On the properties of variational approximations of Gibbs posteriors

Pierre Alquier, James Ridgway Nicolas Chopin

Eline POT

A dissertation submitted as part of the requirements for the degree of Master 2 Apprentissage et Algorithmes at Sorbonne University

April 2024

Supervisors : Badr-Eddine Chérief-Abdellatif and Antoine Godichon-Baggioni

Contents

Introduction

This dissertation is about the article *On the properties of variational approximations of Gibbs posteriors*, written by Pierre Alquier, James Ridgway, and Nicolas Chopin, in 2016 (Alquier, Ridgway, and Chopin 2016).

Gibbs posterior is a usual optimal distribution of estimators in the context of non-asymptotic risk bounds for random estimators. Unfortunately, it is often untractable. This paper studies variational approximations of the Gibbs posterior that fit large datasets and their properties.

In this dissertation, we will first overview PAC Bayesian with some of its theoretical and applied aspects, and introduce the problem of Variational Bayes approximation of the Gibbs posterior. In a second time, we will look into empirical and theoretical bounds under some assumptions. Then, we will apply these results and show some experiments to a classification task with a focus on the convex case, and ranking.

We will examine the principal theorems and their applications from the article while omitting certain aspects. Specifically, we may omit some lengthy proofs available in the paper, in order to concentrate on certain calculations, corollaries whose proofs were not consistently provided, and experiments mentioned in the paper but not conducted.

1 PAC Bayesian Framework

1.1 Framework and notations

We suppose that we observe $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ where each couple is under the same distribution P, unknown. We consider the set of predictors : $\{f_\theta : \mathcal{X} \mapsto \mathbb{R}, \theta \in \Theta\}$.

We also introduce a loss function $l : \mathbb{R}^2 \mapsto \mathbb{R}$. In the following, we will assume that the loss satisfies: $0 \leq l \leq 1$.

The risk R is defined as : $R(\theta) = \mathbb{E}_P[l(Y, f_\theta(X)]$. We denote r_n its empirical counterpart : $r_n(\theta) = \frac{1}{n} \sum_{i=1}^n l(Y_i, f_\theta(Y_i, f_\theta(X_i))$

We note $\bar{\theta} \in \arg\min_\theta R(\theta)$ and $\bar{R} = R(\bar{\theta})$. We note $R^\star$ the risk for the Bayes estimator.

We have a prior π on Θ and we denote $\mathcal{M}_+^1(\Theta)$ the set of probability measures on Θ.

In this dissertation, we will denote $\mathbb{E}$ the expectation taken on $(X, Y) \sim P$. We will denote $\mathbb{E}_\gamma$ the expectation taken on $\theta \sim \gamma$ for any probability measure γ on the space Θ.

Finally, we denote φ the density function, and Φ the cumulative distribution function of the standard Gaussian $\mathcal{N}(0, Id)$.

1.2 Origins of the Gibbs posterior

Before introducing the notion of Gibbs posterior (also known as pseudo-posterior), let us give us more insight about where it comes from and the idea behind it.

This concept can be introduced following the result of Catoni's PAC-Bayes bound (Catoni 2004). This result uses the Donsker and Varadan's variational formula, that we recall here :

Theorem 1 (Donsker and Varadan's variational formula)**.** *For any measurable function $h : \Theta \mapsto \mathbb{R}$ that is bounded, we have :*

$$\log \mathbb{E}_\pi[\exp h(\theta)] = \sup_{\rho \in \mathcal{P}(\Theta)} [\mathbb{E}_\rho[h\theta)] - KL(\rho, \pi)]$$

Moreover, the RHS is reached for the Gibbs measure, whose density with respect to π is defined by :

$$\frac{d\pi_h}{d\pi}(\theta) = \frac{\exp h(\theta)}{\mathbb{E}_\pi[\exp h(\theta)]}$$

Theorem 2 (Catoni's PAC-Bayes bound)**.** *For any $\lambda > 0$, $\varepsilon \in (0, 1)$, and for any prior π over a family of estimators $\mathcal{F}$ we have :*

$$\mathbb{P}\left(\forall \rho \in \mathcal{F}, \mathbb{E}_\rho[R(\theta)] \leq \mathbb{E}_\rho[r_n(\theta)] + \frac{\lambda}{8n} + \frac{KL(\rho, \pi) + \log(1/\varepsilon)}{\lambda}\right) \geq 1 - \varepsilon$$

Proof. The proof relies on 3 results: Donsker and Varadan's variational formula, Markov's inequality, and Hoeffdings's lemma. □

This result motivates the study of a probability measure $\hat{\rho}_\lambda$ that would minimize the quantity $\mathbb{E}_\rho[r_n(\theta)] + \frac{KL(\rho,\pi)}{\lambda}$ in ρ.

1.3 Some definitions

Definition 3 (Gibbs posterior)**.** *For some $\lambda > 0$, we define the Gibbs posterior as :*

$$\hat{\rho}_\lambda(d\theta) = \frac{\exp(-\lambda r_n(\theta))}{\int \exp(-\lambda R) d\pi}\pi(d\theta) = \frac{\exp(-\lambda r_n(\theta))}{\mathbb{E}_\pi[\exp(-\lambda r_n(\theta))]}\pi(d\theta)$$

Definition 4. *The theoretical counterpart of the Gibbs posterior can be written:*

$$\pi_\lambda(d\theta) = \frac{\exp(-\lambda R(\theta))}{\int \exp(-\lambda R) d\pi}\pi(d\theta) = \frac{\exp(-\lambda R(\theta))}{\mathbb{E}_\pi[\exp(-\lambda R(\theta))]}\pi(d\theta)$$

The following definition is an idea coming from the Bayesian Statistic community. Since we want to approximate the Gibbs posterior $\hat{\rho}_\lambda$ which is not reachable, we fix a family of probability distribution and try to approximate the posterior by a distribution within this family, by minimizing the divergence between this distribution and the posterior.

