PAC-Bayesian Theory Meets Bayesian Inference

Pascal Germain, Francis Bach, Alexandre Lacoste, Simon Lacoste-Julien

PAC-Bayesian Theory Meets Bayesian Inference

Pascal Germain[†] **Francis Bach**[†] **Alexandre Lacoste**[‡] **Simon Lacoste-Julien**[†]
[†] INRIA Paris - École Normale Supérieure, firstname.lastname@inria.fr
[‡] Google, allac@google.com

Abstract

We exhibit a strong link between frequentist PAC-Bayesian bounds and the Bayesian marginal likelihood. That is, for the negative log-likelihood loss function, we show that the minimization of PAC-Bayesian generalization bounds maximizes the Bayesian marginal likelihood. This provides an alternative explanation to the Bayesian Occam's razor criteria, under the assumption that the data is generated by a *i.i.d.* distribution. Moreover, as the negative log-likelihood is an unbounded loss function, we motivate and propose a PAC-Bayesian theorem tailored for the sub-Gamma loss family, and we show that our approach is sound on classical Bayesian linear regression tasks.

1 Introduction

Since its early beginning [Shawe-Taylor and Williamson, 1997], the PAC-Bayesian theory claims to provide "PAC guarantees to *Bayesian* algorithms" [McAllester, 1999]. However, despite the amount of work dedicated to this statistical learning theory—many authors improved the initial results[1] and/or generalized them for various machine learning setups[2]—it is mostly used as a *frequentist* method. That is, under the assumptions that the learning samples are *i.i.d.*-generated by a data-distribution, this theory expresses *probably approximately correct* (PAC) bounds on the generalization risk. In other words, with probability $1-\delta$, the generalization risk is at most ε away from the training risk. The *Bayesian* side of PAC-Bayes comes mostly from the fact that these bounds are expressed on the averaging/aggregation/ensemble of multiple predictors (weighted by a *posterior* distribution) and incorporate prior knowledge. Although it is still sometimes referred as a theory that bridges the Bayesian and frequentist approach [*e.g.*, Guyon et al., 2010], it has been merely used to explicitly justify Bayesian methods until now.[3]

In this work, we provide (up to our knowledge) the first direct connection between Bayesian inference techniques [summarized by Ghahramani, 2015] and PAC-Bayesian theory in a general setup. Our study is based on a simple but insightful connection between the Bayesian marginal likelihood and PAC-Bayesian bounds, that we obtain by considering the negative log-likelihood loss function (Section 3). By doing so, we provide an alternative explanation for the Bayesian Occam's razor criteria [Jeffreys and Berger, 1992, MacKay, 1992] in the context of model selection, explained as the complexity-accuracy trade-off appearing in most PAC-Bayesian results. In Section 4, we extend PAC-Bayes theorems to regression problems with unbounded loss, adapted to the negative log likelihood loss function. Finally, we study the Bayesian model selection from a PAC-Bayesian perspective (Section 5), and illustrate our finding on classical Bayesian regression tasks (Section 6).

[1] Seeger [2003], McAllester [2003], Catoni [2007], Lever et al. [2013], Tolstikhin and Seldin [2013], etc.

[2] Langford and Shawe-Taylor [2002], Seldin and Tishby [2010], Seldin et al. [2011, 2012], Bégin et al. [2014], Pentina and Lampert [2014], etc.

[3] Some indirect connections can be found in Seeger [2002], Banerjee [2006], Zhang [2006], Lacoste [2015], Bissiri et al. [2016]. See also Ng and Jordan [2001], Meir and Zhang [2003], Grünwald and Langford [2007], Lacoste-Julien et al. [2011] for other studies drawing links between frequentist statistics and Bayesian inference.

2 PAC-Bayesian Theory

We denote the learning sample $(X,Y)=\{(x_i,y_i)\}_{i=1}^n \in (\mathcal{X}\times\mathcal{Y})^n$, that contains n input-output pairs. The main assumption of frequentist learning theories—including PAC-Bayes—is that (X,Y) is randomly sampled from a data generating distribution that we denote $\mathcal{D}$. Thus, we denote $(X,Y)\sim\mathcal{D}^n$ the *i.i.d.* observation of n elements. From a frequentist perspective, we consider in this work loss functions $\ell : \mathcal{F}\times\mathcal{X}\times\mathcal{Y} \to \mathbb{R}$, where $\mathcal{F}$ is a (discrete or continuous) set of predictors $f : \mathcal{X} \to \mathcal{Y}$, and we write the empirical risk on the sample (X,Y) and the generalization error on distribution $\mathcal{D}$ as

$$\widehat{\mathcal{L}}^{\ell}_{X,Y}(f) \;=\; \frac{1}{n}\sum_{i=1}^{n} \ell(f,x_i,y_i)\,; \quad \mathcal{L}^{\ell}_{\mathcal{D}}(f) \;=\; \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}} \ell(f,x,y)\,.$$

The PAC-Bayesian theory [McAllester, 1999, 2003] studies an averaging of the above losses according to a *posterior* distribution $\hat{\rho}$ over $\mathcal{F}$. That is, it provides *probably approximately correct* generalization bounds on the (unknown) quantity $\mathbf{E}_{f\sim\hat{\rho}}\,\mathcal{L}^{\ell}_{\mathcal{D}}(f) \;=\; \mathbf{E}_{f\sim\hat{\rho}}\,\mathbf{E}_{(x,y)\sim\mathcal{D}}\,\ell(f,x,y)$, given the empirical estimate $\mathbf{E}_{f\sim\hat{\rho}}\,\widehat{\mathcal{L}}^{\ell}_{X,Y}(f)$ and some other parameters. Among these, most PAC-Bayesian theorems rely on the *Kullback-Leibler* divergence $\mathrm{KL}(\hat{\rho}\|\pi) \;=\; \mathbf{E}_{f\sim\hat{\rho}} \ln[\hat{\rho}(f)/\pi(f)]$ between a *prior* distribution π over $\mathcal{F}$—specified before seeing the learning sample X, Y—and the posterior $\hat{\rho}$—typically obtained by feeding a learning process with X, Y.

Two appealing aspects of PAC-Bayesian theorems are that they provide data-driven generalization bounds that are computed on the training sample (*i.e.*, they do not rely on a testing sample) and that are uniformly valid for all $\hat{\rho}$ over $\mathcal{F}$. This explains why many works study them as model selection criteria or as an inspiration for learning algorithm conception. Theorem 1, due to Catoni [2007], has been used to derive or study learning algorithms in Germain et al. [2009], McAllester and Keshet [2011], Hazan et al. [2013], Noy and Crammer [2014].

Theorem 1 (Catoni, 2007)**.** *Given a distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, a hypothesis set $\mathcal{F}$, a loss function $\ell' : \mathcal{F}\times\mathcal{X}\times\mathcal{Y} \to [0,1]$, a prior distribution π over $\mathcal{F}$, a $\delta \in (0,1]$, and a real number $\beta > 0$, with probability at least $1-\delta$ over the choice of $(X,Y)\sim\mathcal{D}^n$, we have*

$$\forall \hat{\rho}\ \textit{on}\ \mathcal{F}\,: \quad \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \mathcal{L}^{\ell'}_{\mathcal{D}}(f) \;\leq\; \frac{1}{1-e^{-\beta}}\left[1 - e^{-\beta\ \mathbf{E}_{f\sim\hat{\rho}}\,\widehat{\mathcal{L}}^{\ell'}_{X,Y}(f) - \frac{1}{n}\left(\mathrm{KL}(\hat{\rho}\|\pi) + \ln\frac{1}{\delta}\right)}\right]. \tag{1}$$

Theorem 1 is limited to loss functions mapping to the range $[0,1]$. Through a straightforward rescaling we can extend it to any bounded loss, *i.e.*, $\ell : \mathcal{F}\times\mathcal{X}\times\mathcal{Y} \to [a,b]$, where $[a,b]\subset\mathbb{R}$. This is done by using $\beta \coloneqq b-a$ and with the *rescaled* loss function $\ell'(f,x,y) \;\coloneqq\; (\ell(f,x,y)-a)/(b-a) \in [0,1]$. After few arithmetic manipulations, we can rewrite Equation (1) as

$$\forall \hat{\rho}\ \text{on}\ \mathcal{F}\,: \quad \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \mathcal{L}^{\ell}_{\mathcal{D}}(f) \;\leq\; a + \tfrac{b-a}{1-e^{a-b}}\left[1 - \exp\left(-\mathop{\mathbf{E}}_{f\sim\hat{\rho}}\widehat{\mathcal{L}}^{\ell}_{X,Y}(f) + a - \tfrac{1}{n}\left(\mathrm{KL}(\hat{\rho}\|\pi) + \ln\tfrac{1}{\delta}\right)\right)\right]. \tag{2}$$

