mathbf{x}), \mathsf{FCN}_2'(\mathbf{x})$. We have

$$\begin{aligned}
|\ell(\mathbf{W}_2, \mathbf{z}) - \ell(\mathbf{W}_2', \mathbf{z})| &\leq \eta \, \|\mathsf{FCN}_2(\mathbf{x}) - \mathsf{FCN}_2'(\mathbf{x})\| \\
&\leq \eta \, \|\sigma_2(W_2\mathsf{FCN}_1(\mathbf{x}) + b_2) - \sigma_2(W_2'\mathsf{FCN}_1'(\mathbf{x}) + b_2')\| \\
&\leq \eta K_\sigma ||W_2\mathsf{FCN}_1(\mathbf{x}) + b_2 - W_2'\mathsf{FCN}_1'(\mathbf{x}) - b_2'|| \\
&\leq \eta K_\sigma \left(||(W_2 - W_2')\mathsf{FCN}_1(\mathbf{x})|| + ||W_2'(\mathsf{FCN}_1(\mathbf{x}) - \mathsf{FCN}_1'(\mathbf{x}))|| + ||b_2 - b_2'||\right).
\end{aligned}$$

Then, we have $||(W_2 - W_2')\mathsf{FCN}_1(\mathbf{x})|| \leq ||(W_2 - W_2')||_F||\mathsf{FCN}_1(\mathbf{x})|| \leq K_\mathbf{x}||(W_2 - W_2')||_F$. The second inequality holding as $\mathsf{FCN}_1(\mathbf{x})$ is a continuous function of the weights. Indeed, as on a compact space, a continuous function reaches its maximum, then its norm is bounded by a certain $K_\mathbf{x}$. Also, as the weights are bounded, any weight matrix has its norm bounded by a certain K_W thus $||W_2'(\mathsf{FCN}_1(\mathbf{x}) - \mathsf{FCN}_1'(\mathbf{x})|| \leq ||W_2'||_F||(\mathsf{FCN}_1(\mathbf{x}) - \mathsf{FCN}_1'(\mathbf{x})|| \leq K_W||\mathsf{FCN}_1(\mathbf{x}) - \mathsf{FCN}_1'(\mathbf{x})||$. Finally, taking $K_{\mathsf{temp}} = \eta K_\sigma \max(K_\mathbf{x}, K_W, 1)$ gives:

$$\begin{aligned}
&|\ell(\mathbf{W}_2, \mathbf{z}) - \ell(\mathbf{W}_2', \mathbf{z})| \\
&\qquad \leq K_{\mathsf{temp}} \left(\|(W_2 - W_2')\|_F + \|b_2 - b_2'\| + \|\mathsf{FCN}_1(\mathbf{x}) - \mathsf{FCN}_1'(\mathbf{x})\|\right).
\end{aligned}$$

Exploiting the recursive assumption that FCN_1 is Lipschitz with respect to its weights $\mathbf{W}_1$ gives $\|\mathsf{FCN}_1(\mathbf{x}) - \mathsf{FCN}_1'(\mathbf{x})\| \leq K_1 ||\mathbf{W}_1 - \mathbf{W}_1'||$.
If we denote by (W_2, b_2) the vector of all concatenated weights, notice that

$$\begin{aligned}
&\|(W_2 - W_2')\|_F + \|b_2 - b_2'\| \\
&\qquad = \sqrt{(\|(W_2 - W_2')\|_F + \|b_2 - b_2'\|)^2} \\
&\qquad \leq \sqrt{2(\|(W_2 - W_2')\|_F^2 + \|b_2 - b_2'\|^2)} \\
&\qquad\qquad = \sqrt{2}\|(W_2, b_2) - (W_2', b_2')\|
\end{aligned}$$

(we used that for any real numbers $a, b, (a+b)^2 \leq 2(a^2 + b^2)$). We then have:

$$\begin{aligned}
&|\ell(\mathbf{W}_2, \mathbf{z}) - \ell(\mathbf{W}_2', \mathbf{z})| \\
&\qquad \leq K_{\mathsf{temp}} \max(\sqrt{2}, K_1) \left(\|(W_2, b_2) - (W_2', b_2')\| + ||\mathbf{W}_1 - \mathbf{W}_1'||\right) \\
&\qquad\qquad \leq \sqrt{2} K_{\mathsf{temp}} \max(\sqrt{2}, K_1) ||\mathbf{W}_2 - \mathbf{W}_2'||.
\end{aligned}$$

The last line holds by reusing the same calculation trick. This concludes the proof for $i = 2$. Then for $i = 1$ the same proof holds by replacing W_2, b_2, FCN_2 by W_1, b_1, FCN_1 and replacing $\mathsf{FCN}_1(\mathbf{x}), \mathsf{FCN}_1'(\mathbf{x})$ by $\mathbf{x}$ (we then do not need to assume a recursive Lipschitz behaviour). Therefore the result holds for $i = 1$.
We then properly apply a recursive argument by assuming the result at rank $i - 1$ reusing the same proof at any rank i by replacing W_2, b_2, FCN_2 by W_i, b_i, FCN_i and $\mathsf{FCN}_1(\mathbf{x}), \mathsf{FCN}_1'(\mathbf{x})$ by $\mathsf{FCN}_{i-1}(\mathbf{x}), \mathsf{FCN}_{i-1}'(\mathbf{x})$. This concludes the proof. ■

E.3.5 Experiments with varying number of priors

The experiments of Section 6.4 rely on data-dependent priors constructed through the procedure PRIORS LEARNING. We fixed a number of priors K equal to $0.2\sqrt{m}$. This number is an empirical tradeoff between the informativeness of our priors and time-efficient computation. However, there is no theoretical intuition for the value of this parameter (the discussion of Section 6.3.1 considered $K = \sqrt{m}$ as a potential tradeoff; see Appendix E.1). Thus, we gather inTables E.1 to E.3 the performance of our learning procedures for $K = \alpha\sqrt{m}$, where $\alpha \in \{0, 0.4, 0.6, 0.8, 1\}$ (the case $\alpha = 0$ being a convention to denote $K = 1$). The experiments are gathered below, and all remaining hyperparameters (except K) are identical to those described in Section 6.4.
Analysis of our results. First, when considering neural networks, note that for any dataset except SEGMENTATION, LETTER, the performances of our methods are similar or better when considering data-dependent priors (*i.e.*, when $\alpha > 0$). A similar remark holds for the linear models for all datasets except for SATIMAGE, SEGMENTATION, and TICTACTOE. This illustrates the relevance of data-dependent priors. We also remark

that there is no value of α, which provides a better performance on all datasets. For instance, considering neural networks, note that $\alpha = 1$ gives the better performance (*i.e.*, the smallest $\mathfrak{R}_{\mathcal{D}}(h)$) for Algorithm 4 ($\frac{1}{\sqrt{m}}$) for the SATIMAGE dataset while, for the same algorithm, the better performance on the SEGMENTATION dataset is attained for $\alpha = 0.8$. Sometimes, the number K does not have a clear influence: on MNIST with NNs, for Algorithm 4 ($\frac{1}{\sqrt{m}}$), our performances are similar, whatever the value of K, but still significantly better than ERM. In any case, note that for every dataset, there exists a value of K and such that our algorithm attains either similar or significantly better performances than ERM on every dataset, which shows the relevance of our learning algorithm to ensure a good generalisation ability. Moreover, there is no obvious choice for the parameters ε. For instance, in Table E.3, for the SEGMENTATION dataset, the parameters $K = 1, \varepsilon = \frac{1}{m}$ are optimal (in terms of test risks) for both models. As $K = 1$ means that our single prior is data-free, this shows that the intrinsic structure of SEGMENTATION makes it less sensitive to both the information contained in the prior ($K = 1$ meaning data-free prior) and the place of the prior itself ($\varepsilon = 1/m$ meaning that we give less weight to the regularisation within our optimisation procedure). On the contrary, the YEAST dataset performs significantly better when $\varepsilon = 1/\sqrt{m}(K = 0.2\sqrt{m})$, exhibiting a positive impact of our data-dependent priors.

