Wasserstein PAC-Bayes Learning: A Bridge Between Generalisation and Optimisation

Maxime Haddouche, Benjamin Guedj

Wasserstein PAC-Bayes Learning: A Bridge Between Generalisation and Optimisation

Maxime Haddouche MAXIME.HADDOUCHE@INRIA.FR
Inria, Université de Lille
and University College London
France

Benjamin Guedj BENJAMIN.GUEDJ@INRIA.FR
Inria and University College London
UK

Abstract

PAC-Bayes learning is an established framework to assess the generalisation ability of learning algorithm during the training phase. However, it remains challenging to know whether PAC-Bayes is useful to understand, before training, why the output of well-known algorithms generalise well. We positively answer this question by expanding the *Wasserstein PAC-Bayes* framework, briefly introduced in Amit et al. (2022). We provide new generalisation bounds exploiting geometric assumptions on the loss function. Using our framework, we prove, before any training, that the output of an algorithm from Lambert et al. (2022) has a strong asymptotic generalisation ability. More precisely, we show that it is possible to incorporate optimisation results within a generalisation framework, building a bridge between PAC-Bayes and optimisation algorithms.

1 Extended introduction

About PAC-Bayes learning. PAC-Bayes learning (see the seminal works of Shawe-Taylor and Williamson, 1997,McAllester, 1998, 1999, 2003b and Catoni, 2003, 2007) is a powerful tool to explain the generalisation ability of learning algorithms in the sense that it gives a control of the generalisation ability of our algorithm of interest during the learning phase. Indeed, PAC-Bayes theory aims to upper bound the gap between the averaged error on a novel unseen datum and the empirical performance on a training set, this being valid at any moment of the optimisation process. Also, PAC-Bayes guarantees classically consists in empirical upper bounds, obtained through various tools such as exponential moments of a Bernoulli (McAllester, 2003b) the log-Laplace transform (Catoni, 2007), Bernstein inequality, (Tolstikhin and Seldin, 2013; Mhammedi et al., 2019) among others. This is crucial as it gives PAC-Bayes a wider range: beyond guarantees for existing algorithms, PAC-Bayes bounds lead to novel learning procedures. Such algorithms have been instantiated in several learning fields *e.g.* deep nets (Dziugaite and Roy, 2017), meta learning (Amit and Meir, 2018), online learning (Haddouche and Guedj, 2022a), reinforcement learning (Fard and Pineau, 2010) or bandits (Seldin et al., 2011).

A theory built around a divergence Most of the PAC-Bayes literature is based on the use of the *Kullback-Leibler (KL) divergence* to control the evolution of a *posterior* distribution (over the predictor space of interest) from a *prior* one, the posterior being classically

data-dependent while the prior is data-free. However, KL divergence is a limited tool 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.

PAC-Bayes learning with Wasserstein distances. A recent line of work led by Amit et al. (2022) investigates on PAC-Bayes generalisation bounds with Wasserstein distance. This idea has been simultaneously developed by Ohana et al. (2022) for sliced adaptive wasserstein distances, which are easier to compute. 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 (but the complexity measure remains a KL divergence).

In this work, we propose a major development of the *Wasserstein PAC-Bayes* (WPB) theory. Indeed, the work of Amit et al. (2022) provide WPB bounds with explicit convergence rates (for bounded losses) only for finite predictor classes or for linear regression problems. We extend those results in a broader framework including uncountable predictor classes and unbounded losses. More precisely, we first propose a novel WPB bound valid on any compact for bounded lipschitz losses. From this, we demonstrate that the WPB framework bypass both the compactness assumption on the predictor class and the bounded loss assumption: Wasserstein PAC-Bayes only requires lipschitz or smooth functions to perform. We obtain explicit bounds for the case of prior and posterior distributions taken within a compact space of Gaussian measures. We also extend our 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.

While our WPB theorems are interesting new results within the generalisation literature (indeed, Wasserstein distances have emerged recently in generalisation bounds see *e.g.* Rodríguez Gálvez et al., 2021), we also show that Wasserstein PAC-Bayes builds a bridge between PAC-bayes and optimisation. More precisely, we show that it is possible to intricate the benefits of an optimisation process within a generalisation framework: to our knowledge it is the first time a PAC-Bayes bound is used like this. We focus on a specific algorithm taken from Lambert et al. (2022) and shows 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 rest of the introduction is structured as follows: we state in Section 1.1 the framework that is used throughout our work. In Section 1.2, we describe how current PAC-Bayes procedures are designed and how their efficiency is evaluated. We discuss the limitations and caveats of this approach. Then in Section 1.3, we describe carefully the main contributions of our work, showing how we progresively construct a WPB theory (using different techniques than Amit et al., 2022) in order to exploit the optimisation results of Lambert et al. (2022).

The rest of the paper details our main results. Section 2 gathers results for compact predictor spaces, Section 3 gives WPB bounds for gaussians prior and posterior, Section 4 furnishes a WPB bound with data-dependent prior for unbounded lipschitz losses and eventually, Section 5 bridges the gap between optimisation and generalisation by exploiting the

optimisation results of Lambert et al. (2022) within a generalisation framework. Appendix A gathers additional background and Appendix B a few additional proofs.

1.1 Framework

We introduce the framework which will be used throughout our work.

Learning theory 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 finite dataset $S = (z_i)_{i\in\{1..m\}} \in \mathcal{Z}^m$ and assume that sequence is iid following the distribution μ. We always assume in this work that $\mathcal{H} \subseteq \mathbb{R}^d$, we denote by $\Sigma_{\mathcal{H}}$ the associated Borel σ-algebra and we refer by $||.||$ the classical euclidean norm. Finally, we denote by $\mathcal{M}_1(\mathcal{H})$ is the set of probabilities on $\mathcal{H}$. We also denote by $\mathcal{P}_1(\mathcal{H})$ (resp. $\mathcal{P}_2(\mathcal{H})$) the subspace of $\mathcal{M}_1(\mathcal{H})$ of with finite order 1 (resp. order 2) moments wrt $||.||$.

Definitions The *generalisation error* R of a predictor $h \in \mathcal{H}$ is $\forall h, R(h) = \mathbb{E}_{z\sim\mu}[\ell(h, z)]$, the *empirical error* of h is $\forall h, R_S(h) = \frac{1}{m}\sum_{i=1}^m \ell(h, z_i)]$. We also denote by *generalisation gap* for any h the quantity $\Delta_S(h) = R(h) - R_S(h)$ and, for any $Q \in \mathcal{M}_1(H)$, $\Delta_S(Q) = \mathbb{E}_{h\sim Q}[\Delta_S(h)]$. In what follows, we refer to $\mathcal{B}(x, r)$ (resp. $\bar{\mathcal{B}}(x, r)$) to denote the ball (resp. closed ball) centered in $x \in \mathbb{R}^d$ of radius r.

Finally, we define the *Gibbs posterior associated to the prior* $P \in \mathcal{M}_1(\mathcal{H})$ which is the measure $P_{-\lambda R_S} \propto \exp(-\lambda R_S(.))dP(.)$.

Additional framework. We then denote by $\mathrm{BW}(\mathbb{R}^d) \subset \mathcal{P}_2(\mathbb{R}^d)$ which consists in 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 $P \in \mathcal{P}_1(\mathbb{R}^d)$ we refer to $T\#P$ to refer to the measure such that for any $B \in \Sigma_{\mathbb{R}^d}, T\#P(B) = P(T^{-1}(B))$. 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 $\mathrm{BW}(\mathbb{R}^d)$, we denote by $C_{\alpha,\beta,M}, 0 \leq \alpha \leq \beta, M \geq 0$ the set:

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

1.2 PAC-Bayes and optimisation: limits and caveats

Optimisation in PAC-Bayes PAC-Bayesian generalisation bounds are dedicated to control how well measures derived from a learning algorithm perform on a novel unseen data. Those bounds involves a complexity term which is classically a Kullback Leibler (KL) divergence. A generic shape of those bounds is described below: with probability $1 - \delta$, for all measure Q:

$$\Delta_S(Q) \leq \sqrt{\frac{\mathrm{COMP}(Q)}{m}}.$$

where COMP is the complexity term taking into account, for instance, a data-free prior P and an approximation term $1 - \delta$.

From an optimisation point of view, this upper bound can be seen as a learning objective, where COMP acts as a regulariser to avoid overfitting on the empirical risk:

$$Q^* := \text{argmin}_{Q \in \mathcal{M}_1(\mathcal{H})} R_S(Q) + \sqrt{\frac{\text{COMP}(Q)}{m}}.$$

Then those algorithms are build to ensure a candidate measure with a good generalisation ability. However the impact of the optimisation process remains unclear: as $\sqrt{\text{COMP}}$ is not necessarily convex, it is unclear whether an optimisation procedure on the previous learning objective will lead to $\hat{Q}$ (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 $[0, 1]$:

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

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

$$dQ^*(h) \propto \exp(-\lambda R_S(h)) dP(h).$$

This optima, while having a closed form, may be hard to compute. A class of methods dedicated to attain this challenging posterior are Markov Chain Monte Carlo (MCMC) methods that rely on carefully constructed Markov chains which (approximately) converge to Q^*. However, MCMC methods may be computationally costful and other methods arose to obtain quickly surrogates of Q^*. In particular, *Variational Inference (VI)* has been developed as a time-efficient solution. More precisely, VI algorithms aims to estimate a surrogate $\hat{Q}$ 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:

$$\hat{Q} = \underset{Q \in \mathcal{C}}{\text{argmin}}\, \text{KL}(Q, Q^*),$$

where $\mathcal{C}$ is a subset of $\mathcal{M}_1(\mathcal{H})$. When $\mathcal{C}$ is the set of Gaussian measures (also known as the *Bures-Wasserstein* manifold), then the convergence of the associated VI algorithm is a contemporary research topic (Altschuler et al., 2021; Lambert et al., 2022).

This candidate $\hat{Q}$ is approximated after N optimisation steps by a measure $\hat{Q}_N$ and is then plotted in the McAllester's bound to assess its efficiency:

$$\Delta_S(\hat{Q}_N) \leq \sqrt{\frac{\text{KL}(\hat{Q}_N, P) + \log(m/\delta)}{2m}}, \tag{1}$$

Role of the prior P. From an optimisation perspective, the conclusion of Equation (1) is that if $\hat{Q}_N$ is a good approximation of $\hat{Q}$ and if the initialisation P is well-chosen, then the generalisation ability $\hat{Q}$ is guaranteed to be high. Assuming such a condition on P may be unrealistic. Furthermore the term $\text{KL}(\hat{Q}_N, P)$ acts as a blackbox as we do not have a theoretical control on how far our $\hat{Q}$ and $\hat{Q}_N$ diverges from the prior. In particular if our prior is ill-chosen, then we could have $\text{KL}(\hat{Q}_N, P) = \mathcal{O}(m)$, making Equation (1) vacuous. Thus, even if the classical PAC-Bayes theorems are giving the meaningful role

of regulariser to P when using those bounds as learning objectives, the role of P in such results is meaningless as it is not a good comparison point to express the benefits of the optimisation procedure.

Data-dependent priors are not enough to explain the generalisation gain throughout optimisation. As shown above, in order to have a sound theoretical control on the generalisation ability of the algorithmic output $\hat{Q}_N$, 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 scenarios:

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

2. The prior is obtained from an algorithmic method on a fraction of training data. Then, such a bound does not inform us whether our optimisation method has been able to reach an optimum during the training phase: similarly to a test bound, it mainly attest the efficiency post-training of the output of the learning algorithm. A relevant example is Table 3 of Perez-Ortiz et al. (2021) 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 wanted pattern to replace Equation (1). In order to attest whether the output of a learning algorithm possess a good generalisation ability, a PAC-Bayes bound should satisfy the following generic pattern:

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

where f is a decreasing function to 0 as N goes to infinity, which comes from the optimisation procedure, D is the way to measure the discrepancy between $P, \hat{Q}$ (classically it would be the KL divergence) and ε is a residual term which could contain for instance the discrepancy $\mathrm{KL}(Q^*, \hat{Q})$ between the approximation and the true minimiser.

Such a guarantee would give theoretical evidence that the generalisation ability of $\hat{Q}_N$ 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 which ensure an optimisation procedure such that $\mathrm{KL}(\hat{Q}_N, \hat{Q}) \leq f(N)\,\mathrm{KL}(P, \hat{Q})$. Such a void is unfortunate, but not surprising as the KL divergence is not a distance: it is not easy to fit optimisation guarantees, often based on geometric properties of the loss, with respect to KL divergence.