Definition 5 (Variational-Bayes (VB) approximation of the Gibbs posterior)**.** *The Variational-Bayes approximation of $\hat{\rho}_\lambda$ is defined by :*

$$\tilde{\rho}_\lambda = \arg\min_{\rho\in\mathcal{F}} KL(\rho, \hat{\rho}_\lambda)$$

where $\mathcal{F}$ is a family of probability measures and KL designed the Kullback Leibleir divergence defined for two measures p, q such that $p << q$ by : $KL(p,q) = \int \log(\frac{dp}{dq})dp$

Proposition 6. *For any probability measure $\rho << \pi$, we have :*

$$\log \mathbb{E}_\pi[\exp(-\lambda r_n(\theta))] = -\lambda \mathbb{E}_\rho[r_n(\theta)] - KL(\rho,\pi) + KL(\rho,\hat{\rho}_\lambda) \tag{1}$$

As $KL(\rho,\hat{\rho}_\lambda)$ is minimized by $\tilde{\rho}_\lambda$ (by definition), and the LHS is independent of ρ, then we have that $-\lambda\mathbb{E}_\rho[r_n(\theta)] - KL(\rho,\pi)$ is maximized by $\tilde{\rho}_\lambda$. In other words, we have to find :

$$\tilde{\rho}_\lambda = \arg\min_{\rho\in\mathcal{F}} \{\lambda\mathbb{E}_\rho[r_n(\theta)] + KL(\rho,\pi)\} \tag{2}$$

Proof. We have :

$$\begin{aligned}
&-\lambda\mathbb{E}_\rho[r_n(\theta)] - KL(\rho,\pi) + KL(\rho,\hat{\rho}_\lambda)\\
=&-\lambda\int r_n(\theta)\rho(d\theta) - \int \log\left(\frac{d\rho}{d\pi}d\rho\right) + \int \log\left(\frac{d\rho}{d\hat{\rho}}\right)d\rho\\
=&-\lambda\int r_n(\theta)\rho(d\theta) - \int \log\left(\frac{d\hat{\rho}}{d\pi}\right)d\rho\\
=&-\lambda\int r_n(\theta)\rho(d\theta) - \int \log\left(\frac{\exp(-\lambda r_n(\theta))}{\int\exp(-\lambda rn)d\pi}\right)d\rho\\
=&\int \log\left(\int \exp(-\lambda r_n)d\pi\right)d\rho\\
=&\log\left(\int \exp(-\lambda r_n)d\pi\right)\\
=&\log\mathbb{E}_\pi[\exp(-\lambda r_n(\theta)]
\end{aligned}$$

□

Definition 7 (Hoeffding assumption)**.** *We said that the Hoeffding assumption is satisfied if there exists a function f, an interval $I \subset \mathbb{R}_+^*$ such that for any $\lambda \in I$ and for any $\theta \in \Theta$, we have :*

$$\begin{aligned} &\mathbb{E}_\theta \mathbb{E}[\exp(\lambda(R(\theta) - r_n(\theta))] \leq \exp f(\lambda, n) \\ \textit{and} \quad &\mathbb{E}_\theta \mathbb{E}[\exp(\lambda(r_n(\theta) - R(\theta))] \leq \exp f(\lambda, n) \end{aligned} \tag{3}$$

Note. The original paper (Alquier, Ridgway, and Chopin 2016) provides results under Hoeffding assumptions and Bernstein assumptions. Here, we will only focus on Hoeffding assumptions.

2 Bounds under Hoeffding assumption

In this section, we suppose that the Hoeffding assumption 3 is satisfied. This will lead us to two results: an empirical bound and an oracle-type inequality.

Theorem 8 (Empirical bounds)**.** *For any $\varepsilon > 0$, we have :*

$$\mathbb{P}\left(\forall \rho \in \mathcal{M}_+^1(\Theta) \quad \mathbb{E}_\rho[R(\theta)] \leq \mathbb{E}_\rho[r_n(\theta)] + \frac{f(\lambda, n) + KL(\rho, \pi) + \log(1/\varepsilon)}{\lambda}\right) \geq 1 - \varepsilon$$

Proof. From the definition of Hoeffding's assumption in 3 and Fubini, we have :

$$\begin{aligned} &\mathbb{E}\left[\int \exp(\lambda(R(\theta) - r_n(\theta)) - f(\lambda, n))\pi(d\theta)\right] \leq 1 \\ i.e. \quad &\mathbb{E}\left[\exp \log \int \exp(\lambda(R(\theta) - r_n(\theta)) - f(\lambda, n))\pi(d\theta)\right] \leq 1 \\ i.e. \quad &\mathbb{E}\left[\exp \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \int \lambda(R - r_n) - KL(\rho, \pi) - f(\lambda, n)\right] \leq 1 \\ i.e. \quad &\mathbb{E}\left[\exp \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \int \lambda(R - r_n) - KL(\rho, \pi) - f(\lambda, n) + \log(\varepsilon)\right] \leq \varepsilon \\ i.e. \quad &\mathbb{P}\left[\sup_{\rho \in \mathcal{M}_+^1(\Theta)} \int \lambda(R - r_n) - KL(\rho, \pi) - f(\lambda, n) + \log(\varepsilon) > 0\right] \leq \varepsilon \end{aligned}$$