E.3.6 Experiments on classical regularisation methods

We perform additional experiments to see the performance of the weight decay, *i.e.*, the L2 regularisation on the weights; the results are presented in Table E.4. Moreover, notice that the 'distance to initialisation' $\|\mathbf{w} - \mathbf{w}_0\|$ (where $\mathbf{w}_0$ is the weights initialized randomly) is a particular case of Algorithm 4 when $K = 1$ (*i.e.*, we treat the data as a single batch, and the prior is the data-free initialisation); the results are in Table E.4.
Analysis of our results. This experiment on the weight decay demonstrates that on a few datasets (namely SENSORLESS and YEAST), when our predictors are neural nets, the weight decay regularisation fails to learn while ours succeeds, as shown in tables below. In general, this table shows that, on most of the datasets, considering data-dependent priors leads to sharper results. This shows the efficiency of our method compared to the 'distance to initialisation' regularisation.

Table E.1. *Performance of Algorithm 4 for neural network models. We consider $\varepsilon = \frac{1}{m}$ and $\varepsilon = \frac{1}{\sqrt{m}}$, with $K = \alpha\sqrt{m}$ and $\alpha \in \{0, 0.4\}$.*

(a) $K = 1$

Dataset	Algo. 4 ($\frac{1}{m}$)		Algo. 4 ($\frac{1}{\sqrt{m}}$)	
	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$
ADULT	0.207	0.207	0.248	0.248
FASHIONMNIST	0.160	0.164	0.158	0.164
LETTER	0.258	0.269	0.268	0.280
MNIST	0.116	0.123	0.085	0.096
MUSHROOMS	0.000	0.000	0.000	0.001
NURSERY	0.705	0.720	0.720	0.736
PENDIGITS	0.704	0.724	0.021	0.037
PHISHING	0.048	0.052	0.038	0.055
SATIMAGE	0.148	0.208	0.147	0.207
SEGMENTATION	0.141	0.176	0.248	0.385
SENSORLESS	0.907	0.911	0.907	0.911
TICTACTOE	0.000	0.042	0.000	0.033
YEAST	0.695	0.712	0.677	0.658

(b) $K = 0.4\sqrt{m}$

Dataset	Algo. 4 ($\frac{1}{m}$)		Algo. 4 ($\frac{1}{\sqrt{m}}$)	
	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$
ADULT	0.167	0.166	0.164	0.164
FASHIONMNIST	0.160	0.164	0.156	0.160
LETTER	0.263	0.275	0.252	0.263
MNIST	0.112	0.120	0.085	0.096
MUSHROOMS	0.000	0.000	0.000	0.000
NURSERY	0.705	0.720	0.706	0.719
PENDIGITS	0.011	0.025	0.010	0.022
PHISHING	0.043	0.053	0.041	0.052
SATIMAGE	0.147	0.178	0.145	0.174
SEGMENTATION	0.345	0.408	0.225	0.416
SENSORLESS	0.075	0.078	0.074	0.077
TICTACTOE	0.000	0.031	0.000	0.019
YEAST	0.450	0.480	0.695	0.712

Table E.2. *Performance of Algorithm 4 compared to ERM on different datasets for linear models. We consider* $\varepsilon = \frac{1}{m}$ *and* $\varepsilon = \frac{1}{\sqrt{m}}$*, with* $K = \alpha\sqrt{m}$ *and* $\alpha \in \{0.6, 0.8\}$.

(a) $K = 0.6\sqrt{m}$

Dataset	Algo. 4 ($\frac{1}{m}$)		Algo. 4 ($\frac{1}{\sqrt{m}}$)	
	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$
ADULT	0.166	0.167	0.166	0.167
FASHIONMNIST	0.127	0.147	0.127	0.150
LETTER	0.288	0.296	0.286	0.296
MNIST	0.067	0.092	0.067	0.093
MUSHROOMS	0.001	0.001	0.001	0.001
NURSERY	0.791	0.802	0.759	0.779
PENDIGITS	0.048	0.061	0.047	0.059
PHISHING	0.062	0.067	0.064	0.068
SATIMAGE	0.146	0.202	0.137	0.199
SEGMENTATION	0.058	0.215	0.058	0.204
SENSORLESS	0.129	0.130	0.130	0.130
TICTACTOE	0.013	0.021	0.013	0.021
YEAST	0.477	0.461	0.478	0.464

(b) $K = 0.8\sqrt{m}$

Dataset	Algo. 4 ($\frac{1}{m}$)		Algo. 4 ($\frac{1}{\sqrt{m}}$)	
	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$
ADULT	0.166	0.167	0.166	0.167
FASHIONMNIST	0.130	0.149	0.128	0.151
LETTER	0.285	0.296	0.288	0.297
MNIST	0.067	0.091	0.067	0.093
MUSHROOMS	0.001	0.001	0.001	0.001
NURSERY	0.771	0.787	0.758	0.778
PENDIGITS	0.047	0.060	0.047	0.059
PHISHING	0.062	0.066	0.065	0.068
SATIMAGE	0.168	0.216	0.137	0.199
SEGMENTATION	0.053	0.212	0.052	0.204
SENSORLESS	0.129	0.130	0.132	0.132
TICTACTOE	0.013	0.021	0.013	0.021
YEAST	0.476	0.461	0.477	0.460

Table E.3. *Performance of Algorithm 4 compared to ERM on different datasets for linear models. We consider* $\varepsilon = \frac{1}{m}$ *and* $\varepsilon = \frac{1}{\sqrt{m}}$*, with* $K = \alpha\sqrt{m}$ *and* $\alpha \in \{1\}$*. We plot the empirical risk* $\mathfrak{R}_{\mathcal{S}}(h)$ *with its associated test risk* $\mathfrak{R}_{\mathcal{D}}(h)$*.*

(a) $K = \sqrt{m}$

	Algo. 4 $(\frac{1}{m})$		Algo. 4 $(\frac{1}{\sqrt{m}})$	
Dataset	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$
ADULT	0.166	0.167	0.166	0.167
FASHIONMNIST	0.354	0.361	0.127	0.151
LETTER	0.287	0.296	0.288	0.298
MNIST	0.068	0.092	0.065	0.092
MUSHROOMS	0.001	0.001	0.001	0.001
NURSERY	0.795	0.805	0.796	0.805
PENDIGITS	0.050	0.062	0.047	0.059
PHISHING	0.062	0.067	0.065	0.067
SATIMAGE	0.143	0.200	0.137	0.201
SEGMENTATION	0.055	0.210	0.055	0.212
SENSORLESS	0.130	0.130	0.131	0.132
TICTACTOE	0.013	0.021	0.392	0.301
YEAST	0.476	0.456	0.476	0.457

(b) *ERM*

Dataset	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$
ADULT	0.166	0.167
FASHIONMNIST	0.139	0.153
LETTER	0.287	0.297
MNIST	0.065	0.091
MUSHROOMS	0.001	0.001
NURSERY	0.794	0.807
PENDIGITS	0.052	0.064
PHISHING	0.064	0.067
SATIMAGE	0.148	0.209
SEGMENTATION	0.087	0.232
SENSORLESS	0.134	0.136
TICTACTOE	0.228	0.238
YEAST	0.470	0.427

