LIER LA THÉORIE PAC-BAYÉSIENNE AUX MINIMA PLATS

Maxime Haddouche[1] & Paul Viallard [2],
& Umut Simsekli[3], & Benjamlin Guedj[4]

[1] *Inria London & Université de Lille, France, maxime.haddouche@inria.fr*
[2] *Inria & Université de Rennes, France, paul.viallard@inria.fr*
[3] *Inria Paris, France, umut.simsekli@inria.fr*
[4] *Inria London & University College London, UK, benjamin.guedj@inria.fr*

L'apprentissage automatique moderne implique généralement des prédicteurs dans un contexte sur-paramétré (nombre de paramètres entraînés supérieur à la taille de l'ensemble de données), et leur apprentissage donne non seulement de bonnes performances sur les données d'entraînement, mais également une bonne capacité de généralisation. Ce phénomène remet en question de nombreux résultats théoriques et reste un problème ouvert. Pour parvenir à une meilleure compréhension, nous présentons dans cet exposé de nouvelles limites de généralisation impliquant des termes de gradient. Pour ce faire, nous combinons la théorie PAC-Bayésienne avec les inégalités de Poincaré et Log-Sobolev, évitant ainsi une dépendance explicite à la dimension de l'espace des prédicteurs. Ces résultats mettent en évidence l'influence positive des *minima plats* (étant des minima avec un voisinage minimisant presque le problème d'apprentissage) sur les performances de généralisation, impliquant directement les bénéfices de la phase d'optimisation.

PAC-Bayes, Statistical Learning, Flat Minima, Poincaré and Log-Sobolev Inequalities ...

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 present in this talk 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. Those 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.

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., 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.

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* $\mathtt{Poinc}(c_P)$*. 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* $\mathtt{L\text{-}Sob}(c_{LS})$*.*

The class of Gaussian distributions is an important particular case of distributions satisfying both Poincaré and log-Sobolev inequalities A gaussian Q with covariance matrix Σ_{op}, Q is $\mathtt{L\text{-}Sob}(c_{LS})$ with constant $c_{LS}(\mathrm{Q}) = 2\|\Sigma\|_{op}$ and also $\mathtt{Poinc}(c_{LS})$ with constant $c_{LS}(\mathrm{Q}) = \|\Sigma\|_{op}$, where $\|.\|_{op}$ denotes the operator norm. Also, we 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]. Theorem 2 shows that when the prior and the loss satisfies a few properties, then the associated Gibbs posterior is $\mathtt{L\text{-}Sob}(c_{LS})$.

Proposition 2 *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* $c_{LS}(\mathrm{Q}) = c_{LS}(\mathrm{P}) \exp\left(4\|\ell_2\|_\infty\right)$.

Theorem 2 applies, *e.g.*, when P is a Gaussian prior $\mathrm{P} = \mathcal{N}(\mu_\mathrm{P}, \Sigma_\mathrm{P})$. Notice that in this case $c_{LS}(\mathrm{P}) = 2\|\Sigma_\mathrm{P}\|_{op}$. This property is a straightforward application of Chafaï [2004, Corollary 2.1] with Guionnet and Zegarlinksi [2003, Property 2.6]. 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].

3 Reaching a flat minimum allows Poincaré posteriors to generalise well

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 $\mathbf{z}$. In this section, we consider posterior distributions Q being $\texttt{Poinc}(c_P)$. This assumption covers the important case of Gaussian measures as well as all measures satisfying a log-Sobolev inequality. 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 3 to later involve flat minima.

Assumption 3 *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*

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

Assumption 3 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, 2023, 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. 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 [2023], Chugg et al. [2023] alongside Poincaré inequality leads to the following.

Theorem 4 *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* **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 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 4 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 4. 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 4 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.

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 4? 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 4 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 4 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 4. 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.

It is possible to go beyond the 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 5 *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 $\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq C$:*

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

Corollary 5 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 $\mathcal{D}$ 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.

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., 2023] among others. However, Theorem 4 and Corollary 5 are not fully empirical and thus, do not have such desirable properties. We circumvent this issue in Theorem 6.

Theorem 6 *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 $\texttt{Poinc}(c_P)$ 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 $\mathbf{z}$,*

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

Here, we showed that to attain fast rates, the 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 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 6 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.

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. Theorem 7 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 7 *For any m, P being L-$Sob(c_{LS})$, any $\ell \geq 0$ such that for any $\mathbf{z}$, $\ell(.,\mathbf{z}) \in \mathrm{H}^1(\mathrm{P})$, we have, for any $\gamma > 0$:*

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

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 8 *For any $C > 0$, any $\gamma > 0$, any prior P being 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 Theorem 2. Then for any $\frac{2}{C} > \lambda > 0$, with probability at least $1-\delta$ over the sample $\mathcal{S}$, for any $m > 0$, such that* Q *is* $\mathtt{QSB}(\ell, C)$ *and $\ell(., \mathbf{z}) \in \mathrm{H}^1(\mathrm{P}_{-\gamma\hat{\mathrm{R}}_{\mathcal{S}_m}})$:*

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

Note that we could have derived analogous to Corollary 5 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 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.

References

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

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 2016.

M. Andriushchenko, F. Croce, M. Müller, M. Hein, and N. Flammarion. A modern look at the relationship between sharpness and generalization. *arXiv preprint arXiv:2302.07011*, 2023.

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

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

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

B. Chugg, H. Wang, and A. 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.

G. K. Dziugaite and D. 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.

G. K. Dziugaite, A. Drouin, B. Neal, N. Rajkumar, E. Caballero, L. Wang, I. Mitliagkas, and D. 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.

P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

I. Gat, Y. Adi, A. G. Schwing, and T. 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.

P. Grunwald, T. Steinke, and L. Zakynthinou. PAC-Bayes, MAC-Bayes and Conditional Mutual Information: Fast rate bounds that handle general VC classes. In M. Belkin and S. 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.

B. 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.

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

F. Hellström, G. Durisi, B. Guedj, and M. Raginsky. Generalization bounds: Perspectives from information theory and PAC-Bayes. *arXiv preprint arXiv:2309.04381*, 2023.

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

S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

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

Z. Mhammedi, P. Grünwald, and B. 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.

B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. 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.

M. Perez-Ortiz, O. Rivasplata, E. Parrado-Hernandez, B. Guedj, and J. Shawe-Taylor. Progress in Self-Certified Neural Networks. In *NeurIPS 2021 Workshop on Bayesian Deep Learning*, 2021.

H. Petzka, M. Kamp, L. Adilova, C. Sminchisescu, and M. Boley. Relative flatness and generalization. *Advances in neural information processing systems*, 34:18420–18432, 2021.

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

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

P. Viallard, M. Haddouche, U. Simsekli, and B. Guedj. Learning via Wasserstein-Based High Probability Generalisation Bounds. *To be published in NeurIPS 2023*, 2023.

K. Wen, Z. Li, and T. 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.

Y. Yue, J. Jiang, Z. Ye, N. Gao, Y. Liu, and K. Zhang. Sharpness-aware minimization revisited: Weighted sharpness as a regularization term. *arXiv preprint arXiv:2305.15817*, 2023.