Where we used the fact that

$$-log \int \exp(-\lambda(R - r_n))d\pi = \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{\int \lambda(R - r_n)d\rho + KL(\rho, \pi)\right\}$$

and that for any random variable $X, \mathbb{E}[\exp(X)] \geq \mathbb{P}(X > 0)$. We switch to the complement to conclude. □

We denote for any set of probability measure $\mathcal{M}$ the following infimum : $\mathcal{B}_\lambda(\mathcal{M}) =$

$\inf_{\rho\in\mathcal{M}}\left\{\mathbb{E}_\rho[R(\theta)]+2\frac{f(\lambda,n)+KL(\rho,\pi)+log(2/\varepsilon)}{\lambda}\right\}$.

The following result allows us to compare $\mathbb{E}_{\hat{\rho}}[R(\theta)]$ to the best possible risk.

Theorem 9 (Oracle-type inequality)**.** *For any $\varepsilon > 0$,*

$$\mathbb{P}\left(\mathbb{E}_{\hat{\rho}}[R(\theta)] \leq \mathcal{B}_\lambda(\mathcal{M}_+^1(\Theta)) \text{ and : } \mathbb{E}_{\tilde{\rho}}[R(\theta)] \leq \mathcal{B}_\lambda(\mathcal{F})\right) \geq 1-\varepsilon$$

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

3 Application to classification

3.1 Framework

We place ourselves in the context of binary classification and linear regression.

In other words, we have $\mathcal{Y}=\{-1,1\}$ and $\forall\theta\in\Theta=\mathcal{X}=\mathbb{R}^d$, and $f_\theta(x)=\mathbb{1}_{\langle\theta,x\rangle>0}$.

We also suppose the couples $\{(X_1,Y_1),\ldots(X_n,Y_n)\}$ iid.

We consider the $0-1$ loss function, so we have $R(\theta)=\mathbb{P}(f_\theta(x)\neq Y)$ and $r_n(\theta)=\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{f_\theta(X_i)\neq Y_i}$.

Lemma 10. *In this framework, the Hoeffding assumption holds with $f(\lambda,n)=\frac{\lambda^2}{2n}$*

In the following, we will consider a Gaussian prior :

$$\pi=\mathcal{N}_d(0,\nu^2 Id)$$

And, we will study three Gaussian families for the variational approximation :

$$\mathcal{F}_1=\{\mathcal{N}(m,\sigma^2 Id), m\in\mathbb{R}^d,\sigma^2\in\mathbb{R}_+^*\} \tag{4}$$
$$\mathcal{F}_2=\{\mathcal{N}(m,\text{diag}(\sigma^2))m\in\mathbb{R}^d,\sigma^2\in(\mathbb{R}_+^*)^d\} \tag{5}$$
$$\mathcal{F}_3=\{\mathcal{N}(m,\Sigma), m\in\mathbb{R}^d,\Sigma\in\mathcal{S}^{d+}\} \tag{6}$$

We note that we have :

$$\mathcal{B}_\lambda(\mathcal{M}_+^1(\Theta))\leq\mathcal{B}_\lambda(\mathcal{F}_3)\leq\mathcal{B}_\lambda(\mathcal{F}_2)\leq\mathcal{B}_\lambda(\mathcal{F}_1) \tag{7}$$

3.2 Theoretical Analysis

If we apply the theorem 8 to this setting, we have the following result. This is a bound for $\mathcal{F}_3$, that can be easily inferred for the others families.

Corollary 11. *For any $\varepsilon > 0$, we have, with probability higher than $1-\varepsilon$, that for any $m \in \mathbb{R}^d$, any $\sigma^2 \in (\mathbb{R}_+)^d$:*

$$\mathbb{E}_{\mathcal{N}(m,\Sigma)}[R(\theta)] \leq \mathbb{E}_{\mathcal{N}(m,\Sigma)}[r_n(\theta)] + \frac{\lambda}{2n} + \frac{\frac{1}{2}(d\log(\nu^2) - \log(|\Sigma|)) + \frac{tr(\Sigma)}{2\nu^2} + \frac{\|m\|^2}{2\nu^2} - \frac{d}{2} + \log(1/\varepsilon)}{\lambda}$$

Proof. We apply 8 with $f(\lambda,n) = \frac{\lambda^2}{2n}$ and $KL(\rho,\pi) = KL(\mathcal{N}(m,\Sigma),\mathcal{N}(0,\nu^2 Id)) = \frac{1}{2}\left[tr((\nu^2)^{-1}\Sigma) - d + m^T(\nu^2)^{-1}m + \log\left(\frac{|\nu^2 Id|}{|\Sigma|}\right)\right]$ □

Definition 12. *We call Assumption $A1$ the assumption that states that there exist a constant $c > 0$ such that for any $(\theta,\theta') \in \Theta^2$ with $\|\theta\| = \|\theta'\| = 1$, we have $\mathbb{P}(\langle X,\theta\rangle\langle X,\theta'\rangle < 0) \leq c\|\theta-\theta'\|$*

Under this assumption, we have the following corollary of the theorem 9.