Table E.4. *Performance of ERM with weight decay (with the L2 regularisation) for linear and neural network models.*

(a) *Linear*

Dataset	L2 Reg. ($\frac{1}{m}$)		L2 Reg. ($\frac{1}{\sqrt{m}}$)	
	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$
ADULT	0.207	0.207	0.248	0.248
FASHIONMNIST	0.141	0.149	0.127	0.150
LETTER	0.285	0.295	0.285	0.296
MNIST	0.067	0.092	0.066	0.092
MUSHROOMS	0.001	0.001	0.000	0.000
NURSERY	0.788	0.799	0.796	0.804
PENDIGITS	0.049	0.060	0.047	0.057
PHISHING	0.063	0.065	0.057	0.062
SATIMAGE	0.144	0.203	0.138	0.200
SEGMENTATION	0.058	0.157	0.075	0.177
SENSORLESS	0.907	0.911	0.907	0.911
TICTACTOE	0.013	0.021	0.013	0.021
YEAST	0.702	0.720	0.693	0.687

(b) *NN*

Dataset	L2 Reg. ($\frac{1}{m}$)		L2 Reg. ($\frac{1}{\sqrt{m}}$)	
	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$	$\mathfrak{R}_{\mathcal{S}}(h)$	$\mathfrak{R}_{\mathcal{D}}(h)$
ADULT	0.207	0.207	0.248	0.248
FASHIONMNIST	0.160	0.166	0.159	0.164
LETTER	0.261	0.275	0.256	0.269
MNIST	0.116	0.125	0.084	0.095
MUSHROOMS	0.000	0.000	0.000	0.000
NURSERY	0.704	0.721	0.770	0.788
PENDIGITS	0.009	0.022	0.012	0.026
PHISHING	0.042	0.050	0.054	0.059
SATIMAGE	0.150	0.215	0.143	0.205
SEGMENTATION	0.141	0.216	0.198	0.371
SENSORLESS	0.907	0.911	0.907	0.911
TICTACTOE	0.000	0.046	0.000	0.021
YEAST	0.662	0.674	0.693	0.683

References

Pierre Alquier. User-friendly Introduction to PAC-Bayes Bounds. *Foundations and Trends® in Machine Learning*. (2024)
—— **Cited on pages 24, 29, 72.**

Pierre Alquier and Gérard Biau. Sparse single-index model. *JMLR*. (2013). URL: `https://dl.acm.org/doi/10.5555/2567709.2502589`
—— **Cited on pages 38, 62.**

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*. (2018)
—— **Cited on pages 32, 33, 38, 40, 107, 124.**

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*. (2016)
—— **Cited on pages 24, 49, 75, 77, 97, 98, 125, 141, 142, 144, 152, 153, 174.**

Jason Altschuler, Sinho Chewi, Patrik R Gerber, and Austin Stromme. Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent. *Advances in Neural Information Processing Systems*. (2021). URL: `https://proceedings.neurips.cc/paper_files/paper/2021/file/b9acb4ae6121c941324b2b1d3fac5c30-Paper.pdf`
—— **Cited on pages 89, 91.**

Ron Amit, Baruch Epstein, Shay Moran, and Ron Meir. Integral Probability Metrics PAC-Bayes Bounds. *Advances in Neural Information Processing Systems (NeurIPS)*. (2022)
—— **Cited on pages 33, 83, 84, 88, 89, 93, 96, 103, 121, 124, 125, 127, 129, 130, 139, 200.**

Ron Amit and Ron Meir. Meta-Learning by Adjusting Priors Based on Extended PAC-Bayes Theory. *International Conference on Machine Learning (ICML)*. (2018)
—— **Cited on pages 30, 32, 103.**

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharp-

ness and generalization. *arXiv preprint arXiv:2302.07011*. (2023)
——— **Cited on page 72.**

Cécile Ane, Sébastien Blachere, Djalil Chafai, Pierre Fougeres, Ivan Gentil, Florent Malrieu, Cyril Roberto, and Gregory Scheffer. Sur les inegalités de Sobolev logarithmiques. Vol. 10. *Societe mathématique de France Paris*. (2000)
——— **Cited on page 175.**

Martin Arjovsky, Soumith Chintala, and Leon Bottou. Wasserstein Generative Adversarial Networks. *International Conference on Machine Learning (ICML)*. (2017)
——— **Cited on page 124.**

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*. (2011). URL: https://doi.org/10.1214/11-AOS918
——— **Cited on pages 32, 40.**

Arindam Banerjee. On Bayesian Bounds. *Proceedings of the 23rd international conference on Machine learning*. (2006)
——— **Cited on page 25.**

Peter Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Conference on Computational Learning Theory (COLT)*. (2001)
——— **Cited on pages 22, 210.**

Peter Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*. (2002)
——— **Cited on pages 22, 210.**

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

William Beckner. A Generalized Poincaré Inequality for Gaussian Measures. *Proceedings of the American Mathematical Society*. (1989). URL: http://www.jstor.org/stable/2046956
——— **Cited on page 75.**

DAVID A BELSLEY, EDWIN KUH, and ROY E WELSCH. Regression diagnostics: Identifying influential data and sources of collinearity. *John Wiley & Sons*. (2005)
——— **Cited on page 66**.

YOSHUA BENGIO, NICOLAS ROUX, PASCAL VINCENT, OLIVIER DELALLEAU, and PATRICE MARCOTTE. Convex Neural Networks. *Advances in Neural Information Processing Systems*. (2005)
——— **Cited on page 122**.

BERNARD BERCU and ABDERRAHMEN TOUATI. Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability*. 18.5. (2008)
——— **Cited on pages 41, 43, 203**.

PETER J BICKEL and KJELL A DOKSUM. Mathematical statistics: basic ideas and selected topics, volumes I-II package. *CRC Press*. (2015)
——— **Cited on page 73**.

FELIX BIGGS and BENJAMIN GUEDJ. Differentiable PAC-Bayes Objectives with Partially Aggregated Neural Networks. *Entropy*. (2021). URL: `https://doi.org/10.3390/e23101280`
——— **Cited on page 29**.

FELIX BIGGS and BENJAMIN GUEDJ. Non-Vacuous Generalisation Bounds for Shallow Neural Networks. *Proceedings of the 39th International Conference on Machine Learning*. *PMLR*. (2022a). URL: `https://proceedings.mlr.press/v162/biggs22a.html`
——— **Cited on page 29**.

FELIX BIGGS and BENJAMIN GUEDJ. On Margins and Derandomisation in PAC-Bayes. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. (2022b). URL: `https://proceedings.mlr.press/v151/biggs22a.html`
——— **Cited on page 29**.

FELIX BIGGS and BENJAMIN GUEDJ. Tighter PAC-Bayes Generalisation Bounds by Leveraging Example Difficulty. *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. *PMLR*. (2023). URL: `https://proceedings.mlr.press/v206/biggs23a.html`
——— **Cited on page 29**.

Gilles Blanchard and François Fleuret. Occam's hammer. *International Conference on Computational Learning Theory*. Springer. (2007)
—— **Cited on page 62**.

L. E. Blumenson. A Derivation of n-Dimensional Spherical Coordinates. *The American Mathematical Monthly*. (1960). URL: http://www.jstor.org/stable/2308932
—— **Cited on page 193**.

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

Stephen Boyd and Lieven Vandenberghe. Convex optimization. *Cambridge University Press*. (2004)
—— **Cited on page 134**.

Herm Jan Brascamp and Elliott H Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of functional analysis*. 22.4. (1976)
—— **Cited on page 75**.

Alexander Camuto, George Deligiannidis, Murat Erdogdu, Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. Fractal structure and generalization properties of stochastic optimization algorithms. *Conference on Neural Information Processing Systems (NeurIPS)*. (2021)
—— **Cited on page 124**.

