On the properties of variational approximations of Gibbs posteriors

Myriam Frikha

April 2024

Contents

1 Introduction

The PAC-Bayesian framework, initially developed in the late 1990s, has emerged as a robust methodology for deriving non-asymptotic risk bounds for random estimators. Originating from foundational works by Shawe-Taylor and Williamson (1997) and McAllester (1998), this framework integrates Bayesian principles with statistical learning theory, offering a probabilistic perspective on learning processes. A significant focus within this area has been on the Gibbs posterior, a distribution that optimally balances empirical risk and model complexity but is often computationally intractable for large datasets.

To address computational challenges, especially relevant in the era of big data, this paper explores variational Bayes (VB) approximations of the Gibbs posterior. Introduced by Jordan et al. (1999), VB methods provide a faster, though approximate, alternative to the traditional Markov chain Monte Carlo sampling techniques. Our comprehensive analysis reveals that VB approximations maintain convergence rates comparable to the original PAC-Bayesian procedures under certain conditions.

We will introduce PAC-Bayesian bounds and Oracle-type inequalities for predictions derived from variational approximations of the Gibbs posterior, based on two distinct sets of assumptions. Additionally, we will explore practical applications of these bounds and present an alternative method for managing *hostile* data.

2 Framework

2.1 Notations

We have access to a sample $(X_1, Y_1), .., (X_n, Y_n)$ taking values in $\mathcal{X} \times \mathcal{Y}$, where the pairs (X_i, Y_i) have the same distribution P. Let's note that we don't necessarily assume that they are independent, this criterion might be used in some specific applications. The label $\mathcal{Y}$ is a subset of $\mathbb{R}$. Besides, a set of predictors is chosen as such $\{f_\theta : \mathcal{X} \to \mathbb{R}, \theta \in \Theta\}$. In addition, we have access to a risk $R(\theta)$. We set $\overline{R} = R(\overline{\theta})$ where $\overline{\theta} \in \underset{\Theta}{argmin}\ R$ which makes $f_{\overline{\theta}}$ the best predictor. We also introduce the empirical analog to the risk function r_n and set $\overline{r}_n = r_n(\overline{\theta})$. Lastly, we define a probability measure $\pi(\theta)$ on the set Θ, as well as the set of all probability measures on Θ, $\mathcal{M}_+^1(\Theta)$.

2.2 Gibbs posterior

Definition 1. *We define, for any $\lambda > 0$, the pseudo-posterior $\hat{\rho}_\lambda$ by*

$$\hat{\rho}_\lambda(d\theta) = \frac{\exp(-\lambda r_n(\theta))}{\int \exp(-\lambda r_n) d\pi} \pi(d\theta).$$

The pseudo-posterior, also known as the Gibbs posterior, plays a major role in the PAC-Bayesian approach. This distribution is derived by minimizing the upper bound of a specific oracle inequality that is applied to stochastic estimators. This gives access to predictors. One specific case worth to note is when $[\exp(-\lambda r_n(\theta))]$ is considered as the likelihood of a certain model, $\hat{\rho}_\lambda$ transforms into a Bayesian posterior distribution.

Indeed, given the framework, the likelihood function is defined as follows [2] :

$$\mathcal{L}(\theta; X_1, \ldots, X_n) = \prod_{i=1}^{n} f_\theta(X_i)$$

Which gives access to the posterior distribution

$$\pi(d\theta | X_1, \ldots, X_n) = \frac{\mathcal{L}(\theta; X_1, \ldots, X_n)\pi(d\theta)}{\int \mathcal{L}(\theta'; X_1, \ldots, X_n)\pi(d\theta')}$$

As

$$\left[\prod_{i=1}^{n} f_\theta(X_i)\right] \pi(d\theta) = \exp\left\{\sum_{i=1}^{n} \log f_\theta(X_i)\right\} \pi(d\theta),$$

We can define the loss $l_i(\theta) = -\log f_\theta(X_i)$ and build an empirical risk around it.

The following definition is a form of the theoretical peer of the Gibbs posterior.

Definition 2. *We define, for any $\lambda > 0$, π_λ as*

$$\pi_\lambda(d\theta) = \frac{\exp[-\lambda R(\theta)]}{\int \exp[-\lambda R]\, d\pi} \pi(d\theta).$$

2.3 Numerical approximations of the Gibbs posterior

Usually, the approximation of Gibbs posteriors is done through Markov Chain Monte Carlo sampling techniques. However, their implementation with large sample sizes tend to be too slow and the complexity of the algorithms is prohibitive. A way to tackle this problem is to use fast deterministic approximations such as Variaional Bayes.

We define a family $\mathcal{F} \in \mathcal{M}_{+}^{1}(\Theta)$ of probability distributions that are considered as tractable. This allows us to define the VB-approximation of the Gibbs posterior : $\overset{\sim}{\rho}_{\lambda}$

Definition 3. *Let*

$$\overset{\sim}{\rho}_{\lambda} = \underset{\rho \in \mathcal{F}}{argmin}\, \mathcal{K}(\rho, \hat{\rho}_{\lambda})$$

where $\mathcal{K}(\rho, \hat{\rho}_{\lambda})$ denotes the KL (Küllback-Leibler) divergence of $\hat{\rho}_{\lambda}$ relative to ρ

A crucial point is to find a family $\mathcal{F}$ which is large enough to establish an adequate minimization and adapted to the computation of $\overset{\sim}{\rho}_{\lambda}$. There are two main types of families.

1. Mean field VB: for a certain decomposition $\Theta = \Theta_1 \times \ldots \times \Theta_d, \mathcal{F}$ is the set of product probability measures

$$\mathcal{F}^{\mathrm{MF}} = \left\{ \rho \in \mathcal{M}_{+}^{1}(\Theta) : \rho(\mathrm{d}\theta) = \prod_{i=1}^{d} \rho_i\left(\, \mathrm{d}\theta_i\right), \forall i \in \{1, \ldots, d\}, \rho_i \in \mathcal{M}_{+}^{1}\left(\Theta_i\right) \right\}.$$

2. Parametric family :

$$\mathcal{F}^{\mathrm{P}} = \left\{ \rho \in \mathcal{M}_{+}^{1}(\Theta) : \rho(d\theta) = f(\theta; m) d\theta, m \in M \right\}$$

   and M is finite-dimentional.