From an algorithm design perspective, Equation (2) suggests optimizing a trade-off between the empirical expected loss and the Kullback-Leibler divergence. Indeed, for fixed π, X, Y, n, and δ, minimizing Equation (2) is equivalent to find the distribution $\hat{\rho}$ that minimizes

$$n \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \widehat{\mathcal{L}}^{\ell}_{X,Y}(f) + \mathrm{KL}(\hat{\rho}\|\pi)\,. \tag{3}$$

As mentioned by Zhang [2006], Catoni [2007], Germain et al. [2009], Lever et al. [2013], Alquier et al. [2015], the *optimal Gibbs posterior* $\hat{\rho}^*$ is given by

$$\hat{\rho}^*(f) \;=\; \tfrac{1}{Z_{X,Y}}\pi(f)\, e^{-\,n\,\widehat{\mathcal{L}}^{\ell}_{X,Y}(f)}\,, \tag{4}$$

where $Z_{X,Y}$ is a normalization term. Notice that the constant β is now absorbed in the loss function as the rescaling factor setting the trade-off between the expected empirical loss and $\mathrm{KL}(\hat{\rho}\|\pi)$.

3 Bridging Bayes and PAC-Bayes

In this section, we show that by choosing the negative-log-likelihood loss function, minimizing the PAC-Bayes bound is equivalent to maximizing the Bayesian marginal likelihood. To obtain this

result, we first consider the Bayesian approach that starts by defining a prior $p(\theta)$ over the set of possible model parameters Θ. This induces a set of probabilistic estimators $f_\theta \in \mathcal{F}$, mapping x to a probability distribution over $\mathcal{Y}$. Then, we can estimate the likelihood of observing y given x and θ, *i.e.*, $p(y|x,\theta) \equiv f_\theta(y|x)$.[4] Using Bayes' rule, we obtain the posterior $p(\theta|X,Y)$:

$$p(\theta|X,Y) \;=\; \frac{p(\theta)\,p(Y|X,\theta)}{p(Y|X)} \;\propto\; p(\theta)\,p(Y|X,\theta)\,, \tag{5}$$

where $p(Y|X,\theta) \;=\; \prod_{i=1}^{n} p(y_i|x_i,\theta)$ and $p(Y|X) = \mathbf{E}_{\theta\sim p(\theta)}\, p(Y|X,\theta)$.

To bridge the Bayesian approach with the PAC-Bayesian framework, we consider the *negative log-likelihood* loss function [see Banerjee, 2006], denoted ℓ_{nll} and defined by

$$\ell_{\text{nll}}(f_\theta, x, y) \;\equiv\; -\ln p(y|x,\theta)\,. \tag{6}$$

Then, we can relate the *empirical loss* $\widehat{\mathcal{L}}^{\ell}_{X,Y}$ of a predictor to its likelihood:

$$\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{nll}}(\theta, x_i, y_i) \;=\; -\frac{1}{n}\sum_{i=1}^{n} \ln p(y_i|x_i,\theta) \;=\; -\frac{1}{n}\ln p(Y|X,\theta)\,,$$

or, the other way around,

$$p(Y|X,\theta) \;=\; e^{-n\,\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta)}. \tag{7}$$

Unfortunately, existing PAC-Bayesian theorems work with bounded loss functions or in very specific context [as Zhang, 2006, Dalalyan and Tsybakov, 2008], and ℓ_{nll} spans the whole real axis in its general form. To this end, in Section 4, we explore PAC-Bayes bounds for unbounded losses. Meanwhile, we consider priors with bounded likelihood. This can be done by assigning a prior of zero to any θ yielding $-\ln p(y|x,\theta) \notin [a,b]$.

Now, using Equation (7) in the optimal posterior (Equation 4) simplifies to:

$$\hat{\rho}^*(\theta) = \frac{\pi(\theta)\,e^{-n\,\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta)}}{Z_{X,Y}} \;=\; \frac{p(\theta)\,p(Y|X,\theta)}{p(Y|X)} = p(\theta|X,Y)\,, \tag{8}$$

where the normalization constant $Z_{X,Y}$ corresponds to the Bayesian *marginal likelihood*:

$$Z_{X,Y} \;\equiv\; p(Y|X) \;=\; \int_\Theta \pi(\theta)\,e^{-n\,\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta)} d\theta\,. \tag{9}$$

This shows that the optimal PAC-Bayes posterior given by the generalization bound of Theorem 1 coincides with the Bayesian posterior, when one chooses ℓ_{nll} as loss function and $\beta \coloneqq b-a$ (as in Equation 2). Moreover, using the posterior of Equation (8) inside Equation (3), we obtain

$$\begin{aligned} & n \mathop{\mathbf{E}}_{\theta\sim\hat{\rho}^*} \widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta) + \mathrm{KL}(\hat{\rho}^*\|\pi) \qquad (10)\\ =\;& n\int_\Theta \frac{\pi(\theta)\,e^{-n\,\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta)}}{Z_{X,Y}}\,\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta)\,d\theta + \int_\Theta \frac{\pi(\theta)\,e^{-n\,\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta)}}{Z_{X,Y}} \ln\left[\frac{\pi(\theta)\,e^{-n\,\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta)}}{\pi(\theta)\,Z_{X,Y}}\right] d\theta \\ =\;& \int_\Theta \frac{\pi(\theta)\,e^{-n\,\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta)}}{Z_{X,Y}} \left[\ln\frac{1}{Z_{X,Y}}\right] d\theta \;=\; \frac{Z_{X,Y}}{Z_{X,Y}}\ln\frac{1}{Z_{X,Y}} \;=\; -\ln Z_{X,Y}\,. \end{aligned}$$

In other words, minimizing the PAC-Bayes bound is equivalent to maximizing the marginal likelihood. Thus, from the PAC-Bayesian standpoint, the latter encodes a trade-off between the averaged negative log-likelihood loss function and the prior-posterior Kullback-Leibler divergence.[5] Although it appears in essence a very different problem, we note that the relation derived in Equation (10) is similar to the one used by variational Bayesian methods, which approximate a hardly computable Bayesian posterior by the "closest" distribution belonging to a parametrized family.

We conclude this section by proposing a compact form of Theorems 1 by expressing it in terms of the marginal likelihood, as a direct consequence of Equation (10).

[4] To stay aligned with the PAC-Bayesian setup, we only consider the discriminative case in this paper. One can extend to the generative setup by considering the likelihood of the form $p(y,x|\theta)$ instead.

[5] To our knowledge, this is the first time this has been reported in a general setup. The thesis of Seeger [2003, Section 3.2] foreseeing this by noticing that "the log marginal likelihood incorporates a *similar trade-off* as the PAC-Bayesian theorem", but using another variant of the PAC-Bayes bound and in the context of classification.

Corollary 2. *Given a data distribution $\mathcal{D}$, a parameter set Θ, a prior distribution π over Θ, a $\delta \in (0,1]$, if ℓ_{nll} lies in $[a,b]$, we have, with probability at least $1-\delta$ over the choice of $(X,Y) \sim \mathcal{D}^n$,*

$$\mathop{\mathbf{E}}_{\theta\sim\hat{\rho}^*} \mathcal{L}^{\ell_{\text{nll}}}_{\mathcal{D}}(\theta) \;\leq\; a + \tfrac{b-a}{1-e^{a-b}} \left[1 - e^a \sqrt[n]{Z_{X,Y}\,\delta}\right], \tag{11}$$

where $\hat{\rho}^*$ is the Gibbs optimal posterior (Eq. 8) and $Z_{X,Y}$ is the marginal likelihood (Eq. 9).

In Section 5, we exploit the link between PAC-Bayesian bounds and Bayesian marginal likelihood to expose similarities between both frameworks in the context of model selection. Beforehand, next Section 4 extends the PAC-Bayesian generalization guarantees to unbounded loss function. This is mandatory to make our study fully valid, as the negative log-likelihood loss function is in general unbounded (as well as other common regression losses).

4 PAC-Bayesian Bounds for Regression

This section aims to extend the PAC-Bayesian results of Section 3 to real valued unbounded loss. These result are used in forthcoming sections to study ℓ_{nll}, but they are valid for broader classes of loss functions. Importantly, our new results are focused on regression problems, as opposed to the usual PAC-Bayesian classification framework.

The new bounds are obtained through a recent theorem of Alquier et al. [2015], stated below (we provide a proof in Appendix A.1 for completeness).