Corollary 13. *Let us assume that we have $\tilde{\rho}_\lambda$, the VB approximation of $\hat{\rho}_\lambda$ on $\mathcal{F}_1, \mathcal{F}_2$ or $\mathcal{F}_3$. We set $\lambda = \sqrt{nd}$ and $\nu = \frac{1}{\sqrt{d}}$. Then, under the hypothesis A1, we have that for any $\varepsilon > 0$, with probability higher than $1-\varepsilon$:*

$$\mathbb{E}_{\hat{\rho}_\lambda}[R(\theta)] \leq \bar{R} + \sqrt{\frac{d}{n}}\log(4ne) + \frac{c}{\sqrt{n}} + \frac{1}{4n}\sqrt{\frac{d}{n}} + \frac{2\log(2/\varepsilon)}{\sqrt{nd}}$$

$$\text{and}\quad \mathbb{E}_{\tilde{\rho}_\lambda}[R(\theta)] \leq \bar{R} + \sqrt{\frac{d}{n}}\log(4ne) + \frac{c}{\sqrt{n}} + \frac{1}{4n}\sqrt{\frac{d}{n}} + \frac{2\log(2/\varepsilon)}{\sqrt{nd}}$$

Proof. Thanks to 7, we only have to prove the result for the family $\mathcal{F}_1$. By applying the theorem 9, we have :

$$\mathcal{B}_\lambda(\mathcal{F}_1) = \inf_{m,\sigma^2}\left\{\mathbb{E}_{\mathcal{N}(m,\sigma^2}R(\theta) + \frac{\lambda}{n} + 2\frac{\frac{d}{2}\left(\log(\nu^2/\sigma^2) + \sigma^2/\nu^2\right) + \frac{m^T m}{2\nu^2} - \frac{d}{2} + \log(2/\varepsilon)}{\lambda}\right\}$$

Since $\bar{\theta}$ (the minimizer of R) is not unique (since f_θ does not depend on $\|\theta\|$, we can take it such as $\|\theta\| = 1$.Then :

$$\begin{aligned} R(\theta) - \bar{R} &= \mathbb{E}[\mathbb{1}_{\langle\theta,X\angle Y<0} - \mathbb{1}_{\langle\bar{\theta},X\angle Y<0}] \\ &\leq \mathbb{E}[\mathbb{1}_{\langle\theta,X\rangle\langle\theta,\bar{}X\rangle<0}] \\ &= \mathbb{P}(\langle\theta,X\rangle\langle\theta,\bar{}X\rangle < 0) \\ &= \mathbb{P}(\langle\frac{\theta}{\|\theta\|},X\rangle\langle\theta,\bar{}X\rangle < 0) \\ &\leq 2c\|\theta,\bar{\theta}\| && \text{by assumption } A1 \end{aligned}$$

Therefore :

$$\mathcal{B}_\lambda(\mathcal{F}_1) \leq \bar{R} + \inf_{m,\sigma^2} \left\{ 2c\mathbb{E}_{\mathcal{N}(m,\sigma^2)}[\|\theta - \bar{\theta}\|] + \frac{\lambda}{n} + 2\frac{\frac{d}{2}\left(\log(\nu^2/\sigma^2) + \sigma^2/\nu^2\right) + \frac{m^T m}{2\nu^2} - \frac{d}{2} + \log(2/\varepsilon)}{\lambda} \right\}$$

Finally, we restrict the infimum to distributions ν such as $m = \bar{\theta}$, and set $\sigma = \frac{1}{2\lambda}$ and $\nu = \frac{1}{\sqrt{d}}$ to conclude. □

3.3 Implementation

For the family F_3, the variational bound to be maximized is :

$$\mathcal{L}_{\lambda,\nu}(m,\Sigma) = \frac{-\lambda}{n}\sum_{i=1}^{n} \Phi\left(-Y_i \frac{X_i m}{\sqrt{X_i \Sigma X_i^T}}\right) - \frac{m^T m}{2\nu^2} + \frac{1}{2}\left(\log(|\Sigma|) - \frac{1}{\nu^2}\mathrm{tr}(\Sigma)\right)$$

Note that the same kind of results for the families $\mathcal{F}_1$ and $\mathcal{F}_2$ can be readily inferred from this case.

Proof. Let us remind our optimization problem as stated in 2 :

$$\tilde{\rho}_\lambda = \arg\min_{\rho \in \mathcal{F}_3} \{\lambda \mathbb{E}_\rho[r_n(\theta)] + KL(\rho, \pi)\}$$

Here, with $\pi = \mathcal{N}(0, \nu^2 Id)$ and $\rho = \mathcal{N}(m, \Sigma)$ (with $m \in R^d$, $\Sigma \in \mathcal{S}^{d+}$), we have that :

$$\begin{aligned} KL(\rho,\pi) &= \frac{1}{2}\left(tr((\nu^2 Id)^{-1}\Sigma) - d + m^T((\nu^2 Id)^{-1})m + \log\left(\frac{det((\nu^2 Id))}{det(\Sigma)}\right)\right) \\ &= \frac{1}{2}\left(\frac{1}{\nu^2}tr(\Sigma) - d + \frac{m^T m}{\nu^2} + \log\left(\frac{\prod_{i=1}^{d}\nu^2}{det(\Sigma)}\right)\right) \\ &= \frac{m^T m}{2\nu^2} + \frac{1}{2}\left(\frac{1}{\nu^2}tr(\Sigma) - \log(det(\Sigma))\right) - \frac{d}{2} + \frac{d}{2}\log(\nu^2) \end{aligned}$$