Our guiding principle. A legitimate question is then the following: is it possible to extend PAC-Bayes theory beyond KL divergence in order to explain before training, with a bound satisfying Equation (2)'s pattern, why the output of optimisation procedure have good generalisation ability? We structured our work in order to provide a positive answer to this question. More precisely we develop a WPB bound satisfying the pattern of Equation (2) for the output of the *Bures-Wasserstein SGD* (Lambert et al., 2022).

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 1 *The 1-Wasserstein distance between $P, Q \in \mathcal{P}_1(\mathcal{H})$ is defined as*

$$W_1(Q,P) = \inf_{\pi \in \Pi(Q,P)} \int_{\mathcal{H}^2} ||x-y|| d\pi(x,y).$$

where $\Pi(Q,P)$ denote the set of probability measures on $\mathcal{H}^2$ whose marginals are Q and P.

We also define the 2-Wasserstein distance on $\mathcal{P}(\mathcal{H})$ as :

$$W_2(Q,P) = \sqrt{\inf_{\pi \in \Pi(Q,P)} \int_{\mathcal{H}^2} ||x-y||^2 d\pi(x,y)}.$$

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 Villani, 2009) of the 1-Wasserstein distance. Our first contribution is to propose a novel PAC-Bayesian exploiting another duality formula (Theorem 5.10 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 defined below.

Definition 2 *We say that a function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ is* uniformly K-Lipschitz *if for any $z \in \mathcal{Z}$, $\ell(.,z)$ is K-Lipschtiz. We also say that a function is* uniformly L-smooth *(or simply smooth) if for any $z \in \mathcal{Z}$, its gradient $\nabla \ell(.,z)$ is L-Lipschitz.*

A WPB bound for compact predictor classes. Our first result succesfully extends the PAC-Bayes philosophy when the discrepancy between measures is weighed through the 1-Wasserstein disrtance. It is stated as follows: for uniformly K-lipschitz functions bounded in $[0,1]$ with $\mathcal{H} \subseteq \mathcal{B}_R := \bar{\mathcal{B}}(0_{\mathbb{R}^d}, R)$. This bounds holds for any prior $P \in \mathcal{M}_1(\mathcal{H})$ with probability at least $1-\delta$, for any posterior distribution $Q \in \mathcal{M}_1(\mathcal{H})$:

$$|\Delta_S(Q)| \leq \mathcal{O}\left(\sqrt{2K(2K+1)\frac{2d \log\left(3\frac{1+2Rm}{\delta}\right)}{m}(1+W_1(Q,P)) + \frac{\log\left(\frac{m}{\delta}\right)}{m}} \right).$$

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 that the lipschitzness (with high probability) of a relevant functional. The duality theorem of Villani (2009, Theorem 5.10) allow us to generate a local change of measure inequality (see *e.g.* Donsker and Varadhan, 1975) required to perform PAC-Bayes learning.

This bounds is stated in Theorem 8 and further discussed in Section 2. However, this result does not cover the celebrated case of PAC-Bayes with Gaussian priors and posteriors. We then developed our next result to encompass this important particular case.

WPB bounds with Gaussians measures for unbounded losses. At the cost of the calculus of the residuals of Euler's Gamma function we obtain in Theorem 9, stated in Section 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 living in a compact $C_{\alpha,\beta,M} \subseteq \mathrm{BW}(\mathbb{R}^d)$ with probability at least $1 - \delta$ for any posterior distribution $Q \in \mathcal{C}$:

$$|\Delta_S(Q)| \leq \mathcal{O}\left(\sqrt{2K(2K+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(1+\sqrt{d/m}+W_1(Q,P)\right)+\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

where $R = \mathcal{O}(\max\sqrt{d\log(d)}, \sqrt{\log(m)})$.

This shows that, using R as an hyperparameter, we are able to nearly maintain the same convergence rate than Theorem 8 at the cost of an extra factor of $\sqrt{\log(dm)}$.

More surprisingly, we are able to remove in Corollary 10 the boundedness assumption to obtain a WPB bound, valid for unbounded uniformly K-lipschitz function with an additional boundedness assumption on $\sup_z \ell(0, z)$. This bound is more sensitive to the dimension of the prolem when few data is available. However the asymptotic dependency remains (nearly) unchanged, at the cost of an extra polynomial factor in $\log(dm)$.

$$|\Delta_S(Q)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1+W_1(Q,P)\right)+\left(1+K^2\log(m)\right)\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right). \tag{3}$$

$\tilde{\mathcal{O}}$ hides a polynomial dependency in $(\log(d), \log(m))$. This result is further discussed ion Section 3.

Our proof technique also allow us to deal with the case of (possibly unbounded) convex smooth loss functions. More details are gathered in Theorem 11 and Corollary 12.

A WPB bound with data-dependent prior. As our guiding priciple is to intricate optimisation results with generalisation ones, we have to overcome the Bayesian paradigm of data-free priors to compare our candidate measure with the optimisation goal. To do so we develop in Section 4 the idea of Dziugaite and Roy (2018) which exploits 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 10 when taking as 'prior', a Gibbs posterior. We propose 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 $Q \in C_{\alpha,\beta,M}$, we have the following asymptotic convergence rate:

$$|\Delta_S(Q)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1+W_1(Q, P_{-\frac{\lambda}{2K}R_S})\right) + (1+K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

We also study non-asymptotic regimes in Theorem 13. While Dziugaite and Roy (2018) exploited differential privacy results results for the Gibbs posterior when the loss is bounded. We succesfully extended them to (possibly unbounded) uniformly lipschitz losses. Ths fact is independent of the WPB framework.

PAC-Bayes provides generalisation guarantees for the Bures-Wasserstein SGD. Until now, we developed a Wasserstein PAC-Bayes theory, this led to a shift with traditional assumptions when using KL divergence. Indeed statiscal 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 WPB theory to provide generalisation guarantees for the Bures-Wasserstein SGD (recalled in Algorithm 1) which approximate the best Gaussian surrogate $\hat{Q}$ of $Q^* := P_{-\frac{\lambda}{2K}R_S}$ (in the sense of the KL divergence, see Section 5 for more explanations).

More precisely, we show that 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 $\hat{Q}_N$, the output of the BW-SGD after N optimisation steps, which is provably close from $\hat{Q}$ with respect to the 2-Wasserstein distance (see Theorem 14). Within the WPB framework, this link is translated in Theorem 16 as a generalisation bound ensuring that asymptotically, the minima reached by the Bures-Wasserstein SGD has a strong generalisation ability.

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

$$|\Delta_S(\hat{Q}_N)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1+W_1(\hat{Q}, Q^*)\right) + (1+K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right),$$

Thus the WPB framework is able to furnish an explicit convergence rate for the generalisation gap avoiding the comparison to an arbitrary prior. Instead our bound shows that a (long enough) run of the Bures-Wasserstein SGD with enough data (or a lipschitz constant small enough) ensure that we will obtain a minimiser with a good generalisation ability.

Furthermore, Theorem 16 is a reformulation of Equation (12), which is to our knowledge, the first PAC-Bayesian bound satisfying the pattern of Equation (2) with $D = \sqrt{dW_2}$ and $\varepsilon = \mathcal{O}(\sqrt{dW_1(\hat{Q}, Q^*)})$. This provides elements of answer to the question asked in the end of Section 1.2 and concludes our work.

Discussion about the assumptions For the sake of clarity, we provide in Figure 1.3 the topography of our main results. We focus on the assumptions required to state each of

our main results and doing so we aim to give to the reader a broader vision of when can our results be applied. We particularly insist that the lipschitzness assumption is at the core of nearly all of our results (except Theorem 11 and Corollary 12). The convexity is required to apply differential privacy and then to obtain Theorem 13. Finally we also precise that while the results of Lambert et al. (2022) are usable with only smoothness and convexity, we are required to add the uniform lipschitz assumption to obtain Theorem 16. The question of whether all of those assumptions are minimal to perform WPB remains open.

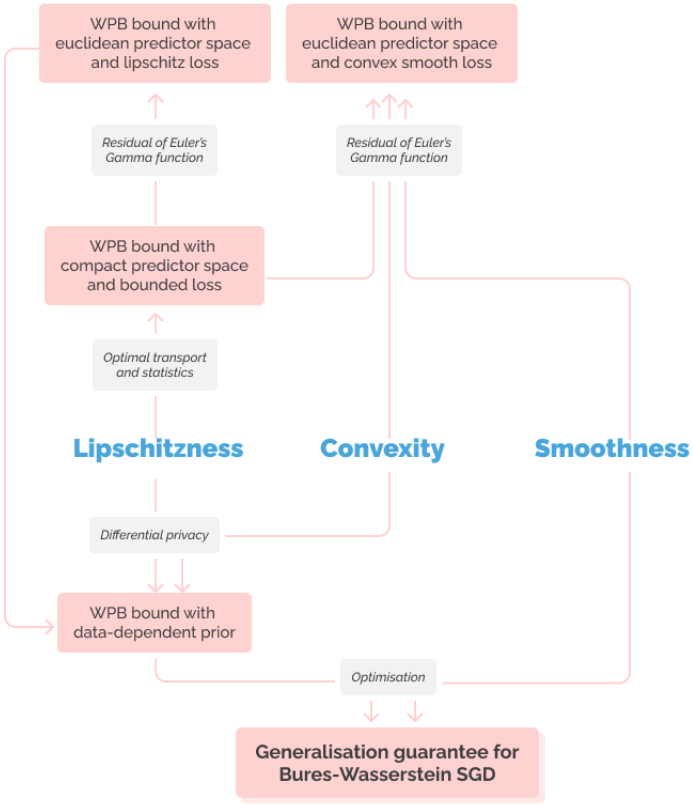

Figure 1: An overwiew of the assumptions required to perform our results. Assumptions are stated in blue, main results lies within pink boxes and the proof technque exploited to obtain results are stated within grey boxes.

2 PAC-Bayesian bounds for compact predictor spaces

We propose in this section 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 route of proof, our goal is to design a surrogate of the change of measure inequality (Donsker and Varadhan, 1975) 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 trasport framework) on the 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 A.1.

Definition 3 (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:*
$N(\mathcal{H}, \varepsilon) := \min\{n \geq 1 \mid \exists \text{ an } \varepsilon\text{-covering of } \mathcal{H} \text{ of size } n\}$

We also define the ε, 1-Wasserstein to be $W_\varepsilon(Q, P) = \varepsilon + W_1(Q, P)$. This cost function is essential to our analysis.

We now state the main results of this section. Additionnal background is gathered in Appendix A.1.

2.1 A Catoni-type bound

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

Theorem 4 *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 $P \in \mathcal{P}_1(\mathcal{H})$ a (data-free) prior distribution and assume that we take 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 $Q \in \mathcal{P}_1(K)$:

$$\Delta_S(Q) \leq 4K\varepsilon + \frac{W_1(Q, P) + 2\varepsilon + \log(2/\delta)}{\lambda} + \frac{\lambda}{2m}.$$

We assumed the loss to be bounded, this assumption can be relaxed to subgaussiannity at no cost. In Theorem 4, 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 replacing the 1-Wasserstein. In WPB this is required to have a control on Δ_S which is exploited in Kantorovich duality (Theorem 18). Furthermore, assuming lipschitzness on a compact space is not that 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 determine the impact of data with regards to the regulariser $W_1(Q, P)$. A small K says that huge variations between data have a small impact on the loss value, then we can give more impact 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 analoguous to the one derived from Alquier et al. (2016, Theorem 4.1) (which yields Gibbs posteriors):

$$\text{argmin}_{Q \in \mathcal{P}_1(\mathcal{H})} \frac{W_1(Q, P)}{\lambda} + \frac{\lambda}{2m}.$$

Theorem 4's proof is stated below and mixes up several arguments from optimal transport with PAC-Bayes learning through covering numbers.

