Journal of Machine Learning Research 17 (2016) 1-41 Submitted 6/15; Revised 2/16; Published 12/16

On the properties of variational approximations of Gibbs posteriors

Pierre Alquier PIERRE.ALQUIER@ENSAE.FR
James Ridgway JAMES.RIDGWAY@ENSAE.FR
Nicolas Chopin NICOLAS.CHOPIN@ENSAE.FR
ENSAE
3 Avenue Pierre Larousse
92245 MALAKOFF, FRANCE

Editor: Yee Whye Teh

Abstract

The PAC-Bayesian approach is a powerful set of techniques to derive non-asymptotic risk bounds for random estimators. The corresponding optimal distribution of estimators, usually called the Gibbs posterior, is unfortunately often intractable. One may sample from it using Markov chain Monte Carlo, but this is usually too slow for big datasets. We consider instead variational approximations of the Gibbs posterior, which are fast to compute. We undertake a general study of the properties of such approximations. Our main finding is that such a variational approximation has often the same rate of convergence as the original PAC-Bayesian procedure it approximates. In addition, we show that, when the risk function is convex, a variational approximation can be obtained in polynomial time using a convex solver. We give finite sample oracle inequalities for the corresponding estimator. We specialize our results to several learning tasks (classification, ranking, matrix completion), discuss how to implement a variational approximation in each case, and illustrate the good properties of said approximation on real datasets.

1. Introduction

A Gibbs posterior, also known as a PAC-Bayesian or pseudo-posterior, is a probability distribution for random estimators of the form:

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

More precise definitions will follow, but for now, θ may be interpreted as a parameter (in a finite or infinite-dimensional space), $r_n(\theta)$ as an empirical measure of risk (e.g. prediction error), and $\pi(\mathrm{d}\theta)$ a prior distribution.

We will follow in this paper the PAC (Probably Approximatively Correct)-Bayesian approach, which originates from machine learning (Shawe-Taylor and Williamson, 1997;

McAllester, 1998; Catoni, 2004); see Catoni (2007) for an exhaustive study, and Jiang and Tanner (2008); Yang (2004); Zhang (2006); Dalalyan and Tsybakov (2008) for related perspectives (such as the aggregation of estimators in the last three papers). There, $\hat{\rho}_\lambda$ appears as the probability distribution that minimizes the upper bound of an oracle inequality on the risk of *random* estimators. The PAC-Bayesian approach offers sharp theoretical guarantees on the properties of such estimators, without assuming a particular model for the data generating process.

The Gibbs posterior has also appeared in other places, and under different motivations: in Econometrics, as a way to avoid direct maximization in moment estimation (Chernozhukov and Hong, 2003); and in Bayesian decision theory, as a way to define a Bayesian posterior distribution when no likelihood has been specified (Bissiri et al., 2013). Another well-known connection, although less directly useful (for Statistics), is with thermodynamics, where r_n is interpreted as an energy function, and λ as the inverse of a temperature.

Whatever the perspective, estimators derived from Gibbs posteriors usually show excellent performance in diverse tasks, such as classification, regression, ranking, and so on, yet their actual implementation is still far from routine. The usual recommendation (Dalalyan and Tsybakov, 2012; Alquier and Biau, 2013; Guedj and Alquier, 2013) is to sample from a Gibbs posterior using MCMC (Markov chain Monte Carlo, see e.g. Green et al., 2015); but constructing an efficient MCMC sampler is often difficult, and even efficient implementations are often too slow for practical uses when the dataset is very large.

In this paper, we consider instead VB (Variational Bayes) approximations, which have been initially developed to provide fast approximations of 'true' posterior distributions (i.e. Bayesian posterior distributions for a given model); see Jordan et al. (1999); MacKay (2002) and Chap. 10 in Bishop (2006).

Our main results are as follows: when PAC-Bayes bounds are available - mainly, when a strong concentration inequality holds - replacing the Gibbs posterior by a variational approximation does not affect the rate of convergence to the best possible prediction, on the condition that the Küllback-Leibler divergence between the posterior and the approximation is itself properly controlled. Furthermore, for convex risks we show that one can obtain polynomial time algorithms based on optimal convex solvers.

We also provide empirical bounds, which may be computed from the data to ascertain the actual performance of estimators obtained by variational approximation. All the results gives strong incentives, we believe, to recommend Variational Bayes as the default approach to approximate Gibbs posteriors. We also provide a R package[1], written in C++ to compute a Gaussian variational approximation in the case of the hinge risk.

The rest of the paper is organized as follows. In Section 2, we present the notations and assumptions. In Section 3, we introduce variational approximations and the corresponding algorithms. The main results are provided in a general form in Section 4: in Subsection 4.1, we give results under the assumption that a Hoeffding type inequality holds (slow rates) and

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

in Subsection 4.2, we give results under the assumption that a Bernstein type inequality holds (fast rates). Note that for the sake of brevity, we will refer to these settings as "Hoeffding assumption" and "Bernstein assumption" even though this terminology is non-standard. We then apply these results in various settings: classification (Section 5), convex classification (Section 6), ranking (Section 7), and matrix completion (Section 8). In each case, we show how to specialise the general results of Section 4 to the considered application, in order to obtain the properties of the VB approximation, and we also discuss its numerical implementation. All the proofs are collected in the Appendix.

2. PAC-Bayesian framework

We observe a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, taking values in $\mathcal{X} \times \mathcal{Y}$, where the pairs (X_i, Y_i) have the same distribution P. We will assume explicitly that the (X_i, Y_i)'s are independent in several of our specialised results, but we do not make this assumption at this stage, as some of our general results, and more generally the PAC-Bayesian theory, may be extended to dependent observations; see e.g. Alquier and Li (2012). The label set $\mathcal{Y}$ is always a subset of $\mathbb{R}$. A set of predictors is chosen by the statistician: $\{f_\theta : \mathcal{X} \to \mathbb{R},\, \theta \in \Theta\}$. For example, in linear regression, we may have: $f_\theta(x) = \langle \theta, x \rangle$, the inner product of $\mathcal{X} = \mathbb{R}^d$, while in classification, one may have $f_\theta(x) = \mathbb{I}_{\langle \theta, x \rangle > 0} \in \{0, 1\}$.

We assume we have at our disposal a risk function $R(\theta)$; typically $R(\theta)$ is a measure of the prediction error. We set $\overline{R} = R(\overline{\theta})$, where $\overline{\theta} \in \arg\min_\Theta R$; i.e. $f_{\overline{\theta}}$ is an optimal predictor. We also assume that the risk function $R(\theta)$ has an empirical counterpart $r_n(\theta)$, and set $\overline{r}_n = r_n(\overline{\theta})$. Often, R and r_n are based on a loss function $\ell : \mathbb{R}^2 \to \mathbb{R}$; i.e. $R(\theta) = \mathbb{E}[\ell(Y, f_\theta(X))]$ and $\overline{r}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f_\theta(X_i))$. (In this paper, the symbol $\mathbb{E}$ will always denote the expectation with respect to the (unknown) law P of the (X_i, Y_i)'s.) There are situations however (e.g. ranking), where R and r_n have a different form.

We define a prior probability measure $\pi(\cdot)$ on the set Θ (equipped with the standard σ-algebra for the considered context), and we let $\mathcal{M}_+^1(\Theta)$ denote the set of all probability measures on Θ.

Definition 2.1 *We define, for any $\lambda > 0$, the pseudo-posterior $\hat{\rho}_\lambda$ by*

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

The pseudo-posterior $\hat{\rho}_\lambda$ (also known as the Gibbs posterior, Catoni (2004, 2007), or the exponentially weighted aggregate, Dalalyan and Tsybakov (2008)) plays a central role in the PAC-Bayesian approach. It is obtained as the distribution that minimizes the upper bound of a certain oracle inequality applied to *random* estimators. Practical estimators (predictors) may be derived from the pseudo-posterior, by e.g. taking the expectation, or sampling from it. Of course, when $\exp[-\lambda r_n(\theta)]$ may be interpreted as the likelihood of a

certain model, $\hat{\rho}_\lambda$ becomes a Bayesian posterior distribution, but we will not restrict our attention to this particular case.

The following 'theoretical' counterpart of $\hat{\rho}_\lambda$ will prove useful to state results.

Definition 2.2 *We define, for any $\lambda > 0$, π_λ as*

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

We will derive PAC-Bayesian bounds on predictions obtained by variational approximations of $\hat{\rho}_\lambda$ under two types of assumptions: a Hoeffding-type assumption, from which we may deduce slow rates of convergence (Subsection 4.1), and a Bernstein-type assumption, from which we may obtain fast rates of convergence (Subsection 4.2).

Definition 2.3 *We say that a Hoeffding assumption is satisfied for prior π when there is a function f and an interval $I \subset \mathbb{R}_+^*$ such that, for any $\lambda \in I$, for any $\theta \in \Theta$,*

$$\left.\begin{array}{l} \pi\left(\mathbb{E}\exp\left\{\lambda[R(\theta) - r_n(\theta)]\right\}\right) \\ \pi\left(\mathbb{E}\exp\left\{\lambda[r_n(\theta) - R(\theta)]\right\}\right) \end{array}\right\} \leq \exp\left[f(\lambda, n)\right]. \tag{1}$$

Inequality (1) can be interpreted as an integrated version (with respect to π) of Hoeffding's inequality, for which $f(\lambda, n) \asymp \lambda^2/n$. In many cases the loss will be bounded uniformly over θ; then Hoeffding's inequality will directly imply (1). The expectation with respect to π in (1) allows us to treat some cases where the loss is not upper bounded by specifying a prior with sufficiently light tails.

Definition 2.4 *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 $\theta \in \Theta$,*

$$\left.\begin{array}{l} \pi\left(\mathbb{E}\exp\left\{\lambda[R(\theta) - \overline{R}] - \lambda[r_n(\theta) - \overline{r}_n]\right\}\right) \\ \pi\left(\mathbb{E}\exp\left\{\lambda[r_n(\theta) - \overline{r}_n] - \lambda[R(\theta) - \overline{R}]\right\}\right) \end{array}\right\} \leq \pi\left(\exp\left[g(\lambda, n)[R(\theta) - \overline{R}]\right]\right). \tag{2}$$

This assumption is satisfied for example by sums of i.i.d. sub-exponential random variables, see Subsection 2.4 p. 27 in Boucheron et al. (2013), when a margin assumption on the function $R(\cdot)$ is satisfied (Tsybakov, 2004). This is discussed in Section 4.2. Again, extensions beyond the i.i.d. case are possible, see e.g. Wintenberger (2010) for a survey and new results. In all these examples, the important feature of the function g that we will use to derive rates of convergence is the fact that there is a constant $c > 0$ such that when $\lambda = cn$, $g(\lambda, n) = g(cn, n) \asymp n$.

As mentioned previously, we will often consider $r_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f_\theta(X_i))$, however, the previous assumptions can also be satisfied when $r_n(\theta)$ is a U-statistic, using Hoeffding's decomposition of U-statistics combined with the corresponding inequality for sums of independent variables (Hoeffding, 1948). This idea comes from Clémençon et al. (2008) and we will use it in our ranking application.

Remark 2.1 *We could consider more generally inequalities of the form*

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

that allow using the more general form of the margin assumption of Mammen and Tsybakov (1999); Tsybakov (2004). PAC-Bayes bounds in this context are provided by Catoni (2007). However, the techniques involved would require many pages to be described so we decided to focus on the cases $\kappa=0$ and $\kappa=1$ to keep the exposition simple.

3. Numerical approximations of the pseudo-posterior

3.1 Monte Carlo

As already explained in the introduction, the usual approach to approximate $\hat{\rho}_\lambda$ is MCMC (Markov chain Monte Carlo) sampling. Ridgway et al. (2014) proposed tempering SMC (Sequential Monte Carlo, e.g. Del Moral et al. (2006)) as an alternative to MCMC to sample from Gibbs posteriors: one samples sequentially from $\hat{\rho}_{\lambda_t}$, with $0=\lambda_0<\cdots<\lambda_T=\lambda$ where λ is the desired temperature. One advantage of this approach is that it makes it possible to contemplate different values of λ, and choose one by e.g. cross-validation. Another advantage is that such an algorithm requires little tuning; see Appendix B for more details on the implementation of tempering SMC. We will use tempering SMC as our gold standard in our numerical studies.

SMC and related Monte Carlo algorithms tend to be too slow for practical use in situations where the sample size is large, the dimension of Θ is large, or f_θ is expensive to compute. This motivates the use of fast, deterministic approximations, such as Variational Bayes, which we describe in the next section.

3.2 Variational Bayes

Various versions of VB (Variational Bayes) have appeared in the literature, but the main idea is as follows. We define a family $\mathcal{F}\subset\mathcal{M}_+^1(\Theta)$ of probability distributions that are considered as tractable. Then, we define the VB-approximation of $\hat{\rho}_\lambda$: $\tilde{\rho}_\lambda$.