Theorem 3 (Alquier et al. [2015])**.** *Given a distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, a hypothesis set $\mathcal{F}$, a loss function $\ell : \mathcal{F}\times\mathcal{X}\times\mathcal{Y} \to \mathbb{R}$, a prior distribution π over $\mathcal{F}$, a $\delta \in (0,1]$, and a real number $\lambda > 0$, with probability at least $1-\delta$ over the choice of $(X,Y) \sim \mathcal{D}^n$, we have*

$$\forall \hat{\rho} \text{ on } \mathcal{F}\text{:} \qquad \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \mathcal{L}^{\ell}_{\mathcal{D}}(f) \;\leq\; \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \widehat{\mathcal{L}}^{\ell}_{X,Y}(f) + \frac{1}{\lambda}\left[\mathrm{KL}(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \Psi_{\ell,\pi,\mathcal{D}}(\lambda,n)\right], \tag{12}$$

$$\text{where} \quad \Psi_{\ell,\pi,\mathcal{D}}(\lambda,n) \;=\; \ln \mathop{\mathbf{E}}_{f\sim\pi} \mathop{\mathbf{E}}_{X',Y'\sim\mathcal{D}^n} \exp\left[\lambda\left(\mathcal{L}^{\ell}_{\mathcal{D}}(f) - \widehat{\mathcal{L}}^{\ell}_{X',Y'}(f)\right)\right]. \tag{13}$$

Alquier et al. [2015] used Theorem 3 to design a learning algorithm for $\{0,1\}$-valued classification losses. Indeed, a bounded loss function $\ell : \mathcal{F}\times\mathcal{X}\times\mathcal{Y} \to [a,b]$ can be used along with Theorem 3 by applying the Hoeffding's lemma to Equation (13), that gives $\Psi_{\ell,\pi,\mathcal{D}}(\lambda,n) \leq \lambda^2(b-a)^2/(2n)$. More specifically, with $\lambda \coloneqq n$, we obtain the following bound

$$\forall \hat{\rho} \text{ on } \mathcal{F}\text{:} \qquad \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \mathcal{L}^{\ell}_{\mathcal{D}}(f) \;\leq\; \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \widehat{\mathcal{L}}^{\ell}_{X,Y}(f) + \tfrac{1}{n}\big[\mathrm{KL}(\hat{\rho}\|\pi) + \ln\tfrac{1}{\delta}\big] + \tfrac{1}{2}(b-a)^2. \tag{14}$$

Note that the latter bound leads to the same trade-off as Theorem 1 (expressed by Equation 3). However, the choice $\lambda \coloneqq n$ has the inconvenience that the bound value is at least $\frac{1}{2}(b-a)^2$, even at the limit $n \to \infty$. Note that another choice that makes the bound converge is $\lambda \coloneqq \sqrt{n}$:

$$\forall \hat{\rho} \text{ on } \mathcal{F}\text{:} \qquad \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \mathcal{L}^{\ell}_{\mathcal{D}}(f) \;\leq\; \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \widehat{\mathcal{L}}^{\ell}_{X,Y}(f) + \tfrac{1}{\sqrt{n}}\big[\mathrm{KL}(\hat{\rho}\|\pi) + \ln\tfrac{1}{\delta} + \tfrac{1}{2}(b-a)^2\big]. \tag{15}$$

A similar result to Equation (15) leads to *long-life learning* algorithms in Pentina and Lampert [2014].

Sub-Gaussian losses. In a regression context, it may be restrictive to consider strictly bounded loss functions. Therefore, we extend Theorem 3 to *sub-Gaussian* loss functions. We say that an unbounded loss function ℓ is sub-Gaussian with a variance factor s^2 under a prior π and a data-distribution $\mathcal{D}$ if it can be described by a sub-Gaussian random variable V, *i.e.*, its moment generating function is upper bounded by the one of a normal distribution of variance s^2 [see Boucheron et al., 2013, Section 2.3]:

$$\psi_{_V}(\lambda) \;=\; \ln \mathbf{E} \exp\Big[\lambda(V - \mathbf{E}\,V)\Big] \;\leq\; \tfrac{\lambda^2 s^2}{2}, \qquad \forall \lambda \in \mathbb{R}, \tag{16}$$

The above sub-Gaussian assumption corresponds to the *Hoeffding assumption* of Alquier et al. [2015], and allows to obtain the following result.

Corollary 4. *Given $\mathcal{D}$, $\mathcal{F}$, ℓ, π and δ defined in Theorem 3 statement, if the loss is sub-Gaussian with variance factor s^2, we have, with probability at least $1-\delta$ over the choice of $(X,Y) \sim \mathcal{D}^n$,*

$$\forall \hat{\rho} \text{ on } \mathcal{F}\text{:} \qquad \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \mathcal{L}^{\ell}_{\mathcal{D}}(f) \;\leq\; \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \widehat{\mathcal{L}}^{\ell}_{X,Y}(f) + \tfrac{1}{n}\big[\mathrm{KL}(\hat{\rho}\|\pi) + \ln\tfrac{1}{\delta}\big] + \tfrac{1}{2}\,s^2.$$

Proof. For $i = 1 \dots n$, we denote ℓ_i a *i.i.d.* realization of the random variable $\ell(f, x, y) - \widehat{\mathcal{L}}^{\ell}_{X',Y'}(f)$.

$\Psi_{\ell,\pi,\mathcal{D}}(\lambda, n) = \ln \mathbf{E} \exp\left[\frac{\lambda}{n} \sum_{i=1}^{n} \ell_i\right] = \ln \prod_{i=1}^{n} \mathbf{E} \exp\left[\frac{\lambda}{n} \ell_i\right] = \sum_{i=1}^{n} \psi_{\ell_i}(\frac{\lambda}{n}) \;\leq\; n \frac{\lambda^2 s^2}{2n^2} \;=\; \frac{\lambda^2 s^2}{2n}$,

where the inequality comes from the sub-Gaussian loss assumption (Equation 16). The result is then obtained from Theorem 3, with $\lambda := n$. □

Sub-Gamma losses. We say that an unbounded loss function ℓ is sub-Gamma with a variance factor s^2 and scale parameter c, under a prior π and a data-distribution $\mathcal{D}$, if it can be described by a re-centered sub-Gamma random variable $V - \mathbf{E}\, V$ [see Boucheron et al., 2013, Section 2.4], that is

$$\psi_V(\lambda) \;\leq\; \frac{s^2}{c^2}(-\ln(1-\lambda c) - \lambda c) \;\leq\; \frac{\lambda^2 s^2}{2(1-c\lambda)}\,, \qquad \forall \lambda \in (0, \tfrac{1}{c})\,.$$

Under this sub-Gamma assumption, we obtain the following new result, which is necessary to study the linear regression in next sections.

Corollary 5. *Given $\mathcal{D}$, $\mathcal{F}$, ℓ, π and δ defined in Theorem 3 statement, if the loss is sub-Gamma with variance factor s^2 and scale $c < 1$, we have, with probability at least $1-\delta$ over $(X, Y) \sim \mathcal{D}^n$,*

$$\forall \hat{\rho} \text{ on } \mathcal{F}\colon \qquad \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \mathcal{L}^{\ell}_{\mathcal{D}}(f) \;\leq\; \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \widehat{\mathcal{L}}^{\ell}_{X,Y}(f) + \tfrac{1}{n}\big[\mathrm{KL}(\hat{\rho}\|\pi) + \ln \tfrac{1}{\delta}\big] + \tfrac{1}{2(1-c)}\, s^2\,. \tag{17}$$

As a special case, with $\ell := \ell_{\text{nll}}$ and $\hat{\rho} := \hat{\rho}^*$ (Eq. 8), we have $\mathop{\mathbf{E}}_{\theta\sim\hat{\rho}^*} \mathcal{L}^{\ell_{\text{nll}}}_{\mathcal{D}}(\theta) \leq \frac{s^2}{2(1-c)} - \frac{1}{n} \ln\left(Z_{X,Y}\, \delta\right)$.