Proof

Step 1: define a good data-dependent function We propose to define, for any sample S and predictor $h \in \mathcal{H}$ we set:

$$f_S(h) = \lambda \Delta_S(h).$$

This function satisfies the following lemma:

Lemma 5 *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:*

$$f_S(h) - f_S(h') \leq 2(1 + 2\lambda K)\varepsilon + ||h - h'||.$$

Proof 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:

$$f_S(h) - f_S(h') = \frac{\lambda}{m}\sum_{i=1}^{m} \mathbb{E}[\ell(h, z) - \ell(h'z)] - \left(\ell(h, z_i) - \ell(h', z_i))\right).$$

We know that for any h, h', z, $|\ell(h, z) - \ell(h', z)| \leq \lambda K||h - h'||$. 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 $(h, h') \in C^2$:

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

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

$$\begin{aligned} f_S(h) - f_S(h') &= f_S(h) - f_S(h_0) + f_S(h_0) - f_S(h'_0) + f_S(h'_0) - f_S(h') \\ &\leq 2\lambda K\left(||h - h_0|| + ||h' - h'_0||\right) + \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\lambda K||h_0 - h'_0|| \\ &\leq 4\lambda K\varepsilon + \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\lambda K||h_0 - h'_0||. \end{aligned}$$

By triangle inequality, $||h_0 - h'_0|| \leq ||h - h'|| + 2\varepsilon$ so we finally have with probability at least $1 - \delta$, for any $h, h' \in K^2$:

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

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_S. 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_S described below:

Lemma 6 *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 S, for any $P \in \mathcal{P}_1(K)$:*

$$\left(\sup_{Q\in\mathcal{P}_1(K)} \mathbb{E}_{h\sim Q}[f_S(h)] - (2(1+\lambda K)\varepsilon - W_1(Q,P)\right) \leq \mathbb{E}_{h\sim P}[f_S(h)].$$

Proof First of all 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 earlier:

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

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 18 as follows: for all $Q, P \in \mathcal{P}_1(\mathcal{H})$:

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

A crucial point is that for a well-chosen λ with high probability, the pair (f_S, f_S) satisfies the condition stated under the last supremum. It is formalised in the lemma below:

Lemma 7 *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 S that, for all measures $Q, P \in \mathcal{P}_1(\mathcal{H})^2$:*

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

Thus, Kantorovich duality (Theorem 18) gives:

$$\left(\sup_{Q\in\mathcal{P}_1(\mathcal{H})} \mathbb{E}_{h\sim Q}[f_S(h)] - W_{\varepsilon'}(Q,P)\right) \leq \mathbb{E}_{h\sim P}[f_S(h)],$$

and using $W_{\varepsilon'} = \varepsilon' + W_1$ and the definition of ε' concludes the proof.

Proof Because our space of predictors $\mathcal{H}$ is compact and that for any $z \in \mathcal{Z}$, the loss function $\ell(., z)$ is K-lipschitz on $\mathcal{H}$, then both the generalisation and empirical risk are continuous on $\mathcal{H}$. Thus $|f_S|$ is also continuous and, by compacity, reaches its maximum M_S on $\mathcal{H}$. Thus for any probability P on $\mathcal{H}, \mathbb{E}_{h\sim P}[|f_S(h)|] \leq M_S < +\infty$ almost surely. This proves the first bullet.

We notice that the second bullet, given our choice of λ, is the exact conclusion of Lemma 5 with probability at least $1-\delta$.

So with probability at least $1-\delta$, Kantorovich duality gives us that for any P, Q with $\varepsilon' = 2(1+\lambda K)\varepsilon$,

$$\mathbb{E}_{h\sim Q}[f_S(h)] - \mathbb{E}_{h\sim P}[f_S(h)] \leq W_{\varepsilon'}(Q, P).$$

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

■

Step 3: The PAC-Bayes route of proof for the 1-Wasserstein distance We start by exploiting Lemma 6: for any prior $P \in \mathcal{P}_1(K)$, 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:

$$\left(\sup_{Q\in\mathcal{P}_1(K)} \mathbb{E}_{h\sim Q}[f_S(h)] - (2(1+2\lambda K)\varepsilon - W_1(Q,P)\right) \leq \mathbb{E}_{h\sim P}[f_S(h)].$$

We then notice that by Jensen Inequality, $\mathbb{E}_{h\sim P}[f_S(h)] \leq \log\left(\mathbb{E}_{h\sim P}[\exp(f_S(h))]\right)$. Then, by Markov's inequality we have with probability $1-\delta/2$:

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

By Fubini and Hoeffding lemma applied m times on the iid sample S, we have

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

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

$$\mathbb{E}_{h\sim Q}[R(h)] \leq \mathbb{E}_{h\sim Q}[R_m(h)] + 4K\varepsilon + \frac{W_1(Q,P) + 2\varepsilon + \log(2/\delta)}{\lambda} + \frac{\lambda}{2m}.$$

Finally, we know that $\mathcal{H}$ is bounded by R so by Proposition 17 we have: $N^2 = N(\bar{\mathcal{B}}(0,R),\varepsilon)^2 \leq (1+2mR)^{2d}$.

Thus, we are allowed us to take λ equal to $\frac{1}{K}\sqrt{\frac{2m}{2d\log(1+\frac{2R}{\varepsilon})+\log(\frac{2}{\delta})}}$. This concludes the proof.

■

2.2 A McAllester-type bound

Our proof technique is rich enough to provide a McAllester type bound, possibly tighter for great values of the 1-Wasserstein.

Theorem 8 *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 $P \in \mathcal{P}_1(\mathcal{H})$ a (data-free) prior distribution. Then, with probability $1-\delta$, for any posterior distribution $Q \in \mathcal{P}_1(\mathcal{H})$:*

$$|\Delta_S(Q)| \leq \sqrt{2K(2K+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(W_1(Q,P)+\varepsilon_m\right) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}},$$

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

We deteriorate the bound of Amit et al. (2022) by transforming a convergence rate of $\sqrt{\frac{W_1(Q,P)}{m}}$ for finite predictor classes onto a $\sqrt{\left(KdW_1(Q,P)+1\right)\frac{\log(m)}{m}}$ 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 $\mathcal{O}\left(\sqrt{\log\left(\frac{m}{\delta}\right)/m}\right)$ which is dimension free and is a consequence of the statistical component of PAC-Bayes learning. Furthermore, note that our proof technique allow us to recover Amit et al. (2022)s rate when considering finite classes.

The proof of Theorem 8 follows the same general route than the one of Theorem 4, with additional calculations. Therefore we defer it to Appendix B.1.

3 PAC-Bayesian bounds for Gaussian distributions

In this section we elaborate 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; Haddouche and Guedj, 2022a). Those bounds are elaborated on Theorem 8 and the overall conclusion is the following: when considering functions with interesting geometric properties (*i.e.* lipschitzness or smothness) on $\mathbb{R}^d$, WPB bounds holds for Gaussian priors and posteriors over $\mathcal{H} = \mathbb{R}^d$ at the cost of small negligible terms (Theorems 9 and 11). More importantly, we whow that in this setup, the assumption of bounded loss is not required anymore to perform WPB: only boundedness on a compact is needed. Thus we are able to propose WPB bounds for unbounded losses (Corollaries 10 and 12)

Our sets of assumption. Until now we assumed working with losses satifying the uniformly Lipschitz condition (Theorem 2) as well as boundedness in $[0,1]$ within a compact bounded by R. To extend our results to unbounded functions on $\mathbb{R}^d$, we provide below two set of assumptions encapsulating the uniformly Lipschitz condition on any compact set.

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

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

Example 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 continously 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 $\ell(h,z) = ||\phi(h) - \psi(z)||^2$ Indeed, $\nabla_h \ell(h,z) = 2(\nabla\phi(h) - \psi(z))$ so on any compact $\mathcal{K}$ bounded by R, $\nabla_h \ell$ is uniformly 2-Lipschitz. Also $\sup_{z\in\mathcal{Z}} ||\nabla_h \ell(0,z)|| \leq 2C$. Note that on $\mathbb{R}^d$, $\ell(.,z)$ is not necessarily Lipschitz for any z (take the case $\phi = Id_{\mathbb{R}^d}$) so **(A1)** is not satisfied.*

A brief summary of our proof technique. To extend Theorem 8 to the case $\mathcal{H} = \mathbb{R}^d$, we use the push-forward distribution $\mathcal{P}_R \# P$ where $P \in C_{\alpha,\beta,M}$ for fixed α, β, M (see Section 1.1 for a recap on those notations). The interest of this manoeuver is to use Theorem 8 by considering projections of our gaussian prior and posterior distributions. When considering Gaussian distributions, the gap between projected distributions and orignal ones is explicitly controlled.

More precisely, for any $R > 0$ large enough, for any $P \in C_{\alpha,\beta,M}$, $W_1(P, \mathcal{P}_R \# P)$ is upper bounded. This is the conclusion of the important technical lemma (Lemma 21), stated with additional background in Appendix A.2. We state below our main WPB results, those holding for lipschitz functions in Section 3.1 and those for smooth functions in Section 3.2.

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 8 on all $\mathbb{R}^d$ when the loss remain bounded. We also show that it is also possible to obtain a WPB bound when our loss function satisfies **(A1)** (*i.e.* with an additional boundedness assumption on $\sup_z \ell(0,z)$) while remaining unbounded (Corollary 10).

Theorem 9 *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 $P \in C_{\alpha,\beta,M}$ a (data-free) prior distribution. Then, with probability $1-\delta$, for any posterior distribution $Q \in C_{\alpha,\beta,M}$:

$$|\Delta_S(Q)| \leq 2\frac{\beta\sqrt{\beta}}{m} + \sqrt{2K(2K+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(W_1(Q,P) + \alpha_m\right) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}}.$$

with $R = \mathcal{O}(\max\sqrt{d\log(d)}, \sqrt{\log(m)})$.
Also $\alpha_m = 2(M+1)\frac{\beta\sqrt{\beta}}{m} + \varepsilon_m = \mathcal{O}\left(1 + \sqrt{d\log(Rm)/m}\right)$ with ε_m defined in Theorem 8.

Theorem 9 shows that, at the cost of additional residual terms, it is possible to maintain the convergence rate of Theorem 8 when considering gaussians 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 A.2.

As in Theorem 8, the idea that a small lipschitz constant tightens the bound is still conveyed here and is of great importance for Corollary 10 which provides a WPB bound for unbounded losses with higher dimension dependecy when few data is accessible.

Proof We take a specific radius R which is the smallest value satisfying `Rad`. The proof starts with a straightforward application of Theorem 8 on the compact $\mathcal{B}(0,R)$, with the prior $\mathcal{P}_R\#P$, and with high probability, for any posterior $\mathcal{P}_R\#Q$ with $Q \in C_{\alpha,\beta,M}$:

$$|\Delta_S(\mathcal{P}_R\#Q)| \leq \sqrt{2K(2K+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(W_1(\mathcal{P}_R\#Q,\mathcal{P}_R\#P)+\varepsilon_m\right)+\frac{\log\left(\frac{3m}{\delta}\right)}{m}}.$$

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

$$|\Delta_S(Q)| \leq |\Delta_S(\mathcal{P}_R\#Q)| + |\Delta_S(Q)-\Delta_S(\mathcal{P}_R\#Q)|,$$

And we also have:

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

The last line holding thanks to Lemma 21 and because $\Delta_S \in [-1,1]$.

Also we have by triangle inequality:

$$W_1(\mathcal{P}_R\#Q,\mathcal{P}_R\#P) \leq W_1(Q,\mathcal{P}_R\#Q) + W_1(Q,P) + W_1(P,\mathcal{P}_R\#P).$$

Because both $Q,P \in C_{\alpha,\beta,M}$, using again Lemma 21 gives:

$$W_1(\mathcal{P}_R\#Q,\mathcal{P}_R\#P) \leq W_1(Q,P) + 2(M+1)\frac{\beta\sqrt{2\beta}}{m}.$$

We then have:

$$|\Delta_S(Q)| \leq 2\frac{\beta\sqrt{2\beta}}{m} + \sqrt{2K(2K+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(W_1(Q,P)+\alpha_m\right)+\frac{\log\left(\frac{3m}{\delta}\right)}{m}}.$$

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

A corollary for unbounded losses. We provably extend Theorem 9 to the case of unbounded Lipschitz losses.

Corollary 10 *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 $P \in C_{\alpha,\beta,M}$ a (data-free) prior distribution. Then, with probability $1-\delta$, for any posterior distribution $Q \in C_{\alpha,\beta,M}$, the three following bounds holds:

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

$$|\Delta_S(Q)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d^{3/2}}{m}\left(\sqrt{\frac{d}{m}} + W_1(Q,P)\right) + (1+K^2 d)\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

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

$$|\Delta_S(Q)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d^{3/2}}{m}\left(1 + W_1(Q,P)\right) + (1+K^2 d)\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

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

$$|\Delta_S(Q)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1 + W_1(Q,P)\right) + (1+K^2 \log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

In all those formulas, $\tilde{\mathcal{O}}$ hides a polynomial dependency in $(\log(d), \log(m))$.

For an explicit formulation of our bounds, we refer to Equation (5).

The message conveyed by this result 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 to perform 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 $\mathcal{O}\left(\sqrt{\log(\frac{m}{\delta})/m}\right)$ of Theorem 8 into $\mathcal{O}\left(\sqrt{(1+K^2 d)\log(\frac{m}{\delta})/m}\right)$ for non-asymptotic regimes and $\mathcal{O}\left(\sqrt{(1+K^2\log(m))\log(\frac{m}{\delta})/m}\right)$ 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 low-data regime, we have the highest dimension dependency and the more data we have, the lesser it is. Note that the dimensionality of the learning problem is controlled by the Lipschitz constant with the limit rate of $\mathcal{O}\left(\sqrt{\log(\frac{m}{\delta})/m}\right)$ 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 losses. Indeed, the existing literature on unbounded losses exploits either general divergence properties (Alquier and Guedj, 2018), 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 (Kuzborskij and Szepesvári, 2019; Rivasplata et al., 2020; Haddouche and Guedj, 2022b; Jang et al., 2023).