We will present PAC-Bayesian bounds on predictions obtained by variational approximations of $\hat{\rho}_{\lambda}$ under two types of assumptions : A Hoeffding-type assumption and a Bernstein-type assumption.

3 Bounds under the Hoeffding and Bernstein assumptions

3.1 Hoeffding and Bernstein assumptions

Definition 4. *We say that a Hoeffding assumption is satisfied for prior π when there is a function f and an interval $I \subseteq \mathbb{R}_{+}^{*}$ such that, for any $\lambda \in I$, for any $\theta \in \Theta$,*

$$\left.\begin{array}{l} \pi\left(\mathbb{E} \exp\left\{\lambda\left[R(\theta) - r_n(\theta)\right]\right\}\right) \\ \pi\left(\mathbb{E} \exp\left\{\lambda\left[r_n(\theta) - R(\theta)\right]\right\}\right) \end{array}\right\} \leq \exp[f(\lambda, n)]$$

This assumption can be linked to the Hoeffding's inequality. Indeed, it can be seen as an integrated version of it, where $f(\lambda, n) \simeq \lambda^2/n$. In addition, it allows to efficiently study cases where the loss is not necessarily upper bounded.

Definition 5. *We say that a Bernstein assumption is satisfied for prior π when there is a function g and an interval $I \subset \mathbb{R}_{+}^{*}$ such that, for any $\lambda \in I$, for any $\theta \in \Theta$,*

$$\left.\begin{array}{l} \pi\left(\mathbb{E} \exp\left\{\lambda[R(\theta) - \bar{R}] - \lambda\left[r_n(\theta) - \bar{r}_n\right]\right\}\right) \\ \pi\left(\mathbb{E} \exp\left\{\lambda\left[r_n(\theta) - \bar{r}_n\right] - \lambda[R(\theta) - \bar{R}]\right\}\right) \end{array}\right\} \leq \pi(\exp[g(\lambda, n)[R(\theta) - \bar{R}]])$$

This assumption is slightly more finessed since it requires knowledge on the optimal risks.

3.2 Empirical bounds

We now present a first empirical bound under previous assumptions.

Theorem 1. *Under the **Hoeffding assumption**, for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$ we have simultaneously for any $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\int R\, d\rho \leq \int r_n\, d\rho + \frac{f(\lambda, n) + K(\rho, \pi) + \log(\frac{1}{\varepsilon})}{\lambda}.$$

This result provides an upper bound on the risk of the Gibbs posterior and its variational approximation. To see this, simply fix the value of ρ to the desired one. The upper bound is an effective way to measure the performance since it can be computed from the data. However, such bound is not sufficient because it doesn't provide the rate of convergence.

Proof. The proof steps are inspired from Catoni (2007) [6], our starting point is definition 1 :

$$\pi\left(\mathbb{E} \exp\left\{\lambda[R(\theta) - r_n(\theta)]\right\}\right) \exp\{f(\lambda, n)\}^{-1} \leq 1$$

Then, applying Fubini's theorem we have :

$$\mathbb{E} \int \exp\left\{\lambda[R(\theta) - r_n(\theta)] - f(\lambda, n)\right\} \pi(d\theta) \leq 1$$

In addition, by definition of the Kullback-Leibler KL, we have that for any $\rho << \pi$:

$$K(\rho, \pi[r_n]) = \lambda \int r_n d\rho + K(\rho, \pi) + \log \int \exp(-h) d\pi \tag{KL}$$

As a consequence, we have :

$$-\log \int \exp(-h) d\pi = \min_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \int h d\rho + K(\rho, \pi) \right\} \tag{1}$$

In particular, we apply (1) with $h(\theta) = \lambda[r_n(\theta) - R(\theta)]$:

$$\int \exp(\lambda[R(\theta) - r_n(\theta)]) d\pi = \exp\left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \int \lambda[R(\theta) - r_n(\theta)] d\rho - K(\rho, \pi) \right\}$$

Which yields :

$$\mathbb{E} \exp \left\{ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \int \lambda[R(\theta) - r_n(\theta)] \rho(d\theta) - K(\rho, \pi) - f(\lambda, n) \right] \leq 1$$

Using $\mathbb{E}[\exp(U)] \geq \mathbb{P}(U > 0)$ for any U, we have:

$$\mathbb{P}\left(\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \left[\int \lambda[R(\theta) - r_n(\theta)] \rho(d\theta) - K(\rho, \pi) - f(\lambda, n) + \log(\epsilon) \right] > 0 \right) \leq \epsilon.$$

Then consider the complementary event:

$$\mathbb{P}\left(\forall \rho \in \mathcal{M}_+^1(\Theta),\ \int R\, d\rho \leq \int r_n\, d\rho + f(\lambda, n) + K(\rho, \pi) + \log\left(\frac{1}{e}\right) \right) \geq 1 - e.$$

□

3.3 Oracle-type inequalities

Another way of exploiting PAC-Bayesian bounds is to compare $\int Rd\hat{\rho}_\lambda$ to the optimal risk.

Theorem 2. *Assume that the* ***Hoeffding assumption*** *is satisfied. For any $\varepsilon > 0$, with probability at least $1-\varepsilon$ we have simultaneously*

$$\int Rd\hat{\rho}_\lambda \leq \; \mathcal{B}_\lambda(\mathcal{M}_+^1(\Theta)) := \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \int Rd\rho + 2\frac{f(\lambda,n) + \mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\}$$

$$\int Rd\tilde{\rho}_\lambda \leq \mathcal{B}_\lambda(\mathcal{F}) := \inf_{\rho \in \mathcal{F}} \left\{ \int Rd\rho + 2\frac{f(\lambda,n) + \mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\}$$

Moreover,

$$\mathcal{B}_\lambda(\mathcal{F}) = \mathcal{B}_\lambda(\mathcal{M}_+^1(\Theta)) + \frac{2}{\lambda} \inf_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{\frac{\lambda}{2}}),$$

In this manner, we can effectively evaluate $\int Rd\hat{\rho}_\lambda$ against the optimal aggregation procedure in $\mathcal{M}_+^1(\Theta)$ and $\int Rd\tilde{\rho}_\lambda$ against the best in $\mathcal{F}$. More critically, this analysis enables us to derive explicit expressions for the right-hand side of these inequalities across various models, thereby facilitating the determination of convergence rates.