Definition 3.1 *Let*

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

where $\mathcal{K}(\rho,\hat{\rho}_\lambda)$ denotes the KL (Küllback-Leibler) divergence of $\hat{\rho}_\lambda$ relative to ρ: $\mathcal{K}(m,\mu)=\int\log[\frac{\mathrm{d}m}{\mathrm{d}\mu}]\mathrm{d}m$ if $m\ll\mu$ (i.e. μ dominates m), $\mathcal{K}(m,\mu)=+\infty$ otherwise.

The difficulty is to find a family $\mathcal{F}$ (a) which is large enough, so that $\tilde{\rho}_\lambda$ may be close to $\hat{\rho}_\lambda$, and (b) such that computing $\tilde{\rho}_\lambda$ is feasible. Moreover, even when there are algorithms for $\tilde{\rho}_\lambda$ that are efficient in practice, we may, depending on the problem at hand, have more

or less strong guarantees on the quality of the optimization. For example, while in Section 6 we consider a setting where an exact upper bound for the optimization error is available, in Section 8 this is no longer the case.

We now review two types of families popular in the VB literature.

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

$$\mathcal{F}^{\mathrm{MF}} = \left\{ \rho \in \mathcal{M}_+^1(\Theta) : \rho(\mathrm{d}\theta) = \prod_{i=1}^{d} \rho_i(\mathrm{d}\theta_i), \forall i \in \{1, \ldots, d\}, \rho_i \in \mathcal{M}_+^1(\Theta_i) \right\}. \quad (3)$$

  The infimum of the KL divergence $\mathcal{K}(\rho, \hat{\rho}_\lambda)$, relative to $\rho = \prod_i \rho_i$ satisfies the following fixed point condition (Parisi, 1988; Bishop, 2006, Chap. 10):

$$\forall j \in \{1, \cdots, d\} \quad \rho_j(\mathrm{d}\theta_j) \propto \exp\left(\int \{-\lambda r_n(\theta) + \log \pi(\theta)\} \prod_{i \neq j} \rho_i(\mathrm{d}\theta_i) \right) \pi(\mathrm{d}\theta_j). \quad (4)$$

  This leads to a natural algorithm were we update successively every ρ_j until stabilization.

- Parametric family:

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

  and M is finite-dimensional; say $\mathcal{F}^{\mathrm{P}}$ is the family of Gaussian distributions (of dimension d). In this case, several methods may be used to compute the infimum. As above, one may used fixed-point iteration, provided an equation similar to (4) is available. Alternatively, one may directly maximize $\int \log[\exp[-\lambda r_n(\theta)] \frac{\mathrm{d}\pi}{\mathrm{d}\rho}(\theta)] \rho(\mathrm{d}\theta)$ with respect to parameter m, using numerical optimization routines. This approach was used for instance in Hoffman et al. (2013) with combination of some stochastic gradient descent to perform inference on a latent Dirichlet allocation model. See also e.g. Khan (2014); Khan et al. (2013) for efficient algorithms for Gaussian variational approximation.

In what follows (Subsections 4.1 and 4.2) we provide tight bounds for the prediction risk of $\tilde{\rho}_\lambda$. This leads to the identification of a condition on $\mathcal{F}$ such that the risk of $\tilde{\rho}_\lambda$ is not worse than the risk of $\hat{\rho}_\lambda$. We will make this condition explicit in various examples, using either mean field VB or parametric approximations.

Remark 3.1 *An useful identity, obtained by direct calculations, is: for any $\rho \ll \pi$,*

$$\log \int \exp\left[-\lambda r_n(\theta)\right] \pi(\mathrm{d}\theta) = -\lambda \int r_n(\theta)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho, \pi) + \mathcal{K}(\rho, \hat{\rho}_\lambda). \quad (5)$$

Since the left hand side does not depend on ρ, one sees that $\tilde{\rho}_\lambda$, which minimizes $\mathcal{K}(\rho, \hat{\rho}_\lambda)$ over $\mathcal{F}$, is also the minimizer of:

$$\tilde{\rho}_\lambda = \arg\min_{\rho \in \mathcal{F}} \left\{ \int r_n(\theta)\rho(\mathrm{d}\theta) + \frac{1}{\lambda}\mathcal{K}(\rho, \pi) \right\}$$

This equation will appear frequently in the sequel in the form of an empirical upper bound.

4. General results

This section gives our general results, under either a Hoeffding Assumption (Definition 2.3) or a Bernstein Assumption (Definition 2.4), on risks bounds for the variational approximation, and how it relates to risks bounds for Gibbs posteriors. These results will be specialised to several learning problems in the following sections.

4.1 Bounds under the Hoeffding assumption

4.1.1 Empirical bounds

Theorem 4.1 *Under the Hoeffding assumption (Definition 2.3), for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$ we have simultaneously for any $\rho \in \mathcal{M}_+^1(\Theta)$,*

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

This result is a simple variant of a result in Catoni (2007) but for the sake of completeness, its proof is given in Appendix A. It gives us an upper bound on the risk of both the pseudo-posterior (take $\rho = \hat{\rho}_\lambda$) and its variational approximation (take $\rho = \tilde{\rho}_\lambda$). These bounds may be be computed from the data, and therefore provide a simple way to evaluate the performance of the corresponding procedure, in the spirit of the first PAC-Bayesian inequalities (Shawe-Taylor and Williamson, 1997; McAllester, 1998, 1999). However, these bounds do not provide the rate of convergence of these estimators. For this reason, we also provide oracle-type inequalities.

4.1.2 Oracle-type inequalities

Another way to use PAC-Bayesian bounds is to compare $\int R\mathrm{d}\hat{\rho}_\lambda$ to the best possible risk, thus linking this approach to oracle inequalities. This is the point of view developed in Catoni (2004, 2007); Dalalyan and Tsybakov (2008).

Theorem 4.2 *Assume that the Hoeffding assumption is satisfied (Definition 2.3). For any $\varepsilon > 0$, with probability at least $1 - \varepsilon$ we have simultaneously*

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

and

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

Moreover,

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

where we remind that π_λ is defined in Definition 2.2.

In this way, we are able to compare $\int R\mathrm{d}\hat{\rho}_\lambda$ to the best possible aggregation procedure in $\mathcal{M}_+^1(\Theta)$ and $\int R\mathrm{d}\tilde{\rho}_\lambda$ to the best aggregation procedure in $\mathcal{F}$. More importantly, we are able to obtain explicit expressions for the right-hand side of these inequalities in various models, and thus to obtain rates of convergence. This will be done in the remaining sections. This leads to the second interest of this result: if there is a $\lambda = \lambda(n)$ that leads to $\mathcal{B}_\lambda(\mathcal{M}_+^1(\Theta)) \leq \overline{R} + s_n$ with $s_n \to 0$ for the pseudo-posterior $\hat{\rho}_\lambda$, then we only have to prove that there is a $\rho \in \mathcal{F}$ such that $\mathcal{K}(\rho, \pi_\lambda)/\lambda \leq cs_n$ for some constant $c > 0$ to ensure that the VB approximation $\tilde{\rho}_\lambda$ also reaches the rate s_n.

We will see in the following sections several examples where the approximation does not deteriorate the rate of convergence. But first let us show the equivalent oracle inequality under the Bernstein assumption.

4.2 Bounds under the Bernstein assumption

In this context the empirical bound on the risk would depend on the minimal achievable risk $\bar{r}_n$, and cannot be computed explicitly. We give the oracle inequality for both the Gibbs posterior and its VB approximation in the following theorem.

Theorem 4.3 *Assume that the Bernstein assumption is satisfied (Definition 2.4). 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:*

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

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

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

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 main difference with Theorem 4.2 is that the function $R(\cdot)$ is replaced by $R(\cdot) - \overline{R}$. This is well known way to obtain better rates of convergence.

5. Application to classification

5.1 Preliminaries

In all this section, we assume that $\mathcal{Y} = \{0, 1\}$ and we consider linear classification: $\Theta = \mathcal{X} = \mathbb{R}^d$, $f_\theta(x) = \mathbf{1}_{\langle \theta, x \rangle \geq 0}$. We put $r_n(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{f_\theta(X_i) \neq Y_i\}}$, $R(\theta) = \mathbb{P}(Y \neq f_\theta(X))$ and assume that the $[(X_i, Y_i)]_{i=1}^n$ are i.i.d. In this setting, it is well-known that the Hoeffding assumption always holds. We state as a reminder the following lemma.

Lemma 1 *Hoeffding assumption* (1) *is satisfied with* $f(\lambda, n) = \lambda^2/(2n)$, $\lambda \in \mathbb{R}_+$.

The proof is given in Appendix A for the sake of completeness.

It is also possible to prove that Bernstein assumption (2) holds in the case where the so-called margin assumption of Mammen and Tsybakov is satisfied. This condition we use was introduced by Tsybakov (2004) in a classification setting, based on a related definition in Mammen and Tsybakov (1999).

Lemma 2 *Assume that Mammen and Tsybakov's margin assumption is satisfied: i.e. there is a constant* C *such that*

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

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

Remark 5.1 *We refer the reader to Tsybakov (2004) for a proof that*

$$\mathbb{P}(0 < |\langle \overline{\theta}, X \rangle| \leq t) \leq C' t$$

for some constant $C' > 0$ *implies the margin assumption. In words, when* X *is not likely to be in the region* $\langle \overline{\theta}, X \rangle \simeq 0$*, where points are hard to classify, then the problem becomes easier and the classification rate can be improved.*

We propose in this context a Gaussian prior: $\pi = \mathcal{N}_d(0, \vartheta^2 I_d)$, and we consider a VB approach based on Gaussian families. The corresponding optimization problem is not convex, but remains feasible as we explain below.

5.2 Three sets of Variational Gaussian approximations

Consider the three following Gaussian families

$$\mathcal{F}_1 = \left\{ \Phi_{\mathbf{m},\sigma^2},\, \mathbf{m} \in \mathbb{R}^d, \sigma^2 \in \mathbb{R}_+^* \right\},$$
$$\mathcal{F}_2 = \left\{ \Phi_{\mathbf{m},\boldsymbol{\sigma}^2},\, \mathbf{m} \in \mathbb{R}^d, \boldsymbol{\sigma}^2 \in (\mathbb{R}_+^*)^d \right\} \text{ (mean field approximation)},$$
$$\mathcal{F}_3 = \left\{ \Phi_{\mathbf{m},\Sigma},\, \mathbf{m} \in \mathbb{R}^d, \Sigma \in \mathcal{S}^{d+} \right\} \text{ (full covariance approximation)},$$

where $\Phi_{\mathbf{m},\sigma^2}$ is Gaussian distribution $N_d(\mathbf{m}, \sigma^2 I_d)$, $\Phi_{\mathbf{m},\boldsymbol{\sigma}^2}$ is $N_d(\mathbf{m}, \mathrm{diag}(\boldsymbol{\sigma}^2))$, and $\Phi_{\mathbf{m},\Sigma}$ is $N_d(\mathbf{m}, \Sigma)$. Obviously, $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \mathcal{M}_+^1(\Theta)$, and

$$\mathcal{B}_\lambda(\mathcal{M}_+^1(\Theta)) \leq \mathcal{B}_\lambda(\mathcal{F}_3) \leq \mathcal{B}_\lambda(\mathcal{F}_2) \leq \mathcal{B}_\lambda(\mathcal{F}_1). \tag{6}$$

Note that, for the sake of simplicity, we will use the following classical notations in the rest of the paper: $\varphi(\cdot)$ is the density of $\mathcal{N}(0,1)$ w.r.t. the Lebesgue measure, and $\Phi(\cdot)$ the corresponding c.d.f. The rest of Section 5 is organized as follows. In Subsection 5.3, we calculate explicitly $\mathcal{B}_\lambda(\mathcal{F}_2)$ and $\mathcal{B}_\lambda(\mathcal{F}_1)$. Thanks to (6) this also gives an upper bound on $\mathcal{B}_\lambda(\mathcal{F}_3)$ and proves the validity of the three types of Gaussian approximations. Then, we give details on algorithms to compute the variational approximation based on $\mathcal{F}_2$ and $\mathcal{F}_3$, and provide a numerical illustration on real data.

5.3 Theoretical analysis

We start with the empirical bound for $\mathcal{F}_2$ (and $\mathcal{F}_1$ as a consequence), which is a direct corollary of Theorem 4.1.

Corollary 5.1 *For any $\varepsilon > 0$, with probability at least $1-\varepsilon$ we have, for any $\mathbf{m} \in \mathbb{R}^d$, $\boldsymbol{\sigma}^2 \in (\mathbb{R}_+)^d$,*

$$\int R \mathrm{d}\Phi_{\mathbf{m},\boldsymbol{\sigma}^2} \leq \int r_n \mathrm{d}\Phi_{\mathbf{m},\boldsymbol{\sigma}^2} + \frac{\lambda}{2n} + \frac{\frac{1}{2}\sum_{i=1}^d \left[\log\left(\frac{\vartheta^2}{\sigma_i^2}\right) + \frac{\sigma_i^2}{\vartheta^2} \right] + \frac{\|\mathbf{m}\|^2}{2\vartheta^2} - \frac{d}{2} + \log\left(\frac{1}{\varepsilon}\right)}{\lambda}.$$

We now want to apply Theorem 4.2 in this context. In order to do so, we introduce an additional assumption.