Proof First of all, we start from Theorem 8 which gives, with probability at least $1-\delta$:

$$|\Delta_S(Q)| \leq \sqrt{2K(2K+1)\frac{\log(\frac{3}{\delta})+2d\log\left(1+2Rm\right)}{m}\left(W_1(Q,P)+\varepsilon_m\right)+\frac{\log\left(\frac{3m}{\delta}\right)}{m}}. \quad (4)$$

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 $P \in C_{\alpha,\beta,M}$. We now assume **(A1)** and consider R to be the smallest value satisfying `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 our compact is bounded by R and the loss is K-Lipschitz) and is K/D_R-Lipschitz.

Applying Equation (4) with ℓ' on $\mathcal{B}(0,R)$ and multiplying by D_R gives, with high probability, for any $Q \in C_{\alpha,\beta,M}$:

$$\begin{aligned}
&|\Delta_S(\mathcal{P}_R\#Q)| \\
&\leq D_R\sqrt{2\frac{K}{D_R}(2\frac{K}{D_R}+1)\frac{\log(\frac{1}{\delta})+2d\log\left(1+2Rm\right)}{m}\left(W_1(\mathcal{P}_R\#Q,\mathcal{P}_R\#P)+\varepsilon_m\right)+\frac{\log\left(\frac{m}{\delta}\right)}{m}} \\
&= \sqrt{2K(2K+D_R)\frac{\log(\frac{1}{\delta})+2d\log\left(1+2Rm\right)}{m}\left(W_1(\mathcal{P}_R\#Q,\mathcal{P}_R\#P)+\varepsilon_m\right)+D_R^2\frac{\log\left(\frac{m}{\delta}\right)}{m}}.
\end{aligned}$$

with $\varepsilon_m = \mathcal{O}\left(1\right)$ defined in Theorem 8.

As in Theorem 9, we have:

$$W_1(\mathcal{P}_R\#Q,\mathcal{P}_R\#P) \leq W_1(Q,P) + 2(M+1)\frac{\beta\sqrt{2\beta}}{m}.$$

We also have:

$$|\Delta_S(Q)| \leq |\Delta_S(\mathcal{P}_R\#Q)| + |\Delta_S(Q)-\Delta_S(\mathcal{P}_R\#Q)|,$$

And we also have:

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

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

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

Finally, applying Lemma 21 gives:

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

Then we have:

$$|\Delta_S(Q)| \leq 2K(M+1)\frac{\beta\sqrt{2\beta}}{m} + \sqrt{2K(2K+D_R)\frac{\log(\frac{1}{\delta})+2d\log\left(1+2Rm\right)}{m}\left(W_1(Q,P)+\alpha_m\right)+D_R^2\frac{\log\left(\frac{3m}{\delta}\right)}{m}}. \quad (5)$$

with $\alpha_m = \mathcal{O}\left(1+\sqrt{d\log(Rm)/m}\right)$ defined in Theorem 9.

Finally we exploit that $R = \mathcal{O}(\sqrt{d\log(d)}, \sqrt{\log(m)})$ (cf Remark 19) and $D_R = \mathcal{O}(1+K^2R)$, to conclude the proof for all the three regimes. ■

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 8 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 11 *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 $P \in C_{\alpha,\beta,M}$ a (data-free) prior distribution. Then, with probability $1-\delta$, for any posterior distribution $Q \in C_{\alpha,\beta,M}$:

$$|\Delta_S(Q)| \leq 2\frac{\beta\sqrt{2\beta}}{m} + \sqrt{2D_R(2D_R+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(W_1(Q,P)+\alpha_m\right)+\frac{\log\left(\frac{3m}{\delta}\right)}{m}}.$$

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 9.

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 than the one of Theorem 9, with additional technical steps. We then defer it to Appendix B.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 12 *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 $\sup_{z\in\mathcal{Z}} ||\ell(0,z)|| = D_\ell < +\infty$.

Let $P \in C_{\alpha,\beta,M}$ *a (data-free) prior distribution. Then, with probability* $1-\delta$ *, for any posterior distribution* $Q \in C_{\alpha,\beta,M}$*, the three following bounds holds:*

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

$$|\Delta_S(Q)| \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{d^{5/2}}{m}\left(\sqrt{\frac{d}{m}} + W_1(Q,P)\right)}\right).$$

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

$$|\Delta_S(Q)| \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{d^{5/2}}{m}\left(1 + W_1(Q,P)\right)}\right).$$

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

$$|\Delta_S(Q)| \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{m}}\left(1 + W_1(Q,P)\right)\right).$$

In all those previous bounds, $\tilde{\mathcal{O}}$ *hides a polynomail factor in* $(\log(d), \log(m))$.

For a complete formulation of our bounds, we refer to Equation (6)

We remark that our bound is mainly interesting in the transitory and asymptotic regime as, contrary to Corollary 10, 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 remain of interest when many data are available as the smoothness assumption is often used in optimisation.

Proof [Proof of Corollary 12]

First of all, we use Theorem 8 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 $1-\delta$:

$$|\Delta_S(Q)| \leq \sqrt{2K(2K+1)\frac{\log(\frac{3}{\delta}) + 2d\log\left(1+2Rm\right)}{m}\left(W_1(Q,P) + \varepsilon_m\right) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}}.$$

Let $P \in C_{\alpha,\beta,M}$. We fix R to be the smallest value satisfying `Rad` and we assume **(A2)**.

On $\mathcal{B}(0,R)$, as seen in the proof of Theorem 11, 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 8 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 $Q \in C_{\alpha,\beta,M}$:

$$|\Delta_S(\mathcal{P}_R\#Q)|$$
$$\leq \sqrt{2D_R(2D_R+C_R)\frac{\log(\frac{3}{\delta}) + 2d\log\left(1+2Rm\right)}{m}\left(W_1(\mathcal{P}_R\#Q, \mathcal{P}_R\#P) + \varepsilon_m\right) + C_R^2\frac{\log\left(\frac{3m}{\delta}\right)}{m}}.$$

with $\varepsilon_m = \mathcal{O}(1)$ defined in Theorem 8.
As in Theorem 9, we have:

$$W_1(\mathcal{P}_R\#Q, \mathcal{P}_R\#P) \leq W_1(Q,P) + 2(M+1)\frac{\beta\sqrt{2\beta}}{m}.$$

We also have:

$$|\Delta_S(Q)| \leq |\Delta_S(\mathcal{P}_R\#Q)| + |\Delta_S(Q) - \Delta_S(\mathcal{P}_R\#Q)|,$$

And we also have:

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

We study the last gap more carefully:

$$|\Delta_S(h) - \Delta_S(\mathcal{P}_R(h))| = \mathbb{E}_z[|\ell(h,z) - \ell(\mathcal{P}_R(h),z)|] + \frac{1}{m}\sum_{i=1}^m |\ell(h,z_i) - \ell(\mathcal{P}_R(h),z_i)|,$$

And we know that for any z, because ℓ is convex smooth:

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

We also have by convexity:

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

In any case, we have for any h, z:

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

Thus:

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

And thanks to Lemma 21, we finally have:

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

Then we have:

$$|\Delta_S(Q)| \leq \left(D_R + \frac{L}{2}(M+1)\right)(M+1)\frac{\beta\sqrt{\beta}}{m} + \sqrt{2D_R(2D_R + C_R)\frac{\log(\frac{3}{\delta}) + 2d\log\left(1+2Rm\right)}{m}\left(W_1(Q,P) + \alpha_m\right) + C_R^2\frac{\log\left(\frac{3m}{\delta}\right)}{m}}. \quad (6)$$

with $\alpha_m = \mathcal{O}\left(1 + \sqrt{d\log(Rm)/m}\right)$ defined in Theorem 9.

Finally we exploit that $R = \mathcal{O}(\sqrt{d\log(d)}, \sqrt{\log(m)})$ (cf Remark 19), that $D_R = \mathcal{O}(1+R)$ and $C_R = \mathcal{O}(1+R^2)$, to conclude the proof for all the three regimes. ■

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 (2018). However, their contribution lies 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; Rogers et al., 2016) which will allow us to exploit differentially private priors when the loss is unbounded, convex and Lipschitz. We recall in Appendix A.3 elements of framework for 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 perpetrates the spirit of Dziugaite and Roy (2018, Theorem 4.2) and is based on the following bound, which is a minor modification of Equation (5), making it valid for any prior (and not only Gaussian ones).

Theorem 13 *Assume that $d \geq 3$, $\mathcal{H} = \mathbb{R}^d$ and that the loss is convex and satisfies **(A1)**. Let $\beta_m = \mathcal{O}(1/\sqrt{m})$ and $\lambda \leq \sqrt{m}$.*

Let $P \in C_{\alpha,\beta,M}$ a (data-free) prior distribution. Then, for any $\beta_m < \delta < 1$, with probability $1-\delta$, for any posterior distribution $Q \in C_{\alpha,\beta,M}$ and the Gibbs prior $P_{-\frac{\lambda}{2K}R_S}$, the following bounds holds:

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

$$|\Delta_S(Q)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d^{3/2}}{m}\left(\sqrt{\frac{d}{m}} + W_1(Q, P_{-\frac{\lambda}{2K}R_S}) + f_R\left(P_{-\frac{\lambda}{2K}R_S}\right)\right) + (1+K^2 d)\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

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

$$|\Delta_S(Q)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d^{3/2}}{m}\left(1 + W_1(Q, P_{-\frac{\lambda}{2K}R_S}) + f_R\left(P_{-\frac{\lambda}{2K}R_S}\right)\right) + (1+K^2 d)\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

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

$$|\Delta_S(Q)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1 + W_1(Q, P_{-\frac{\lambda}{2K}R_S})\right) + (1+K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

In all those formulas, $\tilde{\mathcal{O}}$ hides a polynomial dependency in $(\log(d), \log(m))$.
For an explicit formulation of our bounds, we refer to Equation (11).
Also $R = \mathcal{O}\left(\max\sqrt{d\log(d)}, \sqrt{\log(m)}\right)$, $f_R(P) := W_1(\mathcal{P}_R\#P, P)$.

Note that the asymptotic bound, the condition to get rid of $f_R(P_{-\frac{\lambda}{2K}R_S})$ is that λ is a fixed constant, in particular it does not depend on m. This is essential to apply the law of large number: a fixed learning rate on the Gibs posterior is required for a bound with only explicit terms.

Furthermore, an important message is that lipschitz functions are totally fitted within the PAC-Bayes framework through Wasserstein distances. Indeed, not only are we able to recover McAllester or Catoni-type WPB bounds but we are also able to obtain WPB with data-dependent prior using the same range of techniques than PAC-Bayes learning with KL divergences. Data-dependent WPB bounds have also a supplementary dimension as they provide guarantees for the Bures-Wasserstein SGD of Lambert et al. (2022) as shown in Section 5.

Proof [Proof of Theorem 13]

First of all, we start from a sligthly modified version of Equation (5) which holds for any prior distribution (and not only Gaussian ones). To obtain it we restart from the triangle inequality $W_1(\mathcal{P}_R\#Q, \mathcal{P}_R\#P) \leq W_1(\mathcal{P}_R\#Q, Q) + W_1(Q, P) + f_R(P)$. with $f_R(P) := W_1(\mathcal{P}_R\#P, P)$ and we apply exactly the same route of proof than in Corollary 10. We then obtain, for any data-free prior P, with probability at least $1-\delta$, for any $Q \in C_{\alpha,\beta,M}$:

$$|\Delta_S(Q)| \leq 2K(M+1)\frac{\beta\sqrt{2\beta}}{m} + \sqrt{C_R\frac{\log(\frac{1}{\delta}) + 2d\log\left(1+2Rm\right)}{m}\left(W_1(Q,P) + \alpha_m + f_R(P)\right) + D_R^2\frac{\log\left(\frac{m}{\delta}\right)}{m}}.$$

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

$$|\Delta_S(Q)| > 2K(M+1)\frac{\beta\sqrt{2\beta}}{m} + \sqrt{C_R\frac{\log(\frac{1}{\delta})+2d\log\left(1+2Rm\right)}{m}\left(W_1(Q,P)+\alpha_m+f_R(P)\right)+D_R^2\frac{\log\left(\frac{m}{\delta}\right)}{m}}.$$

And for a given δ', let $\texttt{Ev}(P,\delta') := \{S \in \mathcal{Z}^m \mid \exists Q \in C_{\alpha,\beta,M} \text{ s.t. } \texttt{Bound}(S,P,Q,\delta') \text{ holds}\}$.