Proof. Using the same calculations as in Theorem 1, we have, with probability at least $1-\varepsilon$, simultaneously for all $\rho \in \mathcal{M}_+^1(\Theta)$,

$$\lambda \int R \, \mathrm{d}\rho \leq \lambda \int r_n \, \mathrm{d}\rho + f(\lambda,n) + \mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right) \tag{2}$$

$$\lambda \int r_n \, \mathrm{d}\rho \leq \lambda \int R \, \mathrm{d}\rho + f(\lambda,n) + \mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right) \tag{3}$$

Let's note that another consequence can be derived from the (KL) identity :

$$\pi[h] = arg \min_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \int hd\rho + K(\rho,\pi) \right\}$$

Applying this with $h(\theta) = \lambda r_n(\theta)$ gives :

$$\pi[\lambda r_n] = \hat{\rho}_\lambda = arg \min_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \int \lambda r_n d\rho + K(\rho,\pi) \right\} \tag{4}$$

We hence combine both (2) with $\rho = \hat{\rho}_\lambda$ and (4) to get :

$$\lambda \int R \, \mathrm{d}\hat{\rho}_\lambda \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \lambda \int r_n \, \mathrm{d}\rho + f(\lambda,n) + \mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right) \right\}$$

In addition, taking advantage of (3) gives :

$$\lambda \int R \, \mathrm{d}\hat{\rho}_\lambda \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \lambda \int R \, \mathrm{d}\rho + 2f(\lambda,n) + 2\mathcal{K}(\rho,\pi) + 2\log\left(\frac{2}{\varepsilon}\right) \right\}$$

Which proves the first inequality of the theorem. In order to prove the second inequality, we combine the (KL) identity with $h(\theta) = \lambda r_n$ and (2) in order to get, for any ρ,

$$\lambda \int R \, \mathrm{d}\rho \leq f(\lambda,n) + \mathcal{K}\left(\rho, \hat{\rho}_\lambda\right) - \log \int \exp\left(-\lambda r_n\right) \mathrm{d}\pi + \log\left(\frac{2}{\varepsilon}\right)$$

Which translates to

$$\lambda \int R \, \mathrm{d}\tilde{\rho}_\lambda \leq \inf_{\rho \in \mathcal{F}} \left\{ f(\lambda,n) + \mathcal{K}\left(\rho, \hat{\rho}_\lambda\right) - \log \int \exp\left(-\lambda r_n\right) \mathrm{d}\pi + \log\left(\frac{2}{\varepsilon}\right) \right\}$$

Using the same technique implied previously, in particular combining the (KL) identity and (3), we obtain :

$$\lambda \int R \, \mathrm{d}\tilde{\rho}_\lambda \leq \inf_{\rho \in \mathcal{F}} \left\{ \lambda \int R \, \mathrm{d}\rho + 2f(\lambda, n) + 2\mathcal{K}(\rho, \pi) + 2 \log \left(\frac{2}{\varepsilon} \right) \right\}.$$

Lastly, let's note that

$$\begin{aligned}
\mathcal{B}_\lambda(\mathcal{F}) &= \inf_{\rho \in \mathcal{F}} \left\{ \int R \, \mathrm{d}\rho + \frac{2f(\lambda, n)}{\lambda} + \frac{2\mathcal{K}(\rho, \pi)}{\lambda} + \frac{2 \log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\} \\
&= \inf_{\rho \in \mathcal{F}} \left\{ -\frac{2}{\lambda} \log \int \exp\left(-\frac{\lambda}{2} R\right) \mathrm{d}\pi + \frac{2f(\lambda, n)}{\lambda} + \frac{2\mathcal{K}\left(\rho, \pi_{\frac{\lambda}{2}}\right)}{\lambda} + \frac{2 \log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\} \\
&= -\frac{2}{\lambda} \log \int \exp\left(-\frac{\lambda}{2} R\right) \mathrm{d}\pi + \frac{2f(\lambda, n)}{\lambda} + \frac{2 \log\left(\frac{2}{\varepsilon}\right)}{\lambda} + \frac{2}{\lambda} \inf_{\rho \in \mathcal{F}} \mathcal{K}\left(\rho, \pi_{\frac{\lambda}{2}}\right) \\
&= \mathcal{B}_\lambda\left(\mathcal{M}_+^1(\Theta)\right) + \frac{2}{\lambda} \inf_{\rho \in \mathcal{F}} \mathcal{K}\left(\rho, \pi_{\frac{\lambda}{2}}\right) \qquad \text{, by using (1)}
\end{aligned}$$

□

Theorem 3. *Assume that the* ***Bernstein assumption*** *is satisfied. Assume that $\lambda \in I$ satisfies $\lambda - g(\lambda, n) > 0$. Then for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$ we have simultaneously:*

$$\begin{aligned}
\int R \, \mathrm{d}\hat{\rho}_\lambda - \bar{R} &\leq \overline{\mathcal{B}}_\lambda\left(\mathcal{M}_+^1(\Theta)\right), \\
\int R \, \mathrm{d}\tilde{\rho}_\lambda - \bar{R} &\leq \overline{\mathcal{B}}_\lambda(\mathcal{F}),
\end{aligned}$$

where, for either $\mathcal{A} = \mathcal{M}_+^1(\Theta)$ or $\mathcal{A} = \mathcal{F}$,

$$\overline{\mathcal{B}}_\lambda(\mathcal{A}) = \frac{1}{\lambda - g(\lambda, n)} \inf_{\rho \in \mathcal{A}} \left\{ [\lambda + g(\lambda, n)] \int (R - \bar{R}) \mathrm{d}\rho + 2\mathcal{K}(\rho, \pi) + 2 \log\left(\frac{2}{\varepsilon}\right) \right\}.$$