Proof. Following the same path as in Corollary 4 proof (with $\lambda := n$), we have

$$\Psi_{\ell,\pi,\mathcal{D}}(n, n) = \ln \mathbf{E} \exp\left[\sum_{i=1}^{n} \ell_i\right] = \ln \prod_{i=1}^{n} \mathbf{E} \exp\left[\ell_i\right] = \sum_{i=1}^{n} \psi_{\ell_i}(1) \;\leq\; n \frac{s^2}{2(1-c)} \;=\; \frac{n\, s^2}{2(1-c)}\,,$$

where the inequality comes from the sub-Gamma loss assumption, with $1 \in (0, \frac{1}{c})$. □

Squared loss. The parameters s^2 and c of Corollary 5 rely on the chosen loss function and prior, and the assumptions concerning the data distribution. As an example, consider a regression problem where $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$, a family of linear predictors $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$, with $\mathbf{w} \in \mathbb{R}^d$, and a Gaussian prior $\pi \sim \mathcal{N}(\mathbf{0}, \sigma_\pi^2)$. Let assume that the input examples lie inside a ball of radius γ and the label of $\mathbf{x}$ is given by $y = \mathbf{w}^* \cdot \mathbf{x} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is a Gaussian noise. Under the squared loss function $\ell_{\text{sqr}}(f_{\mathbf{w}}, \mathbf{x}, y) \;=\; (\mathbf{w} \cdot \mathbf{x} - y)^2$, we show in Appendix A.3 that Corollary 5 is valid with $s^2 \geq 2\|\mathbf{w}^*\|^2\gamma^2$ and $c \leq 2(\sigma_\epsilon^2 + \sigma_\pi^2\gamma^2)^2$. The latter term tells us that the bound degrades when the noise increases, as expected. Empirical values of the bound under the squared loss are computed in Section 6.

Regression versus classification. The classical PAC-Bayesian theorems are stated in a classification context and bound the generalization error/loss of the stochastic *Gibbs predictor* $G_{\hat{\rho}}$. In order to predict the label of an example $x \in \mathcal{X}$, the Gibbs predictor first draws a hypothesis $h \in \mathcal{F}$ according to $\hat{\rho}$, and then returns $h(x)$. Maurer [2004] shows that we can generalize PAC-Bayesian bounds on the generalization risk of the Gibbs classifier to any loss function with output between zero and one. Provided that $y \in \{-1, 1\}$ and $h(x) \in [-1, 1]$, a common choice is to use the linear loss function $\ell'_{01}(h, x, y) = \frac{1}{2} - \frac{1}{2} y\, h(x)$. The Gibbs generalization loss is then given by $R_{\mathcal{D}}(G_{\hat{\rho}}) = \mathbf{E}_{(x,y)\sim\mathcal{D}} \mathbf{E}_{h\sim\hat{\rho}}\, \ell'_{01}(h, x, y)$. Many PAC-Bayesian works use $R_{\mathcal{D}}(G_{\hat{\rho}})$ as a surrogate loss to study the zero-one classification loss of the majority vote classifier $R_{\mathcal{D}}(B_{\hat{\rho}})$:

$$R_{\mathcal{D}}(B_{\hat{\rho}}) \;=\; \Pr_{(x,y)\sim\mathcal{D}} \Big(y \mathop{\mathbf{E}}_{h\sim\hat{\rho}} h(x) < 0\Big) \;=\; \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}} I\Big[y \mathop{\mathbf{E}}_{h\sim\hat{\rho}} h(x) < 0\Big], \tag{18}$$

where $I[\cdot]$ being the indicator function. Given a distribution $\hat{\rho}$, an upper bound on the Gibbs risk is converted on an upper bound on the majority vote risk by $R_{\mathcal{D}}(B_{\hat{\rho}}) \leq 2R_{\mathcal{D}}(G_{\hat{\rho}})$ [Langford and Shawe-Taylor, 2002]. In some situations, this *factor of two* may be reached, *i.e.*, $R_{\mathcal{D}}(B_{\hat{\rho}}) \simeq 2R_{\mathcal{D}}(G_{\hat{\rho}})$. In other situations, we may have $R_{\mathcal{D}}(G_{\hat{\rho}}) = 0$ even if $R_{\mathcal{D}}(G_{\hat{\rho}}) = \frac{1}{2} - \epsilon$ (see Germain et al. [2015] for an extensive study). Indeed, these bounds obtained via the Gibbs risk are exposed to be loose and/or unrepresentative of the majority vote generalization error.[6]

[6] It is noteworthy that the best PAC-Bayesian empirical bound values are so far obtained by considering a majority vote of linear classifiers, where prior and posterior are Gaussian [Langford and Shawe-Taylor, 2002, Ambroladze et al., 2006, Germain et al., 2009], similarly to the Bayesian linear regression analyzed in Section 6.

In the current work, we study regression losses instead of classification ones. That is, the provided results express upper bounds on $\mathbf{E}_{f\sim\hat{\rho}}\,\mathcal{L}^{\ell}_{\mathcal{D}}(f)$ for any (bounded, sub-Gaussian, or sub-Gamma) losses. Of course, one may want to bound the regression loss of the averaged regressor $F_{\hat{\rho}}(x) = \mathbf{E}_{f\sim\hat{\rho}}\,f(x)$. In this case, if the loss function ℓ is convex (as the squared loss), Jensen's inequality gives $\mathcal{L}^{\ell}_{\mathcal{D}}(F_{\hat{\rho}}) \leq \mathbf{E}_{f\sim\hat{\rho}}\,\mathcal{L}^{\ell}_{\mathcal{D}}(f)$. Note that a strict inequality replaces the factor two mentioned above for the classification case, due to the non-convex indicator function of Equation (18).

Now that we state generalization bounds for real-valued loss functions, we can continue our study linking PAC-Bayesian results to Bayesian inference. In next section, we focus on model selection.

5 Analysis of Model Selection

Let consider L distinct models $\{\mathcal{M}_i\}_{i=1}^{L}$, each one defined by a set of parameters Θ_i. The PAC-Bayesian theorems naturally suggest selecting the model that is best adapted for the given task by evaluating the bound for each model $\{\mathcal{M}_i\}_{i=1}^{L}$ and selecting the one with the lowest bound [McAllester, 2003, Ambroladze et al., 2006, Zhang, 2006]. This is closely linked with the Bayesian model selection procedure, as we showed in Section 3 that minimizing the PAC-Bayes bound amounts to maximizing the marginal likelihood. Indeed, given a collection of L optimal Gibbs posteriors—one for each mode—given by Equation (8),

$$p(\theta|X,Y,\mathcal{M}_i) \;\equiv\; \hat{\rho}^*_i(\theta) \;=\; \tfrac{1}{Z_{X,Y,i}}\pi_i(\theta)\,e^{n\,\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta)},\;\; \text{for } \theta\in\Theta_i\,, \tag{19}$$

the Bayesian Occam's razor criteria [Jeffreys and Berger, 1992, MacKay, 1992] chooses the one with the higher *model evidence*

$$p(Y|X,\mathcal{M}_i) \;\equiv\; Z_{X,Y,i} \;=\; \int_{\Theta_i}\pi_i(\theta)\,e^{-n\widehat{\mathcal{L}}^{\ell}_{X,Y}(\theta)}\,d\theta\,. \tag{20}$$

Corollary 6 below formally links the PAC-Bayesian and the Bayesian model selection. To obtain this result, we simply use the bound of Corollary 5 L times, together with ℓ_{nll} and Equation (10). From the union bound (*a.k.a.* Bonferroni inequality), it is mandatory to compute each bound with a confidence parameter of δ/L, to ensure that the final conclusion is valid with probability at least $1-\delta$.

Corollary 6. *Given a data distribution $\mathcal{D}$, a family of model parameters $\{\Theta_i\}_{i=1}^{L}$ and associated priors $\{\pi_i\}_{i=1}^{L}$—where π_i is defined over Θ_i—, a $\delta\in(0,1]$, if the loss is sub-Gamma with parameters s^2 and $c<1$, then, with probability at least $1-\delta$ over $(X,Y)\sim\mathcal{D}^n$,*

$$\forall i\in\{1,\ldots,L\}\;:\qquad \mathop{\mathbf{E}}_{\theta\sim\hat{\rho}^*_i}\mathcal{L}^{\ell_{\text{nll}}}_{\mathcal{D}}(\theta) \;\leq\; \tfrac{1}{2(1-c)}\,s^2 - \tfrac{1}{n}\ln\left(Z_{X,Y,i}\,\tfrac{\delta}{L}\right)\,.$$

where $\hat{\rho}^*_i$ is the Gibbs optimal posterior (Eq. 19) and $Z_{X,Y,i}$ is the marginal likelihood (Eq. 20).

Hence, under the uniform prior over the L models, choosing the one with the best model evidence is equivalent to choosing the one with the lowest PAC-Bayesian bound.

Hierarchical Bayes. To perform proper inference of hyperparameters, we have to rely on the *Hierarchical Bayes* approach. This is done by considering an *hyperprior* $p(\eta)$ over the set of hyperparameters H. Then, the prior $p(\theta|\eta)$ can be conditioned on a choice of hyperparameter η. The Bayesian rule of Equation (5) becomes $p(\theta,\eta|X,Y) = \frac{p(\eta)\,p(\theta|\eta)\,p(Y|X,\theta)}{p(Y|X)}$.

