A PAC-Bayesian Link Between Generalisation and Flat Minima

Maxime Haddouche, Paul Viallard, Umut Şimşekli, Benjamin Guedj

A PAC-Bayesian Link Between Generalisation and Flat Minima

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

Paul Viallard* PAUL.VIALLARD@INRIA.FR
Univ Rennes, Inria, CNRS IRISA - UMR 6074, F35000 Rennes, France

Umut Simsekli UMUT.SIMSEKLI@INRIA.FR
Inria, CNRS, Ecole Normale Supérieure, PSL Research University, Paris, France

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

Abstract

Modern machine learning usually involves predictors in the overparametrised setting (number of trained parameters greater than dataset size), and their training yield not only good performances on training data, but also good generalisation capacity. This phenomenon challenges many theoretical results, and remains an open problem. To reach a better understanding, we provide novel generalisation bounds involving gradient terms. To do so, we combine the PAC-Bayes toolbox with Poincaré and Log-Sobolev inequalities, avoiding an explicit dependency on dimension of the predictor space. Our results highlight the positive influence of *flat minima* (being minima with a neighbourhood nearly minimising the learning problem as well) on generalisation performances, involving directly the benefits of the optimisation phase.

Keywords: Generalisation Bounds, PAC-Bayes, Flat Minima, Poincaré, Log-Sobolev Inequalities

1. Introduction

Understanding generalisation in modern machine learning problems has been a major challenge in learning theory. The goal here is to upper-bound the so-called *generalisation error* that is gap between the population and empirical risks, $\mathrm{R}_{\mathcal{D}}(h) - \hat{\mathrm{R}}_{\mathcal{S}_m}(h)$, where $h \in \mathbb{R}^d$ is the parameters of a predictor, $\mathrm{R}_{\mathcal{D}} := \mathbb{E}_{\mathbf{z}\sim\mu}[\ell(h, \mathbf{z})]$ is the population risk, $\mathcal{D}$ is an unknown data distribution, ℓ is a loss function, $\hat{\mathrm{R}}_{\mathcal{S}_m} := \frac{1}{m}\sum_{i=1}^{m} \ell(h, \mathbf{z}_i)$, and finally $\mathcal{S}_m := \{\mathbf{z}_1, \cdots, \mathbf{z}_m\}$ is a dataset with each $\mathbf{z}_i$ is independent and identically distributed (*i.i.d.*) with $\mathcal{D}$.

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

While there have been several attempts to link some form of flatness to generalisation in a mathematically rigorous way (Neyshabur et al., 2017; Petzka et al., 2021; Yue et al., 2023; Andriushchenko et al., 2023), mainly in the framework of 'sharpness aware minimisation' (Foret et al.,

* The work was done when the author was affiliated with Inria Paris.

2020), it has been recently shown that flat minima do not always imply good generalisation. In fact, there exist scenarios such that the flattest minima achieve the worst generalisation performance compared to non-flat ones (Wen et al., 2023).

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

- We show that, when Q satisfies the Poincaré inequality and a technical condition that we identify, we can obtain a 'fast-rate' generalisation bound that diminishes with rate $1/m$ (rather than $1/\sqrt{m}$) and mainly contains two terms:
  - (i) The flatness term: $\mathbb{E}_{h\sim\mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^{m}\|\nabla_h \ell(h,\mathbf{z}_i)\|^2\right]$. This term is directly linked to the Hessian of the loss ℓ, due to the connection between the Fisher information and the Hessian of the loss Bickel and Doksum (2015). For instance, under certain conditions, it can be shown that $\mathrm{trace}(\nabla^2 \hat{\mathrm{R}}_{\mathcal{S}_m}(h)) = \frac{2}{m}\sum_{i=1}^{m}\|\nabla_h \ell(h,\mathbf{z}_i)\|^2$ (Wen et al., 2023, Lemma 4.1).
  - (ii) The classical PAC-Bayesian complexity term $KL(\mathrm{Q},\mathrm{P})$, where KL denotes the Kullback-Leibler divergence and P is data-independent 'prior' distribution.
- We then further analyse the term $KL(\mathrm{Q},\mathrm{P})$. We show that, when Q is a Gibbs distribution, *i.e.*, $\mathrm{Q}(h) \propto \exp(-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}(h))\mathrm{P}(h)$ for some $\gamma > 0$ and P satisfies a log-Sobolev inequality, the generalisation error can be controlled *solely* by the term: $\gamma^2 c_{LS}(\mathrm{P})\,\mathbb{E}_{h\sim\mathrm{Q}}[\|\nabla_h \hat{\mathrm{R}}_{\mathcal{S}_m}(h)\|^2]$, where $c_{LS}(\mathrm{P})$ denotes the log-Sobolev constant of the prior P.
- We finally go beyond the KL divergence to link flat minima to deterministic predictors (*i.e.*, when Q is a Dirac distribution) through a novel Wasserstein-based generalisation bound for gradient Lipschitz loss functions.

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

2. Preliminaries

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

$\mathbb{E}_{h\sim Q}[\mathrm{R}_{\mathcal{D}}(h)]$ and $\hat{\mathrm{R}}_{\mathcal{S}_m}(Q) := \mathbb{E}_{h\sim Q}[\hat{\mathrm{R}}_{\mathcal{S}_m}(h)]$. PAC-Bayes bounds usually aim at controlling the *expected generalisation error (or gap)* for each dataset size m, i.e., $\Delta_{\mathcal{S}_m}(Q) := \mathrm{R}_{\mathcal{D}}(Q) - \hat{\mathrm{R}}_{\mathcal{S}_m}(Q)$.

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

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

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

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

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

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

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

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

The class of Gaussian distributions is an important particular case of distributions satisfying both Poincaré and log-Sobolev inequalities, this is the subject of Proposition 2.

Proposition 2 *For a given pair* (μ, Σ) *of mean and covariance matrix in* $\mathbb{R}^d$, *define* $\mathrm{Q} = \mathcal{N}(\mu, \Sigma)$. *Then we have, for any* $f \in \mathrm{H}^1(\mathrm{Q})$*:*

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

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

In Proposition 2, the first inequality can be derived from the classical log-Sobolev inequality for $\mathcal{N}(\mathbf{0}, \mathrm{Id})$ stated in Gross (1975), with a change of variable. Similarly, the Poincaré inequality can be obtained through a change of variable from the Poincaré inequality for $\mathcal{N}(\mathbf{0}, \mathrm{Id})$ which is a particular case of the Brascamp-Lieb inequality for log-concave probability measures (Brascamp and Lieb, 1976) and is stated explicitly in Beckner (1989, Theorem 1).

We now focus on specific posterior distributions called *Gibbs posteriors, or Gibbs distributions*. For a fixed loss ℓ and dataset $\mathcal{S}_m$, the Gibbs posterior, *w.r.t.* prior $\mathrm{P} \in \mathcal{M}(\mathcal{H})$, risk $\hat{\mathrm{R}}_{\mathcal{S}_m}$ and *inverse temperature* $\gamma > 0$.is defined as $\mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}$ such that $d\mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}(h) \propto \exp(-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}(h)) dP(h)$. Gibbs posteriors are a class of closed-form solutions for relaxation of Catoni (2007, Theorem 1.2.6) stated, for instance, in Alquier et al. (2016, Theorem 4.1). Proposition 3 shows that when the prior and the loss satisfies a few properties, then the associated Gibbs posterior is $\texttt{L-Sob}(c_{LS})$.

Proposition 3 *Assume that* P *is a probability measure on* $\mathbb{R}^d$ *such that* $dP(h) \propto \exp(-V(x))$ *with* V *a smooth function such that* $Hess(V) \succeq \frac{2}{c_{LS}(\mathrm{P})}\mathrm{Id}$. *Assume that* $\ell = \ell_1 + \ell_2$ *with* ℓ_1 *convex, twice differentiable and* ℓ_2 *bounded. Then for any* $\gamma > 0$, *the Gibbs posterior* $\mathrm{Q} = \mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}$ *is* $\texttt{L-Sob}(c_{LS})$ *with constant* $c_{LS}(\mathrm{Q}) = c_{LS}(\mathrm{P})\exp\left(4\|\ell_2\|_\infty\right)$.

Proposition 3 applies, *e.g.*, when P is a Gaussian prior $\mathrm{P} = \mathcal{N}(\mu_\mathrm{P}, \Sigma_\mathrm{P})$. Notice that in this case $c_{LS}(\mathrm{P}) = 2\|\Sigma_\mathrm{P}\|_{op}$. This property is a straightforward application of Chafaï (2004, Corollary 2.1) with Guionnet and Zegarlinksi (2003, Property 2.6) and is stated in Appendix A for completeness. Finally, notice that satisfying a log-Sobolev inequality is stronger than satisfying a Poincaré one. This is stated for instance in Ledoux (2006, Proposition 2.1) and properly recalled in Appendix A.