On the other hand, we have :

$$\begin{aligned}
\mathbb{E}_\rho[r_n(\theta)] &= \mathbb{E}_\rho\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{Y_i\langle\theta,X_i\rangle<0}\right]\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_\rho[\langle\theta,X_i\rangle<0]\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}[Y_i\langle m+\Sigma Z,X_i\rangle<0] \quad \text{where } Z\sim\mathcal{N}(0,Id)\\
&= \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{-Y_i\langle X_i,m\rangle}{\sqrt{\langle X_i,\Sigma X_i\rangle}}\right)
\end{aligned}$$

Combining these results, our optimization problem boils down to:

$$\begin{aligned}
&\underset{m\in R^d,\Sigma\in\mathcal{S}^{d+}}{\arg\min}\left\{\frac{\lambda}{n}\sum_{i=1}^{n}\Phi\left(-Y_i\frac{X_im}{\sqrt{X_i\Sigma X_i^T}}\right)+\frac{m^Tm}{2\nu^2}+\frac{1}{2}\left(\frac{1}{\nu^2}tr(\Sigma)-\log(det(\Sigma))\right)\underbrace{-\frac{d}{2}+\frac{d}{2}\log(\nu^2)}_{\text{const. w.r.t. } m \text{ and } \Sigma}\right\}\\
&=\underset{m\in R^d,\Sigma\in\mathcal{S}^{d+}}{\arg\min}\left\{\frac{\lambda}{n}\sum_{i=1}^{n}\Phi\left(-Y_i\frac{X_im}{\sqrt{X_i\Sigma X_i^T}}\right)+\frac{m^Tm}{2\nu^2}+\frac{1}{2}\left(\frac{1}{\nu^2}tr(\Sigma)-\log(det(\Sigma))\right)\right\}
\end{aligned}$$

□

3.4 Results

We can find a solution to our optimization problem numerically (it is difficult to have any guarantee since the problem is not convex but we can still try).

And we can compute the bound as in 11 from this solution.

If the bound is low, it is good news, and it means that we will not lose anything if we use this method for prediction. On the other hand, if the bound is large, we cannot tell anything.

The authors provided an R package [1] but it seems like it is no longer actively maintained or updated. All the following numerical experiences have been remade from scratch, in Python with `torch` library.

We implemented the algorithm in Python and used optimization techniques provided by the package `torch`. We restrict ourselves in the algorithms to the family $\mathcal{F}_1$ for

[1] PACVB package: https://cran.r-project.org/web/packages/PACVB/index.html

simplicity. The parameters ν and λ are taken as described in the corollary 13.

We applied it to some datasets from UCI [2]. We compared our results with a logistic regression and SVM with a sigmoid kernel.

We compare our classification rates to the logistic regression and the SVM technique. The results are quite satisfying, and we obtained interesting VB bounds. They are displayed in Table 1.

	VB	VB-Bound	LR	SVM
Breast Cancer	0.18	0.04	0.36	0.35
Spect Heart	0.38	0.075	0.36	0.27
Students Dropout	0.34	0.013	0.43	0.44

Table 1: Misclassification rates on several datasets using different techniques: the VB approach, the logistic regression (LR), and an SVM with a sigmoid kernel (SVM). The VB bound is also displayed in the second column.

4 Classification with convexified loss

In this section, we will focus on convex classification that allows us to have an easier optimization problem.

We consider the same setting as before, with an independent Gaussian prior $\pi = \mathcal{N}(0, \nu^2 Id)$

4.1 Hinge Loss

Here we consider the Hinge loss defined by: $h(x) = \max(0, 1 - x)$.

In our context, we will have the empirical and the expected risk written as :

$$
\begin{aligned}
r_n^H(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - Y_i\langle\theta, X_i\rangle) \\
R^H(\theta) &= \mathbb{E}[\max(0, 1 - Y\langle\theta, X\rangle)]
\end{aligned}
$$

We also assume that the X_i's are uniformly bounded almost surely : $\|X_i\| < c_x$ with $c_x > 0$

In this context, we have the following result :

Lemma 14. *The Hoeffding assumption (defined in 3) is verified with* $f(\lambda, n) = \frac{\lambda^2}{4n} - \frac{1}{2}\log\left(1 - \frac{\nu^2\lambda^2 c_x^2}{2n}\right)$ *for* $\lambda < \frac{1}{c_x}\sqrt{\frac{n^2}{\nu}}$

[2] UCI: https://archive.ics.uci.edu/datasets

We note that the bound in the hypothesis of this proposition means that the prior variance ν must not be taken too big relative to λ.

We have the following oracle inequality :

Corollary 15. *Let us assume that we have $\tilde{\rho}_\lambda$, the VB approximation of $\hat{\rho}_\lambda$ on $\mathcal{F}_1, \mathcal{F}_2$ or $\mathcal{F}_3$. We set $\lambda = \sqrt{nd}/c_x$ and $\nu = \frac{1}{\sqrt{d}}$. Then, for any $\varepsilon > 0$, with probability higher than $1 - \varepsilon$:*

$$\mathbb{E}_{\hat{\rho}_\lambda}[R^H(\theta)] \leq \bar{R^H} + \frac{c_x}{2}\sqrt{\frac{d}{n}}\log\left(\frac{n}{d}\right) + c_x\frac{d}{n}\sqrt{\frac{d}{n}} + \frac{1}{\sqrt{nd}}\left(\frac{2c_x^2+1}{2c_x} + 2c_x\log(2/\varepsilon)\right)$$

and $$\mathbb{E}_{\tilde{\rho}_\lambda}[R^H(\theta)] \leq \bar{R^H} + \frac{c_x}{2}\sqrt{\frac{d}{n}}\log\left(\frac{n}{d}\right) + c_x\frac{d}{n}\sqrt{\frac{d}{n}} + \frac{1}{\sqrt{nd}}\left(\frac{2c_x^2+1}{2c_x} + 2c_x\log(2/\varepsilon)\right)$$