Under the negative log-likelihood loss function, we can rewrite the results of Corollary 2 as a generalization bounds on $\mathbf{E}_{\eta\sim\hat{\rho}_0}\,\mathbf{E}_{\theta\sim\hat{\rho}^*_\eta}\,\mathcal{L}^{\ell_{\text{nll}}}_{\mathcal{D}}(\theta)$, where $\hat{\rho}_0(\eta)\propto\pi_0(\eta)\,Z_{X,Y,\eta}$ is the hyperposterior on H and π_0 the hyperprior. Indeed, Equation (11) becomes

$$\mathop{\mathbf{E}}_{\theta\sim\hat{\rho}^*}\mathcal{L}^{\ell_{\text{nll}}}_{\mathcal{D}}(\theta) = \mathop{\mathbf{E}}_{\eta\sim\hat{\rho}^*_0}\mathop{\mathbf{E}}_{\theta\sim\hat{\rho}^*_\eta}\mathcal{L}^{\ell_{\text{nll}}}_{\mathcal{D}}(\theta) \;\leq\; \tfrac{1}{2(1-c)}\,s^2 - \tfrac{1}{n}\ln\left(\mathop{\mathbf{E}}_{\eta\sim\pi_0}Z_{X,Y,\eta}\,\delta\right). \tag{21}$$

To relate to the bound obtained in Corollary 6, we consider the case of a discrete hyperparameter set, $\mathrm{H}=\{\eta_i\}_{i=1}^{L}$, with a uniform prior. Then, Equation (21) becomes

$$\mathop{\mathbf{E}}_{\theta\sim\hat{\rho}^*}\mathcal{L}^{\ell_{\text{nll}}}_{\mathcal{D}}(\theta) = \mathop{\mathbf{E}}_{\eta\sim\hat{\rho}^*_0}\mathop{\mathbf{E}}_{\theta\sim\hat{\rho}^*_\eta}\mathcal{L}^{\ell_{\text{nll}}}_{\mathcal{D}}(\theta) \;\leq\; \tfrac{1}{2(1-c)}\,s^2 - \tfrac{1}{n}\ln\left(\textstyle\sum_{i=1}^{L}Z_{X,Y,\eta_i}\,\tfrac{\delta}{L}\right).$$

This bound is now function of $\sum_{i=1}^{L} Z_{X,Y,\eta_i}$ instead $\max_i Z_{X,Y,\eta_i}$ as in Corollary 6. This yields a tighter bound, corroborating the Bayesian wisdom that model averaging performs best.

When selecting a single hyperparameter η^*, the hierarchical representation is equivalent to choosing a deterministic hyperposterior, satisfying $\hat{\rho}_0(\eta^*) = 1$ and 0 for every other values. We then have

$$\mathrm{KL}(\hat{\rho}||\pi) \;=\; \mathrm{KL}(\hat{\rho}_0||\pi_0) + \mathop{\mathbf{E}}_{\eta\sim\hat{\rho}_0} \mathrm{KL}(\hat{\rho}_\eta||\pi_\eta) \;=\; \ln(L) + \mathrm{KL}(\hat{\rho}_{\eta^*}||\pi_{\eta^*})\,.$$

With the optimal posterior for the selected η^*, we have

$$\begin{aligned} n \mathop{\mathbf{E}}_{\theta\sim\hat{\rho}} \widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta) + \mathrm{KL}(\hat{\rho}||\pi) &= n \mathop{\mathbf{E}}_{\theta\sim\hat{\rho}^*_\eta} \widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\theta) + \mathrm{KL}(\hat{\rho}^*_{\eta^*}||\pi_{\eta^*}) + \ln(L) \\ &= -\ln(Z_{X,Y,\eta^*}) + \ln(L) = -\ln\left(\tfrac{Z_{X,Y,\eta^*}}{L}\right). \end{aligned}$$

Inserting this result into Equation (17), we fall back on the bound obtained in Corollary 6. Hence, by comparing the values of the bounds one can get an estimate on the consequence of performing model selection instead of model averaging.

6 Linear Regression

In this section, we perform *Bayesian linear regression* using the parameterization of Bishop [2006]. The output space is $\mathcal{Y} \coloneqq \mathbb{R}$ and, for an arbitrary input space $\mathcal{X}$, we use a mapping function $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^d$.

The model. Given $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ and model parameters $\theta \coloneqq \langle \mathbf{w}, \sigma \rangle \in \mathbb{R}^d \times \mathbb{R}^+$, we consider the likelihood $p(y|x, \langle \mathbf{w}, \sigma \rangle) = \mathcal{N}(y|\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}), \sigma^2)$. Thus, the negative log-likelihood loss is

$$\ell_{\text{nll}}(\langle\, \mathbf{w}, \sigma \,\rangle, x, y) \;=\; -\ln p(y|x, \langle\, \mathbf{w}, \sigma \,\rangle) \;=\; \tfrac{1}{2}\ln(2\pi\sigma^2) + \tfrac{1}{2\sigma^2}(y - \mathbf{w}\cdot\boldsymbol{\phi}(x))^2 \qquad (22)$$

For a fixed σ, minimizing Equation (22) is equivalent to minimizing the square-loss function $\ell_{\text{sqr}}(\mathbf{w}, x, y) \;=\; (y - \mathbf{w}\cdot\boldsymbol{\phi}(x))^2$. We also consider an isotropic Gaussian prior of mean $\mathbf{0}$ and variance σ_π^2: $p(\mathbf{w}|\sigma_\pi) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \sigma_\pi^2)$. For the sake of simplicity, we consider a fixed noise parameter σ^2 and a fixed prior variance σ_π^2. The Gibbs optimal posterior (see Equation 8) is then given by

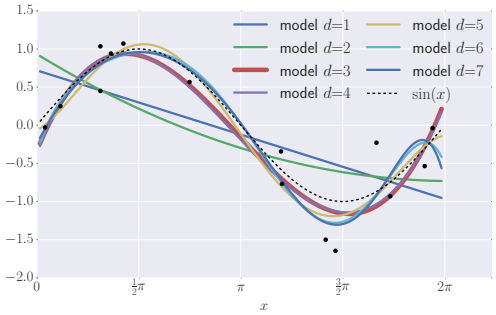
$$\hat{\rho}^*(\mathbf{w}) \equiv p(\mathbf{w}|\sigma, \sigma_\pi) \;=\; \tfrac{p(\mathbf{w}|\sigma,\sigma_\pi)\, p(Y|X,\mathbf{w},\sigma,\sigma_\pi)}{p(Y|X,\sigma,\sigma_\pi)} \;=\; \mathcal{N}(\mathbf{w}\,|\,\widehat{\mathbf{w}}, A^{-1})\,, \qquad (23)$$

where $A \coloneqq \frac{1}{\sigma^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \frac{1}{\sigma_\pi^2}\mathbf{I}$; $\widehat{\mathbf{w}} \coloneqq \frac{1}{\sigma^2}A^{-1}\boldsymbol{\Phi}^T\mathbf{y}$; $\boldsymbol{\Phi}$ is a $n\times d$ matrix such that the i^{th} line is $\boldsymbol{\phi}(x_i)$; $\mathbf{y} \coloneqq [y_1, \ldots y_n]$ is the labels-vector ; and the negative log marginal likelihood is

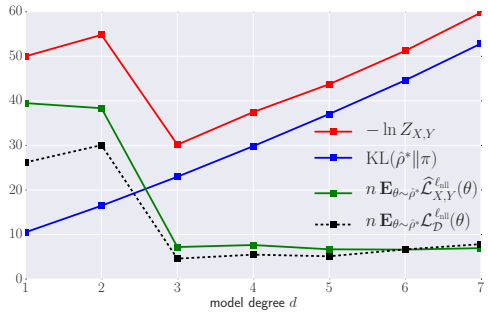
$$\begin{aligned} -\ln\Big(Z_D(\sigma,\sigma_\pi)\Big) &= \tfrac{1}{2\sigma^2}\|\mathbf{y} - \boldsymbol{\Phi}\widehat{\mathbf{w}}\|^2 + \tfrac{n}{2}\ln(2\pi\sigma^2) + \tfrac{1}{2\sigma_\pi^2}\|\widehat{\mathbf{w}}\|^2 + \tfrac{1}{2}\log|A| + d\ln\sigma_\pi \\ &= \underbrace{n\,\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\widehat{\mathbf{w}}) + \tfrac{1}{2\sigma^2}\,\mathrm{tr}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}A^{-1})}_{n\,\mathbf{E}_{\mathbf{w}\sim\hat{\rho}^*}\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\mathbf{w})} + \underbrace{\tfrac{1}{2\sigma_\pi^2}\,\mathrm{tr}(A^{-1}) - \tfrac{d}{2} + \tfrac{1}{2\sigma_\pi^2}\|\widehat{\mathbf{w}}\|^2 + \tfrac{1}{2}\log|A| + d\ln\sigma_\pi}_{\mathrm{KL}\left(\mathcal{N}(\widehat{\mathbf{w}},A^{-1})\,\|\,\mathcal{N}(\mathbf{0},\sigma_\pi^2\mathbf{I})\right)}\,. \end{aligned}$$