In addition,

$$\overline{\mathcal{B}}_\lambda(\mathcal{F}) = \overline{\mathcal{B}}_\lambda\left(\mathcal{M}_+^1(\Theta)\right) + \frac{2}{\lambda - g(\lambda, n)} \inf_{\rho \in \mathcal{F}} \mathcal{K}\left(\rho, \pi_{\frac{\lambda + g(\lambda, n)}{2}}\right).$$

The proof is very similar to the previous Theorem. The reader is reffered to [1] for more details. The main difference with the previous theorem is that the function $R(.)$ is replaced by $R(.) - \overline{R}$ which gives better rates of convergence.

4 Application to classification

4.1 Notations

In this section, we assume that $\mathcal{Y} = \{0, 1\}$ and we consider linear classification : $\Theta = \mathcal{X} = \mathbb{R}^d$, $f_\theta(x) = \mathbb{1}_{<\theta, x> \geq 0}$. We put $r_n(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{f_\theta(X_i) \neq Y_i}$, $R(\theta) = \mathbb{P}(Y \neq f_\theta(X))$. We assume that the $(X_i, Y_i)_{i=1,..,n}$ are i.i.d. Set $\varphi(.)$ as the density of $\mathcal{N}(0, 1)$ w.r.t. the Lebesgue measure and $\phi(.)$ as the corresponding c.d.f measure.

4.2 Bernstein assumption

Since **Theorem3** offers better rates of convergence, let's try to apply it in this framework. This means that we have to prove the Bernstein assumption. In order to do so, we introduce the margin assumption of Mammen and Tsybakov.

Lemma 1. *Assume that Mammen and Tsybakov's margin assumption is satisfied: i.e. there is a constant C such that*

$$\mathbb{E}\left[\left(\mathbf{1}_{f_\theta(X)\neq Y} - \mathbf{1}_{f_{\bar\theta}(X)\neq Y}\right)^2\right] \leq C[R(\theta) - \bar{R}].$$

Then Bernstein assumption is satisfied with $g(\lambda, n) = \frac{C\lambda^2}{2n-\lambda}$.

Proof. This lemma is a direct application of Theorem 2.10 of Boucheron et al [2] (see Appendix for more details). □

4.3 Different sets of Variational Gaussian approximations

We consider a Gaussian prior $\pi = \mathcal{N}_d\left(0, \vartheta^2 I_d\right)$ and a VB approach based on Gaussian families as follows :

$$\begin{aligned}
\mathcal{F}_1 &= \left\{\Phi_{\mathbf{m},\sigma^2}, \mathbf{m} \in \mathbb{R}^d, \sigma^2 \in \mathbb{R}_+^*\right\},\\
\mathcal{F}_2 &= \left\{\Phi_{\mathbf{m},\sigma^2}, \mathbf{m} \in \mathbb{R}^d, \sigma^2 \in \left(\mathbb{R}_+^*\right)^d\right\} \text{ (mean field approximation)}
\end{aligned}$$

where $\Phi_{\mathbf{m},\sigma^2}$ is Gaussian distribution $\mathcal{N}_d\left(\mathbf{m}, \sigma^2 I_d\right), \Phi_{\mathbf{m},\sigma^2}$ is $\mathcal{N}_d\left(\mathbf{m}, \operatorname{diag}\left(\boldsymbol{\sigma}^2\right)\right)$. A crucial observation can be done :

$$\begin{gathered}
\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{M}_+^1(\Theta),\\
\mathcal{B}_\lambda\left(\mathcal{M}_+^1(\Theta)\right) \leq \mathcal{B}_\lambda\left(\mathcal{F}_2\right) \leq \mathcal{B}_\lambda\left(\mathcal{F}_1\right).
\end{gathered} \tag{B}$$

Thanks to (B), it's clear that calculating $\mathcal{B}_\lambda\left(\mathcal{F}_1\right)$ provides an upper bound on $\mathcal{B}_\lambda\left(\mathcal{F}_2\right)$ and justifies the soundness of such a choice.

4.4 Bernstein inequality for Gaussian approximations

We now introduce an application to **Theorem 3**.

Corollary 1. *Assume that the VB approximation is done on either $\mathcal{F}_1, \mathcal{F}_2$. Under Mammen and Tsybakov margin assumption and assumption A1, i.e. there is a constant $c > 0$ such that, for any $(\theta, \theta') \in \Theta^2$ with $\|\theta\| = \|\theta'\| = 1$,*

$$\mathbb{P}\left(\langle X, \theta\rangle\left\langle X, \theta'\right\rangle < 0\right) \leq c\left\|\theta - \theta'\right\|.$$

With $\lambda = \frac{2n}{C+2}$ and $\vartheta > 0$, for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$,

$$\left.\begin{array}{l}\int R \ \mathrm{d}\hat{\rho}_\lambda\\ \int R \ \mathrm{d}\tilde{\rho}_\lambda\end{array}\right\} \leq \bar{R} + \frac{(C+2)(C+1)}{2}\left\{\frac{d\log\frac{n}{\vartheta}}{n} + \frac{d\vartheta}{n^2} + \frac{1}{\vartheta} - \frac{d}{\vartheta n} + \frac{2}{n}\log\frac{2}{\varepsilon}\right\} + \frac{\sqrt{d}2c(2C+1)}{n}.$$

With the right choice of constants, this bound achieves a rate of $d\log(n)/n$. Additionally, knowing that a minimax rate is d/n in this framework, we see that the variational approximation is not significantly outperformed by the best estimator for the problem at hand. Thus, the variational approximation is quite efficient.

