CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

Série des Documents de Travail

n° 2017-39

Concentration of tempered posteriors and of their variational approximations

P. ALQUIER[1]
J. RIDGWAY[2]

[1] CREST-ENSAE, France. E-mail: Pierre.alquier@ensae.fr
[2] INRIA, Lille. E-mail : james.lp.ridgway@gmail.com

Concentration of tempered posteriors and of their variational approximations

Pierre Alquier[(1)] and James Ridgway[(2)]
(1) CREST, ENSAE, Université Paris Saclay*
(2) INRIA Lille

Abstract

While Bayesian methods are extremely popular in statistics and machine learning, their application to massive datasets is often challenging, when possible at all. Indeed, the classical MCMC algorithms are prohibitively slow when both the model dimension and the sample size are large. Variational Bayesian methods aim at approximating the posterior by a distribution in a tractable family. Thus, MCMC are replaced by an optimization algorithm which is orders of magnitude faster. VB methods have been applied in such computationally demanding applications as including collaborative filtering, image and video processing, NLP and text processing... However, despite very nice results in practice, the theoretical properties of these approximations are usually not known. In this paper, we propose a general approach to prove the concentration of variational approximations of fractional posteriors. We apply our theory to two examples: matrix completion, and Gaussian VB.

1 Introduction

1.1 Motivation

In many application of Bayesian statistics, the posterior is not tractable. Markov Chain Monte Carlo algorithms (MCMC) were developped to allow the statistician's to sample from the posterior distribution even in situations where a close form is not available. MCMC methods were successfully used in many applications, and are still one of the most valuable tools in the statistician toolbox. However, many modern applications of statistics and machine learning involve such massive datasets that sampling schemes such as MCMC have become impractical. In order to allow the use of Bayesian approaches with these datasets, it is actually much faster to use optimization algorithms to compute variational approximations of the posterior. Variational Bayes (VB) has indeed become a corner stone algorithm for fast Bayesian inference.

VB has been applied to many challenging problems: matrix completion for collaborative filtering [20], NLP on massive datasets [15], video processing [19], classification with Gaussian processes [13]... For more

*This author gratefully acknowledges financial support from the research programme *New Challenges for New Data* from LCL and GENES, hosted by the *Fondation du Risque* and from Labex ECODEC (ANR-11-LABEX-0047).

background on the methodology the interested reader is referred to the short introduction in Chapter 10 in [5]. A more recent introduction can be found in [6].

Despite its practical success very little attention has been put towards theoretical guaranties for VB. In [29] the authors gave asymptotic results for specific approximations of exponential models. More recently [3] studied VB approximations in machine learning problems. In the distribution-free setting of machine learning, there is actually no likelihood, but a pseudo-likelihood can be defined through a suitable loss function - thus the term pseudo-posterior. Thanks to PAC-Bayesian inequalities from [9, 10], they proved consistency and derived rates of convergence for the approximations in many examples. However, the tools used in [3] are essentially valid for bounded loss functions, so there is no direct way to adapt their method to study VB approximations of posteriors when the log-likelihood is unbounded.

In this paper, we propose a general way to derive concentration rates for approximations of fractional posteriors. Concentration rates are the most natural way to assess "frequentist guarantees for Bayesian estimators": the objective is to prove that the posterior is asymptotically highly concentrated around the true value of the parameter. This approach is now very well understood, we refer the reader to the milestone paper [12]. A recent survey can be found in [23]. Recently, [4] studied the situation where the likelihood $L(\theta)$ is replaced by $L^\alpha(\theta)$ for $0 < \alpha < 1$ in the construction of the posterior, leading to what is usually called a *fractional* or *tempered* posterior. They proved that concentration of the fractional posterior requires actually less hypothesis than concentration of the (true) posterior. Usnig tools from [4], we analyze the concentration of VB approximations of (fractional) posteriors. This allows to obtain consistency and concentration rates in many situations. Especially, we derive a condition for the VB approximation to concentrate at the same rate as the fractional posterior.

1.2 Definitions and notations

We observe a collection of n i.i.d. random variables $(X_1, \cdots, X_n) = X_1^n$ in a measured sample space $(\mathbb{X}, \mathcal{X}, \mathbb{P})$. Let $(P_\theta, \theta \in \Theta)$ be a statistical model, that is, a collection of probability distributions. The objective of the paper will be to estimate the probability distribution of the X_i's. Most of the results will be stated under the assumption that the model is well specified, that is, there exists $\theta_0 \in \Theta$ such that $\mathbb{P} \equiv P_{\theta_0}^{\otimes n}$. However, we will also provide results in the case $\mathbb{P} \equiv (P^*)^{\otimes n}$ where P^* does not belong to the model. In this case, we will only estimate the best approximation of P^* in the model. Still, we will assume in what follows that $\mathbb{P} \equiv P_{\theta_0}^{\otimes n}$, unless explicitely stated otherwise.