3. Reaching a flat minimum allows Poincaré posteriors to generalise well

3.1. Fast rate PAC-Bayes bounds for heavy-tailed losses

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

In this section, we consider posterior distributions Q being $\texttt{Poinc}(c_P)$. This assumption covers the important case of Gaussian measures (Proposition 2) as well as all measures satisfying a log-Sobolev inequality (Proposition 14). We focus on PAC-Bayes bound holding for distributions Q satisfying a particular assumption involving the data distribution $\mathcal{D}$ (contrary to many PAC-Bayes bounds holding for all Q). We then define the *error* of $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ for any datum $\mathbf{z} \in \mathcal{Z}$ as $\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z}) := \mathbb{E}_{h\sim\mathrm{Q}}[\ell(h, \mathbf{z})]$ and identify Assumption 4 to later involve flat minima.

Assumption 4 *We say that* $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ *is* quadratically self-bounded *with respect to* ℓ *and constant* $C > 0$ *(namely* $\texttt{QSB}(\ell, C)$*) if*

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

Assumption 4 is a relaxation of boundedness, as if $\ell \in [0, C]$ then it is $\texttt{QSB}(\ell, C)$. It is an alternative to the bounded expected variance assumption in anytime-valid PAC-Bayes bounds (Haddouche and Guedj, 2023a; Chugg et al., 2023). An issue with such boundedness assumption is that it has to hold for all posteriors, including those providing poor generalisation performances. This is avoided by the $\texttt{QSB}$ assumption which intricate the properties of $\mathcal{D}, \ell$ and Q. Such a design is in line with the conclusions of the recent work of Gastpar et al. (2023), inviting to derive generalisation bounds valid for specific pairs $(\mathrm{Q}, \mathcal{D})$ (and not uniformly valid for all such pairs). Finally, we interpret C as a contraction constant attenuating, on average, the local expansion (governed by variances of Q, and $\mathcal{D}$) of the loss around the mean of Q. Exploiting the PAC-Bayes supermartingales bounds of Haddouche and Guedj (2023a); Chugg et al. (2023) alongside Poincaré inequality leads to the following.

Theorem 5 *For any $C > 0$, any $\frac{2}{C} > \lambda > 0$, any data-free prior P, any $\ell \geq 0$ and any $\delta \in [0, 1]$, we have, with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, any Q being $\texttt{Poinc}(c_P)$, $\texttt{QSB}(\ell, C)$ and $\ell(., \mathbf{z}) \in \mathrm{H}^1(\mathrm{Q})$ for all $\mathbf{z}$,*

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

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

On the role of flat minima in PAC-Bayes learning. Theorem 5 suggests that, in order to attain good generalisation ability, the mean of Q has to be close from two minima: *(i)* on $\hat{\mathrm{R}}_{\mathcal{S}_m}$ in order to make $\hat{\mathrm{R}}_{\mathcal{S}_m}$ small, and *(ii)* on $\mathbb{E}_{\mathbf{z} \sim \mathcal{D}}[\|\nabla_h \ell(h, \mathbf{z})\|^2]$ to make the gradients small. The variance of Q has to fit the flatness of those minima, the flatter they are, the larger the variance in order to shrink the expected terms on the right-hand-side of Theorem 5. Finally, the KL term invites, *e.g.* for Gaussian distributions, to consider high variances, hence flat minima to maintain a small value of the bound.

A focus on C. Taking $\lambda = 1/C$ in Theorem 5 attenuates the impact of the prior distribution and amplifies the gradient term. Then, a small C is desirable when working with flat minima to attenuate an ill-designed prior. Having a small C is reachable in practice: we show in Section 6, for a classification task on MNIST, that the $\texttt{QSB}$ assumption is verified with C strictly smaller than 1 when considering neural networks.

High probability bounds with fast rates, a paradox? Grunwald et al. (2021, page 7) showed that, for a trivial $\mathcal{H} = \{h\} \subset \mathbb{R}^d$, for any loss, any *i.i.d.* dataset $\mathcal{S}_m$ with variance σ^2, we have asymptotically, with probability at least α, for a constant C_α depending on α and $\mathcal{N}(\mathbf{0}, \mathrm{Id})$, we have $\mathrm{R}_{\mathcal{D}}(h) \geq \hat{\mathrm{R}}_{\mathcal{S}_m}(h) + C_\alpha \frac{\sigma^2}{\sqrt{m}}$. Is it paradoxical with Theorem 5? The answer is no: the bound in Grunwald et al. (2021) gives an asymptotic lower bound on the convergence of $\hat{\mathrm{R}}_{\mathcal{S}_m}(h)$ to $\mathrm{R}_{\mathcal{D}}(h)$. Theorem 5 informs us on how $\mathrm{R}_{\mathcal{D}}$ is getting closer from $\frac{1}{1-\lambda/2}\hat{\mathrm{R}}_{\mathcal{S}_m}$ which converges to $\frac{1}{1-\lambda/2}\mathrm{R}_{\mathcal{D}} > \mathrm{R}_{\mathcal{D}}$ as the loss is non-negative. Theorem 5 then show the existence of a 'transition regime' involving a fast rate. Once $\frac{1}{1-\lambda/2}\hat{\mathrm{R}}_{\mathcal{S}_m}$ is reached, the clower bound of Grunwald et al. (2021) ensures an asymptotic regime with slow convergence rate. Note that such transition regimes already appeared in the literature in Tolstikhin and Seldin (2013); Mhammedi et al. (2019) at the cost of additional variance terms compared to Theorem 5. However, such fast rates have never been linked before to flat minima (and optimisation in general), highlighting the potential of our bound to explain the ability of deep neural networks to generalise well in the overparametrised setting (m far smaller than the dimension of $\mathcal{H}$), where flat minima are likely to be reached, as studied, *e.g.*, in Dziugaite et al. (2020), showing correlations between flat minima and generalisation for various learning problems.

Proof [Proof of Theorem 5] We start from Chugg et al. (2023, Corollary 17) instantiated with a single λ, *i.i.d.* data and a prior P. With probability at least $1-\delta$, for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ and $m > 0$:

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

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

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

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

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

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

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

■

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

Corollary 6 *For any $C > 0$, any $\delta \in (0,1)$ any $\frac{2}{C} > \lambda > 0$, any data-free prior P, any $\ell \geq 0$ such that, for any $\mathbf{z} \in \mathcal{Z}$, we have $\ell(.,\mathbf{z}) \in \mathrm{H}^1$ and for any h, the loss function $\ell(h,.)$ is $\mathcal{C}^1$ almost everywhere on $\mathcal{Z}$. If the data distribution $\mathcal{D}$ is $\texttt{Poinc}(c_P)$, then with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, any posterior Q being $\texttt{Poinc}(c_P)$ with $\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq C$:*

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

Proof is deferred to Section C.1. Corollary 6 states that, if Q reached a flat minimum (meaning $\|\nabla_h \ell\|$ is small), and this minimum is robust to the training dataset (meaning $\|\nabla_{\mathbf{z}} \ell\|$ is small), then a fast rate is attainable while only requiring an upper bound on $\mathrm{R}_{\mathcal{D}}(\mathrm{Q})$. This conclusion holds

when $\mathcal{D}$ $\texttt{Poinc}$, encompassing the case of Gaussian mixtures (Schlichting, 2019), which can approximate any smooth density (as recalled in Gat et al., 2022). However, the Poincaré constant of a general mixture is not known, and the upper bound of Schlichting (2019) scales with the number of components, involving potentially high χ^2 divergences.

Comparison with Gat et al. (2022). We compare Corollary 6 with Gat et al. (2022, Theorems 3.5, 3.6). First, our result holds with the assumption that $\mathcal{D}$ follows a Poincaré inequality, which is strictly less restrictive than assuming a log-Sobolev inequality (Proposition 14). Second, they assume a bounded loss and their result holds only for classification problem satisfying a technical assumption on the label repartition (see their Lemma 3.3) while ours holds for any learning problem at the sole assumption of a bounded $\mathrm{R}_{\mathcal{D}}(\mathrm{Q})$, allowing ℓ to be non-negative. Moreover, note that to conclude their proof, Gat et al. (2022) had to use a uniform bound on $\mathbb{E}_{\mathbf{z}}[\|\nabla_{\mathbf{z}}\ell\|]$ in their Theorem 3.5 to have a tractable bound, thus the benefits of gradient norm is unclear. While they overcome this limitation in Gat et al. (2022, Theorem 3.6), the explicit influence of the gradient norm appears within an exponential moment on the losses (attenuated by a logarithm). However, a major limitation is that this exponential moment is averaged *w.r.t.* P, being data-free. Thus, the associated gradients have no apparent reason to be small, and their result cannot be linked to flat minima, contrary to Corollary 6 involving expected gradients *w.r.t.* Q, being the output of an optimisation process.