We know that for a data-free prior P, $\mathbb{P}_{S\in\mu^m}(S \in \texttt{Ev}(P)) \leq \delta$.

To exploit our differential privacy framework, we first assume having a differentially private probability kernel $\mathcal{P}$. We fix $\beta > 0$ and reexploit the idea of Dziugaite and Roy (2018):

$$\mathbb{P}_{S\sim\mu^m}\{S \in \texttt{Ev}(\mathcal{P}(S),\delta')\} \leq \mathrm{e}^{\mathbf{I}_\infty^\beta(\mathcal{P};m)} \mathop{\mathbb{P}}_{(S,S')\sim\mu^{2m}}\left\{S \in \texttt{Ev}\left(\mathcal{P}\left(S'\right)\right)\right\} + \beta \tag{7}$$

$$\leq \mathrm{e}^{\mathbf{I}_\infty^\beta(\mathcal{P};m)}\delta' + \beta = \delta. \tag{8}$$

The last line holds for any $\delta > \beta$ by fixing $\delta' = \mathrm{e}^{-I_\infty^\beta(\mathcal{P};m)}(\delta-\beta)$.

Note that $\log\left(1/\delta'\right) = \log\left(1/\delta-\beta\right) + I_\infty^\beta(\mathcal{P};m)$, this invites to bound the approx-β max-information. To do so, we need to give specific values for the pair (ε,γ). More concretely, let $\varepsilon = \sqrt{\log(m)/m}, \gamma = \varepsilon/m^4$.

Then thanks to Theorem 27, we know that for $\beta_m := \mathcal{O}(1/m)$, we have:

$$I_\infty^\beta(\mathcal{P},m) = O\left(\log(m)\right). \tag{9}$$

The last thing to do is to prove that the probability kernel $\mathcal{P}_0(S) := P_{-\lambda' m R_S}$ is (ε,γ) differentially private. This is true thanks to Theorem 26 which states that $\mathcal{P}_0$ satisfies differetnial privacy as long as $\lambda' \leq \lambda_m$ with:

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

Note that α intervenes because for any prior $P \in C_{\alpha,\beta,M}$, $-\log P(.)$ 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 (7) with $\beta = \beta_m$, that for any $\delta > \beta_m$, $\mathbb{P}_{S\sim\mu^m}\{S \in \texttt{Ev}(\mathcal{P}_0(S),\delta') \leq \delta$ with $\delta' = \mathrm{e}^{-I_\infty^\beta(\mathcal{P};m)}(\delta-\beta)$

Taking the complementary event and recalling that thanks to Equation (9), $\log\left(1/\delta'\right) = \log\left(1/\delta-\beta_m\right) + \mathcal{O}(\log(m))$ gives, for any data-free Gaussian prior P, for any $\delta > \beta_m$, with probability at least $1-\delta$, for any $Q \in C_{\alpha,\beta,M}$:

$$|\Delta_S(Q)| \leq 2K(M+1)\frac{\beta\sqrt{2\beta}}{m}$$
$$+\sqrt{C_R\frac{\log(\frac{1}{\delta-\beta_m})+\mathcal{O}(\log(m))+2d\log\left(1+2Rm\right)}{m}}\left(W_1(Q,P_{-\frac{\lambda}{2K}R_S})+\alpha'_m+f_R(P_{-\frac{\lambda}{2K}R_S})\right)$$
$$+\sqrt{D_R^2\frac{\log\left(\frac{m}{\delta-\beta_m}\right)+\mathcal{O}(\log(m))}{m}}. \quad (11)$$

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 results as it is tighter. where $\alpha'_m = \mathcal{O}(1+{}^{d\log(m)}/_m)$ has the same analytical expression than α_m (defined in Theorem 9) but where all the occurences of δ have been replaced by δ'.

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:

$$f_R(P_{-\frac{\lambda}{2K}R_S}) \leq \mathbb{E}[||X-\mathcal{P}_R(X)||] \underset{m\to\infty}{\to} 0,$$

where X follows the Gibbs distribution $P_{-\frac{\lambda}{2K}R_S}$. The convergence to zero comes from the dominated convergence theorem. Indeed,

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

with $g_m(x) = ||x-\mathcal{P}_R(x)||\frac{\exp(-\lambda R_S(x))}{\mathbb{E}_P[\exp(-\lambda R_S(x))]}$. Thus, bounding crudely gives:

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

We know that $\inf_{m\geq 1}\mathbb{E}_P[\exp\left(-\lambda R_S(x)\right)] := \inf_{m\geq 1}\mathbb{E}_P[f_m(x)] > 0$ 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 $\mathbb{R}$ 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 $\inf_{m\geq 1}\mathbb{E}_P[\exp\left(-\lambda R_S(x)\right)]>0$.

The last thing to do is to use Lemma 21 to ensure that $\int_{\mathbb{R}^d}||x-\mathcal{P}_R(x)||dP(x)\to 0$.

This allow us to get rid of f_R for the asymptotic regime and then, conclude the proof.

■

5 Generalisation ability of the Bures-Wasserstein SGD

For the sake of completeness, we recall (and precise) several elements already defined in Section 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 $\lambda > 0$:

$$\underset{Q \in \mathcal{M}_1(\mathcal{H})}{\operatorname{argmin}} \mathbb{E}_{h\sim Q}[R_S(h)] + \frac{\mathrm{KL}(Q,P)}{\lambda/2K}.$$

We considered the parameter $\lambda/2K$ as it was suggested by Theorem 13.

A closed form solution is given by the Gibbs posterior $Q^* := P_{-\lambda/2K}$ such that $dQ^* \propto \exp(-V_S(h))dh$, with $V_S(h) = \frac{\lambda}{2K} R_S(h) - \log(dP(h))$ and dh 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 inform us when we 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 $\mathcal{M}_1(\mathcal{H})$.

In this section, we focus on the VI approach. As Q^* is the result of an optimal tradeoff between the empirical loss R_S 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:

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

At the cost of this approximation, can we have an optimisation algorithm with convergence guarantees which goes to $\hat{Q}$? Furthermore, if enough data is available, does $\hat{Q}$ possess a good generalisation ability?

We first state the assumptions holding throughout the whole section.

(A3): We assume that $\mathcal{H} = \mathbb{R}^d$:

- There exists $M > 0$ such that $||\hat{m}|| \leq M$ almost surely.
- ℓ is twice differentiable. Also,**(A1), (A2)** holds. 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 $\Sigma = \mathrm{diag}(\alpha), 1 \geq \alpha > 0$. Also, we assume $\lambda \leq 2K$ in the definition of Q^*.

Note that under **(A3)**, we have $0 \prec \alpha I \preceq \nabla^2 V_S \preceq I$. 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 1.

Note that Algorithm 1 is slightly modified with respect to the original version of Lambert et al. (2022). Indeed, in our work, 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 14 as long as we assume **(A3)**.

Algorithm 1: Bures-Wasserstein SGD.

Parameters : Strong convexity parameter $\alpha > 0$, radius $M > 0$; step size $\eta > 0$, initial mean m_0, initial covariance Σ_0

Set up $Q_0 = \mathcal{N}(m_0, \Sigma_0)$.
for $k = 0..N-1$ **do**
  Draw a sample $X_k \sim Q_k$.
  Set $m_k^+ = m_k - \eta \nabla V_S(X_k)$.
  Set $M_k = I - \eta(\nabla V^2(X_k) - \Sigma_k^{-1})$.
  Set $\Sigma_k^+ = M_k \Sigma_k M_k$.
  Set $m_{k+1} = \mathcal{P}_M(m_k^+),\ \Sigma_{k+1} = \mathrm{clip}^{1/\alpha}\, \Sigma_k^+$.
end
Return $(Q_k)_{k=1\dots N}$.

Theorem 14 *Assume **(A3)**. Also, assume that $\eta \leq \frac{\alpha}{6}$ and that we initialize Algorithm 1 at a matrix satisfying $\frac{\alpha}{4} I \preceq \Sigma_0 \preceq \frac{1}{\alpha} I$. Then, for all $k \in \mathbb{N}$,*

$$\mathbb{E} W_2^2 \left(Q_k, \hat{Q}\right) \leq \exp(-\alpha k \eta) W_2^2 \left(Q_0, \hat{Q}\right) + \frac{21 d \eta}{\alpha^2}.$$

In particular, we obtain $\mathbb{E} W_2^2 \left(Q_k, \hat{Q}\right) \leq \varepsilon^2$ provided we set $h \asymp \frac{\alpha^2 \varepsilon^2}{d}$ and the number of iterations to be $k \gtrsim \frac{d}{\alpha^3 \varepsilon^2} \log\left(W_2\left(Q_0, \hat{Q}\right)/\varepsilon\right)$.

We want to incorporate Theorem 14 within our PAC-Bayes theorems. To do so, we need to make sure that not only the outputs of Algorithm 1, but also $\hat{Q}$ all lies within a compact of $BW(\mathbb{R}^d)$. To do so we exploit the following lemma, wich gathers Lemma 6 and the discussion in Section 3.3. of Lambert et al. (2022).

Lemma 15 *Assume **(A3)** and the step-size η of Algorithm 1 is lesser than $1/2$. Also in Algorithm 1, assume that $\alpha/4 I \preceq \Sigma_k$. Then $\alpha/4 I \preceq \Sigma_k^+$, and so, $\alpha/4 I \preceq \Sigma_{k+1} \preceq 1/\alpha I$. Furthermore, $I \preceq \hat{\Sigma} \preceq 1/\alpha I$.*

Thus, if the initialisation of Algorithm 1 is such that $\frac{\alpha}{4} I \preceq \Sigma_0 \preceq \frac{1}{\alpha} I$, then the sequence $(Q_k)_{k \geq 0}$ and Q^* all lies in the compact $C_{\alpha, 1/\alpha, M}$.

Using Lemma 15, we now can apply Theorem 13 and obtain the main result of this section.

Theorem 16 *Assume **(A3)**, also assume that $d \geq 3$. Let $\beta_m = \mathcal{O}(1/\sqrt{m})$ and fix any $\beta_m < \delta < 1$.*

Assume that we perform Algorithm 1, with step size $\eta \asymp \frac{\alpha^2 \delta}{d}$ and the number of iterations to be $N \gtrsim \frac{d}{\alpha^3 \delta} \log\left(W_2\left(Q_0, \hat{Q}\right)/\delta\right)$. We also set the initialisation such that $\frac{\alpha}{4} I \preceq \Sigma_0 \preceq \frac{1}{\alpha} I$, then we can upper bound the generalisation ability of $\hat{Q}_N$ with probability $1 - 2\delta$:

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

$$|\Delta_S(\hat{Q}_N)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K \frac{d}{m}\left(1 + W_1(\hat{Q}, Q^*)\right) + (1 + K^2 \log(m)) \frac{\log\left(\frac{m}{\delta}\right)}{m}}\right),$$

where $\tilde{\mathcal{O}}$ hides a polynomial dependency in $(\log(d), \log(m))$. We refer to Equation (12) for a bound presenting the explicit influence of the Bures-Wasserstein SGD.

Theorem 16 is based on Equation (12) which answers the question stated in the 'Our guiding principle paragraph' of Section 1.2. We successfully designed a bound satisfying the general pattern of Equation (2) by incorporating the optimisation guarantees of Lambert et al. (2022) onto a statistical framework.

This bound builds 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 able to generalise well. Until now PAC-Bayes guarantees were used as a check-up procedure, ensuring that optimisation is performed properly during training, this being done without the insurance that the attained minimiser will generalise well.

Let us analyse the bound: the convergence rate depends on the quality of the approximation $\hat{Q}$ 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 $\hat{Q}$ of Q^*. Again, the limit convergence rate is the statistical ersatz $\mathcal{O}\left(\sqrt{\log(m)/m}\right)$. This roughly says that we cannot hope to converge higher than a Hoeffding test bound in this setting.