Assume that Q is a dominating measure for this family of distributions, and put $p_\theta = \frac{\mathrm{d}P_\theta}{\mathrm{d}Q}(\theta)$. Let $\mathcal{M}_1^+(E)$ the set of all probability distributions on some measurable space $(E, \mathcal{E})$, where the σ-algebra will be given by context.The set Θ is equipped with a σ-algebra $\mathcal{T}$. Let $\pi \in \mathcal{M}_1^+(\Theta)$ denote the prior.

Definition 1.1 (Likelihood) *The likelihood is given by*

$$L_n(\theta) = \prod_{i=1}^n p_\theta(X_i).$$

Put, for any $(\theta, \theta') \in \Theta$*:*

$$r_n(\theta, \theta') = \sum_{i=1}^{n} \log \frac{p_{\theta'}(X_i)}{p_\theta(X_i)}$$

the negative log-likelihood ratio.

In order to manipulate the quantities of the preceding display we need to assume that $(X_1^n, \theta) \mapsto r_n(\theta, \theta_0)$ is measurable for the product σ-field $\mathcal{X} \otimes \mathcal{T}$. This imposes some regularity on $\Theta \times \mathbb{X}$ that will be implicitly assumed in the rest of the paper.

Definition 1.2 (Divergences) *Let* $\alpha \in (0,1)$*. The* α*-Renyi divergence between two probability distributions* P *and* R*:*

$$D_\alpha(P,R) = \begin{cases} \frac{1}{\alpha-1} \log \int \left(\frac{\mathrm{d}P}{\mathrm{d}R}\right)^{\alpha-1} \mathrm{d}P \text{ if } P \ll R \\ +\infty \text{ otherwise.} \end{cases}$$

The Kullback-Leibler (KL) divergence:

$$\mathcal{K}(P,R) = \begin{cases} \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}R}\right) \mathrm{d}P \text{ if } P \ll R \\ +\infty \text{ otherwise.} \end{cases}$$

Remark 1.1 *We remind a few useful properties, proven in [28], the reader already familiar with Renyi divergences can skip this remark. Let* $d_{TV}(P,R)$ *denote the total variation distance. We have* $D_\alpha(P,R) \xrightarrow[\alpha \to 1^-]{} \mathcal{K}(P,R)$ *which gives ground to the notation* $D_1(P,R) = \mathcal{K}(P,R)$*. For any* $\alpha \in (0,1]$*,*

$$\frac{\alpha}{2} d_{TV}^2(P,R) \leq D_\alpha(P,R),$$

for $\alpha = 1$ *this is Pinsker's inequality. The map* $\alpha \mapsto D_\alpha(P,R)$ *is nondecreasing in* $\alpha \in (0,1]$*. Additivity holds: for any* $\alpha \in (0,1]$*,*

$$D_\alpha(P_1 \otimes P_2, R_1 \otimes R_2) = D_\alpha(P_1, R_1) + D_\alpha(P_2, R_2)$$

which is especially useful for i.i.d observations: $D_\alpha(P^{\otimes n}, R^{\otimes n}) = n D_\alpha(P,R)$*. Finally, let* $H(P,R)$ *denote Hellinger's distance, that is* $H^2(P,R) = \int (\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}R})^2$*, we have*

$$H^2(P,R) = 2\left(1 - \mathrm{e}^{-\frac{1}{2} D_{\frac{1}{2}}(P,R)}\right) \leq D_{\frac{1}{2}}(P,R).$$

As mentioned in the introduction we are interested in theoretical guaranties of variational approximations of fractional posteriors. The fractional posterior, that will be our *ideal* estimator, is given by

$$\pi_{n,\alpha}(\mathrm{d}\theta | X_1^n) := \frac{\mathrm{e}^{\alpha r_n(\theta,\theta_0)} \pi(\mathrm{d}\theta)}{\int \mathrm{e}^{\alpha r_n(\theta,\theta_0)} \pi(\mathrm{d}\theta)} \propto L_n^\alpha(\theta) \pi(\mathrm{d}\theta),$$

using the notation of [4]. The variational approximation of the preceding display is defined as the projection in KL divergence onto a predefined family of distributions $\mathcal{F}$.

Definition 1.3 (Variational Bayes) *Let* $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$*. We define the VB approximation* $\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n)$ *by*

$$\tilde{\pi}_{n,\alpha}(\cdot | X_1^n) = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha}(\cdot | X_1^n)).$$

In Section 2 we propose general theorems to prove the concentration of the tempered posterior. One of the key assumptions to prove the concentration of the posterior is that the prior gives enough mass to neighborhoods of the true parameter, see e.g. [4]. Here, an additional, but completely natural assumption is required: it is also required that $\mathcal{F}$ actually contains distributions concentrated around the true parameter. This is formalized in Theorem 2.4. The choice of the family of approximation has then a strong influence on the quality of the approximation. On one end of the spectrum $\mathcal{F} = \mathcal{M}_+^1(\Theta)$ leads to $\tilde{\pi}_{n,\alpha} = \pi_{n,\alpha}$ and in this situation, our result exactly coincides with the one in [4]. This is of course of little interest when $\pi_{n,\alpha}$ is not tractable. On the other end, any family consisting of too few measures will not be rich enough to learn anything.