Olivier Catoni. A PAC-Bayesian approach to adaptive classification. *preprint*. 840. (2003)
—— **Cited on pages 24, 28, 59**.

Olivier Catoni. Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001. Vol. 1851. *Springer Science & Business Media*. (2004)
—— **Cited on pages 27, 38, 39**.

OLIVIER CATONI. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *Institute of Mathematical Statistics*. (2007)
—— **Cited on pages 24, 25, 28, 32, 38, 39, 48, 60, 62, 75, 81, 95, 97, 117, 130, 142, 153, 174.**

OLIVIER CATONI. PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *CoRR*. abs/1603.05229. (2016)
—— **Cited on pages 32, 40, 128.**

OLIVIER CATONI and ILARIA GIULINI. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *CoRR*. abs/1712.02747. (2017)
—— **Cited on pages 32, 128.**

DJALIL CHAFAI. Entropies, convexity, and functional inequalities, On Φ-entropies and Φ-Sobolev inequalities. *Journal of Mathematics of Kyoto University*. (2004)
—— **Cited on pages 75, 173.**

ANDREW CHEE and SEBASTIEN LOUSTAU. Learning with BOT - Bregman and Optimal Transport divergences. (2021)
—— **Cited on pages 124, 125.**

BADR-EDDINE CHÉRIEF-ABDELLATIF, PIERRE ALQUIER, and MOHAMMAD EMTIYAZ KHAN. A generalization bound for online variational inference. *Asian Conference on Machine Learning*. PMLR. (2019)
—— **Cited on pages 61, 65, 66.**

BEN CHUGG, HONGJIAN WANG, and AADITYA RAMDAS. A Unified Recipe for Deriving (Time-Uniform) PAC-Bayes Bounds. *Journal of Machine Learning Research*. (2023). URL: `http://jmlr.org/papers/v24/23-0401.html`
—— **Cited on pages 52, 76, 77, 127, 128, 140, 177, 178, 180, 183, 205, 212, 213.**

EUGENIO CLERICO, TYLER FARGHLY, GEORGE DELIGIANNIDIS, BENJAMIN GUEDJ, and ARNAUD DOUCET. Generalisation under gradient descent via deterministic PAC-Bayes. *arXiv preprint arXiv:2209.02525*. (2022)
—— **Cited on page 30.**

THOMAS M. COVER and JOY A. THOMAS. Elements of Information Theory. *Wiley*. (2001)
—— **Cited on page 22.**

NELLO CRISTIANINI, JOHN SHAWE-TAYLOR, *et al.* An introduction to support vector machines and other kernel-based learning methods. *Cambridge university press.* (2000)
——— **Cited on page 54.**

IMRE CSISZÁR. I-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability.* (1975)
——— **Cited on pages 25, 83, 128, 175.**

VICTOR H DE LA PEÑA, TZE LEUNG LAI, and QI-MAN SHAO. Self-normalized processes: Limit theory and Statistical Applications. Vol. 204. *Springer.* (2009)
——— **Cited on pages 40, 49.**

OFER DEKEL and YORAM SINGER. Data-driven online to batch conversions. *Advances in Neural Information Processing Systems.* (2005)
——— **Cited on page 57.**

NAN DING, XI CHEN, TOMER LEVINBOIM, SEBASTIAN GOODMAN, and RADU SORICUT. Bridging the Gap Between Practice and PAC-Bayes Theory in Few-Shot Meta-Learning. *Advances in Neural Information Processing Systems (NeurIPS).* (2021)
——— **Cited on pages 30, 32.**

M. D. DONSKER and S. R. S. VARADHAN. Asymptotic evaluation of certain Markov process expectations for large time—III. *Communications on Pure and Applied Mathematics.* (1976)
——— **Cited on pages 25, 83, 94, 96, 98, 128, 175.**

JL DOOB. Jean Ville, Étude Critique de la Notion de Collectif. *Bulletin of the American mathematical society.* 45.11. (1939)
——— **Cited on page 41.**

DHEERU DUA and CASEY GRAFF. UCI Machine Learning Repository. (2017)
——— **Cited on page 135.**

BENJAMIN DUPUIS and UMUT ŞIMŞEKLI. Generalization Bounds for Heavy-Tailed SDEs through the Fractional Fokker-Planck Equation. *arXiv preprint arXiv:2402.07723.* (2024)
——— **Cited on page 25.**

RICK DURRETT. Probability: theory and examples. Vol. 49. *Cambridge university press*. (2019)
—— Cited on page 41.

CYNTHIA DWORK, VITALY FELDMAN, MORITZ HARDT, TONI PITASSI, OMER REINGOLD, and AARON ROTH. Generalization in Adaptive Data Analysis and Holdout Reuse. *Advances in Neural Information Processing Systems*. (2015). URL: https://proceedings.neurips.cc/paper/2015/file/bad5f33780c42f2588878a9d07405083-Paper.pdf
—— Cited on page 187.

GINTARE KAROLINA DZIUGAITE, ALEXANDRE DROUIN, BRADY NEAL, NITARSHAN RAJKUMAR, ETHAN CABALLERO, LINBO WANG, IOANNIS MITLIAGKAS, and DANIEL M. ROY. In search of robust measures of generalization. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. (2020). URL: https://proceedings.neurips.cc/paper/2020/hash/86d7c8a08b4aaa1bc7c599473f5dddda-Abstract.html
—— Cited on pages 29, 77, 80.

GINTARE KAROLINA DZIUGAITE, KYLE HSU, WASEEM GHARBIEH, GABRIEL ARPINO, and DANIEL ROY. On the role of data in PAC-Bayes bounds. *International Conference on Artificial Intelligence and Statistics (AISTATS)*. (2021)
—— Cited on page 29.

GINTARE KAROLINA DZIUGAITE and DANIEL ROY. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. *Conference on Uncertainty in Artificial Intelligence (UAI)*. (2017)
—— Cited on pages 29, 30, 32, 39, 62, 79, 85, 92, 103, 134.

GINTARE KAROLINA DZIUGAITE and DANIEL ROY. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors. *Proceedings of the 35th International Conference on Machine Learning*. (2018a). URL: https://proceedings.mlr.press/v80/dziugaite18a.html
—— Cited on page 30.

GINTARE KAROLINA DZIUGAITE and DANIEL M ROY. Data-dependent PAC-Bayes priors via differential privacy. *Advances in Neural Information Processing Systems*. *Curran Associates, Inc.* (2018b). URL: https://proceedings.neurips.cc/paper/

2018/file/9a0ee0a9e7a42d2d69b8f86b3a0756b1-Paper.pdf
——— **Cited on pages 33, 94, 95, 113, 115, 187.**

MAHDI MILANI FARD and JOELLE PINEAU. PAC-Bayesian Model Selection for Reinforcement Learning. *Advances in Neural Information Processing Systems (NIPS).* (2010)
——— **Cited on pages 30, 32, 39.**

ALEC FARID and ANIRUDHA MAJUMDAR. Generalization Bounds for Meta-Learning via PAC-Bayes and Uniform Stability. *Advances in Neural Information Processing Systems (NeurIPS).* (2021)
——— **Cited on pages 30, 32.**

HAMISH FLYNN, DAVID REEB, MELIH KANDEMIR, and JAN PETERS. PAC-Bayesian lifelong learning for multi-armed bandits. *Data Min. Knowl. Discov.* (2022). URL: https://doi.org/10.1007/s10618-022-00825-4
——— **Cited on page 32.**