3.2. Towards fully empirical bound for gradient-Lipschitz functions.

In this section, we assume the loss ℓ is such that, for any $\mathbf{z} \in \mathcal{Z}$, the gradient $\nabla_h \ell(.,\mathbf{z})$ is G-Lipschitz, which is often considered for convergence bounds in optimisation. A large part of high-probability PAC-Bayes bounds are fully empirical: this has numerous advantages including in-training numerical evaluation of generalisation as well as novel PAC-Bayesian algorithms, minimising such empirical bounds; see (Dziugaite and Roy, 2017; Perez-Ortiz et al., 2021; Viallard et al., 2023b) among others. However, Theorem 5 and Corollary 6 are not fully empirical and thus, do not have such desirable properties. We circumvent this issue in Theorem 7.

Theorem 7 *For any $C_1, C_2, c > 0$, any data-free prior P, any $\ell \geq 0$ being $\mathcal{C}^2$ and any $\delta \in [0,1]$, we have, with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, any Q being $\texttt{Poinc}(c_{\mathrm{P}})$ with constant c, $\texttt{QSB}(\ell, C_1)$, $\texttt{QSB}\left(\|\nabla_h \ell\|^2, C_2\right)$ and $\ell(.,\mathbf{z}), \|\nabla_h \ell\|^2(.,\mathbf{z}) \in \mathrm{H}^1(\mathrm{Q})$ for all $\mathbf{z}$,*

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

Proof is deferred to Section C.2. Here, we showed that to attain fast rates, the $\texttt{QSB}$ assumption has to be reached for both the loss and its gradient. This suggests several things on the flat minimum that has to be reached by Q (designed from $\hat{\mathrm{R}}_{\mathcal{S}}$): first, it needs to be close from a flat minimum of $\mathrm{R}_{\mathcal{D}}$ to satisfy the $\texttt{QSB}$ assumption. Second, this minimum also ensures the contraction of the gradients. We then are able to derive an empirical generalisation bound, involving both empirical loss and gradients. Not only Theorem 7 yields, to our knowledge, the first PAC-Bayesian algorithm involving gradient terms, but also can be translated to a generalisation metric in order to understand

generalisation. Such an idea has been exploited recently (Neyshabur et al., 2017; Jiang et al., 2020; Dziugaite et al., 2020). In particular, from $\hat{\mathrm{R}}_{\mathcal{S}}(\mathrm{Q})$, Neyshabur et al. (2017) derived a notion of *sharpness*, stated in Equation (1), aiming to be informative on the flatness of the reached minima for any $\mathrm{Q} = \mathcal{N}(\mu_Q, \sigma^2\mathrm{Id})$. This notion is defined by

$$\mathop{\mathbb{E}}_{\nu\sim\mathcal{N}(\mathbf{0},\sigma^2\mathrm{Id})}\left[\hat{\mathrm{R}}_{\mathcal{S}_m}(\mu_\mathrm{Q}+\nu) - \hat{\mathrm{R}}_{\mathcal{S}_m}(\mu_\mathrm{Q})\right]. \tag{1}$$

Theorem 7 enhance this notion of sharpness by involving the empirical gradients when Q is $\mathtt{QSB}(\ell, C_1)$:

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

where $\text{G-}\hat{\mathrm{R}}_{\mathcal{S}_m}(h) = \frac{1}{m}\sum_{i=1}^m \|\nabla_h \ell(h, \mathbf{z}_i)\|^2$. This gradient term can be seen as an empirical Fisher information, linked to the second-order moment derivative. Thus, (2) involves a notion of flatness on both the loss and its gradient, contrary to (1). For the sake of clarity, we particularise Theorem 7 in Corollary 8 with Gaussian distributions and this novel notion of sharpness.

Corollary 8 *For any $C_1, C_2 > 0$, any fixed variance $\sigma^2 > 0$, any data-free prior $\mathrm{P} = \mathcal{N}(\mu_\mathrm{P}, \sigma^2\mathrm{Id})$, any nonnegative loss ℓ being $\mathcal{C}^2$ and any $\delta \in [0,1]$, we have, with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, any $\mathrm{Q} = \mathcal{N}(\mu_\mathrm{Q}, \sigma^2\mathrm{Id})$ being $\mathtt{QSB}(\ell, C_1)$, $\mathtt{QSB}\left(\|\nabla_h \ell\|^2, C_2\right)$ and $\ell(.,\mathbf{z}), \|\nabla_h\ell\|^2(.,\mathbf{z}) \in \mathrm{H}^1(\mathrm{Q})$ for all $\mathbf{z}$,*

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

4. Generalisation ability of Gibbs distributions with a log-Sobolev prior

One limitation of the results given in Section 3 is that the KL divergence term remains uncontrolled in general as its formulation depends on the nature of P and Q. A close form exists for Gaussian distributions for instance, but this class of distribution is limiting. Perpetrating the spirit of Catoni (2007), we go beyond the Gaussian distributions to focus on the Gibbs posteriors which have naturally appeared in PAC-Bayes through the use of tools from statistical physics. We show that log-Sobolev inequalities allow us to control the KL divergence of such distributions *w.r.t.* their priors.

Controlling the KL divergence when Q is a Gibbs posterior. Lemma 9 exploits the fact that the KL divergence can be formulated as an entropy *w.r.t.* the prior distribution P. It then shows that the KL divergence of the Gibbs posterior $\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}$ *w.r.t.* P is upper bounded by gradient terms as long as P satisfies a log-Sobolev inequality.

Lemma 9 *For any m, P being $\mathtt{L\text{-}Sob}(c_{LS})$, any $\ell \geq 0$ such that for any $\mathbf{z}$, $\ell(.,\mathbf{z}) \in \mathrm{H}^1(\mathrm{P})$, we have, for any $\gamma > 0$:*

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

Proof is deferred to Section C.3. The crucial message of this lemma is that, a flat minimum of $\hat{\mathrm{R}}_{\mathcal{S}}$ allows controlling the KL divergence. This message is new and independent of Section 3 which

focus on flat minima reached for $\mathrm{R}_{\mathcal{D}}$. Note that in this case, the KL divergence has an explicit formulation. However it involves to calculate the exponential moment $\mathbb{E}_{h\sim \mathrm{P}}[\exp(-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m})]$ which is costly in practice. On the contrary, we only need to estimate a second-order moment over $\mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}$.

Generalisation ability of Gibbs posteriors. When Gibbs posteriors are involved, KL divergence is controllable by a gradient term. An ideal way to conclude would be, as in Section 3 to involve Poincaré inequality. However, Gibbs posterior are not necessarily satisfying a Poincaré inequality as in Section 3, we then need to make supplementary assumptions on the loss.

Theorem 10 *For any $C > 0$, any $\gamma > 0$, any prior P being $\texttt{L-Sob}(c_{LS})$, any $\ell \geq 0$ and any $\delta \in [0,1]$, we have the following inequalities. If $\ell \in [0,1]$, then with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, and any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:*

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

If $\ell = \ell_1 + \ell_2$ with ℓ_1 convex, twice differentiable and ℓ_2 bounded, assume that P satisfies the conditions of Proposition 3. Then for any $\frac{2}{C} > \lambda > 0$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, such that Q is $\texttt{QSB}(\ell, C)$ and $\ell(.,\mathbf{z}) \in \mathrm{H}^1(\mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}})$:

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

Proof is deferred to Section C.4. Note that we could have derived analogous to Corollary 6 at the cost of a supplementary Poincaré assumption on $\mathcal{D}$. The influence of the inverse temperature γ is quadratic: this is the price to pay to fit the dataset and reduce the influence of the prior. This dependency is therefore attenuated by a gradient term, small if a flat minimum on the empirical risk has been reached. This suggests that in the case of Gibbs posteriors with log-Sobolev prior, reaching a flat minima on $\hat{\mathrm{R}}_{\mathcal{S}_m}$ controls not only $\hat{\mathrm{R}}_{\mathcal{S}_m}(\mathrm{Q})$, but also the KL divergence and this last point is not reachable when considering Poincaré distributions. The other gradient term comes from Section 3 and requires to be close from a flat minimum on $\mathrm{R}_{\mathcal{D}}$ to attain fast rates.

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

In Sections 3 and 4, we provided various generalisation bounds, benefiting from flat minima. However, our results involve a KL divergence, implying absolute continuity of Q *w.r.t.* P, incompatible with the case of deterministic predictors (Dirac distributions). To circumvent this issue, a recent line of work emerged, involving integral probability metrics, with a particular focus on the 1-Wasserstein distance Amit et al. (2022); Haddouche and Guedj (2023b); Viallard et al. (2023b). The idea behind these works is to replace the change of measure inequality (Csiszar, 1975; Donsker and Varadhan, 1976) by the Kantorovich-Rubinstein duality (Villani, 2009) to trade a KL for a Wasserstein. We go even further here by obtaining the first PAC-Bayesian bound involving directly a 2-Wasserstein distance (see definition 15), trading Lipschitz assumption for gradient-Lipschitz one (well-suited for optimisation). To do so, we first derive a novel change of measure inequality.