In Section 3 and 4 respectively we give two examples of approximations. In Section 3 we study the case of mean field approximations corresponding to the case of block independent distributions

$$\mathcal{F}^{\mathrm{mf}} := \Big\{ \rho(\mathrm{d}\theta) = \bigotimes_{i=1}^{p} \rho_i(\mathrm{d}\theta_i) \in \mathcal{M}_1^+(\Theta), \\ \forall i = 1, \cdots, p \quad \rho_i \in \mathcal{M}_1^+(\Theta_i), \quad \Theta = \Theta_1 \times \cdots \times \Theta_p \Big\},$$

in the context of matrix completion via factorization. In this case, the VB approximation leads to feasible approximation algorithms [20], and our theorem shows that $\tilde{\pi}_{n,\alpha}$ concentrates at the minimax-optimal rate. In Section 4 we study the parametric family of approximation spanned by Gaussian distribution i.e.

$$\mathcal{F}^{\Phi} := \Big\{ \Phi(\mathrm{d}\theta; m, \Sigma), \quad m \in \mathbb{R}^d, \Sigma \in \mathcal{S}_+^d(\mathbb{R}) \Big\}$$

where $\Phi(\mathrm{d}\theta; m, \Sigma)$ is the d dimensional Gaussian measure with mean m and covariance matrix Σ, $\mathcal{S}_+^d(\mathbb{R})$ the cone of $d \times d$ dimensional symmetric positive definite matrices on $\mathbb{R}$. We will show that other approximations are possible, i.e. by constraining the variance of the approximation $\Sigma \in \mathrm{Diag}_+^d(\mathbb{R})$ positive definite $d \times d$ diagonal matrices. Gaussian approximations have been studied in [27, 22]. We specify those results to the case of a logistic regression in Subsection 4.2. In this case, the VB approximation actually turns out to be a convex minimization problem, which can be solved by gradient descent or more sophisticated iterative procedures. This is especially attractive as it allows to prove the concentration of the VB approximation obtained after a finite number of steps. All the proofs are in Section 6.

2 Main results

2.1 A PAC-Bayesian inequality

We start be stating a variant of a result from [4].

Theorem 2.1 *For any $\alpha \in (0,1)$, for any $\varepsilon \in (0,1)$,*

$$\mathbb{P}\Bigg(\forall \rho \in \mathcal{M}_1^+(\Theta), \int D_\alpha(P_\theta, P_{\theta_0}) \rho(\mathrm{d}\theta) \rho(\mathrm{d}\theta) \\ \leq \frac{\alpha}{1-\alpha} \int \frac{r_n(\theta, \theta_0)}{n} \rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho, \pi) + \log\left(\frac{1}{\varepsilon}\right)}{n(1-\alpha)} \Bigg) \geq 1 - \varepsilon.$$

This result is very similar to Theorem 3.5 in [4], the only difference being that in [4] the theorem is stated only for ρ taken as the minimizer of the right-hand side. But the proof of Theorem 2.1 is a relatively straightforward adaptation of the one in [4]. We still provide it in Section 6 for the sake of completeness.

Note that the minimizer of the r.h.s can be explicitly found. Let us remind Donsker and Varadhan's variational inequality (for example Lemma 1.1.3 in Catoni [10]).

Lemma 2.2 *For any probability π on $(\Theta, \mathcal{T})$ and any measurable function $h : \Theta \to \mathbb{R}$ such that $\int \mathrm{e}^h \mathrm{d}\pi < \infty$,*

$$\log \int \mathrm{e}^h \mathrm{d}\pi = \sup_{\rho \in \mathcal{M}_1^+(\Theta)} \int h \mathrm{d}\rho - \mathcal{K}(\rho, \pi),$$

with the convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of μ, the supremum with respect to π on the right-hand side is reached by π_h given by

$$\frac{\mathrm{d}\pi_h}{\mathrm{d}\pi}(\theta) = \frac{\mathrm{e}^{h(\theta)}}{\int \mathrm{e}^h \mathrm{d}\pi}.$$

Using Lemma 2.2 and the definition of $\pi_{n,\alpha}$ we obtain

$$\pi_{n,\alpha}(\cdot|X_1^n) = \operatorname*{arg\,min}_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \alpha \int r_n(\theta, \theta_0) \rho(\mathrm{d}\theta) + \mathcal{K}(\rho, \pi) \right\}$$

so the minimizer of the right hand side of Theorem 2.1 over $\mathcal{M}_+^1(\Theta)$ is simply $\pi_{n,\alpha}(\mathrm{d}\theta|X_1^n)$.