Proof. We begin by applying **Theorem 3**. Let $\lambda < \frac{2n}{C+1}$, we have

$$\begin{aligned}
\overline{\mathcal{B}}_\lambda(\mathcal{F}_1) &= \inf_{\mathbf{m},\sigma^2}\left\{\frac{\lambda + g(\lambda,n)}{\lambda - g(\lambda,n)}\int(R - \bar{R})\mathrm{d}\Phi_{\mathbf{m},\sigma^2} + \frac{1}{\lambda - g(\lambda,n)}\left(2\mathcal{K}\left(\Phi_{\mathbf{m},\sigma^2}, \pi\right) + 2\log\frac{2}{\epsilon}\right)\right\}\\
&= \inf_{\mathbf{m},\sigma^2}\left\{\frac{\lambda + g(\lambda,n)}{\lambda - g(\lambda,n)}\int(R - \bar{R})\mathrm{d}\Phi_{\mathbf{m},\sigma^2} + \frac{1}{\lambda - g(\lambda,n)}\left(2d\left[\frac{1}{2}\log\left(\frac{\vartheta^2}{\sigma^2}\right) + \frac{\sigma^2}{2\vartheta^2}\right] + \frac{\|\mathbf{m}\|^2}{2\vartheta^2} - \frac{d}{2} + 2\log\frac{2}{\epsilon}\right)\right\}
\end{aligned}$$

Since $\bar\theta$ is not unique, we set it such as $\|\bar\theta\| = 1$. This allows the use of the following trick :

$$\begin{aligned}
R(\theta) - \bar{R} &= \mathbb{E}\left[\mathbf{1}_{\langle\theta,X\rangle Y<0} - \mathbf{1}_{\langle\bar\theta,X\rangle Y<0}\right] \leq \mathbb{E}\left[\mathbf{1}_{\langle\theta,X\rangle\langle\bar\theta,X\rangle<0}\right]\\
&= \mathbb{P}(\langle\theta, X\rangle\langle\bar\theta, X\rangle < 0) \leq c\left\|\frac{\theta}{\|\theta\|} - \bar\theta\right\| \leq 2c\|\theta - \bar\theta\|
\end{aligned}$$

Which yields

$$\overline{\mathcal{B}}_\lambda(\mathcal{F}_1) \leq \inf_{(\mathbf{m},\sigma^2)} \left\{ \frac{\lambda + g(\lambda,n)}{\lambda - g(\lambda,n)} 2c \int \|\theta - \bar{\theta}\| \Phi_{\mathbf{m},\sigma^2}(\mathrm{d}\theta) + \frac{1}{\lambda - g(\lambda,n)} \left(2d \left[\frac{1}{2} \log\left(\frac{\vartheta^2}{\sigma^2}\right) + \frac{\sigma^2}{2\vartheta^2} \right] + \frac{\|\mathbf{m}\|^2}{2\vartheta^2} - \frac{d}{2} + 2\log\frac{2}{\epsilon} \right) \right\}$$

We fix $\mathbf{m} = \bar{\theta}$:

$$\overline{\mathcal{B}}_\lambda(\mathcal{F}_1) \leq \inf_{\sigma^2} \left\{ \frac{\lambda + g(\lambda,n)}{\lambda - g(\lambda,n)} 2c\sqrt{d}\sigma + \frac{\lambda}{n} + \frac{1}{\lambda - g(\lambda,n)} \left(2d \left[\frac{1}{2} \log\left(\frac{\vartheta^2}{\sigma^2}\right) + \frac{\sigma^2}{2\vartheta^2} \right] + \frac{1}{2\vartheta^2} - \frac{d}{2} + 2\log\frac{2}{\epsilon} \right) \right\}$$

Finally, taking $\lambda = \frac{2n}{C+2}$ concludes the proof.

□

4.5 Implementation and numerical results

We essentially compare the performance of the mean field VB approximation to a linear SVM (support vector machine) and a radial kernel SVM. We apply classification of two standard Scikit-learn datasets: *Two moons* and *Breast Cancer*. We also include in our study *SPECT* dataset available on the UCI repository[*].

Interestingly, VB outperforms SVM for both *SPECT* and *Two moons* (see Table 1). However, a high error is detected for the *Breast Cancer* dataset. This is a very similar issue encountered in [1]; we used the same dataset as the one mentioned in the article in order to highlight such an observation. This will be discussed further.

Table 1: Missclassification rates for different datasets and for the proposed approximation of the Gibbs posterior. The last two columns are the missclassification rate given by a SVM with a radial kernel and a linear SVM.

Dataset	Covariate 2	Mean Field $\mathcal{F}_2$	SVM radial	SVM linear
SPECT	22	13.8	21.3	26.7
Breast Cancer	30	36.8	6.4	3.5
Two moons	2	1.4	10.0	11.1

In addition, the linear SVM significantly outperforms the radial SVM in classifying the Breast Cancer dataset. This superior performance can be attributed to the data's inherently linear separability (refer to Figure 1). Conversely, datasets like Two Moons and SPECT require a more sophisticated separation approach due to their complex clustering patterns.

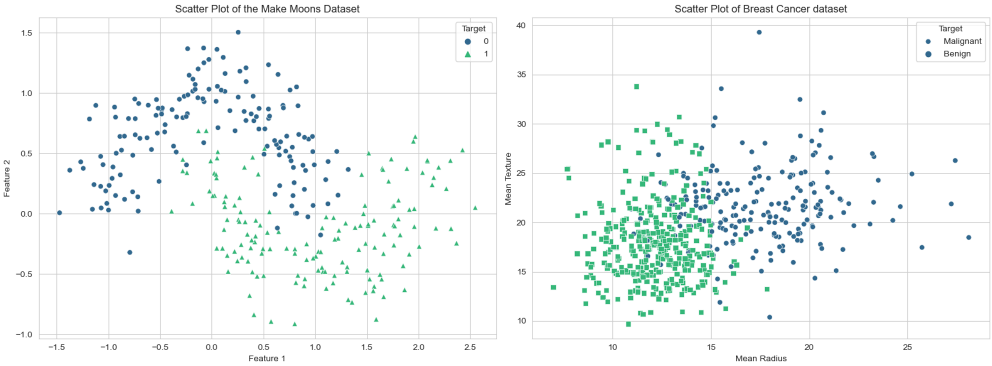

Figure 1: Representation of data for Two Moons and Breast Cancer datasets.