Theorem 11 *Assume $\mathcal{H}$ to have a finite diameter $D > 0$. Then for any function $f : \mathcal{H} \to \mathbb{R}$ with G-Lipschitz gradients, the following holds: for all distributions $\mathrm{P}, \mathrm{Q} \in \mathcal{M}(\mathcal{H})^2$,*

$$\mathbb{E}_{h\sim\mathrm{Q}}[f(h)] \leq \frac{G}{2}W_2^2(Q,P) + \mathbb{E}_{h\sim\mathrm{P}}[f(h)] + D\mathbb{E}_{h\sim\mathrm{Q}}[\|\nabla f(h)\|].$$

Proof is deferred to Section C.5 Theorem 11 shows it is possible when gradients are Lipschitz, to obtain a duality formula involving the gradient of the considered function at the price of a linear dependency on the diameter of $\mathcal{H}$. Theorem 11 is also linked to the change of measure inequality when the prior distribution satisfies a log-Sobolev inequality.

Corollary 12 *Assume that P is such that $d\mathrm{P} \propto \exp(-V)dx$ with V being $\mathcal{C}^2$ and P is $\texttt{L-Sob}(c_{LS})$. Then, for any $R > 0$, any f with gradients G-Lipschitz on $\mathcal{B}(\mathbf{0}, R)$, and any distributions P, Q,*

$$\mathbb{E}_{h\sim\mathrm{Q}}\left[f\left(\mathcal{P}_R(h)\right)\right] \leq \frac{Gc_{LS}(\mathrm{P})}{4}\,\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \mathbb{E}_{h\sim\mathrm{P}}\left[f\left(\mathcal{P}_R(h)\right)\right] + 2R\mathbb{E}_{h\sim\mathrm{Q}}\left[\|\nabla f\left(\mathcal{P}_R(h)\right)\|\right],$$

where $\mathcal{P}_R$ denotes the Euclidean projection on $\mathcal{B}(\mathbf{0}, R)$.

Proof is deferred to Section C.6. Corollary 12 involves a KL divergence and an Euclidean predictor space $\mathcal{H} = \mathbb{R}^d$. This comes at the cost of approximating Q, P by $\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}$. Thus, R is now an hyperparameter which arbitrates a tradeoff between the quality of our approximations and the looseness of the bound (if the gradient norm is large). A notable strength is that the smoothness assumption is relaxed on smoothness over $\mathcal{B}(\mathbf{0}, R)$.

From Theorem 11, we now derive a novel generalisation bound allowing deterministic predictors.

Theorem 13 *Let $\delta \in (0, 1)$ and $P \in \mathcal{M}(\mathcal{H})$ a data-free prior. Assume $\mathcal{H}$ has a finite diameter $D > 0$, $\ell \geq 0$ and that for any m, the generalisation gap $\Delta_{\mathcal{S}_m}$ is G gradient-Lipschitz. Assume that $\mathbb{E}_{h\sim\mathrm{P}}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,z)^2] \leq \sigma^2$, then the following holds with probability at least $1-\delta$, for any $m > 0$ and any Q:*

$$\mathrm{R}_D(\mathrm{Q}) \leq \hat{\mathrm{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{G}{2}W_2^2(\mathrm{Q},\mathrm{P}) + \sqrt{\frac{2\sigma^2\log\left(\frac{1}{\delta}\right)}{m}} + D\mathbb{E}_{h\sim\mathrm{Q}}\left(\left\|\nabla_h\mathrm{R}_{\mathcal{D}}(h) - \nabla_h\hat{\mathrm{R}}_{\mathcal{S}_m}(h)\right\|\right)$$

Proof is deferred to Section C.7. Theorem 13 is not the first generalisation bound to involve a 2-Wasserstein distance (Lugosi and Neu, 2022, 2023). However, those results involve infinitely smooth loss functions. Also, results from Amit et al. (2022); Haddouche and Guedj (2023b); Viallard et al. (2023b) using 1-Wasserstein can be directly relaxed on bounds involving the 2-Wasserstein, while still requiring a Lipschitz loss. On the contrary, our result holds for any nonnegative gradient-Lipschitz $\Delta_{\mathcal{S}_m}$, which is well-suited for optimisation. Theorem 13 involves a slow rate of $1/\sqrt{m}$ as we have to control the generalisation gap *w.r.t.* to P. It is possible to make appear the gradients expected over P using the $\texttt{QSB}$ assumption, but we have no reason to expect those gradients to be small, we then controlled this term uniformly by σ^2. Another restriction of our result compared to previous ones is that it holds for $\mathcal{H}$ having a finite diameter, however, having a small expected $\|\nabla_h\mathrm{R}_{\mathcal{D}} - \nabla_h\hat{\mathrm{R}}_{\mathcal{S}_m}\|$ over Q (which is the case when flat minima on both empirical and true risks are reached) allows taking D large, and thus, having good approximations of measures on a Euclidean space through orthogonal projections as in Corollary 12.

6. An empirical study of Assumption 4 for neural networks

In this section, we check empirically whether the QSB assumption is verified for neural networks. This allows us to verify if Theorem 5 is useful to understand the generalisation ability of neural nets.

Experimental protocol. We consider classification tasks on two datasets: MNIST (LeCun, 1998) and FashionMNIST (Xiao et al., 2017). We kept the original training set $\mathcal{S}_m$ and the original test set denoted by $\mathcal{T}_n$ (of size n). We consider the convolutional neural network of Springenberg et al. (2015) adapted for MNIST and FashionMNIST. The model is composed of 4 layers containing 10 channels with a 5×5-kernel; we set the stride and the padding to 1, except for the second layer, where it is fixed to 2. Each of these (convolutional) layers is followed by a Leaky ReLU activation function. Moreover, an average pooling with a 8×8-kernel is performed before the Softmax activation function. To initialise the weights of the network, we use Glorot and Bengio (2010) uniform initializer, while the biases are initialised in $[-1/\sqrt{250}, +1/1/\sqrt{250}]$ uniformly (except the first layer, the interval is $[-1/5, +1/5]$). Hence, in this case, $\mathcal{H}$ is the set of neural networks with a fixed architecture, and parametrised with a vector $\mathbf{w}$. while the posterior distribution Q is a Gaussian measure $\mathcal{N}(\mathbf{w}, \sigma^2\mathrm{Id})$ centered on the parameters $\mathbf{w}$ associated with the model; σ is set to 10^{-4}. Note that this distribution respects the $\texttt{Poinc}(c_P)$ assumption; see Section 3.1. We train the neural network with the (vanilla) stochastic gradient descent algorithm, where the batch size is equal to 512, and the learning rate is fixed to 10^{-2}. We train for at least 10^4 gradient steps and finish the current epoch when this number of iterations is reached. Our loss ℓ is the bounded cross-entropy loss of Dziugaite and Roy (2017, Section D).
In Figure 1, we report the evolution of three quantities: *(i)* the estimated value of C, *(ii)* the test risk $\hat{\mathrm{R}}_{\mathcal{T}_n}(\mathrm{Q})$ and *(iii)* the test risk with the 01-loss. More precisely, for computational reasons, the risks and C are estimated by sampling one hypothesis $h \sim \mathrm{Q}$ and by computing the values on a mini-batch of $\mathcal{T}_n$ (with 512 examples) at each iteration. Then, Figure 1 represents averaged values on 5 runs, each point of the curve representing the average on 100 iterations of the training process (for 10^4 iterations we only plot 10^2 averaged points for clarity).

Empirical findings. Figure 1 illustrates that, when neural networks are involved for two classification tasks, Q evolves during the optimisation process while maintaining the QSB property with constant $C < 1$. For both MNIST and FashionMNIST, the constant C decreases from approximately 0.55 to 0.45. We deduce two things from this: *(i)* the learning phase, while optimising $\hat{\mathrm{R}}_{\mathcal{S}_m}$ also gain in generalisation ability, shrinking the averaged loss on new data which is translated by a smaller C; and (ii), having a data-free P (0 iteration) being QSB with $C < 1$ suggests that the architecture of our neural network also has an influence on the QSB assumption. As precised in Section 3, having $C < 1$ attenuates the impact of the KL term, thus P. This is desirable as it allows the optimiser to deeply explore the predictor space when P yields poor performances. We also note that the generalisation ability of Q on the training loss nearly matches the performance on the 0-1 loss for MNIST but is deteriorated for FashionMNIST, this invites to study more deeply the design of such surrogates in future work.

Finally, the take-home message of this study is that the QSB assumption is verified for neural networks on MNIST. Such an empirical confirmation is crucial as it is required for our main result (Theorem 5) and thus confirms that, for neural networks, reaching flat minima during the optimisation phase translates in increased generalisation ability.