2.2 Analysis of VB approximations

2.2.1 Concentration of the posterior and its approximation

We specialize the above results to the variational approximation. Elementary calculations show that

$$\begin{aligned}
\tilde{\pi}_{n,\alpha}(\cdot|X_1^n) &= \operatorname*{arg\,min}_{\rho \in \mathcal{F}} \left\{ \alpha \int r_n(\theta, \theta_0) \rho(\mathrm{d}\theta) + \mathcal{K}(\rho, \pi) \right\} \\
&= \operatorname*{arg\,min}_{\rho \in \mathcal{F}} \left\{ -\alpha \int \sum_{i=1}^n \log p_\theta(X_i) \rho(\mathrm{d}\theta) + \mathcal{K}(\rho, \pi) \right\}.
\end{aligned}$$

So, as a consequence, we obtain the following corollary of Theorem 2.1.

Corollary 2.3 *For any $\alpha \in (0,1)$, for any $\varepsilon \in (0,1)$, with probability at least $1 - \varepsilon$,*

$$\begin{aligned}
&\int D_\alpha(P_\theta, P_{\theta_0}) \pi_{n,\alpha}(\mathrm{d}\theta|X_1^n) \\
&\quad \le \inf_{\rho \in \mathcal{M}_1^+(\Theta)} \left\{ \frac{\alpha}{1-\alpha} \int \frac{r_n(\theta, \theta_0)}{n} \rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho, \pi) + \log\left(\frac{1}{\varepsilon}\right)}{n(1-\alpha)} \right\}, \\
&\int D_\alpha(P_\theta, P_{\theta_0}) \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n) \\
&\quad \le \inf_{\rho \in \mathcal{F}} \left\{ \frac{\alpha}{1-\alpha} \int \frac{r_n(\theta, \theta_0)}{n} \rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho, \pi) + \log\left(\frac{1}{\varepsilon}\right)}{n(1-\alpha)} \right\}.
\end{aligned}$$

Obviously, when $\mathcal{F} = \mathcal{M}_1^+(\Theta)$, we have $\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n) = \pi_{n,\alpha}(\mathrm{d}\theta|X_1^n)$ so we can see the first inequality as a special case of the second one. We are now in position to state our main result. It is a general result for $\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)$, keep in mind that in the special case $\mathcal{F} = \mathcal{M}_1^+(\Theta)$ this is also a result on $\pi_{n,\alpha}(\mathrm{d}\theta|X_1^n)$.

Theorem 2.4 *Fix $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$. Assume that $r_n > 0$ is such that there is distribution $\rho_n \in \mathcal{F}$ such that*

$$\int \mathcal{K}(P_{\theta_0}, P_\theta)\rho_n(\mathrm{d}\theta) \leq r_n, \ \int \mathbb{E}\left[\log^2\left(\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}\right)\right]\rho_n(\mathrm{d}\theta) \leq r_n \qquad (1)$$

and

$$\mathcal{K}(\rho_n, \pi) \leq nr_n. \qquad (2)$$

Then, for any $\alpha \in (0,1)$, for any $(\varepsilon, \eta) \in (0,1)^2$,

$$\mathbb{P}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n) \leq \frac{(\alpha+1)r_n + \alpha\sqrt{\frac{r_n}{n\eta}} + \frac{\log\left(\frac{1}{\varepsilon}\right)}{n}}{1-\alpha}\right] \geq 1-\varepsilon-\eta.$$

This theorem is a consequence of Corollary 2.3, its proof is provided in Section 6. Let us now discuss some of the consequences of this theorem.

Note that the assumption involving a distribution ρ_n is not standard. This requires some explanations. Assume in this paragraph that $\mathcal{F} = \mathcal{M}_1^+(\Theta)$. Define $B(r)$, for $r > 0$, as

$$B(r) = \left\{\theta \in \Theta : \mathcal{K}(P_{\theta_0}, P_\theta) \leq r, \mathrm{Var}\left[\log\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}\right] \leq r\right\}.$$

Taking ρ_n as π restricted to $B(r_n)$, $\rho_n = \pi_{|B(r_n)}$ implies that (1) is trivially satisfied. Then, (2) becomes

$$-\log\pi(B(r_n)) \leq nr_n,$$

a usual assumption in the study of concentration of the posterior, see e.g. Theorem 2.1 page 503 in [12] or Subsection 3.2 in [23]. Our main message is that in the previous studies of concentration of the posterior, the choice $\rho_n = \pi_{|B(r_n)}$ was hidden. Other choices might lead to easier calculations in some situations. But, more importantly, in the case where $\mathcal{F} \subsetneq \mathcal{M}_1^+(\Theta)$, it might very well be that $\pi_{|B(r_n)} \notin \mathcal{F}$. In this case, the assumption $-\log\pi(B(r_n)) \leq nr_n$ cannot be used, and (1) and (2) are natural extensions of this assumption for the study of concentration of variational approximations. They provide an explicit condition on the family $\mathcal{F}$ in order to ensure concentration of the approximation.