[*] See notebook for an-in-depth step by step implementation.

4.6 Refinements

The previous application is very simple and could benefit from some improvements. We can for instance consider a convex minimization problem that is better for two main reasons :

- First, the convex nature of the risk function allows for the derivation of finite sample oracle inequalities. These inequalities provide bounds on the estimation error of the estimators derived from the variational approximation, giving theoretical guarantees about their performance.
- Secondly, a variational approximation can be obtained in polynomial time using optimal convex solvers. This is significant because it allows efficient computational solutions that are both feasible and practical for large-scale applications.

A more broad reason is that convexified approaches enable the extension of these methods to various learning tasks like ranking and matrix completion. Formally, we can set $\mathcal{Y} \in \{-1, 1\}$ and change the previous *0-1 loss* into the *hinge loss*. The empirical risk would thus have the form :

$$r_n^H(\theta) = \frac{1}{n}\sum_{i=1}^{n} max(0, 1 - Y_i \langle \theta, X_i \rangle)$$

Another consideration would be to introduce a new form of Gaussian families. Ideally, we would require a family that allows a more detailed and accurate modeling of dependencies between variables. We introduce the *full covariance approximation* (see notebook, section 6 for an in depth implementation) :

$$\mathcal{F}_3 = \left\{\Phi_{\mathbf{m},\Sigma}, \mathbf{m} \in \mathbb{R}^d, \Sigma \in \mathcal{S}^{d+}\right\}$$

The full covariance version may handle the multimodality of the optimization landscape more effectively. This could be particularly beneficial in complex problems where the mean field approximation $\mathcal{F}_2$ might fail to capture significant dependencies between variables due to its more restrictive diagonal covariance structure. This is a good candidate for the previous miscclassification experiment since it is potentially more nuanced than the VB $\mathcal{F}_2$ and still much faster to compute than classical tempered SMC algorithms. For example, the features of the Breast Cancer dataset appear to be much more correlated than the features of the SPECT dataset (see Figure 2). This would explain why $\mathcal{F}_2$ did not perform very well. In this specific case, the $\mathcal{F}_3$ approximation would likely perform much better.

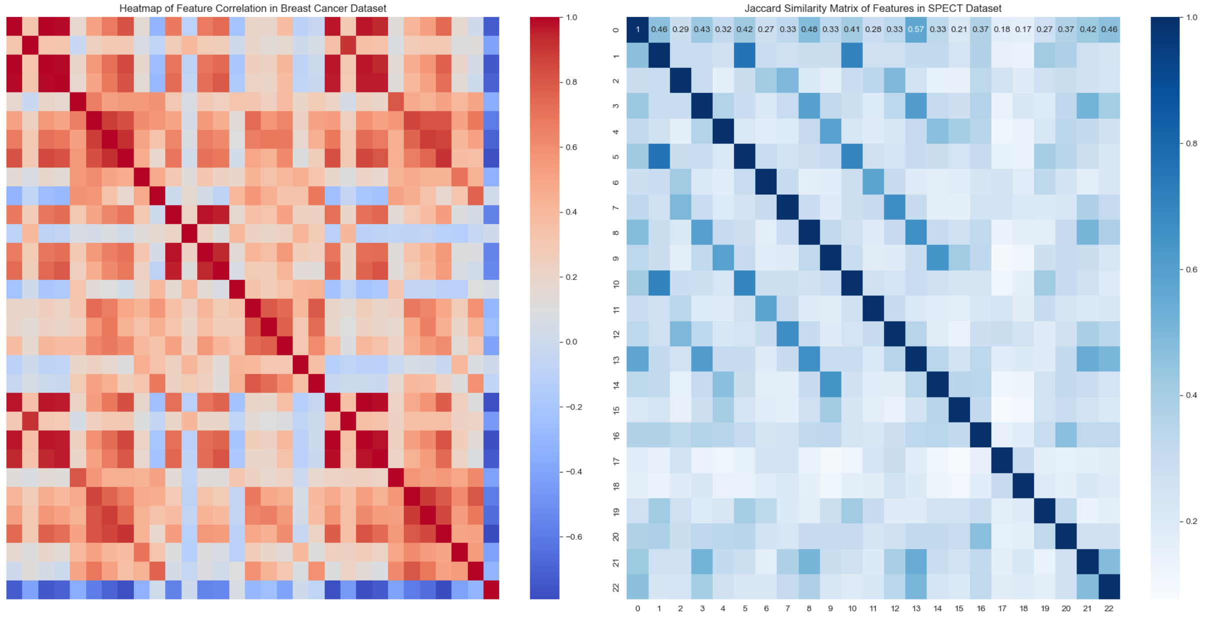

Figure 2: Heat map of feature correlation in Breast cancer dataset (orange) and Jaccard Similarity Matrix of SPECT dataset (blue).

Is it possible to do better than $\mathcal{F}_3$? Full covariance is still relatively slow when working in very large-scale problems. Another approach is to consider a *Decoupled Variational Gaussian Approximation (DVGA).* This technique relies on identifying groups of parameters $G_1, \ldots, G_n$ based on domain knowledge that indicates correlations between subsets of the model's parameters. For each group G_i, we can model the parameters θ_{G_i} with a Gaussian distribution having its own mean vector μ_{G_i} and covariance matrix Σ_{G_i}. DVGA allows for capturing complex dependencies within groups of features, which can be particularly useful in classification tasks where certain groups of features interact in non-trivial ways. This has been, for example, tackled by Khan (2014) [4]. Besides handling complex likelihoods, DVGA can reduce the number of variational parameters, making the computation scalable. This approach also allows for the maximization of the lower bound using a sequence of convex problems, which can be parallelized over data examples and improve computation time. Overall, DVGA can offer a favorable balance between computational feasibility and fidelity to the data structure, which might be compromised in full covariance models due to their intensive computational demands.