Proof We start from Theorem 13, considering the asymptotic case. We have with probability $1-\delta$, for the posterior $\hat{Q}_N$ obtained after N steps of Algorithm 1 distribution $Q \in C_{\alpha,\beta,M}$ and the prior Q^*:

$$|\Delta_S(\hat{Q}_N)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(1 + W_1(\hat{Q}_N, Q^*)\right) + (1 + K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right).$$

Then, the triangle inequality gives that $W_1(Q_k, Q^*) \leq W_1(Q_k, \hat{Q}) + W_1(\hat{Q}, Q^*)$. Finally we exploit Theorem 14 as follows:

$$\begin{aligned}
W_1(Q_k, \hat{Q}) &\leq \sqrt{W_2^2(\hat{Q}_N, \hat{Q})} && \text{by Jensen} \\
&\leq \sqrt{2\frac{\mathbb{E}[W_2^2(Q_k, \hat{Q})]}{\delta}} && \text{by Markov} \\
&\leq \sqrt{2\frac{\exp(-\alpha k\eta)W_2^2\left(Q_0, \hat{Q}\right) + \frac{21d\eta}{\alpha^2}}{\delta}} && \text{by Theorem 14.}
\end{aligned}$$

Note that in the last line, we were able to apply Theorem 14 thanks to Lemma 15. This leads to the following bound:

$$|\Delta_S(\hat{Q}_N)| \leq \tilde{\mathcal{O}}\left(\sqrt{2K\frac{d}{m}\left(f(N,\eta)\sqrt{W_2^2(Q_0,\hat{Q})}+1+\varepsilon\right)+(1+K^2\log(m))\frac{\log\left(\frac{m}{\delta}\right)}{m}}\right). \tag{12}$$

where $f(N,\eta)=\sqrt{\frac{\exp(-\alpha N\eta)W_2^2(Q_0,\hat{Q})}{\delta}}$ and $\varepsilon=\sqrt{\frac{21d\eta}{\alpha^2\delta}}+W_1(\hat{Q},Q^*)$.

Finally using that with step size $\eta \asymp \frac{\alpha^2\delta}{d}$ and the number of iterations to be $N \gtrsim \frac{d}{\alpha^3\delta}\log\left(W_2\left(Q_0,\hat{Q}\right)/\delta\right)$ allow us to bound: $\sqrt{2\frac{\exp(-\alpha k\eta)W_2^2(Q_0,\hat{Q})+\frac{21d\eta}{\alpha^2}}{\delta}} \leq 1$. This concludes the proof. ■

6 Conclusion

We developed the Wasserstein PAC-Bayes theory beyond the results of Amit et al. (2022). We aimed to exploit optimisation results to explain the generalisation ability of existing algorithms and we reached this goal for the Bures-Wasserstein algorithm of Lambert et al. (2022). This work, while providing new tools for generalisation, our works raises unasnwered questions gathered below.

Can we exploit WPB for neural networks? As shown in Figure 1.3, we had to assume, lipschitzness, smoothness and convexity to reach Theorem 16. Those assumptions were required by our framework and the results of Lambert et al. (2022) and thus, does 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 able to attenuate the impact of dimensionality. Whether lipschitzness is enough to reach WPB bound with data-dependent prior (Theorem 13 requires an additional convexity assumption) could involve new differential privacy results and is left as an open question.

Are the classical PAC-Bayesian techniques suited to WPB? In Theorems 4 and 8, we exploited a surrogate of the change of measure inequality to then exploit the PAC-Bayesian heavy machinery. However, those techniques are developed around the control of an exponential moment which appears naturally through the change of measure inequality. However, our surrogate is tighter as it directly involves the true moment with regards to the prior, an interesting lead would be to check wheter tighter concentration bounds(or oter bounds exploiting weaker assumptions than a bounded loss) are accessible. Furthermore, we exploited covering numbers to state that, with high probability, our loss is close from a Lipschitz one. Those covering numbers, while crucials, involve explicitly the dimension of the problem. This is challenging as such a dependency do not appear explicitly in KL-based PAC-Bayes learning (but may be hidden within the KL term). Whether covering numbers are essential to WPB learning is still an open question.

References

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018. ISSN 1573-0565. URL http://dx.doi.org/10.1007/s10994-017-5690-0.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41, 2016. URL http://jmlr.org/papers/v17/15-290.html.

Jason Altschuler, Sinho Chewi, Patrik R Gerber, and Austin Stromme. Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22132–22145. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/b9acb4ae6121c941324b2b1d3fac5c30-Paper.pdf.

Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *International Conference on Machine Learning*, pages 205–214. PMLR, 2018.

Ron Amit, Baruch Epstein, Shay Moran, and Ron Meir. Integral Probability Metrics PAC-Bayes Bounds, 2022. URL https://arxiv.org/abs/2207.00614.

Yoshua Bengio, Nicolas Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex Neural Networks. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper_files/paper/2005/file/0fc170ecbb8ff1afb2c6de48ea5343e7-Paper.pdf.

L. E. Blumenson. A derivation of n-dimensional spherical coordinates. *The American Mathematical Monthly*, 67(1):63–66, 1960. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2308932.

O Catoni. PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. Institute of Mathematical Statistics Lecture Notes—Monograph Series 56. *IMS, Beachwood, OH. MR2483528*, 5544465, 2007.

Olivier Catoni. A PAC-Bayesian approach to adaptive classification. *preprint*, 840, 2003.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in Adaptive Data Analysis and Holdout Reuse. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/bad5f33780c42f2588878a9d07405083-Paper.pdf.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Gintare Karolina Dziugaite and Daniel M Roy. Data-dependent PAC-Bayes priors via differential privacy. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/9a0ee0a9e7a42d2d69b8f86b3a0756b1-Paper.pdf.

M Fard and Joelle Pineau. PAC-Bayesian model selection for reinforcement learning. *Advances in Neural Information Processing Systems*, 23, 2010.

Wolfgang Gabcke. *Neue Herleitung und explizite Restabschätzung der Riemann-Siegel-Formel*. PhD thesis, Georg-August-Universität Göttingen, 1979.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110:393–416, 2021.

Maxime Haddouche and Benjamin Guedj. Online PAC-Bayes Learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25725–25738. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a4d991d581accd2955a1e1928f4e6965-Paper-Conference.pdf.

Maxime Haddouche and Benjamin Guedj. PAC-Bayes with Unbounded Losses through Supermartingales. *arXiv preprint arXiv:2210.00928*, 2022b.

Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, and John Shawe-Taylor. PAC-Bayes unleashed: generalisation bounds with unbounded losses. *Entropy*, 23(10):1330, 2021.

Matthew Holland. PAC-Bayes under potentially heavy tails. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS) 32*, pages 2715–2724. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8539-pac-bayes-under-potentially-heavy-tails.pdf.

Kyoungseok Jang, Kwang-Sung Jun, Ilja Kuzborskij, and Francesco Orabona. Tighter PAC-Bayes Bounds Through Coin-Betting, 2023. URL https://arxiv.org/abs/2302.05829.

Ilja Kuzborskij and Csaba Szepesvári. Efron-Stein PAC-Bayesian Inequalities. *arXiv preprint arXiv:1909.01931*, 2019.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via wasserstein gradient flows. *arXiv preprint arXiv:2205.15902*, 2022.

Sokhna Diarra Mbacke, Florence Clerc, and Pascal Germain. PAC-Bayesian Generalization Bounds for Adversarial Generative Models, 2023.

David McAllester. Simplified PAC-Bayesian margin bounds. In *Learning theory and Kernel machines*, pages 203–215. Springer, 2003a.

David A McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational Learning Theory*, pages 230–234. ACM, 1998.

David A McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational Learning Theory*, pages 164–170. ACM, 1999.

David A McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1): 5–21, 2003b.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes Un-Expected Bernstein Inequality. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS) 32*, pages 12202–12213. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9387-pac-bayes-un-expected-bernstein-inequality.pdf.

Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential Privacy without Sensitivity. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/a7aeed74714116f3b292a982238f83d2-Paper.pdf.

Ruben Ohana, Kimia Nadjahi, Alain Rakotomamonjy, and Liva Ralaivola. Shedding a PAC-Bayesian Light on Adaptive Sliced-Wasserstein Distances, 2022.

Victor M Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

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, 2021.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:16833–16845, 2020.

Borja Rodríguez Gálvez, German Bassi, Ragnar Thobaben, and Mikael Skoglund. Tighter Expected Generalization Error Bounds via Wasserstein Distance. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19109–19121. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9f975093da0252e2c0ae181d74c90dc6-Paper.pdf.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-Information, Differential Privacy, and Post-Selection Hypothesis Testing, 2016. URL https://arxiv.org/abs/1604.03924.

Yevgeny Seldin, François Laviolette, John Shawe-Taylor, Jan Peters, and Peter Auer. PAC-Bayesian Analysis of Martingales and Multiarmed Bandits. *arXiv preprint arXiv:1105.2416*, 2011.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*, pages 2–9. ACM, 1997.

Ilya O Tolstikhin and Yevgeny Seldin. PAC-Bayes-Empirical-Bernstein Inequality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/a97da629b098b75c294dffdc3e463904-Paper.pdf.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Appendix A. Additional background

A.1 Background on optimal transport and covering numbers

We recall a basic property on covering numbers.

Proposition 17 *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 18 (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 $a \in L^1(Q)$ and $b \in L^1(P)$. Then there is duality:

$$\min_{\pi \in \Pi(Q,P)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y) \quad = \sup_{\substack{(\psi,\phi) \in L^1(Q)\times L^1(P) \\ \phi - \psi \leq c}} \left[\int_{\mathcal{Y}} \phi(y) dP(y) - \int_{\mathcal{X}} \psi(x) dQ(x) \right],$$

where $L_1(P)$ refers to the set of all functions integrable with respect to P and the condition $\phi - \psi \leq c$ means that for all $x, y \in \mathcal{X} \times \mathcal{Y}, \phi(y) - \psi(x) \leq c(x,y)$.

A.2 Technical background for Section 3

The theorems of Section 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 $\texttt{Rad}(\alpha, \beta, M, m, d)$ (abbreviated in `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. $R \geq M + \sqrt{2\beta}\sqrt{d \log\left(d \frac{m^{2/d}\sqrt{\beta}}{\sqrt{\pi\alpha}}\right)} = M + \sqrt{2\beta}\sqrt{d \log\left(d \frac{\sqrt{\beta}}{\sqrt{\pi\alpha}}\right) + 2\log(m)}$,
3. $R \geq M + \sqrt{2\beta}\sqrt{1 + \frac{d}{2}}$.

Remark 19 *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 20 (Adapted from Panaretos and Zemel (2020), Equation 2.3) *Let $P \in \mathcal{M}_1^+(\mathbb{R}^d)$ and $R > 0$. The 1-Wasserstein distance between P and $\mathcal{P}_R \# P$ is controlled as follows:*

$$W_1(P, \mathcal{P}_R \# P) \leq \int_{||\mathbf{x}|| > R} ||\mathbf{x} - \mathcal{P}_R(\mathbf{x})|| dP(\mathbf{x}) \leq \int_{||\mathbf{x}|| > R} ||\mathbf{x}|| dP(\mathbf{x}).$$

Lemma 20 invites us to consider projected distributions and to control them through the residual moments of the norm of gaussian vectors. The following results furnishes such an explicit control.

Lemma 21 *For $d \geq 3$, R satisfying `Rad`, any $Q = \mathcal{N}(\mu, \Sigma) \in C_{\alpha,\beta,M}$,*

$$Q(||h|| > R) \leq \frac{\beta\sqrt{2\beta}}{m}.$$

Also, for any $Q \in C_{\alpha,\beta,M}$:

$$W_1(Q, \mathcal{P}_R \# Q) \leq \mathbb{E}_{h \sim Q}\left[||h||\ \mathbb{1}(||h|| > R)\right] \leq (M+1)\frac{\beta\sqrt{2\beta}}{m}.$$

Finally:

$$\mathbb{E}_{h \sim Q}\left[||h||^2\ \mathbb{1}(||h|| > R)\right] \leq (M+1)^2 \frac{\beta\sqrt{2\beta}}{m}.$$

Proof of Lemma 21 is gathered in Appendix B.2.

A.3 Differential privacy background

Definition 22 (Probability kernels) *A* probability kernel $\mathcal{P}$ *from* $\mathcal{Z}^m$ *to* $\mathcal{M}_1(\mathcal{H})$ *is defined as a mapping* $\mathcal{P} : Z^m \to \mathcal{M}_1(\mathcal{H})$.

Definition 23 *A probability kernel* $\mathcal{P} : \mathcal{Z}^m \to T$ *is* (ε, γ)*-differentially private if, for all pairs* $S, S' \in Z^m$ *that differ at only one coordinate, and all measurable subsets* $B \in \Sigma_{\mathcal{H}}$*, we have*

$$\mathbb{P}\{\mathcal{P}(S) \in B\} \leq \mathrm{e}^{\varepsilon}\mathbb{P}\left\{\mathcal{P}\left(S'\right) \in B\right\} + \gamma.$$

Further, ε*-differentially private means* $(\varepsilon, 0)$*-differentially private.*

Remark 24 *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 (2018, footnote 3, Appendix A).*

For our purposes, max-information is the key quantity controlled by differential privacy.