Definition 5.1 *We say that Assumption A1 is satisfied when there is a constant $c > 0$ such that, for any $(\theta, \theta') \in \Theta^2$ with $\|\theta\| = \|\theta'\| = 1$, $\mathbb{P}(\langle X, \theta\rangle \langle X, \theta'\rangle < 0) \leq c\|\theta - \theta'\|$.*

This is not a strong assumption. It is satisfied when X has an isotropic distribution, and more generally when $X/\|X\|$ has a bounded density on the unit sphere[2]. The intuition

2. If the density of $X/\|X\|$ with respect to the uniform measure on the unit sphere is upper bounded by B then $\mathbb{P}(\langle X, \theta\rangle \langle X, \theta'\rangle < 0) \leq \frac{B}{2\pi} \arccos(\langle \theta, \theta'\rangle) \leq \frac{B}{2\pi}\sqrt{5 - 5\langle \theta, \theta'\rangle} \leq \frac{B}{2\pi}\sqrt{\frac{5}{2}}\|\theta - \theta'\|$.

beyond A1 is that for a "typical" X, a very small change in θ will only induce a change in $\mathrm{sign}(\langle X, \theta \rangle)$ with a small probability. When it is not satisfied, two parameters θ and θ' very close to each other can lead to very different predictions, and thus, whatever the accuracy of an approximation of $\bar{\theta}$, it might still lead to poor predictions.

Corollary 5.2 *Assume that the VB approximation is done on either $\mathcal{F}_1$, $\mathcal{F}_2$ or $\mathcal{F}_3$. Take $\lambda = \sqrt{nd}$ and $\vartheta = \frac{1}{\sqrt{d}}$. Under Assumption A1, for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$ we have simultaneously*

$$\left.\begin{array}{l} \int R \mathrm{d}\hat{\rho}_\lambda \\ \int R \mathrm{d}\tilde{\rho}_\lambda \end{array}\right\} \leq \overline{R} + \sqrt{\frac{d}{n}} \log(4ne) + \frac{c}{\sqrt{n}} + \frac{1}{4n}\sqrt{\frac{d}{n}} + \frac{2\log\left(\frac{2}{\varepsilon}\right)}{\sqrt{nd}}.$$

See the appendix for a proof. Note also that the values $\lambda = \sqrt{nd}$ and $\vartheta = \frac{1}{\sqrt{d}}$ allow to derive this almost optimal rate of convergence, but are not necessarily the best choices in practice.

Remark 5.2 *Note that Assumption A1 is not necessary to obtain oracle inequalities on the risk integrated under $\hat{\rho}_\lambda$. We refer the reader to Chapter 1 in Catoni (2007) for such assumption-free bounds. However, it is clear that without this assumption the shape of $\hat{\rho}_\lambda$ and $\tilde{\rho}_\lambda$ might be very different. Thus, it seems reasonable to require that A1 is satisfied for the approximation of $\hat{\rho}_\lambda$ by $\tilde{\rho}_\lambda$ to make sense.*

We finally provide an application of Theorem 4.3. Under the additional constraint that the margin assumption is satisfied, we obtain a better rate.

Corollary 5.3 *Assume that the VB approximation is done on either $\mathcal{F}_1$, $\mathcal{F}_2$ or $\mathcal{F}_3$. Under Assumption A1 (Definition 5.1 page 10), and under Mammen and Tsybakov margin assumption, with $\lambda = \frac{2n}{C+2}$ and $\vartheta > 0$, for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$,*

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

It is possible to minimze the bound with respect to ϑ explicitely, this choice or any constant instead will lead to a rate in $d\log(n)/n$. Note that the rate d/n is minimax-optimal in this context. This is, for example, a consequence of more general results in Lecué (2007) under a general form of the the margin assumption. See the Appendix for a proof.

5.4 Implementation and numerical results

For family $\mathcal{F}_2$ (mean field), the variational lower bound (5) equals

$$\mathcal{L}_{\lambda,\vartheta}(\mathbf{m}, \boldsymbol{\sigma}) = -\frac{\lambda}{n}\sum_{i=1}^{n} \Phi\left(-Y_i \frac{X_i \mathbf{m}}{\sqrt{X_i \mathrm{diag}(\boldsymbol{\sigma}^2) X_i^t}}\right) - \frac{\mathbf{m}^T\mathbf{m}}{2\vartheta} + \frac{1}{2}\sum_{k=1}^{d}\left(\log\sigma_k^2 - \frac{\sigma_k^2}{\vartheta}\right),$$

while for family $\mathcal{F}_3$ (full covariance), it equals

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

Both functions are non-convex, but the multimodality of the latter may be more severe due to the larger dimension of $\mathcal{F}_3$. To address this issue, we recommend using the reparametrization of Opper and Archambeau (2009), which makes the dimension of the latter optimization problem $\mathcal{O}(n)$; see Khan (2014) for a related approach. In both cases, we found that deterministic annealing to be a good approach to optimize such non-convex functions. We refer to Appendix B for more details on deterministic annealing and on our particular implementation.

We now compare the numerical performance of the mean field and full covariance VB approximations to the Gibbs posterior (as approximated by SMC, see Section 3.1) for the classification of standard datasets; see Table 1. The datasets are all available in the UCI repository[3] except for the DNA dataset which is part of the R package mlbench by Leisch and Dimitriadou (2010). When no split between the training sample is provided we split the data in half. The design matrices are centered and scaled before being used. For the Glass dataset we compare the "silicon" class against the other classes.

We also include results for a linear SVM (support vector machine) and a radial kernel SVM; the latter comparison is not entirely fair, since this is a non-linear classifier, while all the other classifiers are linear. Except for the Glass and DNA datasets, the full covariance VB approximation performs as well as or better than both SMC and SVM (while being much faster to compute, especially compared to SMC). Note that some high errors for the VB approximations can be due to the fact that the optimization of the objective is harder (we address this issue in next section).

Interestingly, VB outperforms SMC in certain cases. This might be due to the fact that a VB approximation tends to be more concentrated around the mode than the Gibbs posterior it approximates. Mean field VB does not perform so well on certain datasets (e.g. Indian). This may due either to the approximation family being too small, or to the corresponding optmisation problem to be strongly multi-modal. We address this issue in next section.

6. Application to classification under convexified loss

Compared to the previous section, the advantage of convex classification is that the corresponding variational approximation will amount to minimizing a convex function. This

3. `https://archive.ics.uci.edu/ml/datasets.html`

Dataset	**Covariates**	**Mean Field ($\mathcal{F}_2$)**	**Full cov. ($\mathcal{F}_3$)**	**SMC**	**SVM radial**	**SVM linear**
Pima	7	31.0	21.3	22.3	30.4	21.6
German Credit	60	32.0	33.6	32.0	32.0	33.2
DNA	180	23.6	23.6	23.6	3.5	5.1
SPECTF	22	08.0	06.9	08.5	10.1	21.4
Glass	10	34.6	19.6	23.3	4.7	6.5
Indian	11	48.0	25.5	26.2	26.8	25.3
Breast	10	35.1	1.1	1.1	1.7	1.7

Table 1: Comparison of misclassification rates (%).
Misclassification rates for different datasets and for the proposed approximations of the Gibbs posterior. The last two columns are the missclassification rate given by a SVM with radial kernel and a linear SVM. The hyper-parameters are chosen by cross-validation.

means that (a) the minimization problem will be easier to deal with; and (b) we will be able to compute a bound for the integrated risk after a given number of steps of the minimization procedure.

The setting is the same as in the previous section, except that for convenience we now take $\mathcal{Y} = \{-1, 1\}$, and the risk is based on the hinge loss,

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

We will write R^H for the theoretical counterpart and $\bar{R}^H$ for its minimum in θ. We keep the superscript H in order to allow comparison with the risk R under the 0-1 loss. We assume in this section that the X_i are uniformly bounded, that is, we have almost surely $\|X_i\|_\infty = \max_j |X_{i,j}| < c_x$ for some $c_x > 0$. Note that we do not require an assumption of the form (A1) to obtain the results of this section, as we rely directly on the Lipschitz continuity of the hinge risk.

6.1 Theoretical Results

Contrary to the previous section, the risk is not bounded in θ, and we must specify a prior distribution for the Hoeffding assumption to hold.

Lemma 3 *Under an independent Gaussian prior π such that each component is $N(0, \vartheta^2)$, and for $\lambda < \frac{1}{c_x}\sqrt{\frac{n}{\vartheta}^2}$ and with bounded design $|X_{ij}| < c_x$, Hoeffding assumption (1) is satisfied with $f(\lambda, n) = \lambda^2/(4n) - \frac{1}{2}\log\left(1 - \frac{\vartheta^2\lambda^2 c_x^2}{2n}\right)$.*

The main impact of such a bound is that the prior variance cannot be taken too big relative to λ.

Corollary 6.1 *Assume that the VB approximation is done on either $\mathcal{F}_1$, $\mathcal{F}_2$ or $\mathcal{F}_3$. Take $\lambda = \frac{1}{c_x}\sqrt{\frac{n}{\vartheta^2}}$ and $\vartheta = \frac{1}{\sqrt{d}}$. For any $\varepsilon > 0$, with probability at least $1-\varepsilon$ we have simultaneously*

$$\left.\begin{array}{l} \int R^H \mathrm{d}\hat{\rho}_\lambda \\ \int R^H \mathrm{d}\tilde{\rho}_\lambda \end{array}\right\} \leq \overline{R}^H + \frac{c_x}{2}\sqrt{\frac{d}{n}}\log\frac{n}{d} + c_x\frac{d}{n}\sqrt{\frac{d}{n}} + \frac{1}{\sqrt{nd}}\left(\frac{2c_x^2+1}{2c_x} + 2c_x\log\frac{2}{\epsilon}\right)$$

The oracle inequality in the above corollary enjoys the same rate of convergence as the equivalent result in the preceding section. In the following we link the two results.

Remark 6.1 *As stated in the beginning of the section we can use the estimator specified under the hinge loss to bound the excess risk of the* 0-1 *loss. We write $R^\star$ and $R^{H\star}$ the respective risk for their corresponding Bayes classifiers. From Zhang (2004) (section 3.3) we have the following inequality, linking the excess risk under the hinge loss and the* 0-1 *loss,*

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

for every $\theta \in \mathbb{R}^p$. By integrating with respect to $\tilde{\rho}^H$ (the VB approximation on any $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ of the Gibbs posterior for the hinge risk) and making use of Corollary 6.1 we have with high probability,

$$\tilde{\rho}^H\left(R(\theta)\right) - R^\star \leq \inf_{\theta\in\mathbb{R}^p} R^H(\theta) - R^{H\star} + \mathcal{O}\left(\sqrt{\frac{d}{n}}\log\left(\frac{n}{d}\right)\right).$$

6.2 Numerical application

We have motivated the introduction of the hinge loss as a convex upper bound. In the sequel we show that the resulting VB approximation also leads to a convex optimization problem. This has the advantage of opening a range of possible optimization algorithms (Nesterov, 2004). In addition we are able to bound the error of the approximated measure after a fixed number of iterations (see Theorem 6.2).

Under the model $\mathcal{F}_1$ each individual risk is given by:

$$\rho_{\mathbf{m},\sigma}(r_i(\theta)) = (1-\Gamma_i\mathbf{m})\,\Phi\left(\frac{1-\Gamma_i\mathbf{m}}{\sigma\|\Gamma_i\|_2}\right) + \sigma\|\Gamma_i\|\varphi\left(\frac{1-\Gamma_i\mathbf{m}}{\sigma\|\Gamma_i\|_2}\right) := \Xi_i\left(\begin{pmatrix}\mathbf{m}\\ \sigma\end{pmatrix}\right),$$

writing $\Gamma_i := Y_iX_i$.

Hence the lower bound to be maximized is given by

$$\mathcal{L}(\mathbf{m},\sigma) = -\frac{\lambda}{n}\left\{\sum_{i=1}^{n}(1-\Gamma_i\mathbf{m})\,\Phi\left(\frac{1-\Gamma_i\mathbf{m}}{\sigma\|\Gamma_i\|_2}\right)+\sum_{i=1}^{n}\sigma\|\Gamma_i\|\varphi\left(\frac{1-\Gamma_i\mathbf{m}}{\sigma\|\Gamma_i\|_2}\right)\right\} - \frac{\|\mathbf{m}\|_2^2}{2\vartheta}+\frac{d}{2}\left(\log\sigma^2-\frac{\vartheta}{\sigma^2}\right).$$

It is easy to see that the function is convex in $(\mathbf{m},\sigma)$, first note that the map

$$\Psi:\begin{pmatrix}x\\y\end{pmatrix}\mapsto x\Phi\left(\frac{x}{y}\right)+y\varphi\left(\frac{x}{y}\right),$$

is convex and note that we can write $\Xi_i\left(\begin{pmatrix}\mathbf{m}\\\sigma\end{pmatrix}\right)=\Psi\left(A\begin{pmatrix}x\\y\end{pmatrix}+b\right)$ hence by composition of convex function with linear mappings we have the result. Similar reasoning could be held for the case $\mathcal{F}_2$ and $\mathcal{F}_3$, where in later the parametrization should be done in C such that $\Sigma = CC^t$. The bound is however not universally Lipschitz in σ, this impacts the optimization algorithms. In Theorem 6.2 we define a ball around the optimal value of the objective, containing the initial values. We denote it's radius by M. On this ball the objective is Lipschitz (with coefficient L) and optimal convex solvers can be used (e.g. Nesterov (2004) section 3.2.3).