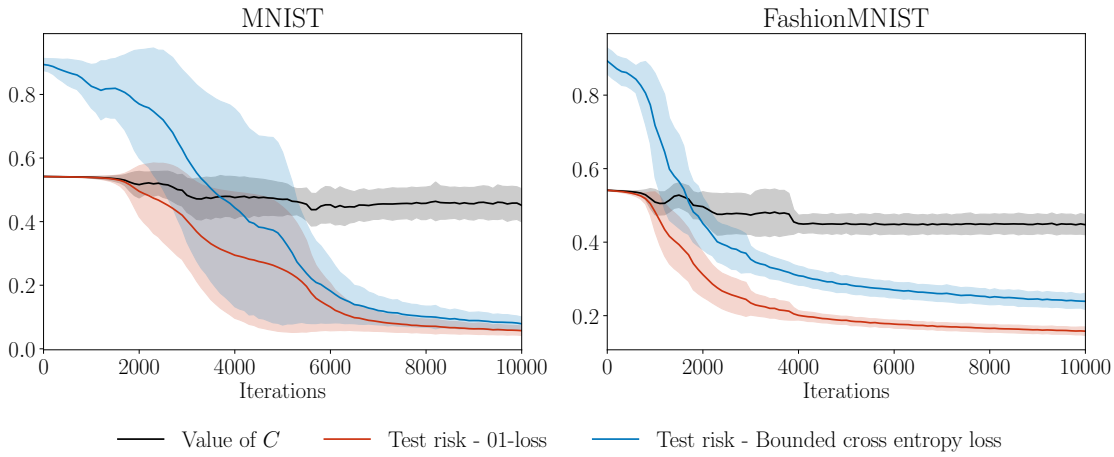

Figure 1: Evolution of the test risks (with the 01-loss and the bounded cross-entropy loss) and the value of C during the training phase.

7. Conclusion

We provide novel PAC-Bayes generalisation bounds, converging faster than $1/\sqrt{m}$ when a low empirical error is reached and that expected gradients are vanishing. This conveys the message that flat minima helps generalisation. However, to complete this analysis, the crucial question is to understand how optimisation algorithms successfully reach flat minima in the overparametrised setting. This important question is left as future work.

Acknowledgments

Paul Viallard and Umut Şimşekli are partially supported by the French program "Investissements d'avenir" ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). Umut Şimşekli is also supported by the European Research Council Starting Grant DYNASTY – 101039676. Benjamin Guedj acknowledges partial support from the French Research Agency through the programme "France 2030" and PEPR IA on grant SHARP ANR-23-PEIA-0008.

References

Pierre Alquier. User-friendly Introduction to PAC-Bayes Bounds. *Foundations and Trends® in Machine Learning*, 2024.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 2016.

Ron Amit, Baruch Epstein, Shay Moran, and Ron Meir. Integral Probability Metrics PAC-Bayes Bounds. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. *arXiv preprint arXiv:2302.07011*, 2023.

Cécile Ané, Sébastien Blachère, Djalil Chafaï, Pierre Fougères, Ivan Gentil, Florent Malrieu, Cyril Roberto, and Grégory Scheffer. *Sur les inégalités de Sobolev logarithmiques*, volume 10. Société mathématique de France Paris, 2000.

William Beckner. A Generalized Poincaré Inequality for Gaussian Measures. *Proceedings of the American Mathematical Society*, 1989. URL http://www.jstor.org/stable/2046956.

Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. CRC Press, 2015.

Herm Jan Brascamp and Elliott H Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of functional analysis*, 22(4):366–389, 1976.

Olivier Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics, 2007.

Djalil Chafaï. Entropies, convexity, and functional inequalities, On Φ-entropies and Φ-Sobolev inequalities. *Journal of Mathematics of Kyoto University*, 44(2):325–363, 2004.

Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A Unified Recipe for Deriving (Time-Uniform) PAC-Bayes Bounds. *Journal of Machine Learning Research*, 2023. URL http://jmlr.org/papers/v24/23-0401.html.

I. Csiszar. I-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, 3, 1975. doi: 10.1214/aop/1176996454. URL https://doi.org/10.1214/aop/1176996454.

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time—III. *Communications on Pure and Applied Mathematics*, 29(4):389–461, 1976. doi: https://doi.org/10.1002/cpa.3160290405. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160290405.

Gintare Karolina Dziugaite and Daniel Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/86d7c8a08b4aaa1bc7c599473f5dddda-Abstract.html.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Michael Gastpar, Ido Nachum, Jonathan Shafer, and Thomas Weinberger. Fantastic Generalization Measures are Nowhere to be Found, 2023.

Itai Gat, Yossi Adi, Alexander G. Schwing, and Tamir Hazan. On the Importance of Gradient Norm in PAC-Bayesian Bounds. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6686e3f2e31a0db5bf90ab1cc2272b72-Abstract-Conference.html.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Leonard Gross. Logarithmic Sobolev Inequalities. *American Journal of Mathematics*, 1975. ISSN 00029327, 10806377. URL http://www.jstor.org/stable/2373688.

Peter Grunwald, Thomas Steinke, and Lydia Zakynthinou. PAC-Bayes, MAC-Bayes and Conditional Mutual Information: Fast rate bounds that handle general VC classes. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2217–2247. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/grunwald21a.html.

Benjamin Guedj. A primer on PAC-Bayesian learning. In *Proceedings of the second congress of the French Mathematical Society*, volume 33, 2019. URL https://arxiv.org/abs/1901.05353.

A. Guionnet and B. Zegarlinksi. *Lectures on Logarithmic Sobolev Inequalities*. Springer Berlin Heidelberg, 2003. doi: 10.1007/978-3-540-36107-7_1. URL https://doi.org/10.1007/978-3-540-36107-7_1.

Maxime Haddouche and Benjamin Guedj. Online PAC-Bayes Learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

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

Maxime Haddouche and Benjamin Guedj. Wasserstein PAC-Bayes Learning: A Bridge Between Generalisation and Optimisation. *CoRR*, abs/2304.07048, 2023b.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, and Maxim Raginsky. Generalization bounds: Perspectives from information theory and PAC-Bayes. *arXiv preprint arXiv:2309.04381*, 2023.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic Generalization Measures and Where to Find Them. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Yann LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Michel Ledoux. Concentration of measure and logarithmic Sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer, 2006.

Gábor Lugosi and Gergely Neu. Generalization Bounds via Convex Analysis. In *Conference on Learning Theory (COLT)*, 2022.

Gábor Lugosi and Gergely Neu. Online-to-PAC Conversions: Generalization Bounds via Regret Analysis, 2023.

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.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring Generalization in Deep Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/10ce03a1ed01077e3e289f3e53c72813-Abstract.html.

F. Otto and C. Villani. Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 173, 2000. URL https://www.sciencedirect.com/science/article/pii/S0022123699935577.

Maria Perez-Ortiz, Omar Rivasplata, Emilio Parrado-Hernandez, Benjamin Guedj, and John Shawe-Taylor. Progress in Self-Certified Neural Networks. In *NeurIPS 2021 Workshop on Bayesian Deep Learning*, 2021.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. *Advances in neural information processing systems*, 34:18420–18432, 2021.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvari, and John Shawe-Taylor. PAC-Bayes Analysis Beyond the Usual Bounds. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, pages 16833–16845. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c3992e9a68c5ae12bd18488bc579b30d-Paper.pdf.

André Schlichting. Poincaré and Log-Sobolev Inequalities for Mixtures. *Entropy*, 2019. doi: 10.3390/e21010089. URL https://doi.org/10.3390/e21010089.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (ICLR) – Workshop Track*, 2015.

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)*. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/a97da629b098b75c294dffdc3e463904-Paper.pdf.

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A general framework for the practical disintegration of PAC-Bayesian bounds. *Machine Learning*, 2023a. ISSN 1573-0565. doi: 10.1007/s10994-023-06391-0. URL https://doi.org/10.1007/s10994-023-06391-0.

Paul Viallard, Maxime Haddouche, Umut Simsekli, and Benjamin Guedj. Learning via Wasserstein-Based High Probability Generalisation Bounds. *To be published in NeurIPS 2023*, 2023b.

Cédric Villani. *Optimal transport: old and new*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, 2009.

Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Dkmpa6wCIx.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms, 2017.

Yun Yue, Jiadi Jiang, Zhiling Ye, Ning Gao, Yongchao Liu, and Ke Zhang. Sharpness-aware minimization revisited: Weighted sharpness as a regularization term. *arXiv preprint arXiv:2305.15817*, 2023.

Appendix A. Supplementary background

A.1. Additional details on Poincaré and Log-Sobolev inequalities.

Proof of proposition 3. **Proof** We define P_1 such that $d\mathrm{P}_1(h) \propto \exp(-V(h) - \gamma\hat{\mathrm{R}}_{\mathcal{S}_m}(h))dh$, then note that, by convexity assumption over ℓ_1, $Hess(V+\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}) \succeq \frac{1}{c_{LS}}\mathrm{Id}$. Then, applying Chafaï (2004, Corollary 2.1), we know that P_1 satisfies a Poincaré inequality with constant $c_{LS}(\mathrm{P})$.