Also note that the choice $\eta = \frac{1}{nr_n}$ and $\varepsilon = \exp(-nr_n)$ leads to the following more standard formulation of concentration results. It shows that, as soon as $(1/n) \ll r_n \ll 1$, the parameter r_n gives a concentration rate for the VB approximation.

Corollary 2.5 *Under the same assumptions as in Theorem 2.4,*

$$\begin{aligned}\mathbb{P}\left[\int D_\alpha(P_\theta, P_{\theta_0})\pi_{n,\alpha}(\mathrm{d}\theta|X_1^n) \leq \frac{2(\alpha+1)}{1-\alpha}r_n\right] &\geq 1 - \frac{1}{nr_n} - \exp(-nr_n) \\ &\geq 1 - \frac{2}{nr_n},\end{aligned}$$

Remark 2.1 *As a special case, when $\alpha = 1/2$, the theorem becomes a concentration result in terms of the more classical Hellinger distance:*

$$\mathbb{P}\left[\int H^2(P_\theta, P_{\theta_0})\tilde{\pi}_{n,1/2}(\mathrm{d}\theta|X_1^n) \leq 3r_n + 2\sqrt{\frac{r_n}{n\eta}} + \frac{\log\left(\frac{1}{\varepsilon}\right)}{n}\right] \geq 1 - \varepsilon - \eta$$

and

$$\mathbb{P}\left[\int H^2(P_\theta, P_{\theta_0})\tilde{\pi}_{n,1/2}(\mathrm{d}\theta|X_1^n) \leq 6r_n\right] \geq 1 - \frac{2}{nr_n}.$$

2.2.2 A simpler result in expectation

Note that it is even possible to simplify the assumptions if we only want a result in expectation.

Theorem 2.6 *Fix $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$. Then*

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)\right] \leq \inf_{\rho\in\mathcal{F}}\left\{\frac{\alpha}{1-\alpha}\int \mathcal{K}(P_{\theta_0}, P_\theta)\rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho,\pi)}{n(1-\alpha)}\right\}.$$

As a special case assume that $r_n > 0$ is such that there is distribution $\rho_n \in \mathcal{F}$ such that

$$\int \mathcal{K}(P_{\theta_0}, P_\theta)\rho_n(\mathrm{d}\theta) \leq r_n \tag{3}$$

and

$$\mathcal{K}(\rho_n, \pi) \leq nr_n. \tag{4}$$

Then, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha}r_n.$$

Remark 2.2 *Special case for the Hellinger:*

$$\mathbb{E}\left[\int H^2(P_\theta, P_{\theta_0})\tilde{\pi}_{n,1/2}(\mathrm{d}\theta|X_1^n)\right] \leq 3r_n.$$

2.2.3 Extension of the result in expectation to the misspecified case

We assume in this section that the true distribution is not in $\{P_\theta, \theta \in \Theta\}$. In order not to change all the notations we define an extented parameter set $\Theta \cup \{\theta_0\}$ where $\theta_0 \notin \Theta$ and define P_{θ_0} as the true distribution. Then we have the following result, which is a direct adaptation of the previous one. Note that in some sense, we expect θ^* to be the minimizer w.r.t θ of $\mathcal{K}(P_{\theta_0}, P_\theta)$ but the following result can be applied even when this minimizer does not exist or is not unique.

Theorem 2.7 *Fix $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$. Then for any $\theta^* \in \Theta$,*

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)\right] \leq \frac{\alpha}{1-\alpha}\mathcal{K}(P_{\theta_0}, P_{\theta^*}) + \inf_{\rho\in\mathcal{F}}\left\{\frac{\alpha}{1-\alpha}\int \mathbb{E}\left[\log\frac{p_{\theta^*}(X_i)}{p_\theta(X_i)}\right]\rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho,\pi)}{n(1-\alpha)}\right\}.$$

As a special case assume that $\theta^* \in \Theta$ and $r_n > 0$ are such that there is distribution $\rho_n \in \mathcal{F}$ such that

$$\int \mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta^*}(X_i)}{p_\theta(X_i)}\right] \rho_n(\mathrm{d}\theta) \leq r_n \tag{5}$$

and

$$\mathcal{K}(\rho_n, \pi) \leq n r_n. \tag{6}$$

Then, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0}) \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n)\right] \leq \frac{\alpha}{1-\alpha}\mathcal{K}(P_{\theta_0}, P_{\theta^*}) + \frac{1+\alpha}{1-\alpha} r_n.$$

This result takes the form of an oracle inequality. However, it is not a proper oracle inequality in the sense that the risk measure used in the left-hand side and the right-hand side are not the same. So, in general, this result might not be very informative. Still, when $\mathcal{K}(P_{\theta_0}, P_{\theta^*})$ is very small then it might be useful.