On the class of function $\mathcal{F}_0 = \left\{\Phi_{\mathbf{m},\frac{1}{n}},\mathbf{m}\in\mathbb{R}^d\right\}$, for which our Oracle inequalities still hold we could get faster numerical algorithms. The objective function has Lipschitz continuous derivatives and we would get a rate of $\frac{L}{(1+k)^2}$.

Other convex loss could be considered which could lead to convex optimization problems. For instance one could consider the exponential loss.

Theorem 6.2 *Assume that the VB approximation is based on either $\mathcal{F}_1,\mathcal{F}_2$ or $\mathcal{F}_3$. Denote by $\tilde{\rho}_k(\mathrm{d}\theta)$ the VB approximated measure after the kth iteration of an optimal convex solver using the hinge loss. Fix $M>0$ large enough so that the optimal approximated mean and variance $\bar{m}$,$\bar{\Sigma}$ are at distance at most M from the initial value used by the solver. Take $\lambda=\sqrt{nd}$ and $\vartheta=\frac{1}{\sqrt{d}}$ then under the hypothesis of Corollary 6.1 with probability $1-\epsilon$*

$$\int R^H\mathrm{d}\tilde{\rho}_k \le \overline{R}^H+\frac{LM}{\sqrt{1+k}}+\frac{c_x}{2}\sqrt{\frac{d}{n}}\log\frac{n}{d}+c_x\frac{d}{n}\sqrt{\frac{d}{n}}+\frac{1}{\sqrt{nd}}\left(\frac{2c_x^2+1}{2c_x}+2c_x\log\frac{2}{\epsilon}\right)$$

where L is the Lipschitz coefficient on a ball of radius M defined above.

Note that this result is stronger and more practical than the previous ones: it ensures a certain error level (with fixed probability $1-\epsilon$) for the k-th iterate of the optimization algorithm, for a known value of k. In contrast, previous results applied to the output of the optimizer "for k large enough".

We find that on average the misclassification error (Table 2) is lower than for the 0-1 loss where we have no guaranties that the maximum is attained.

Dataset	Covariates	Hinge loss	SMC
Pima	7	19.5	22.3
Credit	60	26.2	32.0
DNA	180	4.2	23.6
SPECTF	22	10.1	08.5
Glass	10	2.8	23.3
Indian	11	25.5	25.5
Breast	10	0.5	1.1

Table 2: Comparison of misclassification rates (%).
Misclassification rates for different datasets and for the proposed approximations of the Gibbs posterior. The hyperparameters are chosen by cross-validation. This is to be compared to Table 1. The variational Bayes approximation was computed using the R package we developed (see the introduction for a reference).

7. Application to ranking

7.1 Preliminaries

We now focus on the ranking problem. We follow Clémençon et al. (2008) for the definitions of the basic concepts: $\mathcal{Y} = \{0, 1\}$, $\Theta = \mathcal{X} = \mathbb{R}^d$ and $f_\theta : \mathcal{X}^2 \to \{-1, +1\}$ for $\theta \in \Theta$; $f_\theta(x, x') = 1$ (resp. -1) means that x is more (resp. less) likely to correspond to label 1 than x'. The natural risk function is then

$$R(\theta) = \mathbb{P}\left[(Y_1 - Y_2) f_\theta(X_1, X_2) < 0\right]$$

and the empirical risk

$$r_n(\theta) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \mathbf{1}_{\{(Y_i - Y_j) f_\theta(X_i, X_j) < 0\}}.$$

Again, we recall classical results.

Lemma 4 *The Hoeffding-type assumption is satisfied with* $f(\lambda, n) = \frac{\lambda^2}{n-1}$.

The variant of the margin assumption adapted to ranking was established by Robbiano (2013) and Ridgway et al. (2014).

Lemma 5 *Assume the following margin assumption:*

$$\mathbb{E}[(\mathbf{1}_{f_\theta(X_1,X_2)[Y_1 - Y_2] < 0} - \mathbf{1}_{f_\theta(X_1,X_2)[Y_1 - Y_2] < 0})^2] \leq C[R(\theta) - \overline{R}].$$

Then Bernstein assumption (2) *is satisfied with* $g(\lambda, n) = \frac{C\lambda^2}{n - 1 - 4\lambda}$.

We focus on linear classifiers, $f_\theta(x,x') = -1 + 2 \times \mathbf{1}_{\langle\theta,x\rangle > \langle\theta,x'\rangle}$. Like in the classification setting, $\langle x, \theta\rangle$ is interpreted as a score related to the probability that $Y = 1|X = x$. We consider a Gaussian prior

$$\pi(\mathrm{d}\theta) = \prod_{i=1}^{d} \varphi(\theta_i; 0, \vartheta^2)\mathrm{d}\theta_i$$

and the approximation families will be the same as in Section 5: $\mathcal{F}_1 = \{\Phi_{\mathbf{m},\sigma^2}, \mathbf{m} \in \mathbb{R}^d, \sigma^2 \in \mathbb{R}_+^*\}$, $\mathcal{F}_2 = \{\Phi_{\mathbf{m},\boldsymbol{\sigma}^2}, \mathbf{m} \in \mathbb{R}^d, \boldsymbol{\sigma}^2 \in (\mathbb{R}_+^*)^2\}$ and $\mathcal{F}_3 = \{\Phi_{\mathbf{m},\Sigma}, \mathbf{m} \in \mathbb{R}^d, \Sigma \in \mathcal{S}^{d+}\}$.

7.2 Theoretical study

Here again, we start with the empirical bound.

Corollary 7.1 *For any $\varepsilon > 0$, with probability at least $1-\varepsilon$ we have, for any $\mathbf{m} \in \mathbb{R}^d$, $\sigma^2 \in (\mathbb{R}_+)^d$,*

$$\int R\mathrm{d}\Phi_{\mathbf{m},\boldsymbol{\sigma}^2} \leq \int r_n \mathrm{d}\Phi_{\mathbf{m},\boldsymbol{\sigma}^2} + \frac{\lambda}{n-1} + \frac{\frac{1}{2}\sum_{j=1}^{d}\left[\log\left(\frac{\vartheta^2}{\sigma_i^2}\right) + \frac{\sigma_i^2}{\vartheta^2}\right] + \frac{\|\mathbf{m}\|^2}{2\vartheta^2} - \frac{d}{2} + \log\left(\frac{1}{\varepsilon}\right)}{\lambda}.$$

In order to derive a theoretical bound, we introduce the following variant of Assumption A1.

Definition 7.1 *We say that Assumption A2 is satisfied when there is a constant $c > 0$ such that, for any $(\theta,\theta') \in \Theta^2$ with $\|\theta\| = \|\theta'\| = 1$, $\mathbb{P}(\langle X_1 - X_2, \theta\rangle \langle X_1 - X_2, \theta'\rangle < 0) \leq c\|\theta - \theta'\|$.*

Assumption A2 is just Assumption A1 applied to the distribution of $(X_1 - X_2)$. Intuitively, it means that two parameters close to each other rank X_1 and X_2 in the same way (with large probability).

Corollary 7.2 *Use either $\mathcal{F}_1$, $\mathcal{F}_2$ or $\mathcal{F}_3$. Take $\lambda = \sqrt{\frac{d(n-1)}{2}}$ and $\vartheta = 1$. Under (A2), for any $\varepsilon > 0$, with probability at least $1-\varepsilon$,*

$$\left.\begin{array}{l}\int R\mathrm{d}\hat{\rho}_\lambda \\ \int R\mathrm{d}\tilde{\rho}_\lambda\end{array}\right\} \leq \overline{R} + \sqrt{\frac{2d}{n-1}}\left(1 + \frac{1}{2}\log\left(2d(n-1)\right)\right) + \frac{c\sqrt{2}}{\sqrt{n-1}} + \frac{1}{(n-1)^{3/2}\sqrt{2d}} + \frac{2\sqrt{2}\log\left(\frac{2\mathrm{e}}{\varepsilon}\right)}{\sqrt{(n-1)d}}.$$

Finally, under an additional margin assumption, we have:

Corollary 7.3 *Under Assumption A2 and the margin assumption of Lemma (5), for $\lambda = \frac{n-1}{C+5}$ and $\vartheta > 0$, for any $\varepsilon > 0$, with probability at least $1-\varepsilon$,*

$$\left.\begin{array}{l}\int R\mathrm{d}\hat{\rho}_\lambda \\ \int R\mathrm{d}\tilde{\rho}_\lambda\end{array}\right\} \leq \bar{R} + \frac{(C+5)(C+1)}{2}\left\{\frac{d\log\frac{n}{\vartheta}}{n-1} + \frac{d\vartheta}{n(n-1)} + \frac{1}{\vartheta} - \frac{d}{\vartheta n - 1} + \frac{2}{n-1}\log\frac{2}{\varepsilon}\right\} + \frac{\sqrt{d}4c(C+1)}{n}.$$

It is possible to optimize the bound with respect to ϑ. The proof is similar to the ones of Corollaries 5.2, 5.3 and 7.2.

As in the case of classification, ranking under an AUC loss can be done by replacing the indicator function by the corresponding upper bound given by an hinge loss. In this case we can derive similar results as for the convexified classification in particular we can get a convex minimization problem and obtain result without requiring assumption (A2).

7.3 Algorithms and numerical results

As an illustration we focus here on family $\mathcal{F}_2$ (mean field). In this case the VB objective to maximize is given by:

$$\mathcal{L}(\mathbf{m}, \sigma^2) = -\frac{\lambda}{n_+ n_-} \sum_{i:y_i=1, j:y_j=0} \Phi\left(-\frac{\Gamma_{ij}\mathbf{m}}{\sqrt{\sum_{k=1}^d (\gamma_{ij}^k)^2 \sigma_k^2}}\right) - \frac{\|\mathbf{m}\|_2^2}{2\vartheta} + \frac{1}{2}\sum_{k=1}^d \left[\log \sigma_k^2 - \frac{\sigma_k^2}{\vartheta}\right], \tag{7}$$

where $\Gamma_{ij} = X_i - X_j$, $n_+ = \text{card}\{1 \leq i \leq n : Y_i = 1\}$, $n_- = n - n_+ = \text{card}\{1 \leq i \leq n : Y_i = 0\}$ and where $(\gamma_{ij}^k)_k$ are the elements of Γ.

This function is expensive to compute, as it involves $n_+ n_-$ terms, the computation of which is $\mathcal{O}(p)$.

We propose to use a stochastic gradient descent in the spirit of Hoffman et al. (2013). The model we consider is not in an exponential family, meaning we cannot use the trick developed by these authors. We propose instead to use a standard descent.

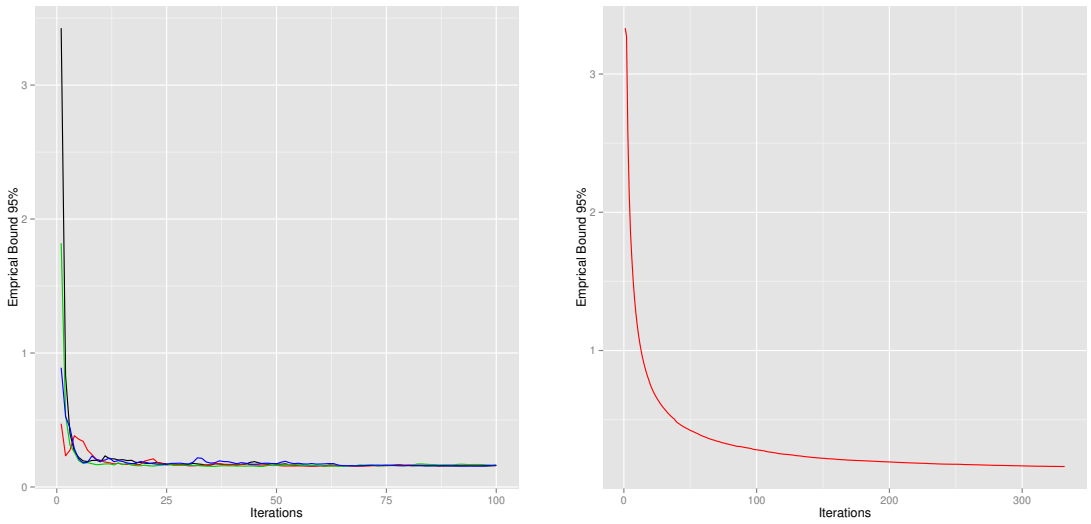
The idea is to replace the gradient by a unbiased version based on a batch of size B as described in Algorithm 4 in the Appendix. Robbins and Monro (1951) show that for a step-size $(\lambda_t)_t$ such that $\sum_t \lambda_t^2 < \infty$ and $\sum_t \lambda_t = \infty$ the algorithm converges to a local optimum.