Finally, defining P_2 such that $dP_2(h) \propto \exp(-\frac{\gamma}{m}\sum_{i=1}^m \ell_2(h, \mathbf{z}_i))$, thanks to the boundedness of ℓ_2, we use Guionnet and Zegarlinksi (2003, Property 2.6), which ensure that $P_2 = \mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}} d\mathrm{P}_1(h)$ satisfies a Log-Sobolev inequality with constant $2c_{LS}(\mathrm{P})\exp(4\|\ell_2\|_\infty)c_P(\mathrm{P})$. Noting that $\mathrm{P}_2 = \mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}}}$ concludes the proof. ■

Proof of Ledoux (2006, Propsition 2.1) We prove here Proposition 14, stated below, showing that Log-Sobolev implies Poincaré.

Proposition 14 *If* Q *is* $\texttt{L-Sob}(c_{LS})$*, then it is also* $\texttt{Poinc}(c_P)$*.*
We then have $c_P(\mathrm{Q}) = \frac{c_{LS}(\mathrm{Q})}{2}$.

We provide the proof for completeness.
Proof Let $f \in \mathrm{H}^1(\mathrm{Q})$, such that $\mathbb{E}_Q[f] = 0$ and $\mathbb{E}[f^2] = 1$. For any $\varepsilon > 0$, $1 + \varepsilon f \in \mathrm{H}^1$. We then apply the Log-Sobolev inequality on $1 + \varepsilon f$:

$$\mathbb{E}_Q\left[(1+\varepsilon f)^2\left(2\log(1+\varepsilon f) - \log(1+\varepsilon^2)\right)\right] \leq c_{LS}(\mathrm{Q})\varepsilon^2\mathbb{E}_\mathrm{Q}\left[\|\nabla f\|^2\right].$$

Note that, by a Taylor expansion, $\log(1+\varepsilon f) = \varepsilon f - \frac{(\varepsilon f)^2}{2} + o\left(\varepsilon^2\right)$ and also that $\log(1+\varepsilon^2) = \varepsilon^2 + o(\varepsilon^2)$. Then, plugging this into the previous equation gives:

$$\mathbb{E}_Q\left[2\varepsilon f + 3(\varepsilon f)^2 - \varepsilon^2 + o(\varepsilon^2)\right] \leq c_{LS}(\mathrm{Q})\varepsilon^2\mathbb{E}_\mathrm{Q}\left[\|\nabla f\|^2\right].$$

We use that $\mathbb{E}[f] = 0$ and we then divide by ε^2. Taking the limit $\varepsilon \to 0$ gives:

$$\mathbb{E}_Q\left[3f^2 - 1\right] \leq c_{LS}(\mathrm{Q})\mathbb{E}_\mathrm{Q}\left[\|\nabla f\|^2\right].$$

Using that $\mathbb{E}[f^2] = 1$ gives:

$$1 \leq \frac{c_{LS}(\mathrm{Q})}{2}\mathbb{E}_\mathrm{Q}\left[\|\nabla f\|^2\right]$$

Then, for any $g \in \mathrm{H}^1(\mathrm{Q})$ applying this proof on $f = \frac{g - \mathbb{E}_\mathrm{Q}[g]}{\sqrt{\mathrm{Var}_\mathrm{Q}(g)}}$ concludes the proof. ■

A.2. Wasserstein distances

We recall here the definition of Wasserstein distances, valid for any Polish space $\mathcal{H}$ equipped with a distance d.

Definition 15 *The 1-Wasserstein distance between $P, Q \in \mathcal{M}(\mathcal{H})^2$ is defined as*

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

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

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

Appendix B. PAC-Bayes bounds for Lipschitz losses through log-Sobolev inequalities

Extending Catoni's bound to Lipschitz losses. A well-known relaxation of Catoni (2007, Theorem 1.2.6) (see *e.g.* Alquier et al., 2016, Theorem 4.1) holding for subgaussian losses has been widely used in practice as a tractable PAC-Bayesian algorithm exhibiting a linear dependency on the KL divergence. We exploit below a consequence of the Herbst argument as stated, *e.g.*, in Ledoux (2006, Section 2.3), stating that a L-Lipschitz function of a random variable following a distribution $\mathcal{D}$ being $\mathtt{L-Sob}(c_{LS})$ is $L\sqrt{c_{LS}(\mathcal{D})}$ subgaussian. This yields the following corollary.

Corollary 16 *Let $\lambda > 0$, $m \geq 1$ and a data-free prior P. Assume that for any $h \in \mathcal{H}$, $\ell(h,.)$ is L-Lipschitz and that the data distribution $\mathcal{D}$ is $\mathtt{L-Sob}(c_{LS})$. Then for with probability at least $1-\delta$ over $\mathcal{S}$, for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$,*

$$\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq \hat{\mathrm{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{\lambda} + \frac{2\lambda^2 L^2 c_{LS}(\mathcal{D})}{m}.$$

Proof We take $f(h) = \lambda \Delta_S(h) := \lambda(\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) - \hat{\mathrm{R}}_{\mathcal{S}_m}(\mathrm{Q}))$ first use the change of measure inequality (Csiszar, 1975; Donsker and Varadhan, 1976) to state that, for any Q,

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

Markov's inequality alongside Fubini's theorem gives, with probability at least $1-\delta$,

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

Now, we use that f is L-Lipschitz for all h, then the function $\mathcal{S} \to \Delta_{\mathcal{S}}(h)$ is $\frac{2}{\sqrt{m}}$-Lipschitz on $\mathcal{S}$ for each h. As $\mathcal{D}$ is $\mathtt{L-Sob}(c_{LS})$ inequality, $\mathcal{D}^{\otimes m}$ is also $\mathtt{L-Sob}(c_{LS})$ with identical constant (Ané et al., 2000, Corollary 3.2.3). Then, using Herbst argument similarly as in Ledoux (2006, Section 2.3) allow us to conclude that f is $2L\sqrt{c_{LS}(\mathcal{D})}$-subgaussian, thus,

$$\log\left(\mathbb{E}_{h\sim\mathrm{P}}\mathbb{E}_{\mathcal{S}}\left[\exp\left(f(h)\right)\right]\right) \leq \frac{2\lambda^2 L^2}{m}.$$

This concludes the proof. ■

Disintegrated PAC-Bayes bounds Numerical estimation of PAC-Bayes bounds is usually challenging as it often involves Monte-Carlo approximations of the expectation over the posterior Q. A recent line of work (Rivasplata et al., 2020; Haddouche and Guedj, 2022; Viallard et al., 2023a) studies *disintegrated PAC-Bayes bounds e.g.*, bounds holding with high-probability on both the dataset $\mathcal{S}$ and a single predictor h drawn from the posterior Q. Those bounds are relevant for practitioners as they require little computational time. However, a drawback of these bounds is that existing disintegrated bounds do not allow the KL divergence to be used as a complexity measure. Either disintegrated KL (Rivasplata et al., 2020) or Rényi divergences (Viallard et al., 2023a), which can be seen as a relaxation of the KL one, are considered.

Using again the subgaussianity behavior of Lipschitz losses, it is possible to attain PAC-Bayesian disintegrated bounds as long as the posterior distribution satisfies a log-Sobolev inequality with sharp constant (achievable for instance for Gaussian distribution with small operator norm).

Lemma 17 *Assume that for any* $\mathbf{z}, \ell(., \mathbf{z})$ *is* L*-Lipschitz and that* Q *is* $\texttt{Poinc}(c_P)$ *with* $c_P(\mathrm{Q}) \leq 1/m$*. Then, with probability* $1-\delta$ *over the draw of* $h \sim \mathrm{Q}$*:*

$$\Delta_{\mathcal{S}_m}(h) \leq \Delta_{\mathcal{S}_m}(\mathrm{Q}) + \sqrt{\frac{2L^2 \log(1/\delta)}{m}}.$$

This lemma states that, as long as we assume our loss to be Lipschitz *w.r.t.* h, then it is possible to easily derive disintegrated PAC-Bayesian bounds. Also notice that Lemma 17 is easily completed by Corollary 16 which makes appear a KL divergence as complexity. Note also that as the loss is Lipschitz, it is also possible to make appear 1-Wasserstein distance through the bounds of Haddouche and Guedj (2023b); Viallard et al. (2023b). Thus Having a Log-Sobolev assumption with sharp constant on the posterior distribution is enough to provide disintegrated PAC-Bayesian bounds involving KL or Wasserstein terms instead of Rényi divegerences or disintegrated KL.

Appendix C. Proofs

C.1. Proof of Corollary 6

To start this proof, we first state an important intermediary theorem, holding at no assumption on the data-distribution.