Remark 16 (Link between the excess risk under the hinge loss and the 0-1 loss)**.** *For any $\theta \in \mathbb{R}^d$, we have:*

$$R(\theta) - R^\star \leq R^H(\theta) - R^{H\star}$$

In other words, the estimator computed under the hinge loss bounds the excess risk of the 0-1 loss.

The following result is strong and interesting as it allows us access to an error level for "any" iteration of our algorithm (and not at the end or after a sufficient amount of iterations) :

Theorem 17. *Let us presume that the same hypotheses as those employed in the preceding corollary remain valid. Let us denote $\tilde{\rho}_k(d\theta)$ the VB approximation measure at the k^{th} iteration of the solver (using the hinge loss). Let us set $M > 0$ large enough so that the approximated mean of variance $\bar{m}, \bar{\Sigma}$ are at a distance at most M from the initial values of the solver. We also set $\lambda = \sqrt{nd}$ and $\nu = \frac{1}{\sqrt{d}}$.*

Then, with probability higher than $1 - \varepsilon$, we have :

$$\mathbb{E}_{\tilde{\rho}_k}[R^H(\theta)] \leq \bar{R}^H + \frac{LM}{\sqrt{1+k}} + \frac{c_x}{2}\frac{d}{n}\log\left(\frac{n}{d}\right) + c_x\frac{d}{n}\sqrt{\frac{d}{n}} + \frac{1}{\sqrt{nd}}\left(\frac{2c_x^2+1}{2c_x} + 2c_x\log(2/\varepsilon)\right)$$

Implementation and results

In this part, we will only consider the family $\mathcal{F}_1$, but the reasoning would be the same for the two others. Let us note $\rho_{m,\sigma}$ a distribution in $\mathcal{F}_1$.

For a fixed individual (X_i, Y_i), the hinge risk is given by :

$$
\begin{aligned}
\mathbb{E}_{\rho_{m,\sigma}}[r_i(\theta)] &= \mathbb{E}_{\rho_{m,\sigma}}[\max(0, 1 - Y_i\langle\theta, X_i\rangle] \\
&= \mathbb{E}_{\rho_{m,\sigma}}[\mathbb{1}_{1 \leq Y_i\langle\theta, X_i\rangle}(1 - Y_i\langle\theta, X_i\rangle)] \\
&= \int_{\mathbb{R}^d} \mathbb{1}_{1 \leq Y_i\langle\theta, X_i\rangle}(1 - Y_i\langle\theta, X_i\rangle)\varphi\left(\frac{\theta - m}{\sigma}\right)d\theta \\
&= \int_{\mathbb{R}^d} \mathbb{1}_{1 \leq Y_i\langle m + \sigma z, X_i\rangle}(1 - Y_i\langle m + \sigma z, X_i\rangle)\varphi(z))dz \\
&= (1 - Y_iX_im)\int_{\frac{1 - Y_iX_im}{\sigma\|Y_iX_i\|}}^{\infty} \varphi(z)dz - \sigma\|Y_iX_i\| \int_{\frac{1 - Y_iX_im}{\sigma\|Y_iX_i\|}}^{\infty} \overbrace{z\varphi(z)}^{=\varphi'(z)}dz \\
&= (1 - Y_iX_im)\varPhi\left(\frac{1 - Y_iX_im}{\sigma\|Y_iX_i\|}\right) + \sigma\|Y_iX_i\|\varphi\left(\frac{1 - Y_iX_im}{\sigma\|Y_iX_i\|}\right)
\end{aligned}
$$

Therefore, our optimization problem (which has the nice property of being convex) is :

$$
\begin{aligned}
\operatorname*{arg\,min}_{m \in \mathbb{R}^d, \sigma \in \mathbb{R}_+^\star} \Bigg\{ &\frac{\lambda}{n}\sum_{i=1}^{n}(1 - Y_iX_im)\varPhi\left(\frac{1 - Y_iX_im}{\sigma\|Y_iXi_{\|}}\right) + \sum_{i=1}^{n}\left[\sigma\|Y_iXi_{\|}\varphi\left(\frac{1 - Y_iX_im}{\sigma\|Y_iX_i\|}\right)\right] \\
&+ \frac{m^Tm}{2\nu^2} - \frac{d}{2}\left(\log(\sigma^2) - \frac{\sigma^2}{\nu^2}\right)\Bigg\}
\end{aligned}
$$

Numerical Results

The misclassification rates and hinge losses are displayed in 2 and 3 respectively.