Definition 25 (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* $\mathcal{P} : \mathcal{Z}^m \to \mathcal{M}_1(\mathcal{Z})$*, the* β*-approx. max-information of* $\mathcal{P}$*, denoted* $I_\infty^\beta(\mathcal{P}, m)$*, is the least value* k *such that, for all* $\mu \in \mathcal{M}_1(\mathcal{Z}), I_\infty^\beta(S; \mathcal{P}(S)) \leq k$ *when* $S \sim \mu^m$*. The max-information of* $\mathcal{P}$ *is defined similarly.*

Dziugaite and Roy (2018) exploited a boundedness assumption to control the exponential mechanism of McSherry and Talwar (2007). This ensures that the Gibbs posterior $\mathcal{P}(S) = P_{-\lambda m R_S}$ is ε-diffrentially private for ε given in Dziugaite and Roy (2018, 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 26 (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*

$$\lambda \leq \frac{\varepsilon}{2K}\sqrt{\frac{m_P}{1 + 2\log\left(1/\gamma\right)}}$$

Then the probability kernel $\mathcal{P} : S \to P_{-\lambda m R_S}$ *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 $m_P \geq \alpha$.

Our final preliminary result is Theorem 3.1 of Rogers et al. (2016) which upper bounds the β-approx. max-information of any (ε, γ) differentially private probability kernel.

Proposition 27 *Let $\mathcal{P} : \mathcal{Z}^n \to \mathcal{M}_1(\mathcal{H})$ 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*

$$I_\infty^\beta(\mathcal{P}, m) = O\left(\epsilon^2 m + m\sqrt{\frac{\gamma}{\epsilon}}\right).$$

Appendix B. Additional proofs

B.1 Proof of Theorem 8

Proof We fix $\lambda > 0$.

Step 1: define a good data-dependent function We propose to define, for any sample S and predictor $h \in \mathcal{H}$ we set:

$$f_S(h) = \lambda \Delta_S^2(h).$$

This function satisfies the following lemma:

Lemma 28 *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 $h, h' \in \mathcal{H}$:

$$f_S(h) - f_S(h') \leq \varepsilon_m + ||h - h'||.$$

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

Proof We rename $N := N(\mathcal{H}, \varepsilon)$. For any $h, h' \in \mathcal{H}^2$, we have:

$$f_S(h) - f_S(h') = \lambda\left(\Delta_S(h) - \Delta_S(h')\right).\left(\Delta_S(h) + \Delta_S(h')\right)$$

The proof of Lemma 5 gives us with probability at least $1 - \delta$, for any $h, h' \in \mathcal{H}^2$,

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

Thus with probability $1 - \delta$:

$$\begin{aligned} f_S(h) - f_S(h') &\leq \left(4\lambda K\varepsilon + \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\lambda K\left(2\varepsilon + ||h - h||\right)\right).\left(2\sup_{h\in K}\Delta_S(h)\right) \\ &= \lambda\left(2K\varepsilon\left(2 + \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\right) + K\sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}||h - h'||\right).\left(2\sup_{h\in K}\Delta_S(h)\right) \end{aligned}$$

Because $\mathcal{H}$ is compact and ℓ is K-lipschitz, Δ_S is continuous so there exists h_S such that $\sup_{h\in\mathcal{H}} \Delta_S(h) = \Delta_S(h_S)$.

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 $||h_S - h_0|| \leq \varepsilon$. 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 $h \in C$:

$$\Delta_S(h) \leq \sqrt{\frac{\log(\frac{N}{\delta})}{2m}} \leq \sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}}$$

Finally using that Δ_S is $2K$-lipschitz gives us with probability at least $1-\delta$:

$$\begin{aligned}\sup_{h\in K} \Delta_S(h) = \Delta_S(h_S) &= \Delta_S(h_0) + (\Delta_S(h_S) - \Delta_S(h_0)) \\ &\leq \sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}} + 2K\varepsilon\end{aligned}$$

So finally, with probability $1-2\delta$, we have, for any $h, h' \in \mathcal{H}^2$:

$$\begin{aligned}&\frac{1}{\lambda}\left(f_S(h) - f_S(h')\right) \\ &\quad\leq \left(2K\varepsilon\left(2 + \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\right) + K\sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}||h-h'||\right) \times 2\left(\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}} + 2K\varepsilon\right)\end{aligned}$$

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:

$$\begin{aligned}f_S(h) - f_S(h') &\leq \frac{2\varepsilon\left(2 + \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\right)}{\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}}\left(\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}} + 2K\varepsilon\right)} + ||h-h'|| \\ &\leq \frac{4m\varepsilon}{\log\left(\frac{N^2}{\delta}\right)}\left(2 + \sqrt{\frac{\log\left(\frac{N^2}{\delta}\right)}{2m}}\right) + ||h-h'|| \\ &\leq \frac{4}{\log(\frac{1}{\delta})}\left(2 + \sqrt{\frac{\log\left(\frac{1}{\delta}\right) + 2d\log(1+2Rm)}{2m}}\right) + ||h-h'||\end{aligned}$$

The last line holding because $N \geq 1$ and that $N \leq N(\bar{\mathcal{B}}(0,R)), \varepsilon) \leq \left(1 + \frac{2R}{\varepsilon}\right)^d$ thanks to Proposition 17 ($\varepsilon = 1/m$). This proves the lemma.

■

Step 2 A probabilistic change of measure inequality for f_S. 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_S described below:

Lemma 29 *For any λ, ε_m defined as in Lemma 28, any $\delta > 0$, we have with probability $1-2\delta$ over the sample S, for any $P \in \mathcal{P}_1(\mathcal{H})$:*

$$\left(\sup_{Q \in \mathcal{P}_1(\mathcal{H})} \mathbb{E}_{h\sim Q}[f_S(h)] - \varepsilon_m - W_1(Q,P) \right) \leq \mathbb{E}_{h\sim P}[f_S(h)].$$

Proof 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 Theorem 1 the same way we did in Lemma 6. This leads to

$$W_\varepsilon(Q,P) = \sup_{\substack{(\psi,\phi)\in L^1(Q)\times L^1(P) \\ \psi - \phi \leq c_\varepsilon}} \left[\mathbb{E}_{h\sim Q}[\psi(h)] - \mathbb{E}_{h\sim P}[\phi(h)] \right].$$

A crucial point is that for a well-chosen λ with high probability, the pair (f_S, f_S) satisfies the condition stated under the last supremum. It is formalised in the lemma below:

Lemma 30 *Given our choices of λ, ε_m, we have with probability at least $1-2\delta$ over the sample S that, for all measures $Q, P \in \mathcal{P}_1(\mathcal{H})^2$:*

- *$f_S \in L_1(Q), L_1(P)$,*
- *for all $h, h' \in \mathcal{H}^2$, $f_S(h) - f_S(h') \leq c_{\varepsilon_m}(h,h')$.*

Thus, Kantorovich duality gives us:

$$\left(\sup_{Q \in \mathcal{P}_1(\mathcal{H})} \mathbb{E}_{h\sim Q}[f_S(h)] - W_{\varepsilon_m}(Q,P) \right) \leq \mathbb{E}_{h\sim P}[f_S(h)],$$

and using $W_{\varepsilon_m} = \varepsilon_m + W_1$ concludes the proof.

Proof Because our space of predictors is compact and that for any $z \in \mathcal{Z}$, the loss function $\ell(.,z)$ is K-lipschitz on $\mathcal{H}$, then both the generalisation and empirical risk are continuous on $\mathcal{H}$. Thus $|f_S|$ is also continuous and, by compacity, reaches its maximum M_S on $\mathcal{H}$. Thus for any probability P on $K, \mathbb{E}_{h\sim P}[|f_S(h)|] \leq M_S < +\infty$ almost surely. This proves the first bullet.

We notice that the second bullet, given our choice of λ, is the exact conclusion of Lemma 28 with probability at least $1-2\delta$.

So with probability at least $1-2\delta$, Kantorovich duality gives us that for any P, Q

$$\mathbb{E}_{h\sim Q}[f_S(h)] - \mathbb{E}_{h\sim P}[f_S(h)] \leq W_{\varepsilon_m}(Q,P).$$

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

■

Step 3: The PAC-Bayes route of proof for the 1-Wasserstein distance We start by exploiting Lemma 29: for any prior $P \in \mathcal{P}_1(\mathcal{H})$, for λ, ε_m defined as in Lemma 28, with probability at least $1-2\delta$ we have:

$$\left(\sup_{Q \in \mathcal{P}_1(\mathcal{H})} \mathbb{E}_{h\sim Q}[f_S(h)] - \varepsilon_m - W_1(Q,P) \right) \leq \mathbb{E}_{h\sim P}[f_S(h)].$$

We then notice that by Jensen Inequality:
$\mathbb{E}_{h\sim P}[f_S(h)] \leq \frac{\lambda}{2(m-1)} \log\left(\mathbb{E}_{h\sim P}[\exp(2(m-1)\Delta_S^2(h))]\right)$.
Then, by Markov's inequality we have with probability $1-\delta$:

$$\mathbb{E}_{h\sim P}[f_S(h)] \leq \frac{\lambda}{2(m-1)} \log\left(\frac{\mathbb{E}_S \mathbb{E}_{h\sim P}\left[\exp\left(2(m-1)\Delta_S^2(h)\right)\right]}{\delta} \right).$$

By Fubini and Lemma 5 of McAllester (2003a), we have

$$\mathbb{E}_S \mathbb{E}_{h\sim P}\left[\exp(f_S(h))\right] \leq m.$$

Taking an union bound and dividing by λ gives us with probability $1-3\delta$, for any posterior Q:

$$\mathbb{E}_{h\sim Q}[\Delta_S^2(h)] \leq \frac{W_1(Q,P)+\varepsilon_m}{\lambda} + \frac{\log\left(\frac{m}{\delta}\right)}{2(m-1)}.$$

We also remark that we can upper bound λ:

$$\begin{aligned}
\lambda^{-1} &= 2K\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}}\left(\sqrt{\frac{\log(\frac{N^2}{\delta})}{2m}} + \frac{2K}{m}\right) \\
&\leq 2K(2K+1)\frac{\log(\frac{1}{\delta}) + 2d\log\left(1+2Rm\right)}{2m}.
\end{aligned}$$

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 17.

Then; bounding $1/2m, 1/(2m-1)$ by $1/m$ gives us, with probability at least $1-3\delta$, for any posterior Q:

$$\mathbb{E}_{h\sim Q}[\Delta_S^2(h)] \leq 2K(2K+1)\frac{\log(\frac{1}{\delta}) + 2d\log\left(1+2Rm\right)}{m}\left(W_1(Q,P)+\varepsilon_m\right) + \frac{\log\left(\frac{m}{\delta}\right)}{m}$$

We finally exploit Jensen's inequality to remark that for any Q, $\mathbb{E}_{h\sim Q}[\Delta_S^2(h)] \geq \left(\mathbb{E}_{h\sim Q}[\Delta_S(h)]\right)^2$.

Then, with probability at least $1-3\delta$, for any posterior Q:

$$|\Delta_S(Q)| \leq \sqrt{2K(2K+1)\frac{2d\log\left(\frac{1+2Rm}{\delta}\right)}{m}\left(W_1(Q,P)+\varepsilon_m\right)+\frac{\log\left(\frac{m}{\delta}\right)}{m}}$$

Taking $\delta' = \delta/3$ concludes the proof. ■

B.2 Proof of Lemma 21

We recall our lemma of interest:

Lemma 31 *For R satisfying the* **Rad** *assumption, any $Q = \mathcal{N}(\mu,\Sigma) \in C_{\alpha,\beta,M}$,*

$$Q(||h|| > R) \leq \frac{\beta\sqrt{2\beta}}{m}.$$

Also, for any $Q \in C_{\alpha,\beta,M}$:

$$W_1(Q, \mathcal{P}_R\#Q) \leq \mathbb{E}_{h\sim Q}\left[||h||\ \mathbb{1}(||h|| > R)\right] \leq (M+1)\frac{\beta\sqrt{2\beta}}{m}.$$

Finally:

$$\mathbb{E}_{h\sim Q}\left[||h||^2\ \mathbb{1}(||h|| > R)\right] \leq (M+1)^2\frac{\beta\sqrt{2\beta}}{m}.$$

Proof [Proof of Lemma 21]

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:

$$\sqrt{(2\pi)^d|\Sigma|}.Q(||x|| > R) = \int_{||\mathbf{x}||>R} f_{\mu,\Sigma}(\mathbf{x})d\mathbf{x} \leq \int_{||\mathbf{x}||>R-M} f_{0,\Sigma}(\mathbf{x})d\mathbf{x}$$

where $|\Sigma|$ the determinant of Σ.