Theorem 18 *For any* $C > 0$*, any* $\frac{2}{C} > \lambda > 0$*, any data-free prior* P*, any nonnegative loss function* ℓ *such that, for any* $\mathbf{z} \in \mathcal{Z}, \ell(., \mathbf{z}) \in \mathrm{H}^1$*, and any* $\delta \in [0,1]$*, we have, with probability at least* $1-\delta$ *over the sample* $\mathcal{S}$*, for any* $m > 0$*, any posterior* Q *being* $\texttt{Poinc}(c_P)$*, such that* $\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq C$ *and such that for any* $\mathbf{z}$*,* $\ell(., \mathbf{z}) \in \mathrm{H}^1(\mathrm{Q})$*:*

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

Theorem 18 exhibits the influence of the gradient norm of $\nabla_h \ell$ on the generalisation ability: small gradients makes the bound vanish, the remaining variance term is not treated for now and can

be assumed bounded, but we cannot then recover a fast rate. We show next that assuming additional assumption over the data distribution circumvent this issue.

Proof We re-start from Chugg et al. (2023, Corollary 17), for any $\lambda > 0$, with probability at least $1-\delta$, for any $m > 0$, any posterior Q:

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

Then, the last term is controlled as follows,

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

We then make appear a supplementary variance term:

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

Note that by Fubini, the last term on the right-hand side is exactly $\mathrm{R}_{\mathcal{D}}(\mathrm{Q})^2$, then using that the averaged true risk is lesser than C, and re-organising the terms in Chugg et al. (2023, Corollary 17) gives, for $\lambda \in \left(0, \frac{2}{C}\right)$:

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

■

Now Theorem 18 is proven, we only need to exploit the Poincaré assumption on the data distribution on the variance term to obtain Corollary 6.

C.2. Proof of Theorem 7

Proof We start again from Theorem 5, with $\lambda = 1/C_1$ then have with probability $1-\delta/2$:

$$\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\left(\hat{\mathrm{R}}_{\mathcal{S}_m}(\mathrm{Q}) + 2C_1\frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{m}\right) + \frac{c_P(\mathrm{Q})}{C_1}\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h\ell(h,\mathbf{z})\|^2\right)\right]. \tag{3}$$

We now remark that $g(h,\mathbf{z}) := \|\nabla_h\ell(h,\mathbf{z})\|^2$ is nonnegative. Then, given our assumptions, we apply the route of proof of Theorem 5 on g *i.e.* we start again from the (Chugg et al., 2023, Corollary

17), apply Poincaré's inequality on Q and use the QSB assumption on g. We then have for any $\lambda > 0$, with probability at least $1 - \delta/2$, any Q being $\texttt{Poinc}(c_P)$, $\texttt{QSB}\,(g, C_2)$ and $g(., \mathbf{z}) \in \mathrm{H}^1(\mathrm{Q})$ for all $\mathbf{z}$:

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right] \leq \mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^{m}\|\nabla_h \ell(h,\mathbf{z}_i)\|^2\right] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{\lambda m} + \frac{\lambda c_P(\mathrm{Q})}{2}\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h g(h,\mathbf{z})\|^2\right)\right] + \frac{\lambda C_2}{2}\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right]. \quad (4)$$

Finally, notice that, by definition of g, $\nabla_h g(h,\mathbf{z}) = 2Hess_h(\ell)(h,\mathbf{z})\nabla_h \ell(h,\mathbf{z})$, where $Hess_h(\ell)$ denotes the Hessian of ℓ. Thus, using that $\ell(.,\mathbf{z})$ is G gradient Lipschitz for any $\mathbf{z}$ gives, for any $(h,\mathbf{z})$ that $\|\nabla_h g(h,\mathbf{z})\| \leq 2G\|\nabla_h \ell(h,z)\|$. Plugging this in (4) gives:

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right] \leq \mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^{m}\|\nabla_h \ell(h,\mathbf{z}_i)\|^2\right] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{\lambda m} + \frac{\lambda}{2}\left(4c_P(\mathrm{Q})G^2 + C_2\right)\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right]. \quad (5)$$

Finally, using that $c_P(\mathrm{Q}) = c$, taking $\lambda = \frac{1}{4cG^2 + C_2}$ and re-organising the terms in (5) gives:

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right] \leq 2\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\frac{1}{m}\sum_{i=1}^{m}\|\nabla_h \ell(h,\mathbf{z}_i)\|^2\right] + 2(4cG^2 + C_2)\frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{m} \quad (6)$$

Finally, taking an union bound and plugging (6) in (3) concludes the proof. ■

C.3. Proof of Lemma 9

Proof For conciseness, we rename $\mathrm{Q} := \mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}$. We first notice that, denoting by $\frac{d\mathrm{Q}}{d\mathrm{P}}$ the Radon-Nikodym derivative of Q with respect to P:

$$\begin{aligned}\mathrm{KL}\left(\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}, \mathrm{P}\right) &= \mathbb{E}_{h\sim\mathrm{Q}}\left[\log\left(\frac{d\mathrm{Q}}{d\mathrm{P}}(h)\right)\right] \\ &= \mathrm{Ent}_{\mathrm{P}}\left(\frac{d\mathrm{Q}}{d\mathrm{P}}\right) = \mathrm{Ent}_P[g^2],\end{aligned}$$

where $g = \sqrt{\frac{d\mathrm{Q}}{d\mathrm{P}}}$.

Recall that $\frac{d\mathrm{Q}}{d\mathrm{P}}(h) = \frac{1}{Z}\exp\left(-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}(h)\right)$ where $Z = \mathbb{E}_{h\sim\mathrm{P}}\left[\exp\left(-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}(h)\right)\right]$.

Then, $g(h) = \frac{1}{\sqrt{Z}}\exp(-\frac{\gamma}{2}\hat{\mathrm{R}}_{\mathcal{S}_m}(h))$ belongs in $\mathrm{H}^1(\mathrm{P})$ as long as $\ell \in \mathrm{H}^1$. Indeed, as exp is infinitely smooth, $g \in D_1(\mathbb{R}^d)$, also as the loss is nonnegative, then $g \leq \frac{1}{\sqrt{Z}}$ thus $g \in L^2(\mathrm{P})$.

Finally, $\nabla g = -\frac{\gamma}{2} g(h) \nabla \hat{\mathrm{R}}_{\mathcal{S}_m}(h)$. As $g(h) \leq \frac{1}{\sqrt{K}}$, we only need to bound $\|\nabla \hat{\mathrm{R}}_{\mathcal{S}_m}(h)\|^2$ to ensure that $g \in \mathrm{H}^1(\mathrm{P})$:

$$\begin{aligned}\|\nabla \hat{\mathrm{R}}_{\mathcal{S}_m}(h)\|^2 &= \frac{1}{m^2} \sum_{1 \leq i,j \leq m} \langle \nabla \ell(h, \mathbf{z}_i), \nabla \ell(h, \mathbf{z}_j) \rangle \\ &\leq \frac{1}{2m^2} \sum_{1 \leq i,j \leq m} \|\nabla \ell(h, \mathbf{z}_i)\|^2 + \|\nabla \ell(h, \mathbf{z}_j)\|^2\end{aligned}$$

As we assumed $\|\nabla \ell(., \mathbf{z})\|^2 \in L^2(\mathrm{P})$ for all $\mathbf{z}$, we conclude that $g \in \mathrm{H}^1$.
We then can apply the log-Sobolev inequality to conlude that

$$\begin{aligned}\mathrm{KL}\left(\mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}, \mathrm{P}\right) &\leq c_{LS}(\mathrm{P}) \mathop{\mathbb{E}}_{h \sim \mathrm{P}} [\|\nabla g(h)\|^2] \\ &= \frac{\gamma^2 c_{LS}(\mathrm{P})}{4} \mathop{\mathbb{E}}_{h \sim \mathrm{P}} \left[\|\nabla_h \hat{\mathrm{R}}_{\mathcal{S}_m}(h)\|^2 g^2(h)\right] \\ &= \frac{\gamma^2 c_{LS}(\mathrm{P})}{4} \mathop{\mathbb{E}}_{h \sim \mathrm{P}} \left[\|\nabla_h \hat{\mathrm{R}}_{\mathcal{S}_m}(h)\|^2 \frac{d\mathrm{Q}}{d\mathrm{P}}(h)\right] \\ &= \frac{\gamma^2 c_{LS}(\mathrm{P})}{4} \mathop{\mathbb{E}}_{h \sim \mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}} \left[\|\nabla_h \hat{\mathrm{R}}_{\mathcal{S}_m}(h)\|^2\right]\end{aligned}$$

■

C.4. Proof of Theorem 10

Proof We start again from Chugg et al. (2023, Corollary 17) instantiated with a single λ, *i.i.d.* data and a prior P. Then with probability at least $1 - \delta$, for any posterior Q and $m > 0$:

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

where $\mathbf{z} \sim \mathcal{D}$ is independent of $\mathcal{S}$.

For the first inequality, we just take $\lambda = 1$, we use that $\ell(h, \mathbf{z})^2 \leq \ell(h, \mathbf{z})$ and re-organise the terms. Finally, we upper bound the KL term thanks to Lemma 9.