	VB	LR	SVM
Breast Cancer	0.017	0.36	0.35
Spect Heart	0.14	0.36	0.27
Students Dropout	0.15	0.43	0.44

Table 2: Misclassification rates on several datasets using different techniques: the VB approach, the logistic regression (LR), and a SVM with a sigmoid kernel (SVM)

	VB	LR	SVM
Breast Cancer	0.035	0.72	0.71
Spect Heart	0.29	0.73	0.55
Students Dropout	0.30	0.86	0.88

Table 3: Hinge Losses on several datasets using different techniques: the VB approach, the logistic regression (LR), and an SVM with a sigmoid kernel (SVM)

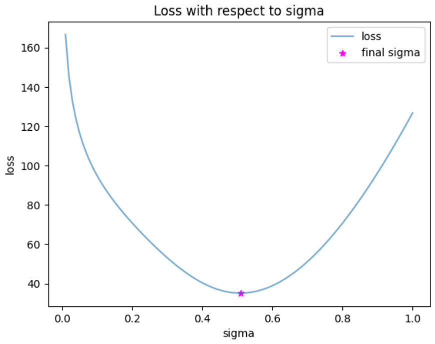

Figure 1: Illustration of the convexity of the optimization problem. This is the loss w.r.t. σ on the Breast Dataset.

4.2 Exponential Loss

The authors of the papers also mentioned the exponential loss to have a convex optimization problem but did not explicit the calculations and implementation details. We do it in this part.

Let us consider the exponential loss: $l(\hat{y}, y) = \exp(-y\hat{y})$.

Then, we have the empirical risk associated defined as: $r_n^E(\theta) = \frac{1}{n}\sum_{i=1}^n \exp(-Y_i\langle\theta, X_i\rangle)$ and the expected risk: $R^E(\theta) = \mathbb{E}[\exp(Y\langle\theta, X\rangle)]$.

Here again, we will consider the family $\mathcal{F}_1$, and for any probability measure $\rho_{m,\sigma} \in \mathcal{F}_1$, we have the empirical of each individual :

$$\begin{aligned}
\mathbb{E}_{\rho_{m,\sigma}}[r_i^E(\theta)] &= \mathbb{E}_{\rho_{m,\sigma}}[\exp(-Y_i\langle\theta, X_i\rangle)] \\
&= \mathbb{E}[\exp(-Y_iX_im - \sigma\|Y_iX_i\|Z)] \quad \text{where } Z \sim \mathcal{N}(0, Id) \\
&= \exp(-Y_iX_im)\mathbb{E}[\exp(\sigma\|Y_iX_i\|Z)] \\
&= \exp\left(-Y_iX_im + \frac{1}{2}\sigma^2\|Y_iX_i\|^2\right)
\end{aligned}$$

(Where we used the fact that the moment of a Gaussian $X \sim \mathcal{N}(\mu, \Sigma)$ can be written as : $\mathbb{E}[e^{tX}] = \exp(\mu^T t + \frac{1}{2}t^T\Sigma t))$

Finally, our optimization problem is :

$$\underset{m\in\mathbb{R}^d, \sigma\in\mathbb{R}_+^\star}{\arg\min}\left\{\frac{\lambda}{n}\sum_{i=1}^n\left(\exp\left(-Y_iX_im + \frac{1}{2}\sigma^2\|Y_iX_i\|^2\right)\right) + \frac{m^Tm}{2\nu^2} - \frac{d}{2}\left(\log(\sigma^2) - \frac{\sigma^2}{\nu^2}\right)\right\}$$

Numerical Results

The numerical results of our experiments are displayed in Tables 4 and 5.

	VB	LR	SVM
Breast Cancer	0.087	0.36	0.35
Spect Heart	0.12	0.36	0.27
Students Dropout	0.30	0.43	0.44

Table 4: Misclassification rates on several datasets using different techniques: the VB approach, the logistic regression (LR), and an SVM with a sigmoid kernel (SVM)

	VB	LR	SVM
Breast Cancer	0.57	1.21	1.20
Spect Heart	0.68	1.23	1.02
Students Dropout	1.12	1.38	1.40

Table 5: Exponential losses on several datasets using different techniques: the VB approach, the logistic regression (LR), and an SVM with a sigmoid kernel (SVM)

5 Application to ranking

Now, let us consider the ranking problem. Here, we have $\mathcal{Y} = \{0, 1\}$, $\Theta = \mathcal{X} = \mathbb{R}^d$ and we consider linear classifiers : $f_\theta(x, x') = -1 + 2\mathbb{1}_{\langle\theta,x\rangle>\langle\theta,x'\rangle}$.

In the following, we will denote : $n_1 = \#\{i \in \{1, \ldots, n\} : Y_i = 1\}$ and $n_0 = \#\{i \in \{1, \ldots, n\} : Y_i = 0\}$

We consider a Gaussian prior π of the form : $\pi(d\theta) = \prod_{i=1}^{d} \varphi(\theta_i; 0, \nu^2)d\theta_i$

Then, the empirical risk and the expected risk can be written as :

$$\begin{aligned} r_n(\theta) &= \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{1}_{(Y_i - Y_j)f_\theta(X_i - X_j)<0} \\ R(\theta) &= \mathbb{P}((Y_1 - Y_2)f_\theta(X_1, X_2) < 0) \end{aligned}$$

Proposition 18. *In this setting, Hoeffding's assumption is verified with $f(\lambda, n) = \frac{\lambda^2}{n-1}$.*

If we apply the theorem 8, we have the following empirical bound :

Corollary 19. *For any $\varepsilon > 0$, with probability higher than $1 - \varepsilon$, we have that for any $m \in \mathbb{R}^d$, any $\sigma^2 \in \mathbb{R}_+^d$:*

$$\mathbb{E}_{\mathcal{N}(m,\sigma^2)}[R(\theta)] \leq \mathbb{E}_{\mathcal{N}(m,\sigma^2)}[r_n(\theta)] + \frac{\lambda}{n-1} + \frac{\frac{1}{2}\sum_{i=1}^{d}\left(\log\left(\frac{\nu^2}{\sigma_i^2}\right) + \frac{\sigma_i^2}{\nu^2}\right) + \frac{m^T m}{2\nu^2} - \frac{d}{2} + \log(1/\varepsilon)}{\lambda}$$