PIERRE FORET, ARIEL KLEINER, HOSSEIN MOBAHI, and BEHNAM NEYSHABUR. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412.* (2020)
——— **Cited on page 72.**

WOLFGANG GABCKE. Neue Herleitung und explizite Restabschätzung der Riemann-Siegel-Formel. PhD thesis. Georg-August-Universität Göttingen, (1979)
——— **Cited on page 193.**

MICHAEL GASTPAR, IDO NACHUM, JONATHAN SHAFER, and THOMAS WEINBERGER. Fantastic Generalization Measures are Nowhere to be Found. (2023)
——— **Cited on page 76.**

ITAI GAT, YOSSI ADI, ALEXANDER G. SCHWING, and TAMIR HAZAN. On the Importance of Gradient Norm in PAC-Bayesian Bounds. *NeurIPS.* (2022)
——— **Cited on pages 75, 79.**

PASCAL GERMAIN, FRANCIS BACH, ALEXANDRE LACOSTE, and SIMON LACOSTE-JULIEN. PAC-Bayesian theory meets Bayesian inference. *Advances in Neural Information Processing Systems.* 29. (2016)
——— **Cited on page 58.**

PASCAL GERMAIN, ALEXANDRE LACASSE, FRANÇOIS LAVIOLETTE, and MARIO MARCHAND. PAC-Bayesian learning of linear classifiers. *International Conference on Machine Learning (ICML)*. (2009)
——— **Cited on pages 26, 39, 40.**

MANUEL GIL, FADY ALAJAJI, and TAMAS LINDER. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*. (2013)
——— **Cited on page 168.**

XAVIER GLOROT and YOSHUA BENGIO. Understanding the difficulty of training deep feedforward neural networks. *International Conference on Artificial Intelligence and Statistics (AISTATS)*. (2010)
——— **Cited on page 85.**

HENRY GOUK, EIBE FRANK, BERNHARD PFAHRINGER, and MICHAEL J CREE. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*. (2021)
——— **Cited on page 122.**

ANIL GOYAL, EMILIE MORVANT, PASCAL GERMAIN, and MASSIH-REZA AMINI. PAC-Bayesian Analysis for a Two-Step Hierarchical Multiview Learning Approach. *Machine Learning and Knowledge Discovery in Databases - European Conference (ECML-PKDD)*. (2017)
——— **Cited on page 124.**

LEONARD GROSS. Logarithmic Sobolev Inequalities. *American Journal of Mathematics*. (1975)
——— **Cited on page 74.**

PETER GRUNWALD, THOMAS STEINKE, and LYDIA ZAKYNTHINOU. PAC-Bayes, MAC-Bayes and Conditional Mutual Information: Fast rate bounds that handle general VC classes. *Proceedings of Thirty Fourth Conference on Learning Theory*. (2021). URL: `https://proceedings.mlr.press/v134/grunwald21a.html`
——— **Cited on pages 22, 27, 77.**

BENJAMIN GUEDJ. A Primer on PAC-Bayesian Learning. *Proceedings of the second congress of the French Mathematical Society*. (2019)
——— **Cited on pages 24, 25, 72.**

BENJAMIN GUEDJ and PIERRE ALQUIER. PAC-Bayesian estimation and prediction in sparse additive models. *Electron. J. Statist.* (2013). URL: https://doi.org/10.1214/13-EJS771
—— **Cited on pages 38, 39, 62.**

BENJAMIN GUEDJ and SYLVAIN ROBBIANO. PAC-Bayesian high dimensional bipartite ranking. *Journal of Statistical Planning and Inference*. 196. (2018). ISSN: 0378-3758. DOI: https://doi.org/10.1016/j.jspi.2017.10.010. URL: http://www.sciencedirect.com/science/article/pii/S0378375817301945
—— **Cited on pages 26, 39.**

A. GUIONNET and B. ZEGARLINKSI. Lectures on Logarithmic Sobolev Inequalities. *Séminaire de Probabilités XXXVI*. Ed. by JACQUES AZÉMA, MICHEL ÉMERY, MICHEL LEDOUX, and MARC YOR. *Springer Berlin Heidelberg*. (2003). DOI: 10.1007/978-3-540-36107-7_1. URL: https://doi.org/10.1007/978-3-540-36107-7_1
—— **Cited on pages 75, 173.**

MERT GÜRBÜZBALABAN, UMUT ŞIMŞEKLI, and LINGJIONG ZHU. The heavy-tail phenomenon in SGD. *International Conference on Machine Learning (ICML)*. (2021)
—— **Cited on pages 31, 37, 140.**

MAHDI HAGHIFAM, BORJA RODRÍGUEZ-GÁLVEZ, RAGNAR THOBABEN, MIKAEL SKOGLUND, DANIEL M. ROY, and GINTARE KAROLINA DZIUGAITE. Limitations of Information-Theoretic Generalization Bounds for Gradient Descent Methods in Stochastic Convex Optimization. *Proceedings of The 34th International Conference on Algorithmic Learning Theory*. (2023). URL: https://proceedings.mlr.press/v201/haghifam23a.html
—— **Cited on page 31.**

URI HASSON, SAMUEL A NASTASE, and ARIEL GOLDSTEIN. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*. (2020)
—— **Cited on pages 15, 19.**

ELAD HAZAN. Introduction to online convex optimization. *Foundations and Trends® in Optimization*. 2.3-4. (2016)
—— **Cited on pages 54, 58, 152.**

ELAD HAZAN, AMIT AGARWAL, and SATYEN KALE. Logarithmic regret algorithms for online convex optimization. *Machine Learning*. (2007)
—— **Cited on page 54**.

FREDRIK HELLSTRÖM and GIUSEPPE DURISI. Generalization Bounds via Information Density and Conditional Information Density. (2020). DOI: 10.1109/JSAIT.2020.3040992
—— **Cited on page 27**.

FREDRIK HELLSTRÖM and GIUSEPPE DURISI. A New Family of Generalization Bounds Using Samplewise Evaluated CMI. *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. (2022). URL: http://papers.nips.cc/paper%5C_files/paper/2022/hash/41b6674c28a9b93ec8d22a53ca25bc3b-Abstract-Conference.html
—— **Cited on page 27**.

FREDRIK HELLSTRÖM, GIUSEPPE DURISI, BENJAMIN GUEDJ, and MAXIM RAGINSKY. Generalization bounds: Perspectives from information theory and PAC-Bayes. *arXiv preprint arXiv:2309.04381*. (2023)
—— **Cited on pages 24, 72**.

SEPP HOCHREITER and JÜRGEN SCHMIDHUBER. Flat minima. *Neural computation*. 9.1. (1997)
—— **Cited on page 72**.

DIRK VAN DER HOEVEN, TIM VAN ERVEN, and WOJCIECH KOTŁOWSKI. The Many Faces of Exponential Weights in Online Learning. *Proceedings of the 31st Conference On Learning Theory. PMLR*. (2018). URL: https://proceedings.mlr.press/v75/hoeven18a.html
—— **Cited on pages 64, 68**.

MATTHEW HOLLAND. PAC-Bayes under potentially heavy tails. *Advances in Neural Information Processing Systems (NeurIPS) 32*. Ed. by H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D ALCHÉ-BUC, E. FOX, and R. GARNETT. *Curran Associates, Inc.* (2019). URL: http://papers.nips.cc/paper/8539-pac-bayes-under-potentially-heavy-tails.pdf
—— **Cited on pages 26, 32, 38, 40, 107**.

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

Kyoungseok Jang, Kwang-Sung Jun, Ilja Kuzborskij, and Francesco Orabona. Tighter PAC-Bayes Bounds Through Coin-Betting. *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*. *Proceedings of Machine Learning Research*. (2023). URL: `https://proceedings.mlr.press/v195/jang23a.html`
—— **Cited on pages 52, 107, 127**.

Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*. (2017)
—— **Cited on page 72**.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic Generalization Measures and Where to Find Them. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. (2020). URL: `https://openreview.net/forum?id=SJgIPJBFvH`
—— **Cited on pages 29, 80**.

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization. *Advances in Neural Information Processing Systems (NIPS)*. (2008)
—— **Cited on page 40**.

Leonid Vitalievitch Kantorovitch. Mathematical Methods of Organizing and Planning Production. *Management Science*. (1960)
—— **Cited on page 127**.

Hoel Kervadec, Jose Dolz, Jing Yuan, Christian Desrosiers, Eric Granger, and Ismail Ben Ayed. Constrained deep networks: Lagrangian optimization via log-barrier extensions. *European Signal Processing Conference (EUSIPCO)*. (2022)
—— **Cited on page 134**.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. *High dimensional probability II*. (2000)
—— **Cited on page 210**.

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

ILJA KUZBORSKIJ, KWANG-SUNG JUN, YULIAN WU, KYOUNGSEOK JANG, and FRANCESCO ORABONA. Better-than-KL PAC-Bayes Bounds. *arXiv preprint arXiv:2402.09201*. (2024)
—— **Cited on page 52.**

ILJA KUZBORSKIJ and CSABA SZEPESVÁRI. Efron-Stein PAC-Bayesian Inequalities. *arXiv*. abs/1909.01931. (2019)
—— **Cited on pages 32, 38, 40–42, 49, 107.**

MARC LAMBERT, SINHO CHEWI, FRANCIS R. BACH, SILVÈRE BONNABEL, and PHILIPPE RIGOLLET. Variational inference via Wasserstein gradient flows. *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. Ed. by SANMI KOYEJO, S. MOHAMED, A. AGARWAL, DANIELLE BELGRAVE, K. CHO, and A. OH. (2022). URL: `http://papers.nips.cc/paper%5C_files/paper/2022/hash/5d087955ee13fe9a7402eedec879b9c3-Abstract-Conference.html`
—— **Cited on pages 89, 91, 92, 96, 114, 118–122.**

YANN LECUN. The MNIST database of handwritten digits. (1998)
—— **Cited on pages 85, 135.**

MICHEL LEDOUX. Concentration of measure and logarithmic Sobolev inequalities. *Seminaire de probabilites XXXIII*. *Springer*. (2006)
—— **Cited on pages 75, 173, 175.**

GAEL LETARTE, PASCAL GERMAIN, BENJAMIN GUEDJ, and FRANÇOIS LAVIOLETTE. Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by HANNA M. WALLACH, HUGO LAROCHELLE, ALINA BEYGELZIMER, FLORENCE D'ALCHÉ-BUC, EMILY B. FOX, and ROMAN GARNETT. (2019). URL: `https://proceedings.neurips.cc/paper/2019/hash/`

7ec3b3cf674f4f1d23e9d30c89426cce-Abstract.html
—— **Cited on page 29**.

Guy Lever, François Laviolette, and John Shawe-Taylor. Distribution-dependent PAC-Bayes priors. *International Conference on Algorithmic Learning Theory*. Springer. (2010)
—— **Cited on pages 32, 39**.

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*. 473. (2013)
—— **Cited on pages 32, 39**.

Le Li, Benjamin Guedj, and Sébastien Loustau. A quasi-Bayesian perspective to online clustering. *Electronic Journal of Statistics*. (). URL: https://doi.org/10.1214/18-EJS1479
—— **Cited on page 61**.

Keqin Liu and Qing Zhao. Multi-Armed Bandit Problems with Heavy Tail Reward Distributions. *Allerton Conference on Communication, Control, and Computing*. (2011)
—— **Cited on page 127**.

Ben London. A PAC-Bayesian Analysis of Randomized Learning with Application to Stochastic Gradient Descent. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. (2017). URL: https://proceedings.neurips.cc/paper/2017/hash/7fea637fd6d02b8f0adf6f7dc36aed93-Abstract.html
—— **Cited on page 30**.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal Algorithms for Lipschitz Bandits with Heavy-tailed Rewards. *International Conference on Machine Learning (ICML)*. (2019)
—— **Cited on page 127**.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The Expressive Power of Neural Networks: A View from the Width. (2017). URL: https://proceedings.neurips.cc/paper/2017/hash/32cbf687880eb1674a07bf717761dd3a-Abstract.html
—— **Cited on pages 15, 19**.

GABOR LUGOSI and GERGELY NEU. Online-to-PAC Conversions: Generalization Bounds via Regret Analysis. (2023)
——— **Cited on pages 84, 140.**

GÁBOR LUGOSI and GERGELY NEU. Generalization Bounds via Convex Analysis. *Conference on Learning Theory (COLT)*. (2022)
——— **Cited on pages 84, 124.**

ANDREAS MAURER. A note on the PAC-Bayesian theorem. *arXiv*. cs/0411099. (2004)
——— **Cited on pages 24, 26.**

SOKHNA DIARRA MBACKE, FLORENCE CLERC, and PASCAL GERMAIN. PAC-Bayesian Generalization Bounds for Adversarial Generative Models. (2023). arXiv: `2302.08942 [cs.LG]`. URL: `https://arxiv.org/abs/2302.08942`
——— **Cited on page 88.**

DAVID MCALLESTER. Simplified PAC-Bayesian margin bounds. *Learning theory and Kernel machines*. *Springer*. (2003a)
——— **Cited on page 192.**

DAVID A MCALLESTER. Some PAC-Bayesian theorems. *Proceedings of the eleventh annual conference on Computational Learning Theory*. ACM. (1998)
——— **Cited on page 23.**

DAVID A MCALLESTER. PAC-Bayesian model averaging. *Proceedings of the twelfth annual conference on Computational Learning Theory*. ACM. (1999)
——— **Cited on page 24.**

DAVID A MCALLESTER. PAC-Bayesian Stochastic Model Selection. *Machine Learning*. (2003)
——— **Cited on pages 24, 38.**

FRANK MCSHERRY and KUNAL TALWAR. Mechanism design via differential privacy. *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE. (2007)
——— **Cited on page 187.**

ZAKARIA MHAMMEDI, PETER GRÜNWALD, and BENJAMIN GUEDJ. PAC-Bayes Un-Expected Bernstein Inequality. *Advances in Neural Information Processing Systems (NeurIPS) 32*. (2019). URL: `http://papers.nips.cc/paper/9387-pac-bayes-`

un-expected-bernstein-inequality.pdf
——— **Cited on pages 26, 32, 33, 39, 77.**

Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential Privacy without Sensitivity. *Advances in Neural Information Processing Systems*. (2016)
——— **Cited on pages 113, 187, 188.**

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*. (1781)
——— **Cited on page 124.**

Alfred Müller. Integral Probability Metrics and Their Generating Classes of Functions. *Advances in Applied Probability*. 29.2. (1997)
——— **Cited on page 124.**

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

Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*. (2015)
——— **Cited on page 61.**

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M. Roy. Information-Theoretic Generalization Bounds for Stochastic Gradient Descent. *Proceedings of Thirty Fourth Conference on Learning Theory*. (2021). url: https://proceedings.mlr.press/v134/neu21a.html
——— **Cited on page 30.**

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring Generalization in Deep Learning. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. (2017). url: https://proceedings.neurips.cc/paper/2017/hash/10ce03a1ed01077e3e289f3e53c72813-Abstract.html
——— **Cited on pages 29, 72, 80.**