Last equality comes from $\frac{1}{2\sigma^2}\,\mathrm{tr}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}A^{-1}) + \frac{1}{\sigma_\pi^2}\,\mathrm{tr}(A^{-1}) = \mathrm{tr}(\frac{1}{2\sigma^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi}A^{-1} + \frac{1}{\sigma_\pi^2}A^{-1}) = \mathrm{tr}(A^{-1}A) = d$ (see Appendix A.4 for complete calculations). This exhibits how the Bayesian regression optimization problem can be express by the minimization of a PAC-Bayesian bound, expressed by a trade-off between $\mathbf{E}_{\mathbf{w}\sim\hat{\rho}^*}\,\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\mathbf{w})$ and $\mathrm{KL}\big(\mathcal{N}(\widehat{\mathbf{w}}, A^{-1})\,\|\,\mathcal{N}(\mathbf{0}, \sigma_\pi^2\,\mathbf{I})\big)$.

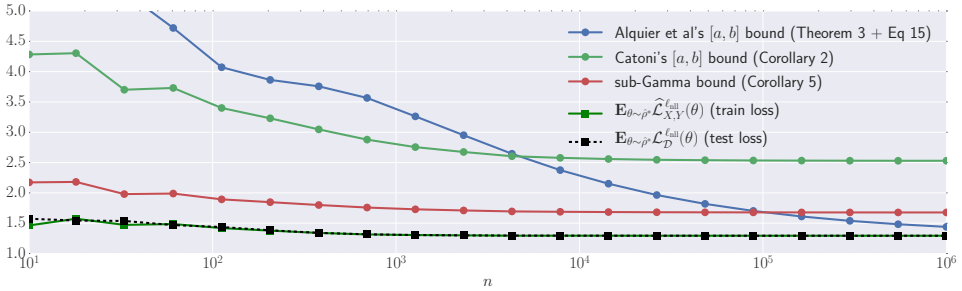
Model selection experiment. To produce Figures 1a and 1b, we reimplemented the toy experiment of Bishop [2006, Subsection 3.5.1]. That is, we generated a learning sample of 15 data points according to $y = \sin(x) + \epsilon$, where x is uniformly sampled in the interval $[0, 2\pi]$ and $\epsilon \sim \mathcal{N}(0, \frac{1}{4})$ is a Gaussian noise. We then learn seven different polynomial models with the regression given by Equation (23). More precisely, for a polynomial model of degree d, we map input $x \in \mathbb{R}$ to a vector $\boldsymbol{\phi}(x) = [1, x^1, x^2, \ldots, x^d] \in \mathbb{R}^{d+1}$, and we fix parameters $\sigma_\pi^2 = \frac{1}{0.005}$ and $\sigma^2 = \frac{1}{2}$. Figure 1a illustrates the seven learned models. Figure 1b shows the marginal likelihood computed for each polynomial model, and is designed to reproduce Bishop [2006, Figure 3.14], where it is explained that the marginal likelihood correctly indicates that the polynomial model of degree $d = 3$ is "the simplest model which gives a good explanation for the observed data". We show that this claim is well quantified by the trade-off intrinsic to our PAC-Bayesian approach: the complexity KL term keeps increasing with the parameter $d \in \{1, 2, \ldots, 7\}$, while the empirical risk drastically decreases from $d = 2$ to $d = 3$, and only slightly afterward. Moreover, we show the generalization risk (computed on a test sample of size 1000) tends to increase with complex models (for $d \geq 4$).



(a) Predicted models. Black dots are the 15 training samples.



(b) Decomposition of the marginal likelihood into the empirical loss and KL-divergence.



(c) Bound values on a synthetic dataset according to the number of training samples.

Figure 1: Model selection experiment (a-b); and comparison of bounds values (c).

Empirical comparison of bound values. Figure 1c compares the values of the PAC-Bayesian bounds presented in this paper on a synthetic dataset where the inputs points are randomly generated to a Gaussian $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$ in $\mathbb{R}^{20}$, and the outputs are given by $y = \mathbf{w}^* \cdot \mathbf{x} + \epsilon$, with $\|\mathbf{w}^*\|{=}1$ and $\epsilon \sim \mathcal{N}(0, \frac{1}{3})$. We perform Bayesian linear regression in the input space, *i.e.*, $\boldsymbol{\phi}(\mathbf{x}){=}\mathbf{x}$, fixing $\sigma_\pi^2{=}\frac{1}{100}$ and $\sigma^2{=}2$. That is, we compute the optimal posterior of Equation (23) for training samples of sizes from 10 to 10^6. For each learned model, we compute empirical negative loss likelihood of Equation (22), and the three PAC-Bayes bounds, with confidence parameter of $\delta{=}\frac{1}{20}$. We estimate the bounds parameters a, b, s, c from observed samples.

For small and medium sized training samples ($n \lesssim 10^4$), the bound of Corollary 5, that we have developed for (unbounded) sub-Gamma losses, gives as far the tighter guarantees than the two other results for $[a, b]$-bounded losses. However, our new bound always maintains a gap of $s^2/2(1{-}c)$ between its value and the expected loss. The result of Corollary 2 [adapted from Catoni, 2007] from bounded losses suffer from a similar gap, while having higher values that our sub-Gaussian result. Finally, the result of Theorem 3 [Alquier et al., 2015], combined with $\lambda = 1/\sqrt{n}$ (Eq 15), converges to the expected loss, but it provides good guarantees only for large training sample ($n \gtrsim 10^5$). Note that the latter bound is not directly minimized by our "optimal posterior", as opposed to the one with $\lambda = 1/n$ (Eq 14), for which we observe values in $[19.2, 19.8]$ (not displayed on Figure 1c).

7 Conclusion

The first contribution of this paper is to bridge the concepts underlying the Bayesian and the PAC-Bayesian approaches. This was done by showing that, under proper parametrization, the minimization of the PAC-Bayesian bound minimizes the marginal likelihood. This study, that relies on the real-valued negative log-likelihood loss function, motivates the second contribution of this paper, which is to prove PAC-Bayesian generalization bounds for regression with unbounded sub-Gamma loss functions, that provides generalization guarantees for the squared loss in regression tasks.

In this work, we studied model selection techniques. On a broader perspective, we would like to suggest both Bayesian and PAC-Bayesian frameworks may have more to learn from each other than what has been done lately. As future work, we plan to study other Bayesian techniques through the light of PAC-Bayesian tools, such as *variational Bayes* and *empirical Bayes* methods.

References

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *ArXiv e-prints*, 2015. URL `http://arxiv.org/abs/1506.04091`.

A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In *NIPS*, 2006.

Arindam Banerjee. On Bayesian bounds. In *ICML*, pages 81–88, 2006.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian theory for transductive learning. In *AISTATS*, pages 105–113, 2014.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities : a nonasymptotic theory of independence*. Oxford university press, 2013. ISBN 978-0-19-953525-5.

Olivier Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*, volume 56. Inst. of Mathematical Statistic, 2007.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *ICML*, pages 353–360, 2009.

Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *JMLR*, 16:787–860, 2015.

Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521:452–459, 2015.

Peter Grünwald and John Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66(2-3):119–149, 2007.

Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin C. Cawley. Model selection: Beyond the Bayesian/frequentist divide. *JMLR*, 11:61–87, 2010.

Tamir Hazan, Subhransu Maji, Joseph Keshet, and Tommi S. Jaakkola. Learning efficient random maximum a-posteriori predictors with non-decomposable loss functions. In *NIPS*, pages 1887–1895, 2013.

William H. Jeffreys and James O. Berger. Ockham's razor and Bayesian analysis. *American Scientist*, 1992.

Alexandre Lacoste. *Agnostic Bayes*. PhD thesis, Université Laval, 2015.

Simon Lacoste-Julien, Ferenc Huszar, and Zoubin Ghahramani. Approximate inference for the loss-calibrated Bayesian. In *AISTATS*, pages 416–424, 2011.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *NIPS*, pages 423–430, 2002.