In our case we propose to sample pairs of data with replacement and use the unbiased version of the derivative of the risk component. We use a simple gradient descent without any curvature information. One could also use recent research on stochastic quasi Newton-Raphson (Byrd et al., 2014).

For illustration, we consider a small dataset (Pima), and a larger one (Adult). Both datasets are available in the UCI repository[4]. As for the previous experiment the data is scaled and centered. The latter is already quite challenging with $n_+ n_- = 193,829,520$ pairs to compare. In both cases with different size of batches convergence is obtained with a few iterations only and leads to acceptable bounds.

In Figure 1 we show the empirical bound on the AUC risk as a function of the iteration of the algorithm, for several batch sizes. The bound is taken for 95% probability, the batch sizes are taken to be $B = 1, 10, 20, 50$ for the Pima dataset, and 50 for the Adult dataset. The figure shows an additional feature of VB approximation in the context of

4. https://archive.ics.uci.edu/ml/datasets.html



(a) Pima (b) adult

Figure 1: Error bound at each iteration, stochastic descent, Pima and Adult datasets.

Stochastic VB with fixed temperature $\lambda = 100$ for Pima and $\lambda = 1000$ for adult. The left panel shows several curves that correspond to different batch sizes; these curves are hard to distinguish. The right panel is for a batch size of 50. The adult dataset has $n = 32556$ observation and $n_+ n_- = 193829520$ possible pairs. The convergence is obtained in order of seconds. The bounds are the empirical bounds obtained in Corollary 7.1 for a probability of 95%.

Gibbs posterior: namely the possibility of computing the empirical upper bound given by Corollary 7.1. That is we can check the quality of the bound at each iteration of the algorithm, or for different values of the hyperparameters.

8. Application to matrix completion

The matrix completion problem has received increasing attention recently, partly due to spectacular theoretical results (Candès and Tao, 2010), and to challenging applications like the Netflix challenge (Bennett and Lanning, 2007). In the perspective of this paper, the specific interest of this application is twofold. First, this is a case where the family of approximations is not parametric, but rather of the form (3), i.e. the family of products of independent components. Then, there is no known theoretical result for the Gibbs estimator in the considered model, yet we can still directly bound the loss induced by the variational approximation.

We observe i.i.d. pairs $((X_i, Y_i))_{i=1}^n$ where $X_i \in \{1, \dots, m_1\} \times \{1, \dots, m_2\}$, and we assume that there is a $m_1 \times m_2$-matrix M such that $Y_i = M_{X_i} + \varepsilon_i$ and the ε_i are centred. Assuming that X_i is uniform on $\{1, \dots, m_1\} \times \{1, \dots, m_2\}$, that $f_\theta(X_i) = \theta_{X_i}$, and taking the quadratic risk, $R(\theta) = \mathbb{E}\left[(Y_i - \theta_{X_i})^2\right]$, we have that

$$R(\theta) - \overline{R} = \frac{1}{m_1 m_2} \|\theta - M\|_F^2$$

where $\|\cdot\|_F$ stands for the Frobenius norm.

A common way to parametrize the problem is

$$\Theta = \{\theta = UV^T, U \in \mathbb{R}^{m_1 \times K}, V \in \mathbb{R}^{m_2 \times K}\}$$

where K is large; e.g. $K = \min(m_1, m_2)$. Following Salakhutdinov and Mnih (2008), we define the following prior distribution: $U_{\cdot,j} \sim \mathcal{N}(0, \gamma_j I)$, $V_{\cdot,j} \sim \mathcal{N}(0, \gamma_j I)$ where the γ_j's are i.i.d. from an inverse gamma distribution, $\gamma_j \sim \mathcal{IG}(a, b)$.

Note that VB algorithms were used in this context by Lim and Teh (2007) (with a slightly simpler prior however: the γ_j's are fixed rather than random). Since then, this prior and variants were used in several papers (e.g. Lawrence and Urtasun, 2009; Zhou et al., 2010). Until now, no theoretical results were proved to the best of our knowledge. Two papers prove minimax-optimal rates for slightly modified estimators (by truncation), for which efficient algorithms are unknown (Mai and Alquier, 2015; Suzuki, 2014). However, using Theorems 4.2 and 4.3 we are able to prove the following: *if* there is a PAC-Bayesian bound leading to a rate for $\hat{\rho}_\lambda$ in this context, then the same rate holds for $\tilde{\rho}_\lambda$. In other words: if someone proves the conjecture that the Gibbs estimator is minimax-optimal (up to log terms) in this context, then the VB approximation will enjoy automatically the same property.

We propose the following approximation:

$$\mathcal{F} = \left\{ \rho(\mathrm{d}(U, V)) = \prod_{i=1}^{m_1} u_i(\mathrm{d}U_{i,\cdot}) \prod_{j=1}^{m_2} v_j(\mathrm{d}V_{j,\cdot}) \right\}.$$

Theorem 8.1 *Assume that $M = UV^T$ with $|U_{i,k}|, |V_{j,k}| \leq C$. Assume that $\mathrm{rank}(M) = r$ so that we can assume that $U_{\cdot,r+1} = \cdots = U_{\cdot,K} = V_{\cdot,r+1} = \cdots = V_{\cdot,K} = 0$ (note that the prior π does not depend on the knowledge of r though). Choose the prior distribution on the hyper-parameters γ_j as inverse gamma $\mathrm{Inv}{-}\Gamma(a, b)$ with $b \leq 1/[2\beta(m_1 \vee m_2)\log(2K(m_1 \vee m_2))]$. Then there is a constant $\mathcal{C}(a, C)$ such that, for any $\beta > 0$,*

$$\inf_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_\beta) \leq \mathcal{C}(a, C) \left\{ r(m_1 + m_2) \log\left[\beta b (m_1 + m_2) K\right] + \frac{1}{\beta} \right\}.$$

See the Appendix for a proof.

For instance, in Theorem 4.3, in classification and ranking we had λ, $\lambda - g(\lambda, n)$ and $\lambda + g(\lambda, n)$ of order $\mathcal{O}(n)$. In this case we would have:

$$\frac{2}{\lambda - g(\lambda, n)} \inf_{\rho \in \mathcal{F}} \mathcal{K}\left(\rho, \pi_{\frac{\lambda + g(\lambda,n)}{2}}\right) = \mathcal{O}\left(\frac{\mathcal{C}(a, C) r(m_1 + m_2) \log\left[nb(m_1 + m_2)K\right]}{n}\right),$$

and note that in this context it is know that the minimax rate is at least $r(m_1 + m_2)/n$ (Koltchinskii et al., 2011).

8.1 Algorithm

As already mentioned, the approximation family is not parametric in this case, but rather of type mean field. The corresponding VB algorithm amounts to iterating equation (4), which takes the following form in this particular case:

$$u_j(\mathrm{d}U_{j,\cdot}) \propto \exp\left\{-\frac{\lambda}{n}\sum_i \mathbb{E}_{V,U_{-j}}\left[(Y_{X_i} - (UV^T)_{X_i})^2\right] - \sum_{k=1}^K \mathbb{E}_{\gamma_j}\left[\frac{1}{2\gamma_k}\right] U_{jk}^2\right\}$$

$$v_j(\mathrm{d}V_{j,\cdot}) \propto \exp\left\{-\frac{\lambda}{n}\sum_i \mathbb{E}_{V_{-j},U}\left[(Y_{X_i} - (UV^T)_{X_i})^2\right] - \sum_{k=1}^K \mathbb{E}_{\gamma_j}\left[\frac{1}{2\gamma_k}\right] V_{jk}^2\right\}$$

$$p(\gamma_k) \propto \exp\left\{-\frac{1}{2\gamma_k}\left(\sum_j \mathbb{E}_U U_{kj}^2 + \sum_i \mathbb{E}_V V_{ik}^2\right) + (\alpha + 1)\log\frac{1}{\gamma_k} - \frac{\beta}{\gamma_k}\right\}$$

where the expectations are taken with respect to the thus defined variational approximations. One recognises Gaussian distributions for the first two, and an inverse Gamma distribution for the third. We refer to Lim and Teh (2007) for more details on this algorithm and for a numerical illustration. However, we point out that in this case, while the algorithm seems to work well in practice, there is no theoretical guarantee that it will converge to the global minimum of the problem.

9. Discussion

We showed in several important scenarios that approximating a Gibbs posterior through VB (Variational Bayes) techniques does not deteriorate the rate of convergence of the corresponding procedure. We also described practical algorithms for fast computation of these VB approximations, and provided empirical bounds that may be computed from the data to evaluate the performance of the so-obtained VB-approximated procedure. We believe these results provide a strong incentive to recommend VB as the default approach to approximate Gibbs posteriors, in lieu of Monte Carlo methods. We also developed a R package[5] for convexified losses (classification and bipartite ranking), applying the ideas of Section 6.

5. PACVB package: `https://cran.r-project.org/web/packages/PACVB/index.html`

We hope to extend our results to other applications beyond those discussed in this paper, such as regression. One technical difficulty with regression is that the risk function is not bounded, which makes our approach a bit less direct to apply. In many papers on PAC-Bayesian bounds for regression, the noise can be unbounded (usually, it is assumed to be sub-exponential), but one assumes that the predictors are bounded, see e.g. Alquier and Biau (2013). However, using the robust loss function of Audibert and Catoni, it is possible to relax this assumption (Audibert and Catoni, 2011; Catoni, 2012). This requires a more technical analysis, which we leave for further work.

Appendix A. Proofs

A.1 Preliminary remarks

Direct calculation yields, for any $\rho \ll \pi$ with $\int r_n \mathrm{d}\rho < \infty$,

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

Two well known consequences are

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

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

We will use these inequalities many times in the followings. The most frequent application will be with $h(\theta) = \lambda r_n(\theta)$ (in this case $\pi[\lambda r_n] = \hat{\rho}_\lambda$) or $h(\theta) = \pm\lambda[r_n(\theta) - R(\theta)]$, the first case leads to

$$\mathcal{K}(\rho, \hat{\rho}_\lambda) = \lambda \int r_n \mathrm{d}\rho + \mathcal{K}(\rho, \pi) + \log \int \exp(-\lambda r_n) \mathrm{d}\pi, \tag{8}$$

$$\hat{\rho}_\lambda = \arg \min_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \lambda \int r_n \mathrm{d}\rho + \mathcal{K}(\rho, \pi) \right\}, \tag{9}$$

$$-\log \int \exp(-\lambda r_n) \mathrm{d}\pi = \min_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \lambda \int r_n \mathrm{d}\rho + \mathcal{K}(\rho, \pi) \right\}. \tag{10}$$

We will use (8), (9) and (10) several times in this appendix.

A.2 Proof of the theorems in Subsection 4.1

Proof of Theorem 4.1. This proof follows the standard PAC-Bayesian approach (see Catoni (2007)). Apply Fubini's theorem to the first inequality of (1):

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

then apply the preliminary remark with $h(\theta) = \lambda[r_n(\theta) - R(\theta)]$:

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

Multiply both sides by ε and use $\mathbb{E}[\exp(U)] \geq \mathbb{P}(U > 0)$ for any U to obtain:

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

Then consider the complementary event:

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

□

Proof of Theorem 4.2. Using the same calculations as above, we have, with probability at least $1-\varepsilon$, simultaneously for all $\rho \in \mathcal{M}_+^1(\Theta)$,

$$\lambda\int R\mathrm{d}\rho\leq\lambda\int r_n\mathrm{d}\rho+f(\lambda,n)+\mathcal{K}(\rho,\pi)+\log\left(\frac{2}{\varepsilon}\right) \tag{11}$$

$$\lambda\int r_n\mathrm{d}\rho\leq\lambda\int R\mathrm{d}\rho+f(\lambda,n)+\mathcal{K}(\rho,\pi)+\log\left(\frac{2}{\varepsilon}\right). \tag{12}$$

We use (11) with $\rho = \hat{\rho}_\lambda$ and (9) to get

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

and plugging (12) into the right-hand side, we obtain

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

Now, we work with $\tilde{\rho}_\lambda = \arg\min_{\rho\in\mathcal{F}}\mathcal{K}(\rho,\hat{\rho}_\lambda)$. Plugging (8) into (11) we get, for any ρ,

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

By definition of $\tilde{\rho}_\lambda$, we have:

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

and, using (8) again, we obtain:

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

We plug (12) into the right-hand side to obtain:

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

This proves the second inequality of the theorem. In order to prove the claim

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

note that

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

This ends the proof. □

A.3 Proof of Theorem 4.3 (Subsection 4.2)

Proof of Theorem 4.3. As in the proof of Theorem 4.1, we apply Fubini, then (10) to the first inequality of (2) to obtain

$$\mathbb{E}\exp\left\{ \sup_\rho \int \left[\lambda[R(\theta)-\overline{R}] - \lambda[r_n(\theta)-\overline{r}_n] - g(\lambda,n)[R(\theta)-\overline{R}]\right]\rho(\mathrm{d}\theta) - \mathcal{K}(\rho,\pi) \right\} \leq 1$$

and we multiply both sides by $\varepsilon/2$ to get

$$\mathbb{P}\left\{ \sup_\rho \left[[\lambda - g(\lambda,n)] \left[\int R\mathrm{d}\rho - \overline{R} \right] \geq \lambda \left[\int r_n \mathrm{d}\rho - \overline{r}_n \right] + \mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right) \right] \right\} \leq \frac{\varepsilon}{2}. \quad (13)$$