Ruben Ohana, Kimia Nadjahi, Alain Rakotomamonjy, and Liva Ralaivola. Shedding a PAC-Bayesian Light on Adaptive Sliced-Wasserstein Distances. *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. (2023). URL: `https://proceedings.mlr.press/v202/ohana23a.html`
—— **Cited on page 88.**

Yuki Ohnishi and Jean Honorio. Novel Change of Measure Inequalities with Applications to PAC-Bayesian Bounds and Monte Carlo Estimation. *International Conference on Artificial Intelligence and Statistics (AISTATS)*. (2021)
—— **Cited on pages 33, 40, 124.**

Luca Oneto, Davide Anguita, and Sandro Ridella. PAC-Bayesian analysis of distribution dependent priors: Tighter risk bounds and stability analysis. *Pattern Recognition Letters*. (2016)
—— **Cited on pages 32, 39.**

Francesco Orabona and Dávid Pál. Coin Betting and Parameter-Free Online Learning. *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. (2016). URL: `https://proceedings.neurips.cc/paper/2016/hash/320722549d1751cf3f247855f937b982-Abstract.html`
—— **Cited on page 52.**

Francesco Orabona and Tatiana Tommasi. Training Deep Networks without Learning Rates Through Coin Betting. *Advances in Neural Information Processing Systems (NIPS)*. (2017)
—— **Cited on page 135.**

F. Otto and C. Villani. Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*. (2000)
—— **Cited on page 183.**

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*. 33.3. (1997)
—— **Cited on page 66.**

Victor M Panaretos and Yoav Zemel. An invitation to statistics in Wasserstein space. *Springer Nature*. (2020)
—— **Cited on page 186.**

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

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-bayes bounds with data dependent priors. *Journal of Machine Learning Research*. (2012)
—— **Cited on pages 32, 39.**

F. Pedregosa *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12. (2011)
—— **Cited on page 66.**

Anastasia Pentina and Christoph H. Lampert. A PAC-Bayesian bound for Lifelong Learning. *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. Vol. 32. *JMLR Workshop and Conference Proceedings*. *JMLR.org*. (2014). URL: http://proceedings.mlr.press/v32/pentina14.html
—— **Cited on page 32.**

Maria Perez-Ortiz, Omar Rivasplata, Benjamin Guedj, Matthew Gleeson, Jingyu Zhang, John Shawe-Taylor, Miroslaw Bober, and Josef Kittler. Learning PAC-Bayes Priors for Probabilistic Neural Networks. (2021a)
—— **Cited on pages 29, 33, 62, 92.**

Maria Perez-Ortiz, Omar Rivasplata, Emilio Parrado-Hernandez, Benjamin Guedj, and John Shawe-Taylor. Progress in Self-Certified Neural Networks. *NeurIPS 2021 Workshop on Bayesian Deep Learning*. (2021b)
—— **Cited on pages 29, 62, 79.**

Maria Perez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvari. Tighter Risk Certificates for Neural Networks. *Journal of Machine Learning Research*. (2021)
—— **Cited on pages 29, 62.**

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative Flatness and Generalization. *Advances in neural*

information processing systems. (2021)
—— **Cited on page 72.**

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*. 11.5-6. (2019)
—— **Cited on page 127.**

Antoine Picard-Weibel and Benjamin Guedj. On change of measure inequalities for f-divergences. *arXiv*. abs/2202.05568. (2022)
—— **Cited on pages 33, 40, 107, 124.**

Alexander Rakhlin and Karthik Sridharan. Online Learning with Predictable Sequences. *Proceedings of the 26th Annual Conference on Learning Theory. Proceedings of Machine Learning Research. PMLR*. (2013a). URL: `https://proceedings.mlr.press/v30/Rakhlin13.html`
—— **Cited on page 54.**

Sasha Rakhlin and Karthik Sridharan. Optimization, Learning, and Games with Predictable Sequences. *Advances in Neural Information Processing Systems*. (2013b). URL: `https://proceedings.neurips.cc/paper/2013/file/f0dd4a99fba6075a9494772b58f95280-Paper.pdf`
—— **Cited on page 54.**

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

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. *Advances in Neural Information Processing Systems (NeurIPS)*. (2020)
—— **Cited on pages 32, 40, 54, 56, 62, 64, 65, 67–69, 107, 157, 158, 161, 175.**

Omar Rivasplata, Vikram M. Tankasali, and Csaba Szepesvari. PAC-Bayes with Backprop. *arXiv*. (2019)
—— **Cited on pages 29, 62.**

Borja Rodriguez-Galvez, German Bassi, Ragnar Thobaben, and Mikael Skoglund. Tighter Expected Generalization Error Bounds via Wasserstein Distance. *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual.*

(2021)
—— **Cited on pages 22, 89, 124.**

Borja Rodriguez-Galvez, Ragnar Thobaben, and Mikael Skoglund. More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime-validity. *CoRR*. (2023)
—— **Cited on page 52.**

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-Information, Differential Privacy, and Post-Selection Hypothesis Testing. (2016)
—— **Cited on pages 113, 188.**

Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. PACOH: Bayes-optimal meta-learning with PAC-guarantees. *International Conference on Machine Learning (ICML)*. (2021)
—— **Cited on pages 30, 32.**

Jonas Rothfuss, Martin Josifoski, Vincent Fortuin, and Andreas Krause. PAC-Bayesian Meta-Learning: From Theory to Practice. *arXiv*. abs/2211.07206. (2022)
—— **Cited on pages 30, 32.**

Daniel Russo and James Zou. Controlling Bias in Adaptive Data Analysis Using Information Theory. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*. (2016). URL: `http://proceedings.mlr.press/v51/russo16.html`
—— **Cited on page 27.**

Daniel Russo and James Zou. How Much Does Your Data Exploration Overfit? Controlling Bias via Information Usage. *IEEE Transactions on Information Theory*. 66.1. (2020)
—— **Cited on page 124.**

Otmane Sakhi, Pierre Alquier, and Nicolas Chopin. PAC-Bayesian Offline Contextual Bandits With Guarantees. *International Conference on Machine Learning (ICML)*. (2023)
—— **Cited on page 30.**

André Schlichting. Poincaré and Log-Sobolev Inequalities for Mixtures. *Entropy*. (2019). DOI: 10.3390/e21010089. URL: `https://doi.org/10.3390/e21010089`
—— **Cited on page 79.**

YEVGENY SELDIN, NICOLÒ CESA-BIANCHI, PETER AUER, FRANÇOIS LAVIOLETTE, and JOHN SHAWE-TAYLOR. PAC-Bayes-Bernstein Inequality for Martingales and its Application to Multiarmed Bandits. *Proceedings of the Workshop on Online Trading of Exploration and Exploitation 2. PMLR.* (2012a). URL: https://proceedings.mlr.press/v26/seldin12a.html
—— **Cited on pages 32, 39, 42–44, 49–51, 141–143.**

YEVGENY SELDIN, FRANÇOIS LAVIOLETTE, NICOLÒ CESA-BIANCHI, JOHN SHAWE-TAYLOR, and PETER AUER. PAC-Bayesian Inequalities for Martingales. *IEEE Transactions on Information Theory*. (2012)
—— **Cited on pages 30, 32, 39, 42, 131, 132.**

YEVGENY SELDIN, FRANÇOIS LAVIOLETTE, JOHN SHAWE-TAYLOR, JAN PETERS, and PETER AUER. PAC-Bayesian Analysis of Martingales and Multiarmed Bandits. *arXiv*. (2011)
—— **Cited on pages 30, 32, 39, 42.**

SHAI SHALEV-SHWARTZ. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*. (2012)
—— **Cited on pages 54, 61, 152.**