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theor. Comput. Sci.*, 473:4–28, 2013.

David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

Andreas Maurer. A note on the PAC-Bayesian theorem. *CoRR*, cs.LG/0411099, 2004.

David McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

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

David A. McAllester and Joseph Keshet. Generalization bounds and consistency for latent structural probit and ramp loss. In *NIPS*, pages 2205–2212, 2011.

Ron Meir and Tong Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *NIPS*, pages 841–848. MIT Press, 2001.

Asf Noy and Koby Crammer. Robust forward algorithms via PAC-Bayes and Laplace distributions. In *AISTATS*, 2014.

Anastasia Pentina and Christoph H. Lampert. A PAC-Bayesian bound for lifelong learning. In *ICML*, 2014.

Matthias Seeger. PAC-Bayesian generalization bounds for gaussian processes. *JMLR*, 3:233–269, 2002.

Matthias Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, 2003.

Yevgeny Seldin and Naftali Tishby. PAC-Bayesian analysis of co-clustering and beyond. *JMLR*, 11, 2010.

Yevgeny Seldin, Peter Auer, François Laviolette, John Shawe-Taylor, and Ronald Ortner. PAC-Bayesian analysis of contextual bandits. In *NIPS*, pages 1683–1691, 2011.

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. In *UAI*, 2012.

John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayesian estimator. In *COLT*, 1997.

Ilya O. Tolstikhin and Yevgeny Seldin. PAC-Bayes-empirical-Bernstein inequality. In *NIPS*, 2013.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Information Theory*, 52(4):1307–1321, 2006.

A Supplementary material

A.1 Proof of Theorem 3

Proof. From Donsker-Varadhan's change of measure, with $\phi(f) \coloneqq \lambda\left(\mathcal{L}^{\ell}_{\mathcal{D}}(f) - \widehat{\mathcal{L}}^{\ell}_{X,Y}(f)\right)$, we have $\forall\, \hat{\rho}$ on $\mathcal{F}$:

$$\begin{aligned}\lambda\big(\mathop{\mathbf{E}}_{f\sim\hat{\rho}} \mathcal{L}^{\ell}_{\mathcal{D}}(f) - \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \widehat{\mathcal{L}}^{\ell}_{X,Y}(f)\big) &= \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \lambda\left(\mathcal{L}^{\ell}_{\mathcal{D}}(f) - \widehat{\mathcal{L}}^{\ell}_{X,Y}(f)\right) \\ &\leq \mathrm{KL}(\hat{\rho}\|\pi) + \ln\left(\mathop{\mathbf{E}}_{f\sim\pi} e^{\lambda\left(\mathcal{L}^{\ell}_{\mathcal{D}}(f) - \widehat{\mathcal{L}}^{\ell}_{X,Y}(f)\right)}\right).\end{aligned}$$

Now, we apply apply Markov's inequality on the random variable $\mathbf{E}_{f\sim\pi}\, e^{\lambda\left(\mathcal{L}^{\ell}_{\mathcal{D}}(f) - \widehat{\mathcal{L}}^{\ell}_{X,Y}(f)\right)}$:

$$\Pr_{X,Y\sim\mathcal{D}^n}\left(\zeta_\pi(X,Y) \,\leq\, \frac{1}{\delta} \mathop{\mathbf{E}}_{X',Y'\sim\mathcal{D}^n} \zeta_\pi(X',Y')\right) \,\geq\, 1-\delta\,.$$

This implies that with probability at least $1-\delta$ over the choice of $X, Y \sim \mathcal{D}^n$, we have $\forall\, \hat{\rho}$ on $\mathcal{F}$:

$$\mathop{\mathbf{E}}_{f\sim\hat{\rho}} \mathcal{L}^{\ell}_{\mathcal{D}}(f) \,\leq\, \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \widehat{\mathcal{L}}^{\ell}_{X,Y}(f) + \frac{1}{\lambda}\left[\mathrm{KL}(\hat{\rho}\|\pi) + \ln \frac{\mathop{\mathbf{E}}_{X',Y'\sim\mathcal{D}^n} \Psi_{\ell,\pi,\mathcal{D}}(\lambda, n)}{\delta}\right].$$

□

A.2 Proof of Equation (14) and (15)

Proof. For $i = 1 \ldots n$, we denote ℓ_i a realization of the random variable $\ell(f,x,y) - \widehat{\mathcal{L}}^{\ell}_{X',Y'}(f)$. Each ℓ_i is *i.i.d.*, zero-mean, and bounded by $a-b$ and $b-a$, as $\ell(f,x,y) \in [a,b]$. Thus,

$$\begin{aligned}\mathop{\mathbf{E}}_{f\sim\pi} \mathop{\mathbf{E}}_{X',Y'\sim\mathcal{D}^n} \exp\left[\lambda\left(\mathcal{L}^{\ell}_{\mathcal{D}}(f) - \widehat{\mathcal{L}}^{\ell}_{X',Y'}(f)\right)\right] &= \mathbf{E}\exp\left[\frac{\lambda}{n}\sum_{i=1}^{n}\ell_i\right] \\ &= \prod_{i=1}^{n}\mathbf{E}\exp\left[\frac{\lambda}{n}\ell_i\right] \\ &\leq \prod_{i=1}^{n}\exp\left[\frac{\lambda^2(a-b-(b-a))^2}{8n^2}\right] \\ &= \prod_{i=1}^{n}\exp\left[\frac{\lambda^2(b-a)^2}{2n^2}\right] \\ &= \exp\left[\frac{\lambda^2(b-a)^2}{2n}\right],\end{aligned}$$

where the Inequality comes from Hoeffding's lemma.

With $\lambda \coloneqq n$, Equation (12) becomes Equation (14) :

$$\begin{aligned}\mathop{\mathbf{E}}_{f\sim\hat{\rho}} \mathcal{L}^{\ell}_{\mathcal{D}}(f) &\leq \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \widehat{\mathcal{L}}^{\ell}_{X,Y}(f) + \frac{1}{n}\left[\mathrm{KL}(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \frac{n^2(b-a)^2}{2n}\right] \\ &= \mathop{\mathbf{E}}_{f\sim\hat{\rho}} \widehat{\mathcal{L}}^{\ell}_{X,Y}(f) + \frac{1}{n}\left[\mathrm{KL}(\hat{\rho}\|\pi) + \ln\frac{1}{\delta}+\right] + \frac{1}{2}(b-a)^2\,.\end{aligned}$$

Similarly, with $\lambda \coloneqq \sqrt{n}$, Equation (12) becomes Equation (15) . □

A.3 Study of the Squared Loss

Assume that:

- $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_\pi^2\,\mathbf{I})\,,\ (\mathbf{w} \in \mathbb{R}^d)$

- $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$
- $y = \mathbf{w}^* \cdot \mathbf{x} + \epsilon$
- $\forall \mathbf{x} \in \mathcal{X}\ ,\ \gamma \geq \sup \|\mathbf{x}\|$

We have $y|\mathbf{x} \sim \mathcal{N}(\mathbf{x} \cdot \mathbf{w}^*, \sigma_\epsilon^2)$. Thus, $z|\mathbf{x} = (y - \mathbf{w} \cdot \mathbf{x})|\mathbf{x} \sim \mathcal{N}(\mathbf{x} \cdot \mathbf{w}^*, \sigma_\epsilon^2 + \sigma_\pi^2 \|\mathbf{x}\|^2)$

$$
\begin{aligned}
e^{\psi} &= \mathop{\mathbf{E}}_{\mathbf{x}} \mathop{\mathbf{E}}_{y|\mathbf{x}} \mathop{\mathbf{E}}_{\mathbf{w}} \exp(\lambda[(y - \mathbf{w} \cdot \mathbf{x})^2 - \mathop{\mathbf{E}}_{\mathbf{x}} \mathop{\mathbf{E}}_{y|\mathbf{x}} \mathop{\mathbf{E}}_{\mathbf{w}} (y - \mathbf{w} \cdot \mathbf{x})^2]) \\
&= \mathop{\mathbf{E}}_{\mathbf{x}} \mathop{\mathbf{E}}_{z|\mathbf{x}} \exp(\lambda[z^2 - \mathop{\mathbf{E}}_{\mathbf{x}} \mathop{\mathbf{E}}_{z|\mathbf{x}} z^2]) \\
&\leq \mathop{\mathbf{E}}_{\mathbf{x}} \mathop{\mathbf{E}}_{z|\mathbf{x}} \exp(\lambda[z^2]) \\
(\blacklozenge) \qquad &= \mathop{\mathbf{E}}_{\mathbf{x}} \frac{1}{\sqrt{1 - 2\lambda(\sigma_\epsilon^2 + \sigma_\pi^2 \|\mathbf{x}\|^2)^2}} \exp\left(\frac{\lambda(\mathbf{w}^* \cdot \mathbf{x})^2}{1 - 2\lambda(\sigma_\epsilon^2 + \sigma_\pi^2 \|\mathbf{x}\|^2)^2}\right) \\
&\leq \mathop{\mathbf{E}}_{\mathbf{x}} \frac{1}{\sqrt{1 - 2\lambda(\sigma_\epsilon^2 + \sigma_\pi^2 \gamma^2)^2}} \exp\left(\frac{\lambda(\mathbf{w}^* \cdot \mathbf{x})^2}{1 - 2\lambda(\sigma_\epsilon^2 + \sigma_\pi^2 \gamma^2)^2}\right) \\
&\leq \frac{1}{\sqrt{1 - 2\lambda(\sigma_\epsilon^2 + \sigma_\pi^2 \gamma^2)^2}} \exp\left(\frac{\lambda \|\mathbf{w}^*\|^2 \gamma^2}{1 - 2\lambda(\sigma_\epsilon^2 + \sigma_\pi^2 \gamma^2)^2}\right).
\end{aligned}
$$