Remark 2.3 *Special case for the Hellinger:*

$$\mathbb{E}\left[\int H^2(P_\theta, P_{\theta_0}) \tilde{\pi}_{n,1/2}(\mathrm{d}\theta | X_1^n)\right] \leq \mathcal{K}(P_{\theta_0}, P_{\theta^*}) + 3r_n.$$

3 Application to matrix completion

3.1 Context

Matrix completion problems received an increasing attention in the past few years due to the possible application to collaborative filtering. Let us describe briefly the model: in this case, our parameter θ is a matrix $M \in \mathbb{R}^{m \times p}$, with $m, p \geq 1$. Under P_M, the observations are random entries of this matrix with possible noise:

$$Y_i = M_{i_k, j_k} + \varepsilon_k$$

where the (i_k, j_k) are i.i.d $\mathcal{U}(\{1, \dots, m\} \times \{1, \dots, p\})$. For the sake of simplicity we will assume that the ε_k are i.i.d $\mathcal{N}(0, \sigma^2)$, and that σ^2 is known, so we only have to estimate M. Elementary calculations show that

$$\mathcal{K}(P_M, P_N) = \frac{1}{mp}\sum_{i=1}^{m}\sum_{j=1}^{p} \frac{(M_{i,j} - N_{i,j})^2}{2\sigma^2} = \frac{\|M - N\|_F^2}{2\sigma^2 mp}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. In the noiseless case $\sigma^2 = 0$, [8] proved that it is possible to recover exactly M under the assumption that its rank is small enough. Various extensions to noisy settings, and approximately low-rank matrices, can be found in [7, 17, 16]. The main messages of this papers is that the minimax rate of convergence is $(m+p)\mathrm{rank}(M)/n$, possibly up to log terms. Bayesian estimators were proposed in [24, 18, 30, 1] using factorized Gaussian priors. Convergence of the posterior mean was proved in [21], but only for a bounded prior, so this does not cover the Gaussian prior used in practice. Similarly, [26] proves concentration of a truncated version of the posterior. Finally, note that for very large datasets the MCMC algorithm proposed in [24] is too slow, a fast VB approximation was proposed in [20].

First, let us remind the classical so-called "factorized Gaussian" prior used in the literature, and the VB approximation of [20]. We will then prove the concentration of the fractional posterior and of its VB approximation, both at the minimax rate.

3.2 Definition of the prior and of the VB approximation

Fix $K \in \{1, \dots, m \wedge p\}$. The main idea of factorized priors is that, when $\text{rank}(M) \leq K$ then we have

$$M = UV^T$$

for some matrices U of dimension $p \times K$ and V of dimension $m \times K$. Thus, we can define a prior on M by specifying priors on U and V. A usual choice is that the entries $U_{i,k}$ and $V_{j,k}$ are independent $\mathcal{N}(0, \gamma_k)$ and finally γ_k is inverse gamma, that is $1/\gamma_k \sim \Gamma(a, b)$. The reason beyond these choices is conjugacy: put $\gamma = (\gamma_1, \dots, \gamma_K)$, it it then possible to compute the conditional posteriors of $U|V, \gamma$, of $V|U, \gamma$ and $\gamma|U, V$. So we can use the Gibbs sampler, what was done in [24]. However, when m and p are very large, each step takes hours, so we cannot simulate thousands of MCMC steps. This motivated the introduction of VB approximations in [20]. We define the family $\mathcal{F}$ as factorized approximations: any distribution in $\mathcal{F}$ is under the form

$$\rho(\mathrm{d}U, \mathrm{d}V, \mathrm{d}\gamma) = \bigotimes_{i=1}^{m} \rho_{U_i}(\mathrm{d}U_{i,\cdot}) \bigotimes_{j=1}^{p} \rho_{V_j}(\mathrm{d}V_{j,\cdot}) \bigotimes_{k=1}^{K} \rho_{\gamma_k}(\gamma_k).$$

The minimization of the VB program is shown in many cited papers, see [1] and all the references therein. We remind the main result: ρ_{U_i} is $\mathcal{N}(\mathbf{m}_{i,\cdot}^T, \mathcal{V}_i)$, ρ_{V_j} is $\mathcal{N}(\mathbf{n}_{j,\cdot}^T, \mathcal{W}_j)$ and ρ_{γ_k} is $\Gamma(a + (m_1 + m_2)/2, \beta_k)$ for some $m \times K$ matrix $\mathbf{m}$ whose rows are denoted by $\mathbf{m}_{i,\cdot}$ and some $p \times K$ matrix $\mathbf{n}$ whose rows are denoted by $\mathbf{n}_{j,\cdot}$ and some vector $\beta = (\beta_1, \dots, \beta_K)$. The parameters are updated iteratively through the formulae

1. moments of U:

$$\mathbf{m}_{i,\cdot}^T := \frac{2\alpha}{n} \mathcal{V}_i \sum_{k: i_k = i} Y_{i_k, j_k} \mathbf{n}_{j_k,\cdot}^T$$

$$\mathcal{V}_i^{-1} := \frac{2\alpha}{n} \sum_{k: i_k = i} \left[\mathcal{W}_{j_k} + \mathbf{n}_{j_k,\cdot} \mathbf{n}_{j_k,\cdot}^T \right] + \left(a + \frac{m_1 + m_2}{2} \right) \mathbf{diag}(\beta)^{-1}$$