We now consider the second inequality in (2):

$$\mathbb{E}\exp\left\{\lambda[r_n(\theta)-\overline{r}_n] - \lambda[R(\theta)-\overline{R}] - g(\lambda,n)[R(\theta)-\overline{R}]\right\} \leq 1.$$

The same derivation leads to

$$\mathbb{P}\left\{\sup_{\rho}\left[[\lambda - g(\lambda,n)]\left[\int r_n \mathrm{d}\rho - \overline{r}_n\right] \geq \lambda\left[\int R\mathrm{d}\rho - \overline{R}\right] + \mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right]\right\} \leq \frac{\varepsilon}{2}. \quad (14)$$

We combine (13) and (14) by a union bound argument, and we consider the complementary event: with probability at least $1-\varepsilon$, simultaneously for all $\rho \in \mathcal{M}_+^1(\Theta)$,

$$[\lambda - g(\lambda,n)]\left[\int R\mathrm{d}\rho - \overline{R}\right] \leq \lambda\left[\int r_n \mathrm{d}\rho - \overline{r}_n\right] + \mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right), \quad (15)$$

$$\lambda\left[\int r_n \mathrm{d}\rho - \overline{r}_n\right] \leq [\lambda + g(\lambda,n)]\left[\int R\mathrm{d}\rho - \overline{R}\right] + \mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right). \quad (16)$$

We now derive consequences of these two inequalities (in other words, we focus on the event where these two inequalities are satisfied). Using (9) in (15) yields

$$[\lambda - g(\lambda,n)]\left[\int R\mathrm{d}\hat{\rho}_\lambda - \overline{R}\right] \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)}\left\{\lambda\left[\int r_n \mathrm{d}\rho - \overline{r}_n\right] + \mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right\}.$$

We plug (16) into the right-hand side to obtain:

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

Now, we work with $\tilde{\rho}_\lambda$. Plugging (8) into (13) we get

$$[\lambda - g(\lambda,n)]\left[\int R\mathrm{d}\rho - \overline{R}\right] \leq \mathcal{K}(\rho,\hat{\rho}_\lambda) - \log\int \exp[-\lambda(r_n - \overline{r}_n)]\mathrm{d}\pi + \log\left(\frac{2}{\varepsilon}\right).$$

By definition of $\tilde{\rho}_\lambda$, we have:

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

Then, apply (8) again to get:

$$[\lambda - g(\lambda,n)]\left[\int R\mathrm{d}\tilde{\rho}_\lambda - \overline{R}\right] \leq \inf_{\rho\in\mathcal{F}}\left\{\lambda\int (r_n - \overline{r}_n)\mathrm{d}\rho + \mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)\right\}.$$

Plug (16) into the right-hand side to get

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

□

A.4 Proofs of Section 5

Proof of Lemma 1. Combine Theorem 2.1 p. 25 and Lemma 2.2 p. 27 in Boucheron et al. (2013). □

Proof of Lemma 2. Apply Theorem 2.10 in Boucheron et al. (2013), and plug the margin assumption. □

Proof of Corollary 5.2. We remind that thanks to (6) it is enough to prove the claim for $\mathcal{F}_1$. We apply Theorem 4.2 to get:

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

Note that the minimizer of R, $\overline{\theta}$, is not unique (because $f_\theta(x)$ does not depend on $\|\theta\|$) and we can chose it in such a way that $\|\overline{\theta}\| = 1$. Then

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

So:

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

We now restrict the infimum to distributions ν such that $\mathbf{m} = \overline{\theta}$:

$$\mathcal{B}(\mathcal{F}_1) \leq \overline{R} + \inf_{\sigma^2} \left\{ 2c\sqrt{d}\sigma + \frac{\lambda}{n} + \frac{d\log\left(\frac{\vartheta^2}{\sigma^2}\right) + \frac{d\sigma^2}{\vartheta^2} + \frac{1}{\vartheta^2} - d + 2\log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\}.$$

We put $\sigma = \frac{1}{2\lambda}$ and substitute $\frac{1}{\sqrt{d}}$ for ϑ to get

$$\mathcal{B}(\mathcal{F}_1) \leq \overline{R} + \frac{\lambda}{n} + \frac{c\sqrt{d} + d\log(4\frac{\lambda^2}{d}) + \frac{d^2}{4\lambda^2} + 2\log\left(\frac{2}{\varepsilon}\right)}{\lambda}.$$

Substitute $\sqrt{nd}$ for λ to get the desired result. □

Proof of Corollary 5.3. We apply Theorem 4.3:

$$\int (R - \overline{R})\mathrm{d}\tilde{\rho}_\lambda \leq \inf_{\mathbf{m},\sigma^2} \left\{ \frac{\lambda + g(\lambda,n)}{\lambda - g(\lambda,n)} \int (R - \bar{R})\mathrm{d}\Phi_{\mathbf{m},\sigma^2} + \frac{1}{\lambda - g(\lambda,n)} \left(2\mathcal{K}(\Phi_{\mathbf{m},\sigma^2}, \pi) + 2\log\frac{2}{\epsilon} \right) \right\}$$

where $\lambda < \frac{2n}{C+1}$. Computations similar to those in the the proof of Corollary 5.2 lead to

$$\int R d\tilde{\rho}_\lambda \leq \overline{R} + \inf_{\mathbf{m},\sigma^2} \left\{ 2c\frac{\lambda + g(\lambda,n)}{\lambda - g(\lambda,n)} \int \|\theta - \overline{\theta}\| \Phi_{\mathbf{m},\sigma^2}(\mathrm{d}\theta) + 2\frac{\frac{1}{2}\sum_{j=1}^d \left[\log\left(\frac{\vartheta^2}{\sigma^2}\right) + \frac{\sigma^2}{\vartheta^2} \right] + \frac{\|\mathbf{m}\|^2}{2\vartheta^2} - \frac{d}{2} + \log\left(\frac{2}{\varepsilon}\right)}{\lambda - g(\lambda,n)} \right\}.$$

taking $\mathbf{m} = \bar{\theta}$ and $\lambda = \frac{2n}{C+2}$, we get the result. □

A.5 Proofs of Section 6

Proof of Lemma 3. For fixed θ we can upper bound the individual risk such that:

$$0 \leq \max(0, 1 - < \theta, X_i > Y_i) \leq 1 + |< \theta, X_i >|$$

such that we can apply Hoeffding's inequality conditionally on X_i and fixed θ.

We get,

$$\begin{aligned} \mathbb{E}\left[\exp\left(\lambda(R^H - r_n^H)\right) | X_1, \cdots, X_n\right] &\leq \exp\left\{ \frac{\lambda^2}{8n^2} \sum_{i=1}^n (1 + |< \theta, X_i >|)^2 \right\} \\ &\leq \exp\left\{ \frac{\lambda^2}{4n} + \frac{\lambda^2 c_x^2}{4n} \|\theta\|^2 \right\} \end{aligned}$$

where the last inequality stems from the fact that $(a+b)^2 \leq 2\left(a^2 + b^2\right)$ and the fact that we have supposed the X_i to be bounded. We can take the expectation of this term with

respect to the X_i's and with respect to our Gaussian prior.

$$\pi\left\{\mathbb{E}\left[\exp\left(\lambda(R^H - r_n^H)\right)\right]\right\} \leq \frac{\exp\left(\frac{\lambda^2}{4n}\right)}{(2\pi)^{\frac{d}{2}}\sqrt{\vartheta^2}} \int \exp\left(\frac{\lambda^2 c_x^2}{4n}\|\theta\|^2 - \frac{1}{2\vartheta^2}\|\theta\|^2\right) \mathrm{d}\theta$$
$$\leq \frac{\exp\left(\frac{\lambda^2}{4n}\right)}{(2\pi)^{\frac{d}{2}}\sqrt{\vartheta^2}} \int \exp\left(-\frac{1}{2}\left[\frac{1}{\vartheta^2} - \frac{\lambda^2 c_x^2}{2n}\right]\|\theta\|^2\right) \mathrm{d}\theta$$

The integral is a properly defined Gaussian integral under the hypothesis that $\frac{1}{\vartheta^2} - \frac{\lambda^2 c_x^2}{2n} > 0$ hence $\lambda < \frac{1}{c_x}\sqrt{\frac{n}{\vartheta}^2}$. The integral is proportional to a Gaussian and we can directly write:

$$\pi\left\{\mathbb{E}\left[\exp\left(\lambda(R^H - r_n^H)\right)\right]\right\} \leq \frac{\exp\left(\frac{\lambda^2}{4n}\right)}{\sqrt{1 - \frac{\vartheta^2\lambda^2 c_x^2}{2n}}}$$

writing everything in the exponential gives the desired result. □

Proof of Corollary 6.1. We apply Theorem 4.2 to get:

$$\mathcal{B}_\lambda(\mathcal{F}_1) = \inf_{(\mathbf{m},\sigma^2)} \left\{ \int R^H \mathrm{d}\Phi_{\mathbf{m},\sigma^2} + \frac{\lambda}{2n} - \frac{1}{\lambda}\log\left(1 - \frac{\vartheta^2\lambda^2 c_x^2}{2n}\right) + 2\frac{\mathcal{K}(\Phi_{\mathbf{m},\sigma^2},\pi) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\}$$
$$= \inf_{(m,\sigma^2)} \left\{ \int R^H \mathrm{d}\Phi_{\mathbf{m},\sigma^2} + \frac{\lambda}{2n} - \frac{1}{\lambda}\log\left(1 - \frac{\vartheta\lambda^2 c_x^2}{2n}\right) \right.$$
$$\left. + 2\frac{\frac{1}{2}\sum_{j=1}^d \left[\log\left(\frac{\vartheta^2}{\sigma^2}\right) + \frac{\sigma^2}{\vartheta^2}\right] + \frac{\|m\|^2}{2\vartheta^2} - \frac{d}{2} + \log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\}.$$

We use the fact that the hinge loss is Lipschitz and that the (X_i) are uniformly bounded $\|X\|_\infty < c_x$. We get $R^H(\theta) \leq \bar{R}^H + c_x\sqrt{d}\|\theta - \bar{\theta}\|$ and restrict the infemum to distributions ν such that $m = \overline{\theta}$:

$$\mathcal{B}(\mathcal{F}_1) \leq \overline{R}^H + \inf_{\sigma^2} \left\{ c_x d\sigma^2 + \frac{\lambda}{2n} - \frac{1}{\lambda}\log\left(1 - \frac{\vartheta^2\lambda^2 c_x^2}{2n}\right) + \frac{d\log\left(\frac{\vartheta^2}{\sigma^2}\right) + \frac{d\sigma^2}{\vartheta^2} + \frac{1}{\vartheta^2} - d + 2\log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\}.$$

We specify $\sigma^2 = \frac{1}{\sqrt{dn}}$ and $\lambda = c_x\sqrt{\frac{n}{\vartheta^2}}$ such that we get:

$$\mathcal{B}(\mathcal{F}_1) \leq \overline{R^H} + c_x\sqrt{\frac{d}{n}} + \frac{\sqrt{\vartheta^2}}{2c_x\sqrt{n}} - c_x\sqrt{\frac{\vartheta^2}{n}}\log\left(1 - \frac{1}{2}\right) + d\frac{c_x\vartheta}{\sqrt{n}}\log\left(\vartheta^2\sqrt{nd}\right) + c_x\vartheta\frac{\frac{d}{n\vartheta^2} + \frac{1}{\vartheta^2} - d + 2\log\left(\frac{2}{\varepsilon}\right)}{\sqrt{n}}.$$

To get the correct rate we take the prior variance to be $\vartheta^2 = \frac{1}{d}$ by replacing in the above equation we get the desired result.