$$
\begin{aligned}
\psi &\leq \frac{\lambda \|\mathbf{w}^*\|^2 \gamma^2}{1 - 2\lambda(\sigma_\epsilon^2 + \sigma_\pi^2 \gamma^2)^2} - \frac{1}{2} \ln\left(1 - 2\lambda(\sigma_\epsilon^2 + \sigma_\pi^2 \gamma^2)^2\right) \\
&\leq \frac{\lambda \|\mathbf{w}^*\|^2 \gamma^2}{1 - 2\lambda(\sigma_\epsilon^2 + \sigma_\pi^2 \gamma^2)^2} \\
&= \frac{2\lambda \|\mathbf{w}^*\|^2 \gamma^2}{2(1 - 2\lambda(\sigma_\epsilon^2 + \sigma_\pi^2 \gamma^2)^2)} \\
&\leq \frac{2\lambda^2 s^2}{2(1 - c\lambda)}\,,
\end{aligned}
$$

with $s^2 \geq 2\frac{\|\mathbf{w}^*\|^2 \gamma^2}{\lambda}$ and $c \leq 2(\sigma_\epsilon^2 + \sigma_\pi^2 \gamma^2)^2$. Note that the Equality $(\blacklozenge)$ is only valid for $\lambda < \frac{1}{c}$.

A.4 Linear Regression

We defined $A \coloneqq \frac{1}{\sigma^2} \mathbf{\Phi}^T \mathbf{\Phi} + \frac{1}{\sigma_\pi^2} \mathbf{I}$; $\widehat{\mathbf{w}} \coloneqq \frac{1}{\sigma^2} A^{-1} \mathbf{\Phi}^T \mathbf{y}$; $\mathbf{\Phi}$ as a $n \times d$ matrix such that the i^{th} line is $\boldsymbol{\phi}(x_i)$; $\mathbf{y} \coloneqq [y_1, \dots y_n]$ as the labels-vector ; and we decompose of the marginal likelihood into the PAC-Bayesian trade-off:

$$
\begin{aligned}
-\ln\Big(Z_D(\sigma, \sigma_\pi)\Big) &= \tfrac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{\Phi}\widehat{\mathbf{w}}\|^2 + \tfrac{n}{2} \ln(2\pi\sigma^2) + \tfrac{1}{2\sigma_\pi^2} \|\widehat{\mathbf{w}}\|^2 + \tfrac{1}{2} \log|A| + d \ln \sigma_\pi \\
&= \underbrace{n\, \widehat{\mathcal{L}}_{X,Y}^{\ell_{\text{nll}}}(\widehat{\mathbf{w}}) + \tfrac{1}{2\sigma^2} \operatorname{tr}(\mathbf{\Phi}^T \mathbf{\Phi} A^{-1})}_{n\, \mathbf{E}_{\mathbf{w} \sim \hat{\rho}^*} \widehat{\mathcal{L}}_{X,Y}^{\ell_{\text{nll}}}(\mathbf{w})} + \underbrace{\tfrac{1}{2\sigma_\pi^2} \operatorname{tr}(A^{-1}) - \tfrac{d}{2} + \tfrac{1}{2\sigma_\pi^2} \|\widehat{\mathbf{w}}\|^2 + \tfrac{1}{2} \log|A| + d \ln \sigma_\pi}_{\mathrm{KL}\left(\mathcal{N}(\widehat{\mathbf{w}}, A^{-1}) \,\|\, \mathcal{N}(\mathbf{0}, \sigma_\pi^2 \mathbf{I})\right)}\,,
\end{aligned}
$$

which is based on the following three equalities:

$$
\begin{aligned}
n \mathop{\mathbf{E}}_{\mathbf{w} \sim \hat{\rho}} \widehat{\mathcal{L}}_{X,Y}^{\ell_{\text{nll}}}(\mathbf{w}) &= \mathop{\mathbf{E}}_{\mathbf{w} \sim \hat{\rho}} \sum_i -\ln p(y_i | x_i, \mathbf{w}) \\
&= \mathop{\mathbf{E}}_{\mathbf{w} \sim \hat{\rho}} \left(\frac{n}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_i (y_i - \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i))^2\right) \\
&= \frac{n}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \mathop{\mathbf{E}}_{\mathbf{w} \sim \hat{\rho}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|^2
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\,\mathop{\mathbf{E}}_{\mathbf{w}\sim\hat{\rho}}\left(\|\mathbf{y}\|^2 - 2\mathbf{y}\mathbf{\Phi}\mathbf{w} + \mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w}\right)\\
&= \frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left(\|\mathbf{y}\|^2 - 2\mathbf{y}\mathbf{\Phi}\widehat{\mathbf{w}} + \mathop{\mathbf{E}}_{\mathbf{w}\sim\hat{\rho}}\mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w}\right)\\
&= \frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left(\|\mathbf{y}\|^2 - 2\mathbf{y}\mathbf{\Phi}\widehat{\mathbf{w}} + \operatorname{tr}\left(\mathbf{\Phi}^T\mathbf{\Phi}A^{-1}\right) + \widehat{\mathbf{w}}^T\mathbf{\Phi}^T\mathbf{\Phi}\widehat{\mathbf{w}}\right)\\
&= \frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{\Phi}\widehat{\mathbf{w}}\right\|^2 + \frac{1}{2\sigma^2}\operatorname{tr}\left(\mathbf{\Phi}^T\mathbf{\Phi}A^{-1}\right)\\
&= n\,\widehat{\mathcal{L}}^{\ell_{\text{nll}}}_{X,Y}(\widehat{\mathbf{w}}) + \frac{1}{2\sigma^2}\operatorname{tr}\left(\mathbf{\Phi}^T\mathbf{\Phi}A^{-1}\right)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{KL}\big(\mathcal{N}(\widehat{\mathbf{w}}, A^{-1})\,\|\,\mathcal{N}(\mathbf{0}, \sigma_\pi^2\mathbf{I})\big) &= \frac{1}{2}\left(\operatorname{tr}\left((\sigma_\pi^2\mathbf{I})^{-1}A^{-1}\right) + \frac{1}{\sigma_\pi^2}\|\widehat{\mathbf{w}}\|^2 - d + \log\frac{|\sigma_\pi^2\mathbf{I}|}{|A|}\right)\\
&= \frac{1}{2}\left(\frac{1}{\sigma_\pi^2}\operatorname{tr}\left(A^{-1}\right) + \frac{1}{\sigma_\pi^2}\|\widehat{\mathbf{w}}\|^2 - d + \log|A^{-1}| + d\ln\sigma_\pi^2\right)
\end{aligned}
$$

$$
\begin{aligned}
\tfrac{1}{2\sigma^2}\operatorname{tr}\left(\mathbf{\Phi}^T\mathbf{\Phi}A^{-1}\right) + \tfrac{1}{\sigma_\pi^2}\operatorname{tr}\left(A^{-1}\right) &= \operatorname{tr}\left(\tfrac{1}{2\sigma^2}\mathbf{\Phi}^T\mathbf{\Phi}A^{-1} + \tfrac{1}{\sigma_\pi^2}A^{-1}\right)\\
&= \operatorname{tr}\left(A^{-1}(\tfrac{1}{2\sigma^2}\mathbf{\Phi}^T\mathbf{\Phi} + \tfrac{1}{\sigma_\pi^2}\mathbf{I})\right)\\
&= \operatorname{tr}\left(A^{-1}A\right)\\
&= d\,.
\end{aligned}
$$