We now use that because $Q \in C_{\alpha,\beta,M}$, $\alpha Id \preceq \Sigma \preceq \beta Id$. 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:

$$Q(||h|| > R) = \frac{1}{\sqrt{2\pi\alpha}^d}\int_{||\mathbf{x}||>(R-M)} \exp\left(\frac{1}{2\beta}||\mathbf{x}||^2\right)d\mathbf{x}$$

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 $\frac{(R-M)^2}{2\beta} \geq 1 + \frac{d}{2}$):

$$\Gamma\left(1+\frac{d}{2}, \frac{(R-M)^2}{2\beta}\right) \leq \frac{d+2}{2} \exp\left(-\frac{(R-M)^2}{2\beta}\right)\left(\frac{(R-M)^2}{2\beta}\right)^{d/2}.$$

We now control this quantity through the following lemma:

Lemma 32 *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 :*

$$f(r) \leq -\frac{d}{2}\log\left(\sqrt{\frac{\beta}{\pi\alpha}}\right) - \log(m) - \log\left(\frac{d+2}{2}\right).$$

Proof First of all, f is decreasing on $[d/2, +\infty)$. Notice that if $r_0 = d\log\left(d\frac{m^{2/d}\sqrt{\beta}}{\sqrt{\pi\alpha}}\right)$, then $r_0 \geq d/2$ 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:

$$\log(r_0) = \log(d) + \log\left(\log\left(dm^{2/d}\sqrt{\frac{\beta}{\pi\alpha}}\right)\right)$$

So, multiplying by $d/2$ gives:

$$\frac{d}{2}\log(r_0) = -\frac{d}{2}\log\left(m^{2/d}\sqrt{\frac{\beta}{\pi\alpha}}\right) + \frac{r_0}{2} + \frac{d}{2}\log\log\left(dm^{2/d}\sqrt{\frac{\beta}{\pi\alpha}}\right)$$

Finally:

$$f(r_0) = -\frac{d}{2}\log\left(m^{2/d}\sqrt{\frac{\beta}{\pi\alpha}}\right) - \frac{r_0}{2} + \frac{d}{2}\log\log\left(dm^{2/d}\frac{\beta}{\pi\alpha}\right)$$

We conclude the proof by proving

$$-\frac{r_0}{2} + \frac{d}{2}\log\log\left(dm^{2/d}\sqrt{\frac{\beta}{\pi\alpha}}\right) \leq -\log\left(\frac{d+2}{2}\right)$$

. Note that this sentence is equivalent to:

$$dm^{2/d}\sqrt{\frac{\beta}{\pi\alpha}} - \left(1+\frac{d}{2}\right)^{2/d}\log\left(dm^{2/d}\sqrt{\frac{\beta}{\pi\alpha}}\right) \geq 0.$$

This last sentence is true because for $d \geq 3$, $\left(1+\frac{d}{2}\right)^{2/d} \leq 2$ and the function $\mathbb{R}^+$, $x \to x - 2\log(x)$ is positive. This concludes the proof. ■

We then have

$$\exp\left(-\frac{(R-M)^2}{2\beta}\right)\left(\frac{(R-M)^2}{2\beta}\right)^{d/2} = \exp\left(f\left(\frac{(R-M)^2}{2\beta}\right)\right) \leq \sqrt{\frac{\pi\alpha}{\beta}}^d \times \frac{2}{d+2} \times \frac{1}{m}$$

Hence the final bound:

$$Q(||h|| > R) \leq \frac{\beta\sqrt{2\beta}}{m}$$

Second bound. We use Lemma 20 to have

$$W_1(Q, \mathcal{P}_R\#Q) \leq \int_{||\mathbf{x}||>R} ||\mathbf{x} - \mathcal{P}_R(\mathbf{x})|| dP(\mathbf{x})$$

By definition of the projection on a closed convex, $||\mathbf{x} - \mathcal{P}_R(\mathbf{x})|| \leq ||\mathbf{x}||$. Thus:

$$\begin{aligned}
&\leq \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{||\mathbf{x}||>R} ||\mathbf{x}|| f_{\mu,\Sigma}(\mathbf{x})d\mathbf{x} \\
&\leq \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{||\mathbf{x}||>R} ||\mathbf{x}-\mu|| f_{\mu,\Sigma}(\mathbf{x})d\mathbf{x} + MQ(||h|| > R) \\
&\leq \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{||\mathbf{x}||>R} ||\mathbf{x}-\mu|| f_{\mu,\Sigma}(\mathbf{x})d\mathbf{x} + \frac{M\beta\sqrt{2\beta}}{m},
\end{aligned}$$

The last line holding thanks to first part of the proof, then using again that $||\mu|| \leq M$ gives:

$$W_1(Q, \mathcal{P}_R\#Q) \leq \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{||\mathbf{x}||>R-M} ||\mathbf{x}|| f_{0,\Sigma}(\mathbf{x})d\mathbf{x} + \frac{M\beta\sqrt{2\beta}}{m}.$$

Then using the same arguments than in the first part of the proof gives:

$$W_1(Q, \mathcal{P}_R \# Q) \leq \frac{1}{\sqrt{2\pi\alpha}^d} \int_{||\mathbf{x}||>R-M} ||\mathbf{x}|| \exp\left(-\frac{||\mathbf{x}||^2}{2\beta}\right) d\mathbf{x} + \frac{M\beta\sqrt{2\beta}}{m}$$

We use the hyperspherical coordinate to obtain:

$$\begin{aligned}
\int_{||\mathbf{x}||>R-M} ||\mathbf{x}|| \exp\left(-\frac{||\mathbf{x}||^2}{2\beta}\right) d\mathbf{x} &= \int_{R-M}^{+\infty} r^d \exp\left(-\frac{r^2}{2\beta}\right) dr \\
&\leq \int_{R-M}^{+\infty} r^{d+1} \exp\left(-\frac{r^2}{2\beta}\right) dr \\
&= \beta\sqrt{2\beta}^{d+1} \int_{\frac{(R-M)^2}{2\beta}}^{+\infty} r^{d/2} \exp^{-r} dr \\
&= \beta\sqrt{2\beta}^{d+1} \Gamma\left(\frac{d+1}{2}, \frac{(R-M)^2}{2\beta}\right).
\end{aligned}$$

The second line holding because $R - M \geq 1$. Then applying again Theorem 32 gives:

$$\begin{aligned}
\mathbb{E}_{h\sim Q}\left[||h||\ \mathbb{1}(||h|| > R)\right] &\leq \beta\sqrt{2\beta}\sqrt{\frac{\beta}{\pi\alpha}}^d \times \frac{d+2}{2}\sqrt{\frac{\pi\alpha}{\beta}}^d \times \frac{2}{d+2} \times \frac{1}{m} + \frac{M\beta\sqrt{2\beta}}{m} \\
&= (M+1)\frac{\beta\sqrt{2\beta}}{m}.
\end{aligned}$$

Hence the final bound:

$$W_1(Q, \mathcal{P}_R \# Q) \leq \mathbb{E}_{h\sim Q}\left[||h||\ \mathbb{1}(||h|| > R)\right] \leq (M+1)\frac{\beta\sqrt{2\beta}}{m}$$

Third bound. We start again as follows:

$$\begin{aligned}
\mathbb{E}_{h\sim Q}[||h||^2 \mathbb{1}(||h|| > R)] &= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \int_{||\mathbf{x}||>R} ||\mathbf{x}||^2 f_{\mu,\Sigma}(\mathbf{x}) d\mathbf{x} \\
&= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \int_{||\mathbf{x}||>R} ||\mathbf{x}-\mu||^2 + 2\langle\mu, \mathbf{x}-\mu\rangle + ||\mu||^2 f_{\mu,\Sigma}(\mathbf{x}) d\mathbf{x}.
\end{aligned}$$

Then, using that μ is the mean of Q and that $||\mu|| \leq M$ gives:

$$\begin{aligned}
\mathbb{E}_{h\sim Q}[||h||^2 \mathbb{1}(||h|| > R)] \leq &\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \int_{||\mathbf{x}||>R} ||\mathbf{x}-\mu||^2 f_{\mu,\Sigma}(\mathbf{x}) d\mathbf{x} \\
&+ 2M\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \int_{||\mathbf{x}||>R} ||\mathbf{x}-\mu|| f_{\mu,\Sigma}(\mathbf{x}) d\mathbf{x} + M^2 Q(||h|| > R),
\end{aligned}$$

Then, the first and second bounds of Lemma 21 gives:

$$\mathbb{E}_{h\sim Q}[||h||^2 \mathbb{1}(||h|| > R)] \leq \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \int_{||\mathbf{x}||>R} ||\mathbf{x}-\mu||^2 f_{\mu,\Sigma}(\mathbf{x}) d\mathbf{x} + (M^2+2M)\frac{\beta\sqrt{2\beta}}{m}$$

Finally:

$$\mathbb{E}_{h\sim Q}[||h||^2\mathbb{1}(||h|| > R)] \leq \frac{1}{\sqrt{2\pi\alpha}^d}\int_{||\mathbf{x}||>R-M} ||\mathbf{x}||^2 \exp\left(-\frac{||\mathbf{x}||^2}{2\beta}\right) d\mathbf{x} + (M^2+2M+2)\frac{\beta\sqrt{2\beta}}{m}$$

We use the hyperspherical coordinate to obtain:

$$\begin{aligned}\int_{||\mathbf{x}||>R-M} ||\mathbf{x}||^2 \exp\left(-\frac{||\mathbf{x}||^2}{2\beta}\right) d\mathbf{x} &= \int_{R-M}^{+\infty} r^{d+1}\exp\left(-\frac{r^2}{2\beta}\right) dr \\ &= \beta\sqrt{2\beta}^{d+1}\int_{\frac{(R-M)^2}{2\beta}}^{+\infty} r^{d/2}\exp^{-r} dr \\ &= \beta\sqrt{2\beta}^{d+1}\Gamma\left(\frac{d+1}{2}, \frac{(R-M)^2}{2\beta}\right).\end{aligned}$$

Then applying again Theorem 32 gives:

$$\begin{aligned}\mathbb{E}_{h\sim Q}\left[||h||^2\ \mathbb{1}(||h|| > R)\right] &\leq \frac{\beta\sqrt{2\beta}}{m} + (M^2+2M)\frac{\beta\sqrt{2\beta}}{m} \\ &= (M+1)^2\frac{\beta\sqrt{2\beta}}{m}.\end{aligned}$$

This concludes the proof. ∎

B.3 Proof of Theorem 11

Proof [Proof of Theorem 11]

We take a specific radius R which is the smallest value satisfying `Rad`.

We first notice that because for all z, $\ell(.,z)$ is L-smooth, then on $\mathcal{B}(0,R)$, the gradients of $\ell(.,z)$ 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 8 on the compact $\mathcal{B}(0,R)$, with the prior $\mathcal{P}_R\#P$, and with high probability, for any posterior $\mathcal{P}_R\#Q$ with $Q \in C_{\alpha,\beta,M}$:

$$|\Delta_S(\mathcal{P}_R\#Q)| \leq \sqrt{2D_R(2D_R+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(W_1(\mathcal{P}_R\#Q, \mathcal{P}_R\#P) + \varepsilon_m\right) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}}.$$

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

$$|\Delta_S(Q)| \leq |\Delta_S(\mathcal{P}_R\#Q)| + |\Delta_S(Q) - \Delta_S(\mathcal{P}_R\#Q)|$$

And we also have as in the proof of Theorem 9:

$$|\Delta_S(Q) - \Delta_S(\mathcal{P}_R\#Q)| \leq 2Q(||h|| > R) \leq 2\frac{\beta\sqrt{2\beta}}{m}$$

Also we have by triangle inequality:

$$W_1(\mathcal{P}_R\#Q, \mathcal{P}_R\#P) \leq W_1(Q, \mathcal{P}_R\#Q) + W_1(Q, P) + W_1(P, \mathcal{P}_R\#P).$$

Because both $Q, P \in C_{\alpha,\beta,M}$, using again Lemma 21 gives:

$$W_1(\mathcal{P}_R\#Q, \mathcal{P}_R\#P) \leq W_1(Q, P) + 2(M+1)\frac{\beta\sqrt{2\beta}}{m}.$$

We then have:

$$|\Delta_S(Q)| \leq 2\frac{\beta\sqrt{2\beta}}{m} + \sqrt{2D_R(2D_R+1)\frac{2d\log\left(3\frac{1+2Rm}{\delta}\right)}{m}\left(W_1(Q,P) + \alpha_m\right) + \frac{\log\left(\frac{3m}{\delta}\right)}{m}}.$$

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