□

Proof of Theorem 6.2. From Nesterov (2004) (th. 3.2.2) we have the following bound on the objective function minimized by VB, (the objective is not uniformlly Lipschitz)

$$\rho^k(r_n^H) + \frac{1}{\lambda}\mathcal{K}(\rho^k, \pi) - \inf_{\rho \in \mathcal{F}_1}\left\{\rho(r_n^H) + \frac{1}{\lambda}\mathcal{K}(\rho, \pi)\right\} \leq \frac{LM}{\sqrt{1+k}}. \tag{17}$$

We have from equation (11) specified for measures ρ^k probability $1-\varepsilon$,

$$\lambda \int r_n^H \mathrm{d}\rho^k \leq \lambda \int R^H \mathrm{d}\rho^k + f(\lambda, n) + \mathcal{K}(\rho^k, \pi) + \log\left(\frac{1}{\varepsilon}\right)$$

Combining the two equations yields,

$$\int R^H \mathrm{d}\rho^k \leq \frac{LM}{\sqrt{1+k}} + \frac{1}{\lambda}f(n,\lambda) + \inf_{\rho \in \mathcal{F}_1}\left\{\rho(r_n^H) + \frac{1}{\lambda}\mathcal{K}(\rho, \pi)\right\} + \frac{1}{\lambda}\log\frac{1}{\varepsilon}$$

We can therefore write for any $\rho \in \mathcal{F}_1$,

$$\int R^H \mathrm{d}\rho^k \leq \frac{LM}{\sqrt{1+k}} + \frac{1}{\lambda}f(n,\lambda) + \rho(r_n^H) + \frac{1}{\lambda}\mathcal{K}(\rho, \pi) + \frac{1}{\lambda}\log\frac{1}{\varepsilon}$$

Using equation (11) a second time we get with probability $1-\varepsilon$

$$\int R^H \mathrm{d}\rho^k \leq \frac{LM}{\sqrt{1+k}} + \frac{2}{\lambda}f(n,\lambda) + \rho(R^H) + \frac{2}{\lambda}\mathcal{K}(\rho, \pi) + \frac{2}{\lambda}\log\frac{2}{\varepsilon}$$

Because this is true for any $\rho \in \mathcal{F}_1$ in $1-\varepsilon$ we can write the bound for the smallest measure in $\mathcal{F}_1$.

$$\int R^H \mathrm{d}\rho^k \leq \frac{LM}{\sqrt{1+k}} + \frac{2}{\lambda}f(n,\lambda) + \inf_{\rho \in \mathcal{F}_1}\left\{\rho(R^H) + \frac{2}{\lambda}\mathcal{K}(\rho, \pi)\right\} + \frac{2}{\lambda}\log\frac{2}{\varepsilon}$$

By taking the Gaussian measure with variance $\frac{1}{n}$ and mean $\overline{\theta}$ in the infimum and taking $\lambda = \frac{1}{c_x}\sqrt{nd}$ and $\vartheta = \frac{1}{d}$, we can use the results of Corollary 6.1 to get the result.□

A.6 Proofs of Section 7

Proof of Lemma 4. The idea of the proof is to use Hoeffding's decomposition of U-statistics combined with Hoeffding's inequality for iid random variables. This was done in ranking by Clémençon et al. (2008), and later in Robbiano (2013); Ridgway et al. (2014) for ranking via aggregation and Bayesian statistics. The proof is as follows: we define

$$q_{i,j}^{\theta} = \mathbf{1}_{(Y_i - Y_j)f_\theta(X_i, X_j) < 0} - R(\theta)$$

so that

$$U_n := \frac{1}{n(n-1)} \sum_{i,j} q_{i,j}^{\theta} = r_n(\theta) - R(\Theta).$$

From Hoeffding (1948) we have

$$U_n = \frac{1}{n!} \sum_{\pi} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} q^{\theta}_{\pi(i),\pi(i+\lfloor \frac{n}{2} \rfloor)}$$

where the sum is taken over all the permutations π of $\{1, \ldots, n\}$. Jensen's inequality leads to

$$\begin{aligned}
\mathbb{E} \exp[\lambda U_n] &= \mathbb{E} \exp \left[\lambda \frac{1}{n!} \sum_{\pi} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} q^{\theta}_{\pi(i),\pi(i+\lfloor \frac{n}{2} \rfloor)} \right] \\
&\leq \frac{1}{n!} \sum_{\pi} \mathbb{E} \exp \left[\frac{\lambda}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} q^{\theta}_{\pi(i),\pi(i+\lfloor \frac{n}{2} \rfloor)} \right].
\end{aligned}$$

We now use, for each of the terms in the sum we use the same argument as in the proof of Lemma 1 to get

$$\mathbb{E} \exp[\lambda U_n] \leq \frac{1}{n!} \sum_{\pi} \exp \left[\frac{\lambda^2}{2 \lfloor \frac{n}{2} \rfloor} \right] \leq \exp \left[\frac{\lambda^2}{n-1} \right]$$

(in the last step, we used $\lfloor \frac{n}{2} \rfloor \geq (n-1)/2$). We proceed in the same way to upper bound $\mathbb{E} \exp[-\lambda U_n]$. $\square$

Proof of Lemma 5. As already done above, we use Bernstein inequality and Hoeffding decomposition. Fix θ. We define this time

$$q_{i,j}^{\theta} = \mathbf{1}\{\langle \theta, X_i - X_j \rangle (Y_i - Y_j) < 0\} - \mathbf{1}\{\langle \overline{\theta}, X_i - X_j \rangle (Y_i - Y_j) < 0\} - R(\theta) + \overline{R}$$

so that

$$U_n := r_n(\theta) - \overline{r}_n - R(\theta) + \overline{R} = \frac{1}{n(n-1)} \sum_{i \neq j} q_{i,j}^{\theta}.$$

Then,

$$U_n = \frac{1}{n!} \sum_{\pi} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} q^{\theta}_{\pi(i),\pi(i+\lfloor \frac{n}{2} \rfloor)}.$$

Jensen's inequality:

$$\mathbb{E}\exp[\lambda U_n] = \mathbb{E}\exp\left[\lambda\frac{1}{n!}\sum_{\pi}\frac{1}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} q^{\theta}_{\pi(i),\pi(i+\lfloor\frac{n}{2}\rfloor)}\right]$$
$$\leq \frac{1}{n!}\sum_{\pi}\mathbb{E}\exp\left[\frac{\lambda}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} q^{\theta}_{\pi(i),\pi(i+\lfloor\frac{n}{2}\rfloor)}\right].$$

Then, for each of the terms in the sum, use Bernstein's inequality:

$$\mathbb{E}\exp\left[\frac{\lambda}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} q^{\theta}_{\pi(i),\pi(i+\lfloor\frac{n}{2}\rfloor)}\right] \leq \exp\left[\frac{\mathbb{E}((q^{\theta}_{\pi(1),\pi(1+\lfloor\frac{n}{2}\rfloor)})^2)\frac{\lambda^2}{\lfloor\frac{n}{2}\rfloor}}{2\left(1-2\frac{\lambda}{\lfloor\frac{n}{2}\rfloor}\right)}\right].$$

We use again $\lfloor\frac{n}{2}\rfloor \geq (n-1)/2$. Then, as the pairs (X_i, Y_i) are iid, we have $\mathbb{E}((q^{\theta}_{\pi(1),\pi(1+\lfloor\frac{n}{2}\rfloor)})^2) = \mathbb{E}((q^{\theta}_{1,2})^2)$ and then $\mathbb{E}((q^{\theta}_{1,2})^2) \leq C[R(\theta) - \overline{R}]$ thanks to the margin assumption. So

$$\mathbb{E}\exp\left[\frac{\lambda}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} q^{\theta}_{\pi(i),\pi(i+\lfloor\frac{n}{2}\rfloor)}\right] \leq \exp\left[\frac{C[R(\theta) - \overline{R}]\frac{\lambda^2}{n-1}}{\left(1-\frac{4\lambda}{n-1}\right)}\right].$$

This ends the proof of the proposition. □

Proof of Corollary 7.2. The calculations are similar to the ones in the proof of Corollary 5.2 so we don't give the details. Note that when we reach

$$\mathcal{B}_{\lambda}(\mathcal{F}_1) \leq \overline{R} + \frac{2\lambda}{n-1} + \frac{c\sqrt{d} + d\log(2\lambda) + \frac{d}{4\lambda^2} + 2\log\left(\frac{2\mathrm{e}}{\varepsilon}\right)}{\lambda},$$

an approximate minimization with respect to λ leads to the choice $\lambda = \sqrt{\frac{d(n-1)}{2}}$. □

A.7 Proofs of Section 8

Proof. First, note that, for any ρ,

$$\mathcal{K}(\rho, \pi_{\beta}) = \beta\int(R-\overline{R})\mathrm{d}\rho + \mathcal{K}(\rho,\pi) + \log\int\exp\left[-\beta(R-\overline{R})\right]\mathrm{d}\pi$$
$$\leq \beta\int(R-\overline{R})\mathrm{d}\rho + \mathcal{K}(\rho,\pi).$$

Now, we define a subset of $\mathcal{F}$ that will be used for the calculation of the bound. We define for $\delta > 0$ the probability distribution $\rho_{U,V,\delta}(\mathrm{d}\theta)$ as π conditioned to $\theta = \mu\nu^T$ with μ is

uniform on $\{\forall(i,\ell), |\mu_{i,\ell} - U_{i,\ell}| \leq \delta\}$ and ν is uniform on $\{\forall(j,\ell), |\nu_{i,\ell} - V_{j,\ell}| \leq \delta\}$. Note that

$$\begin{aligned}\int (R-\overline{R})\mathrm{d}\rho_{M,N,\delta} &= \int \mathbb{E}((\theta_X - M_X)^2)\rho_{U,V,\delta}(\mathrm{d}\theta) \\ &\leq \int 3\mathbb{E}(((UV^T)_X - M_X)^2)\rho_{U,V,\delta}(\mathrm{d}(\mu,\nu)) \\ &\quad + 3\int \mathbb{E}(((U\nu^T)_X - (UV^T)_X)^2)\rho_{U,V,\delta}(\mathrm{d}(\mu,\nu)) \\ &\quad + 3\int \mathbb{E}(((\mu\nu^T)_X - (U\nu^T)_X)^2)\rho_{U,V,\delta}(\mathrm{d}(\mu,\nu)).\end{aligned}$$

By definition, the first term is $= 0$. Moreover:

$$\begin{aligned}&\int \mathbb{E}(((U\nu^T)_X - (UV^T)_X)^2)\rho_{U,V,\delta}(\mathrm{d}(\mu,\nu)) \\ &= \int \frac{1}{m_1 m_2}\sum_{i,j}\left[\sum_k U_{i,k}(\nu_{j,k} - V_{j,k})\right]^2 \rho_{U,V,\delta}(\mathrm{d}(\mu,\nu)) \\ &\leq \int \frac{1}{m_1 m_2}\sum_{i,j}\left[\sum_k U_{i,k}^2\right]\left[\sum_k (\nu_{j,k} - V_{j,k})^2\right]\rho_{U,V,\delta}(\mathrm{d}(\mu,\nu)) \\ &\leq KrC^2\delta^2.\end{aligned}$$

In the same way,

$$\begin{aligned}\int \mathbb{E}(((\mu\nu^T)_X - (U\nu^T)_X)^2)\rho_{U,V,\delta}(\mathrm{d}(\mu,\nu)) &\leq \int \|\mu - U\|_F^2 \|\nu\|_F^2 \rho_{U,V,\delta}(\mathrm{d}(\mu,\nu)) \\ &\leq Kr(C+\delta)^2\delta^2.\end{aligned}$$

So:

$$\int (R-\overline{R})\mathrm{d}\rho_{M,N,\delta} \leq 2Kr\delta^2(C+\delta^2).$$

Now, let us consider the term $\mathcal{K}(\rho_{U,V,\delta}, \pi)$. An explicit calculation is possible but tedious. Instead, we might just introduce the set $\mathcal{G}_\delta = \{\theta = \mu\nu^T, \|\mu - U\|_F \leq \delta, \|\nu - V\|_F \leq \delta\}$ and note that $\mathcal{K}(\rho_{U,V,\delta}, \pi) \leq \log \frac{1}{\pi(\mathcal{G}_\delta)}$. An upper bound for $\mathcal{G}_\delta$ is calculated page 317-320 in Alquier (2014) and the result is given by (10) in this reference:

$$\begin{aligned}\mathcal{K}(\rho_{U,V,\delta}, \pi) \leq\ & 4\delta^2 + 2\|U\|_F^2 + 2\|N\|_F^2 + 2\log(2) \\ &+ (m_1+m_2)r\log\left(\frac{1}{\delta}\sqrt{\frac{3\pi(m_1 \vee m_2)K}{4}}\right) + 2K\log\left(\frac{\Gamma(a)3^{a+1}\exp(2)}{b^{a+1}2^a}\right)\end{aligned}$$

as soon as the restriction $b \leq \frac{\delta^2}{2m_1 K \log(2m_1 K)}, \frac{\delta^2}{2m_2 K \log(2m_2 K)}$ is satisfied. So we obtain:

$$\mathcal{K}(\rho_{U,V,\delta}, \pi_\beta) \leq \beta 2Kr\delta^2(C+\delta^2) + 4\delta^2 + 2\|U\|_F^2 + 2\|N\|_F^2 + 2\log(2) \\ + (m_1+m_2)r \log\left(\frac{1}{\delta}\sqrt{\frac{3\pi(m_1 \vee m_2)K}{4}}\right) + 2K\log\left(\frac{\Gamma(a)3^{a+1}\exp(2)}{b^{a+1}2^a}\right).$$

Note that $\|U\|_F^2 \leq C^2 r m_1$, $\|V\|_F^2 \leq C^2 r m_2$ and $K \leq m_1 + m_2$ so it is clear that the choice $\delta = \sqrt{\frac{1}{\beta}}$ and $b \leq \frac{1}{2\beta(m_1 \vee m_2)\log(2K(m_1 \vee m_2))}$ leads to the existence of a constant $\mathcal{C}(a, C)$ such that

$$\mathcal{K}(\rho_{U,V,\delta}, \pi_\beta) \leq \mathcal{C}(a,C)\left\{r(m_1+m_2)\log\left[\beta b(m_1+m_2)K\right] + \frac{1}{\beta}\right\}.$$

□

Appendix B. Implementation details

B.1 Sequential Monte Carlo

Tempering SMC approximates iteratively a sequence of distribution ρ_{λ_t}, with

$$\rho_{\lambda_t}(\mathrm{d}\theta) = \frac{1}{Z_t}\exp\left(-\lambda_t r_n(\theta)\right)\pi(\mathrm{d}\theta),$$

and temperature ladder $\lambda_0 = 0 < \ldots < \lambda_T = \lambda$. The pseudo-code below is given for an adaptive sequence of temperatures.