2. moments of V:

$$\mathbf{n}_{j,\cdot}^T := \frac{2\alpha}{n} \mathcal{W}_j \sum_{k: j_k = j} Y_{i_k, j_k} \mathbf{m}_{i_k,\cdot}^T$$

$$\mathcal{W}_j^{-1} := \frac{2\alpha}{n} \sum_{k: j_k = j} \left[\mathcal{V}_{i_k} + \mathbf{m}_{i_k,\cdot} \mathbf{m}_{i_k,\cdot}^T \right] + \left(a + \frac{m_1 + m_2}{2} \right) \mathbf{diag}(\beta)^{-1}$$

3. moments of γ:

$$\beta_k := \frac{1}{2} \left[\sum_{i=1}^{m_1} \left(\mathbf{m}_{i,k}^2 + (\mathcal{V}_i)_{k,k} \right) + \sum_{j=1}^{m_2} \left(\mathbf{n}_{j,k}^2 + (\mathcal{V}_j)_{k,k} \right) \right]$$

(where $(\mathcal{V}_i)_{k,k}$ denotes the (k,k)-th entry of the matrix $\mathcal{V}_i$ and $(\mathcal{W}_j)_{k,k}$ denotes the (k,k)-th entry of the matrix $\mathcal{W}_j$).

3.3 Concentration of the posteriors

For $r \geq 1$ and $B > 0$ we define $\mathcal{M}(r, B)$ as the set of pairs of matrices $(\bar{U}, \bar{V})$ with dimensions $m \times K$ and $p \times K$ respectively, with $\bar{U} = (\bar{U}_{1,\cdot}|\ldots|\bar{U}_{r,\cdot}|0|\ldots|0)$ and $\bar{V} = (\bar{V}_{1,\cdot}|\ldots|\bar{V}_{r,\cdot}|0|\ldots|0)$ and $\|U\|_\infty, \|V\|_\infty < B$ where for any matrix A we use the notation $\|A\|_\infty$ to denote $\max_{i,j} |A_{i,j}|$ (the sup norm of the vectorized matrix).

Theorem 3.1 *Fix* $b = \frac{B^2}{512(nmp)^4[(m\vee p)K]^2}$ *and take* a *as any constant. Then*

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)\right]$$
$$\leq \inf_{1\leq r\leq K}\inf_{(\bar{U},\bar{V})\in\mathcal{M}(r,B)}\left\{\frac{\alpha}{1-\alpha}\frac{\left[\|M-\bar{U}\bar{V}^T\|_F+\frac{\sqrt{B}}{n}\right]^2}{2\sigma^2 mp}\right.$$
$$\left.+\frac{2(1+\alpha)(1+2a)r(m+p)\left[\log(nmp)+\mathcal{C}(a)\right]}{n(1-\alpha)}\right\}$$

where the constant $\mathcal{C}(a) = \log(8\sqrt{\pi}\Gamma(a)2^{10a+1}) + 3$.

Note as a special case that when $M = \bar{U}\bar{V}^T$ for $(\bar{U}, \bar{V}) \in \mathcal{M}(r, B)$ then we have exactly

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)\right]$$
$$\leq \frac{2(1+\alpha)(1+2a)r(m+p)\left[\log(nmp)+\mathcal{C}(a)+\frac{\alpha B}{2\sigma^2 mp}\right]}{n(1-\alpha)},$$

we have the minimax-rate rate $\mathcal{O}(r(m+p)/n)$ up to log terms. Still assuming that $M = \bar{U}\bar{V}^T$ for $(\bar{U}, \bar{V}) \in \mathcal{M}(r, B)$ it is also possible to state a proper concentration result as an application of Corollary 2.5. We omit the proof as it is exactly similar to the one of Theorem 3.1.

Theorem 3.2 *Assume* $M = \bar{U}\bar{V}^T$ *for* $(\bar{U}, \bar{V}) \in \mathcal{M}(r, B)$ *and take* b *as in Theorem 3.1. Then*

$$\mathbb{P}\left[\int D_\alpha(P_\theta, P_{\theta_0})\pi_{n,\alpha}(\mathrm{d}\theta|X_1^n) \leq \frac{2(\alpha+1)}{1-\alpha}r_n\right] \geq 1 - \frac{2}{nr_n}$$

where

$$r_n = \frac{\mathcal{D}(a, \sigma^2, B)r(m+p)\log(nmp)}{n}$$

for some (known) constant $\mathcal{D}(a, \sigma^2, B)$.