5 Bounds for hostile data

We previously presented efficient bounds that hold under strong exponential moment assumptions. We would like to explore, in this section, principal ideas for a relaxation of these constraints. In particular, we try to present PAC-Bayesian bounds for dependent, heavy-tailed observations, which we usually refer to as *hostile data.* This has important implications since it allows to construct bounds for other supervised learning techniques, such as regression. This idea was presented by Alquier and Guedj (2019) [5].

5.1 PAC-Bayesian bound

Let's first present two central quantities :

Definition 6. *For any function g, let*

$$\mathcal{M}_{g,n} = \int \mathbb{E}\left[g\left(|r_n(\theta) - R(\theta)|\right)\right]\pi(\mathrm{d}\theta).$$

Definition 7. *Let f be a convex function with $f(1) = 0$. The f-divergence between two distributions ρ and π is defined by*

$$D_f(\rho,\pi) = \int f\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}\pi$$

when ρ is absolutely continous with respect to π, and $D_f(\rho,\pi) = +\infty$ otherwise.

We use the following notation: $\phi_p(x) = x^p$.In addition, we denote the Kullback-Leibler divergence by $\mathcal{K}(\rho,\pi) = D_f(\rho,\pi)$ when $f(x) = x\log(x)$, and the chi-square divergence $\chi^2(\rho,\pi) = D_{\phi_2 - 1}(\rho,\pi)$.

Theorem 4. *Fix $p > 1$, put $q = \frac{p}{p-1}$ and fix $\delta \in (0,1)$. With probability at least $1-\delta$ we have for any aggregation distribution ρ*

$$\left|\int R\ \mathrm{d}\rho - \int r_n\ \mathrm{d}\rho\right| \leq \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}}\left(D_{\phi_p - 1}(\rho,\pi) + 1\right)^{\frac{1}{p}}$$

Theorem 4 primarily conveys that it's possible to evaluate the observable risk in relation to R by examining two specific elements: the moment $\mathcal{M}_{\phi_q,n}$ that is influenced by the data distribution, and the divergence $D_{\phi_p - 1}(\rho,\pi)$ which emerges as an indicator of the complexity inherent in the set Θ.

Proof. Introduce $\Delta_n(\theta) := |r_n(\theta) - R(\theta)|$.

$$\begin{aligned}\left|\int R\ \mathrm{d}\rho - \int r_n\ \mathrm{d}\rho\right| &\leq \int \Delta_n\ \mathrm{d}\rho = \int \Delta_n \frac{\mathrm{d}\rho}{\mathrm{d}\pi}\mathrm{d}\pi \\ &\leq \left(\int \Delta_n^q\ \mathrm{d}\pi\right)^{\frac{1}{q}}\left(\int\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)^p\ \mathrm{d}\pi\right)^{\frac{1}{p}} \text{ (Hölder inequality)} \\ &\leq \left(\frac{\mathbb{E}\int \Delta_n^q\ \mathrm{d}\pi}{\delta}\right)^{\frac{1}{q}}\left(\int\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)^p\ \mathrm{d}\pi\right)^{\frac{1}{p}} \text{ (Markov inequality, w. prob. 1- } \delta \text{)} \\ &= \left(\frac{\mathcal{N}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}}\left(D_{\phi_p - 1}(\rho,\pi) + 1\right)^{\frac{1}{p}}.\end{aligned}$$

□

As a consequence of Theorem 4, we have, with probability at least $1-\delta$ for any probability measure ρ,

$$\int R\ \mathrm{d}\rho \leq \int r_n\ \mathrm{d}\rho + \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}}\left(D_{\phi_p - 1}(\rho,\pi) + 1\right)^{\frac{1}{p}}$$

Thus, the aggregation $\hat{\rho}_n$ distribution can be defined as the minimizer of the left hand side. This would be the focus of the next section.

5.2 Oracle inequality

We start this section by introducing a formal definition of $\hat{\rho}_n$.

Definition 8. *We define $\bar{r}_n = \bar{r}_n(\delta,p)$ as*

$$\bar{r}_n = \min\left\{u \in \mathbb{R}, \int \left[u - r_n(\theta)\right]_+^q \pi(\mathrm{d}\theta) = \frac{\mathcal{M}_{\phi_q,n}}{\delta}\right\}.$$

Note that such a minimum always exists as the integral is a continuous function of u, is equal to 0 when $u = 0$ and $\to \infty$ when $u \to \infty$ We then define

$$\frac{\mathrm{d}\hat{\rho}_n}{\mathrm{d}\pi}(\theta) = \frac{[\bar{r}_n - r_n(\theta)]_+^{\frac{1}{p-1}}}{\int [\bar{r}_n - r_n]_+^{\frac{1}{p-1}} \ \mathrm{d}\pi}$$

We know express an Oracle-type inequality.

Theorem 5. *Under the assumptions of* ***Theorem 4****, with probability at least* $1 - \delta$*,*

$$\int R \ \mathrm{d}\hat{\rho}_n \leq \bar{r}_n \leq \inf_{\rho} \left\{ \int R \ \mathrm{d}\rho + 2 \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta} \right)^{\frac{1}{q}} \left(D_{\phi_p - 1}(\rho, \pi) + 1 \right)^{\frac{1}{p}} \right\}$$

Proof. First, we combine **Theorem 4** and the definition of $\hat{\rho}_n$:

$$\int R \ \mathrm{d}\hat{\rho}_n \leq \int r_n \ \mathrm{d}\hat{\rho}_n + \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta} \right)^{\frac{1}{q}} \left(D_{\phi_p - 1}\left(\hat{\rho}_n, \pi\right) + 1 \right)^{\frac{1}{p}} = \inf_{\rho} \left\{ \int r_n \ \mathrm{d}\rho + \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta} \right)^{\frac{1}{q}} \left(D_{\phi_p - 1}(\rho, \pi) + 1 \right)^{\frac{1}{p}} \right\}$$