J. SHAWE-TAYLOR and R. C. WILLIAMSON. A PAC analysis of a Bayes estimator. *Proceedings of the 10th annual conference on Computational Learning Theory*. ACM. (1997)
—— **Cited on page 23.**

UMUT ŞIMŞEKLI, LEVENT SAGUN, and MERT GÜRBÜZBALABAN. A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks. *International Conference on Machine Learning (ICML)*. (2019)
—— **Cited on page 31.**

ALEKSANDRS SLIVKINS. Introduction to Multi-Armed Bandits. *Foundations and Trends in Machine Learning*. (2019)
—— **Cited on page 128.**

JACK W SMITH, JAMES E EVERHART, WC DICKSON, WILLIAM C KNOWLER, and ROBERT SCOTT JOHANNES. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the annual symposium on computer application in medical care*. American Medical Informatics Association. (1988)
—— **Cited on page 66.**

JOST TOBIAS SPRINGENBERG, ALEXEY DOSOVITSKIY, THOMAS BROX, and MARTIN RIEDMILLER. Striving for Simplicity: The All Convolutional Net. *International Conference on Learning Representations (ICLR) – Workshop Track*. (2015)
——— **Cited on page 85.**

THOMAS STEINKE and LYDIA ZAKYNTHINOU. Reasoning About Generalization via Conditional Mutual Information. *Proceedings of Thirty Third Conference on Learning Theory. PMLR*. (2020). URL: `https://proceedings.mlr.press/v125/steinke20a.html`
——— **Cited on page 27.**

W NICK STREET, WILLIAM H WOLBERG, and OLVI L MANGASARIAN. Nuclear feature extraction for breast tumor diagnosis. *Biomedical image processing and biomedical visualization*. SPIE. (1993)
——— **Cited on page 66.**

RICHARD S SUTTON and ANDREW G BARTO. Reinforcement Learning: An introduction. *MIT press*. (2018)
——— **Cited on page 50.**

DANIEL SVOZIL, VLADIMIR KVASNICKA, and JIRI POSPICHAL. Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*. (1997)
——— **Cited on page 54.**

TERENCE TAO. An introduction to measure theory. Vol. 126. *American Mathematical Society Providence*. (2011)
——— **Cited on page 46.**

NIKLAS THIEMANN, CHRISTIAN IGEL, OLIVIER WINTENBERGER, and YEVGENY SELDIN. A strongly quasiconvex PAC-Bayesian bound. *International Conference on Algorithmic Learning Theory*. PMLR. (2017)
——— **Cited on pages 26, 58.**

ILYA O. TOLSTIKHIN and YEVGENY SELDIN. PAC-Bayes-Empirical-Bernstein Inequality. *Advances in Neural Information Processing Systems (NeurIPS)*. (2013)
——— **Cited on pages 26, 39, 77.**

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

VLADIMIR VAPNIK and ALEXEY CHERVONENKIS. On the uniform convergence of relative frequencies of events to their probabilities. *Doklady Akademii Nauk USSR*. (1968)
——— **Cited on page 210**.

VLADIMIR VAPNIK and ALEXEY CHERVONENKIS. Theory of pattern recognition. (1974)
——— **Cited on page 210**.

VLADIMIR NAUMOVICH VAPNIK. The Nature of Statistical Learning Theory, Second Edition. *Statistics for Engineering and Information Science*. *Springer*. (2000)
——— **Cited on pages 22, 23**.

PAUL VIALLARD, PASCAL GERMAIN, AMAURY HABRARD, and EMILIE MORVANT. A general framework for the practical disintegration of PAC-Bayesian bounds. *Machine Learning*. (2023)
——— **Cited on pages 32, 33, 62–65, 67, 158, 167, 175, 176**.

CÉDRIC VILLANI. Optimal transport: old and new. *Grundlehren der mathematischen Wissenschaften* 338. *Springer*. (2009)
——— **Cited on pages 83, 93, 94, 96, 127, 128, 181, 185, 202, 204, 210, 212**.

HAO WANG, MARIO DIAZ, JOSE CZNDIDO SILVEIRA SANTOS FILHO, and FLAVIO P. CALMON. An Information-Theoretic View of Generalization via Wasserstein Distance. *IEEE*. (2019). URL: https://doi.org/10.1109/ISIT.2019.8849359
——— **Cited on pages 22, 124, 200**.

KAIYUE WEN, ZHIYUAN LI, and TENGYU MA. Sharpness Minimization Algorithms Do Not Only Minimize Sharpness To Achieve Better Generalization. *Thirty-seventh Conference on Neural Information Processing Systems*. (2023). URL: https://openreview.net/forum?id=Dkmpa6wCIx
——— **Cited on pages 72, 73**.

OLIVIER WINTENBERGER. Stochastic Online Convex Optimization; Application to probabilistic time series forecasting. *arXiv preprint arXiv:2102.00729*. (2021)
——— **Cited on pages 54, 55, 57, 140, 157**.

Yi-Shan Wu and Yevgeny Seldin. Split-kl and PAC-Bayes-split-kl Inequalities for Ternary Random Variables. *Advances in Neural Information Processing Systems*. (2022). URL: `https://proceedings.neurips.cc/paper_files/paper/2022/file/49ffa271264808cf500ea528ed8ec9b3-Paper-Conference.pdf`
—— **Cited on pages 26, 39.**

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. (2017)
—— **Cited on pages 85, 135.**

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. (2017). URL: `https://proceedings.neurips.cc/paper/2017/hash/ad71c82b22f4f65b9398f76d8be4c615-Abstract.html`
—— **Cited on pages 27, 124.**

Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. *International Conference on Machine Learning*. PMLR. (2016)
—— **Cited on page 54.**

Yun Yue, Jiadi Jiang, Zhiling Ye, Ning Gao, Yongchao Liu, and Ke Zhang. Sharpness-Aware Minimization Revisited: Weighted Sharpness as a Regularization Term. *arXiv preprint arXiv:2305.15817*. (2023)
—— **Cited on page 72.**

Jingwei Zhang, Tongliang Liu, and Dacheng Tao. An Optimal Transport View on Generalization. *arXiv*. abs/1811.03270. (2018)
—— **Cited on page 124.**

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J. Reddi, Sanjiv Kumar, and Suvrit Sra. Why are Adaptive Methods Good for Attention Models? *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. (2020). URL: `https://proceedings.neurips.cc/paper/2020/hash/b05b57f6add810d3b7490866d74c0053-Abstract.html`
—— **Cited on page 31.**

Lijun Zhang, Tianbao Yang, rong jin, and Zhi-Hua Zhou. Dynamic Regret of Strongly Adaptive Methods. *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. (N.d.). URL:

https://proceedings.mlr.press/v80/zhang18o.html
——— **Cited on page 54.**

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic Regret of Convex and Smooth Functions. *Advances in Neural Information Processing Systems.* (2020). URL: https://proceedings.neurips.cc/paper/2020/file/939314105ce8701e67489642ef4d49e8-Paper.pdf
——— **Cited on page 54.**

Sijia Zhou, Yunwen Lei, and Ata Kaban. Toward Better PAC-Bayes Bounds for Uniformly Stable Algorithms. *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.* (2023). URL: http://papers.nips.cc/paper%5C_files/paper/2023/hash/5e8309c9ca683e11672e3dbcd4b87776-Abstract-Conference.html
——— **Cited on page 31.**

Vincent Zhuang and Yanan Sui. No-Regret Reinforcement Learning with Heavy-Tailed Rewards. *International Conference on Artificial Intelligence and Statistics (AISTATS).* (2021)
——— **Cited on page 127.**

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. *Proceedings of the 20th international conference on machine learning (icml-03).* (2003)
——— **Cited on pages 54, 66, 131, 151.**