Algorithm 1 Tempering SMC

Input N (number of particles), $\tau \in (0,1)$ (ESS threshold), $\kappa > 0$ (random walk tuning parameter)

Init. Sample $\theta_0^i \sim \pi_\xi(\theta)$ for $i = 1$ to N, set $t \leftarrow 1$, $\lambda_0 = 0$, $Z_0 = 1$.

Loop a. Solve in λ_t the equation

$$\frac{\{\sum_{i=1}^N w_t(\theta_{t-1}^i)\}^2}{\sum_{i=1}^N \{w_t(\theta_{t-1}^i))^2\}} = \tau N, \quad w_t(\theta) = \exp[-(\lambda_t - \lambda_{t-1}) r_n(\theta)] \tag{18}$$

using bisection search. If $\lambda_t \geq \lambda_T$, set $Z_T = Z_{t-1} \times \left\{ \frac{1}{N} \sum_{i=1}^N w_t(\theta_{t-1}^i) \right\}$, and stop.

b. Resample: for $i = 1$ to N, draw A_t^i in $1, \ldots, N$ so that $\mathbb{P}(A_t^i = j) = w_t(\theta_{t-1}^j) / \sum_{k=1}^N w_t(\theta_{t-1}^k)$; see Algorithm 2 in the appendix.

c. Sample $\theta_t^i \sim M_t(\theta_{t-1}^{A_t^i}, \mathrm{d}\theta)$ for $i = 1$ to N where M_t is a MCMC kernel that leaves invariant π_t; see comments below.

d. Set $Z_t = Z_{t-1} \times \left\{ \frac{1}{N} \sum_{i=1}^N w_t(\theta_{t-1}^i) \right\}$.

The algorithm outputs a weighted sample (w_T^i, θ_T^i) approximately distributed as target posterior, and an unbiased estimator of the normalizing constant Z_{λ_T}.

Step **b.** of algorithm B.1 depends of a resampling algorithm. We choose to use Systematic resampling, see Algorithm 2.

Algorithm 2 Systematic resampling

Input: Normalised weights $W_t^j := w_t(\theta_{t-1}^j)/\sum_{i=1}^N w_t(\theta_{t-1}^i)$.

Output: indices $A^i \in \{1, \ldots, N\}$, for $i = 1, \ldots, N$.

a. Sample $U \sim \mathcal{U}([0, 1])$.

b. Compute cumulative weights as $C^n = \sum_{m=1}^n NW^m$.

c. Set $s \leftarrow U$, $m \leftarrow 1$.

d. **For** $n = 1 : N$

While $C^m < s$ **do** $m \leftarrow m + 1$.

$A^n \leftarrow m$, and $s \leftarrow s + 1$.

End For

For the MCMC step, we used a Gaussian random-walk Metropolis kernel, with a covariance matrix for the random step that is proportional to the empirical covariance matrix of the current set of simulations.

B.2 Optimizing the bound

A natural idea to find a global optimum of the objective is to try to solve a sequence of local optimization problems with increasing temperatures. For $\gamma = 0$ the problem can be solved exactly (as a KL divergence between two Gaussians). Then, for two consecutive temperatures, the corresponding solutions should be close enough.

This idea has been coined under several names. It has a long history in variational inference under the name 'deterministic annealing'; see e.g. Yuille (2010) for an application to Markov random fields. In addition the intermediate results can be of interest in our case for selecting the temperature. One can compute the bound at almost no additional cost as a function of the current risk. In turns this can be used to monitor the bound.

Algorithm 3 Deterministic annealing

Input $(\lambda_t)_{t\in[0,T]}$ a sequence of temperature

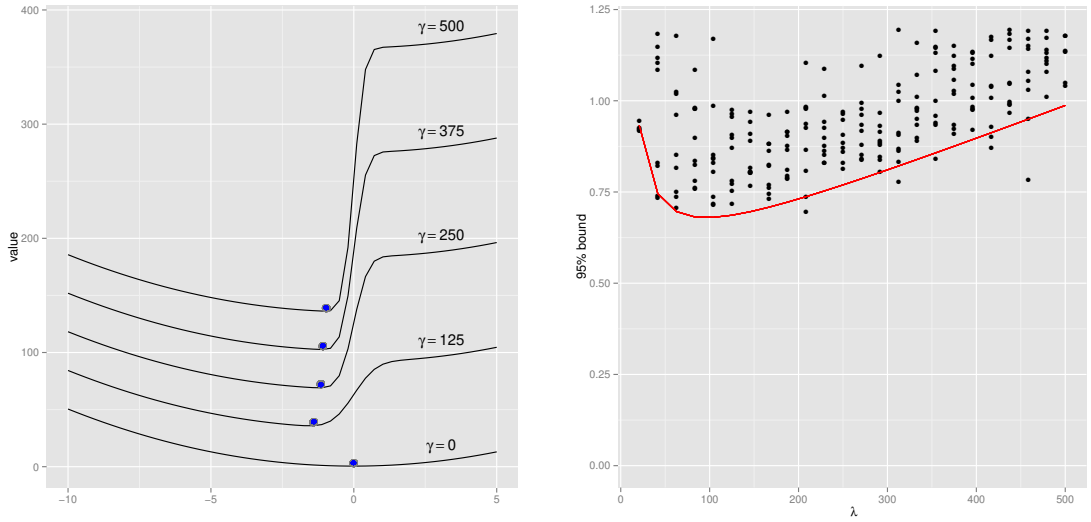
Init. Set $m = 0$ and $\Sigma = \vartheta I_d$, the values minimizing KL-divergence for $\lambda = 0$

Loop t=1,...,T

a. $m^{\lambda_t}, \Sigma^{\lambda_t}$ = Minimize $\mathcal{L}^{\lambda_t}(m, \Sigma)$ using some local optimization routine with initial points $m^{\lambda_{t-1}}, \Sigma^{\lambda_{t-1}}$

b. Break if the empirical bound increases.

End Loop



(a) A one dimensional problem

(b) Empirical bound

Figure 2: Deterministic annealing on a Pima Indians with one covariate and full model resp.
The right panel gives the empirical bound obtained for the DA method (in red). The dots are the results of direct global optimization based on L-BFGS algorithms (with starting values drawn from the prior). Each optimization problem is repeated 20 times.

We find that using a deterministic annealing algorithm with a limited amount of steps helps in finding a high enough optimum. On the left panel of Figure 2, we can see the one

dimensional case where the initial problem $\gamma = 0$ corresponds to a convex minimization problem and where the increasing temperature gradually complexifies the optimization problem. Figure 2 shows that the solution given by DA is in average lower than randomly initialized optimization.

Appendix C. Stochastic gradient descent

The stochastic gradient descent algorithm used in Section 7 is described as Algorithm 4.

Algorithm 4 Stochastic Gradient Descent

Input B a batch size, an unbiased estimator of the gradient $\hat{\nabla}_B f$, $\eta \in (0,1)$ and c

While $\neg$converged

a. $x_{t+1} = x_t - \lambda_t \hat{\nabla}_B f(x_t)$

b. Update $\lambda_{t+1} = \frac{1}{(t+c)^\eta}$

End Loop

In all our experiments we take $c = 1$ and $\eta = 0.9$.

References

P. Alquier. Bayesian methods for low-rank matrix estimation: short survey and theoretical study. In S. Jain, R. Munos, F. Stephan, and T. Zeugmann, editors, *Algorithmic Learning Theory*. Springer - Lecture Notes in Artificial Intelligence, 2014.

P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14(1):243–280, 2013.

P. Alquier and X. Li. Prediction of quantiles by statistical learning and application to GDP forecasting. In J.-G. Ganascia, P. Lenca, and J.-M. Petit, editors, *Discovery Science*. Springer - Lecture Notes in Artificial Intelligence, 2012.

J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *Ann. Statist.*, 39 (5):2766–2794, 10 2011. doi: 10.1214/11-AOS918. URL `http://dx.doi.org/10.1214/11-AOS918`.

J. Bennett and S. Lanning. The netflix prize. In *Proceedings of KDD Cup and Workshop 07*, 2007.

C. M. Bishop. *Pattern Recognition and Machine Learning*, chapter 10. Springer, 2006.

P. Bissiri, C. Holmes, and S. Walker. A general framework for updating belief distributions. *arXiv preprint arXiv:1306.6430*, 2013.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities*. Oxford University Press, 2013.

R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer. A stochastic quasi-Newton method for large-scale optimization. *arXiv preprint arXiv:1401.7020*, 2014.

E. J. Candès and T. Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2044061. URL `http://dx.doi.org/10.1109/TIT.2010.2044061`.

O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

O. Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007.

O. Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. H. Poincaré Probab. Statist.*, 48(4):1148–1185, 11 2012. doi: 10.1214/11-AIHP454. URL `http://dx.doi.org/10.1214/11-AIHP454`.

V. Chernozhukov and H. Hong. An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346, 2003.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *Ann. Stat.*, 36(2):844–874, 2008.

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Science*, 78(5):1423–1443, 2012.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, 68(3):411–436, 2006. ISSN 1467-9868.

P. J. Green, K. Latuszynski, M. Pereyra, and C. P. Robert. Bayesian computation: a perspective on the current state, and sampling backwards and forwards. Preprint arXiv:1502.01148, 2015.

B. Guedj and P. Alquier. PAC-Bayesian estimation and prevision in sparse additive models. *Electronic Journal of Statistics*, 7:264–291, 2013.

W. Hoeffding. Probability inequalities for sums of random variables. *Annals of Mathematical Statistics*, 10:293–325, 1948.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

W. Jiang and M. A. Tanner. Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36(5):2207–2231, 2008.

M. I. Jordan, Z. Ghahrapani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, (37):183–233, 1999.

M. E. Khan. Decoupled variational Gaussian inference. In *Advances in Neural Information Processing Systems*, pages 1547–1555, 2014.

M. E. Khan, A. Aravkin, M. Friedlander, and M. Seeger. Fast dual variational inference for non-conjugate latent gaussian models. In *Proceedings of The 30th International Conference on Machine Learning*, pages 951–959, 2013.

V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

N. D. Lawrence and R. Urtasun. Non-linear matrix factorization with Gaussian processes. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 601–608. ACM, 2009.

G. Lecué. Méthodes d'aggrégation: optimalité et vitesses rapides. Ph.D. thesis, Université Paris 6, 2007.

F. Leisch and E. Dimitriadou. mlbench: Machine learning benchmark problems. *R package version*, 2:1–1, 2010.

Y. J. Lim and Y. W. Teh. Variational Bayesian approach to movie rating prediction. *Proceedings of KDD Cup and Workshop*, 7:15–21, 2007.

D. J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2002.

T. T. Mai and P. Alquier. A Bayesian approach for matrix completion: optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 9:823–841, 2015.

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of of the Twelth Annual Conference On Computational Learning Theory, Santa Cruz, California (Electronic)*, pages 164–170. ACM, New-York, 1999.

D.A McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234. ACM, New York, 1998.

Y. Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.

M. Opper and C. Archambeau. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

G. Parisi. *Statistical field theory*. Addison-Wesley, New-York, 1988.

J. Ridgway, P. Alquier, N. Chopin, and F. Liang. PAC-Bayesian AUC classification and scoring. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 658–666. Curran Associates Inc., 2014.

S. Robbiano. Upper bounds and aggregation in bipartite ranking. *Electronic Journal of Statistics*, 7:1249–1271, 2013.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887. ACM, 2008.

J. Shawe-Taylor and R.C. Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9. ACM, 1997.

T. Suzuki. Convergence rate of Bayesian tensor estimator: Optimal rate without restricted strong convexity. arXiv preprint arXiv:1408.3092 (accepted by ICML2015), 2014.

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

O. Wintenberger. Deviation inequalities for sums of weakly dependent time series. *Electronic Communications in Probability*, 15:489–503, 2010.

Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10:25–47, 2004.

A. Yuille. Belief propagation, mean-field and the Bethe approximation. Technical report, Dept. Statistics UCLA, 2010.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.

T. Zhang. Information theoretical upper and lower bounds for statistical estimation. *IEEE Transaction on Information Theory*, 52:1307–1321, 2006.

M. Zhou, C. Wang, M. Chen, J. Paisley, D. Dunson, and L. Carin. Nonparametric bayesian matrix completion. *Proc. IEEE SAM*, 2010.