Now, we prove that the right hand side is $\bar{r}_n$. For any ρ, we have :

$$\begin{aligned}
\bar{r}_n - \int r_n \ \mathrm{d}\rho &= \int [\bar{r}_n - r_n] \, \mathrm{d}\rho \\
&= \int [\bar{r}_n - r_n]_+ \, \mathrm{d}\rho - \int [\bar{r}_n - r_n]_- \ \mathrm{d}\rho \\
&\leq \int [\bar{r}_n - r_n]_+ \, \mathrm{d}\rho = \int [\bar{r}_n - r_n]_+ \, \frac{\mathrm{d}\rho}{\mathrm{d}\pi} \mathrm{d}\pi \\
&\leq \left(\int [\bar{r}_n - r_n]_+^q \ \mathrm{d}\pi \right)^{\frac{1}{q}} \left(\int \left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi} \right)^p \ \mathrm{d}\pi \right)^{\frac{1}{p}} \quad \text{(by Hölder inequality)} \\
&\leq \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta} \right)^{\frac{1}{q}} \left(D_{\phi_p - 1}(\rho, \pi) + 1 \right)^{\frac{1}{p}} \qquad \text{(by definition of } \bar{r}_n\text{)}
\end{aligned}$$

Finally, applying **Definition 8**, we have :

$$\begin{aligned}
\bar{r}_n - \int r_n \ \mathrm{d}\hat{\rho}_n = \int [\bar{r}_n - r_n]_+ \, \mathrm{d}\hat{\rho}_n &= \frac{\int [\bar{r}_n - r_n]_+ [\bar{r}_n - r_n]_+^{\frac{1}{p-1}} \ \mathrm{d}\pi}{\int [\bar{r}_n - r_n]_+^{\frac{1}{p-1}} \ \mathrm{d}\pi} = \frac{\int [\bar{r}_n - r_n]_+^q \ \mathrm{d}\pi}{\int [\bar{r}_n - r_n]_+^{\frac{1}{p-1}} \ \mathrm{d}\pi} \\
&= \frac{\left(\int [\bar{r}_n - r_n]_+^q \ \mathrm{d}\pi \right)^{\frac{1}{p} + \frac{1}{q}}}{\int [\bar{r}_n - r_n]_+^{\frac{1}{p-1}} \ \mathrm{d}\pi} = \left(\int [\bar{r}_n - r_n]_+^q \ \mathrm{d}\pi \right)^{\frac{1}{q}} \frac{\left(\int [\bar{r}_n - r_n]_+^{\frac{p}{p-1}} \ \mathrm{d}\pi \right)^{\frac{1}{p}}}{\int [\bar{r}_n - r_n]_+^{\frac{1}{p-1}} \ \mathrm{d}\pi} \\
&= \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta} \right)^{\frac{1}{q}} \left(\int \left(\frac{\mathrm{d}\hat{\rho}_n}{\mathrm{d}\pi} \right)^p \ \mathrm{d}\pi \right)^{\frac{1}{p}} = \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta} \right)^{\frac{1}{q}} \left(D_{\phi_p - 1}\left(\hat{\rho}_n, \pi\right) + 1 \right)^{\frac{1}{p}}
\end{aligned}$$

□

This new bound is thus very efficient and uses Csiszár's f -divergence to generalize the Kullback-Leibler divergence mentioned in previous sections. It can for example derive a risk bound for auto-regression with heavy-tailed time series, which was explored by Alquier and Guedj (2019).

6 Conclusion

PAC-Bayesian framework continues to be a vibrant area of research, with ongoing developments in theory and applications. Recent papers have shown a focus on extending this framework to domain adaptation [7] and meta-learning [8].

A Boucheron et al. - Theorem 2.10

This section presents a Theorem that serves as a guideline for the demonstration of **Lemma 1**

Theorem 6. *Let $X_1, \ldots, X_n$ be independent realvalued random variables. Assume that there exist positive numbers v and c such that $\sum_{i=1}^n \boldsymbol{E}\left[X_i^2\right] \leq v$ and*

$$\sum_{i=1}^n \boldsymbol{E}\left[(X_i)_+^q\right] \leq \frac{q!}{2} v c^{q-2} \quad \text{for all integers } q \geq 3.$$

If $S = \sum_{i=1}^n (X_i - \boldsymbol{E}X_i)$, then for all $\lambda \in (0, 1/c)$ and $t > 0$,

$$\psi_S(\lambda) \leq \frac{v\lambda^2}{2(1 - c\lambda)}$$

and

$$\psi_S^*(t) \geq \frac{v}{c^2} h_1\left(\frac{ct}{v}\right),$$

where $h_1(u) = 1 + u - \sqrt{1 + 2u}$ for $u > 0$. In particular, for all $t > 0$,

$$\boldsymbol{P}\{S \geq \sqrt{2vt} + ct\} \leq e^{-t}.$$

The Mammen and Tsybakov's margin assumption allows to fix the value of v and c mentioned in this theorem.

References

[1] Alquier Pierre, Ridgway James, Chopin Nicolas, (2016). *On the properties of variational approximations of Gibbs posteriors.*

[2] Alquier Pierre, (2021). *User-friendly introduction to PAC-Bayes bounds.*

[3] Boucheron Stéphane, Lugosi Gabor, Massart Pascal, (2012). *Concentration inequalities, a non asymptotic theory of independence.*

[4] Khan Mohammad Emtiyaz, (2014). *Decoupled Variational Gaussian Inference.*

[5] Alquier Pierre, Guedj Benjamin, (2019). *Simpler PAC-Bayesian bounds for hostile data.*

[6] Catoni Olivier, (2007). *PAC-Bayesian supervised classication: the thermodynamics of statistical learning.*

[7] Sicilia Anthony, Atwell Katherine, Alikhani Malihe, Hwang Seong Jae (2022). *PAC-Bayesian domain adaptation bounds for multi-class learners.*

[8] Riou Charles, Alquier Pierre, Chérief-Abdellatif Badr-Eddine (2023). *Bayes meets Bernstein at the Meta level: an analysis of fast rates in Meta-learning with PAC-Bayes.*