Here again, we will need an assumption to get a theoretical bound. It is the assumption $A1$ applied to $(X_1 - X_2)$:

Definition 20. *We call assumption $A2$ the assumption that states that there exists* *a constant* $c > 0$ *such that for any* $(\theta, \theta') \in \Theta^2$ *with* $\|\theta\| = \|\theta'\| = 1$, *we have* $\mathbb{P}(\langle X_1, X_2, \theta\rangle\langle X_1, X_2, \theta'\rangle < 0) \leq c\|\theta - \theta'\|$

Under this assumption, we have the following corollary of the theorem 9 :

Corollary 21. *Let us consider the family* $\mathcal{F}_1, \mathcal{F}_2$ *to* $\mathcal{F}_3$, *and set* $\lambda = \frac{\sqrt{d(n-1}}{2}$ *and* $\nu = 1$. *Then, under hypothesis* $A2$, *we have for any* ε, *with probability higher than* $1 - \varepsilon$ *:*

$$\mathbb{E}_{\hat{\rho}_\lambda}[R(\theta)] \leq \bar{R} + \sqrt{\frac{2d}{n-1}}\left(1 + \frac{1}{2}\log(2d(n-1))\right) + \frac{c\sqrt{2}}{\sqrt{n-1}} + \frac{1}{(n-1)^{\frac{3}{2}}\sqrt{2d}} + \frac{2\sqrt{2}\log(\frac{2e}{\varepsilon})}{\sqrt{(n-1)d}}$$

$$\mathbb{E}_{\tilde{\rho}_\lambda}[R(\theta)] \leq \bar{R} + \sqrt{\frac{2d}{n-1}}\left(1 + \frac{1}{2}\log(2d(n-1))\right) + \frac{c\sqrt{2}}{\sqrt{n-1}} + \frac{1}{(n-1)^{\frac{3}{2}}\sqrt{2d}} + \frac{2\sqrt{2}\log(\frac{2e}{\varepsilon})}{\sqrt{(n-1)d}}$$

5.1 Implementation

In this context, and considering the family $\mathcal{F}_2$, our optimization problem boils down to:

$$\underset{m\in\mathbb{R}^d, \sigma\in\mathbb{R}_+^{\star d}}{\arg\min}\left\{\frac{\lambda}{n_1 n_0}\sum_{\substack{i:y_i=1\\ j:y_j=0}} \Phi\left(-\frac{(Xi - X_j)m}{\sqrt{\sum_{k=1}^d (X_i - X_j)_k^2 \sigma_k^2}}\right) + \frac{m^T m}{2\nu^2} - \frac{1}{2}\sum_{i=1}^d (\log\sigma_i^2 - \frac{\sigma_i^2}{\nu^2})\right\}$$

Again, we can apply optimization algorithms such as a Gradient Descent or Stochastic Descent, and find the optimal m and σ^2.

6 Conclusion and discussions

In this work, we derived practical algorithms for swiftly computing these Variational Bayes approximations. Also, we got empirical measures derived from the data to evaluate the performance of the resulting VB-approximated procedure. We also saw that approximating a Gibbs posterior using VB techniques does not worsen the convergence rate.

Our experiments focused on the classification tasks since the reasoning for ranking would have been the same. However, a few things could have been done in our experiments such as comparison of performance between families. Besides, it would have been interesting to compare the performance of VB approximation to Markov Chain Monte Carlo algorithms such as Sequential Monte Carlo as suggested in the original paper.

Bibliography

Alquier, Pierre (Mar. 2023). *User-friendly introduction to PAC-Bayes bounds*. en. arXiv:2110.11216 [cs, math, stat]. URL: http://arxiv.org/abs/2110.11216.

Alquier, Pierre and Karim Lounici (Jan. 2011). "PAC-Bayesian bounds for sparse regression estimation with exponential weights". en. In: *Electronic Journal of Statistics* 5.none. ISSN: 1935-7524. DOI: 10.1214/11-EJS601. URL: https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-5/issue-none/PAC-Bayesian-bounds-for-sparse-regression-estimation-with-exponential-weights/10.1214/11-EJS601.full.

Alquier, Pierre, James Ridgway, and Nicolas Chopin (Dec. 2016). "On the properties of variational approximations of Gibbs posteriors". en. In.

Catoni, Olivier (2004). "A PAC-Bayesian approach to adaptive classification". en. In.

Dziugaite, Gintare Karolina and Daniel M. Roy (Oct. 2017). *Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data*. en. arXiv:1703.11008 [cs]. URL: http://arxiv.org/abs/1703.11008.

Jiang, Wenxin and Martin A. Tanner (Oct. 2008). "Gibbs posterior for variable selection in high-dimensional classification and data mining". en. In: *The Annals of Statistics* 36.5. ISSN: 0090-5364. DOI: 10.1214/07-AOS547. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-36/issue-5/Gibbs-posterior-for-variable-selection-in-high-dimensional-classification-and/10.1214/07-AOS547.full.

Khan, Mohammad Emtiyaz E (n.d.). "Decoupled Variational Gaussian Inference". en. In: ().