For the second inequality, we exploit Proposition 3 to use the fact that $\mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}$ is $\texttt{L-Sob}(c_{LS})$ alongside Proposition 14 which ensures that $\mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}$ is $\texttt{Poinc}(c_P)$ with constant equal to $c_{LS}\left(\mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}\right)/2$.

We then apply a route of proof similar to Theorem 5. We have :

$$\mathop{\mathbb{E}}_{h \sim \mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}} \left[\mathbb{E}_{\mathbf{z} \sim \mathcal{D}}[\ell(h, \mathbf{z})^2]\right] = \mathop{\mathbb{E}}_{\mathbf{z} \sim \mathcal{D}} \left[\mathrm{Var}_{h \sim \mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}} (\ell(h, \mathbf{z})) + \left(\mathop{\mathbb{E}}_{h \sim \mathrm{P}_{-\gamma \hat{\mathrm{R}}_{\mathcal{S}_m}}} [\ell(h, \mathbf{z})] \right)^2 \right]$$

Applying Poincaré's inequality then gives:

$$\leq \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[c_P(\mathrm{P})e^{4\|\ell_2\|_\infty}\mathop{\mathbb{E}}_{h\sim\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}}\left(\|\nabla_h\ell(h,\mathbf{z})\|^2\right)+\left(\mathop{\mathbb{E}}_{h\sim\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}}[\ell(h,\mathbf{z})]\right)^2\right].$$

Finally, using that $\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}}$ is $\mathtt{QSB}(\ell, C)$ allow us to re-organise the terms as in Theorem 5. Combining this with Lemma 9 to bound the KL divergence concludes the proof.

■

C.5. Proof of Theorem 11

Proof We assume first $G = 1$. We start from the Kantorovich duality formula (Villani, 2009, Theorem 5.10) instantiated with the cost function $c(x, y) = \|x - y\|^2$. We have for any Q, P, because W_2 is a distance:

$$W^2(\mathrm{Q},\mathrm{P}) = W^2(\mathrm{P},\mathrm{Q}) = \sup_{\phi,\psi}\ \mathbb{E}_{h\sim\mathrm{Q}}[\phi(h)] - \mathbb{E}_{h\sim\mathrm{P}}[\psi(h)], \tag{7}$$

where the supremum is taken over the functions $\phi, \psi \in L^1(\mathrm{Q}) \times L^1(\mathrm{P})$ such that for all $h, h' \in \mathcal{H}^2$, $\phi(h) - \psi(h') \leq \|h - h'\|^2$.

We claim that if $\phi(h) = f(h) - D\|\nabla f(h)\|$ and $\psi(h') = f(h')$ then the pair Φ, Ψ satisfies $\phi(h) - \psi(h') \leq \frac{\|h-h'\|^2}{2}$.

Indeed,

$$\begin{aligned}\phi(h) - \psi(h') &= f(h) - f(h') - D\|\nabla f(h)\| \\ &= f\circ g(1) - f\circ g(0) - D\|\nabla f(h)\|,\end{aligned}$$

where $g(t) = th + (1-t)h'$. Then, by the fundamental theorem of calculus, we have

$$\begin{aligned}\phi(h) - \psi(h') &= \int_0^1 (f\circ g)'(t)dt - D\|\nabla f(h)\| \\ &= \int_0^1 \left\langle \nabla f\left(th + (1-t)h'\right), h - h'\right\rangle dt - D\|\nabla f(h)\|.\end{aligned}$$

We now control the last term using that $\|h - h'\| \leq D$ and Cauchy-Schwarz:

$$\begin{aligned}\phi(h) - \psi(h') &\leq \int_0^1 \left\langle \nabla f\left(th + (1-t)h'\right), h - h'\right\rangle dt - \left\langle \nabla f(h), h - h'\right\rangle \\ &= \int_0^1 \left\langle \nabla f\left(th + (1-t)h'\right) - \nabla f(h), h - h'\right\rangle dt.\end{aligned}$$

Then by Cauchy-Schwarz alongside Lipschitz gradient,

$$\begin{aligned}\phi(h) - \psi(h') &\leq \|h - h'\| \int_0^1 \left\|\nabla f\left(th + (1-t)h'\right) - \nabla f(h)\right\| dt \\ &\leq \|h - h'\| \int_0^1 (1-t)dt \left\|h - h'\right\| dt \\ &= \frac{\|h - h'\|^2}{2}.\end{aligned}$$

We then conclude by applying (7) to the pair $(2\phi, 2\psi)$. The general case with $G \neq 1$ is immediate when considering the pair $(\frac{2}{G}\phi, \frac{2}{G}\psi)$. ■

C.6. Proof of Corollary 12

Proof We fix $R > 0$ and we start from Theorem 11 with predictor space $\mathcal{H}_0 = \mathcal{B}(\mathbf{0}, R)$, f being gradient-Lipschitz on this ball and prior and posterior $\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}$,

$$\mathbb{E}_{h\sim\mathrm{Q}}\left[f\left(\mathcal{P}_R(h)\right)\right] \leq \frac{G}{2} W_2^2\left(\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}\right) + \mathbb{E}_{h\sim\mathrm{P}}\left[f\left(\mathcal{P}_R(h)\right)\right] + 2R\mathbb{E}_{h\sim\mathrm{Q}}\left[\left\|\nabla f\left(\mathcal{P}_R(h)\right)\right\|\right].$$

We first prove that $W_2^2\left(\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}\right) \leq W_2^2(\mathrm{Q}, \mathrm{P})$. Let $\pi \in \Gamma(\mathrm{Q}, \mathrm{P})$ being the optimal transport coupling from P to Q, *i.e.*

$$W_2^2(\mathrm{Q}, \mathrm{P}) = \mathbb{E}_{(X,Y)\sim\pi}\left[\|X - Y\|^2\right].$$

Then notice that if we denote by $\pi_1 = (\mathcal{P}_R, \mathcal{P}_R)\#\pi$, then $\pi_1 \in \Gamma\left(\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}\right)$ and so:

$$\begin{aligned}W_2^2\left(\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}\right) &\leq \mathbb{E}_{(X,Y)\sim\pi_1}\left[\|X - Y\|^2\right] \\ &= \mathbb{E}_{(X,Y)\sim\pi_1}\left[\|\mathcal{P}_R(X) - \mathcal{P}_R(Y)\|^2\right].\end{aligned}$$

Using that $\mathcal{P}_R$ is 1-Lipschitz gives,

$$\begin{aligned}W_2^2\left(\mathcal{P}_R\#\mathrm{Q}, \mathcal{P}_R\#\mathrm{P}\right) &\leq \mathbb{E}_{(X,Y)\sim\pi_1}\left[\|X - Y\|^2\right] \\ &= W_2^2(\mathrm{Q}, \mathrm{P}).\end{aligned}$$

Then we need to control $W_2^2(\mathrm{Q}, \mathrm{P})$. To do so, we use the fact that P is $\mathtt{L\text{-}Sob}(c_{LS})$ to affirm, through Otto-Villani's theorem (Otto and Villani, 2000, Theorem 1) that the following holds: $W_2^2(\mathrm{Q}, \mathrm{P}) \leq \frac{c_{LS}(\mathrm{P})}{2}\mathrm{KL}(\mathrm{Q}, \mathrm{P})$. This concludes the proof. ■

C.7. Proof of Theorem 13

Proof We start from Theorem 11, using that $\Delta_{\mathcal{S}_m}$ is G-gradient-Lipschitz for any m to obtain:

$$\mathbb{E}_{h\sim\mathrm{Q}}[\Delta_{\mathcal{S}_m}(h)] \leq \frac{G}{2}W_2^2(Q,P) + \mathbb{E}_{h\sim\mathrm{P}}[\Delta_{\mathcal{S}_m}(h)] + D\mathbb{E}_{h\sim\mathrm{Q}}[\|\nabla\Delta_{\mathcal{S}_m}(h)\|]$$

The only thing left to control is $\mathbb{E}_{h\sim\mathrm{P}}[\Delta_{\mathcal{S}_m}(h)]$. For this, we use that P alongisde the supermartingale concentration inequality of Chugg et al. (2023, Corollary 17) instantiated with prior equal to posterior, *i.i.d.* data showing that, for any $\lambda > 0$, with probability at least $1-\delta$,

$$\mathbb{E}_{h\sim\mathrm{P}}[\Delta_{\mathcal{S}_m}(h)] \leq \frac{\log(1/\delta)}{\lambda} + \frac{\lambda}{2}\mathbb{E}_{h\sim\mathrm{P}}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,z)^2].$$

The last term on the right-hand side is bounded by σ^2 by assumption, then, taking $\lambda = \sqrt{\frac{2\log(1/\delta)}{\sigma^2}}$ gives finally $\mathbb{E}_{h\sim\mathrm{P}}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,z)^2] \leq \sqrt{2\log(1/\delta)/m}$, concludes the proof. ■