4 Gaussian variational Bayes

In this section we consider $\Theta \subset \mathbb{R}^d$ and the class of Gaussian approximations $\mathcal{F}^\Phi := \left\{\Phi(d\theta; m, \Sigma), \quad m \in \mathbb{R}^d, \Sigma \in \mathcal{S}_+^d(\mathbb{R})\right\}$, thus the algorithm will consist in projecting onto the closest Gaussian distribution in KL sense. Depending on the hypotheses made on the covariance matrix we can build different approximations. For instance define:

$$\mathcal{F}_{diag}^\Phi := \left\{\Phi(d\theta; m, \Sigma), \quad m \in \mathbb{R}^d, \Sigma \in \mathrm{Diag}_d(\mathbb{R})\right\}$$

$$\mathcal{F}_{id}^{\Phi} := \left\{ \Phi(d\theta; m, \sigma^2 I_d), \quad m \in \mathbb{R}^d, \sigma^2 \in \mathbb{R}_{+\star} \right\}.$$

We have by definition $\mathcal{F}_{id}^{\Phi} \subseteq \mathcal{F}_{diag}^{\Phi} \subseteq \mathcal{F}^{\Phi}$.

The remarkable fact of Gaussian VB is that it allows to recast integration as a finite dimension optimization problem. The choice of a specific Gaussian is a trade off between accuracy and computational complexity. We will show in the following that, under some assumption on the likelihood, the integrated $\alpha-$Reyni divergence is convergent for most of the approximations.

To simplify the exposition of the results we will restrict our study to the case of Gaussian priors: $\pi = \mathcal{N}(0, \vartheta^2 I_p)$. One can readily see that in Theorem 2.4 the prior appears only in the condition $\frac{1}{n}\mathcal{K}(\rho, \pi) \leq r_n$, many other distribution could be used, providing different rates.

In the rest of the section we assume some regularity on the log density.

Assumption 4.1 *We assume that for a model $\{p_\theta, \theta \in \Theta\}$ there exists a measurable real valued function $M(\cdot)$ and $p \in \mathbb{N}^\star \cup \{\frac{1}{2}\}$*

$$|\log p_\theta(X_1) - \log p_{\theta'}(X_1)| \leq M(X_1) \left\|\theta - \theta'\right\|_2^{2p}$$

Furthermore we assume that $\mathbb{E}M(X_1) =: B_1, \quad \mathbb{E}M^2(X_1) =: B_2 < \infty$.

The assumption allows us to consider the case of Lipschitz log-densities ($p = 1/2$) as well as Gaussian distributions ($p = 1$). In Section 4.2 we specify our result to the case of logistic regression.

Theorem 4.1 *Let the family of approximation be $\mathcal{F}$ with $\mathcal{F}_{id}^{\Phi} \subset \mathcal{F}$ as defined above and that the model satisfies Assumption 4.1. We put*

$$r_n = \frac{B_1}{n} \vee \frac{B_2}{n^2} \vee \left\{ \frac{d}{n} \left[\frac{1}{2} \log\left(\vartheta^2 n^2 C_{p,d}\right) + \frac{1}{n\vartheta^2} \right] + \frac{\|\theta_0\|^2}{n\vartheta^2} - \frac{d}{2n} \right\}$$

where

$$C_{p,d} = \begin{cases} \sqrt{d} & \text{if } p = \frac{1}{2}, \\ 2^{2p} \frac{\Gamma(2p + \frac{1}{2}d)}{\Gamma(\frac{1}{2}d)} & \text{otherwise.} \end{cases}$$

Then for any $\alpha \in (0,1)$, for any η, ϵ

$$\mathbb{P}\left[\int D_\alpha(P_\theta, P_{\theta_0}) \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n) \leq \frac{(\alpha+1) r_n + \alpha\sqrt{\frac{r_n}{n\eta}} + \frac{\log\left(\frac{1}{\varepsilon}\right)}{n}}{1-\alpha} \right] \geq 1 - \varepsilon - \eta.$$

4.1 Stochastic variational Bayes

In many cases the model is not conjugate, i.e. the VB objective does not have a close form solution. We can however use a full Gaussian approximation and a stochastic gradient descent on the objective function defined by the KL divergence. This approach has been studied in [27].

We may write our variational bound as the following minimisation problem

$$\min_{\rho \in \mathcal{F}_\Phi} \int \rho(d\theta) \log \frac{d\rho(\theta)}{d\pi_{n,\alpha}(\theta | X^n)} \tag{7}$$

or after droping the constants,

$$\min_{m \in \mathbb{R}^d, \Sigma \in \mathcal{S}_+^d} \left\{ -\alpha \int \log p_\theta(Y^n) \Phi(d\theta; m, \Sigma) + \int \log \frac{d\Phi(\theta; m, \Sigma)}{d\pi(\theta)} \Phi(d\theta; m, \Sigma) \right\}. \tag{8}$$

In [27] the authors suggest using a parametrization of the problem where we replace the optimization over Σ by a minimization over the matrix C where $CC^t = \Sigma$. To simplify the notations in this section define

$$F : x = (m, C) \in \mathbb{R}^d \times \mathbb{R}^{d\times d} \mapsto \mathbb{E}\left[f(x, \xi)\right]$$

to be the objective of the minimization problem (8), where $\xi \sim \mathcal{N}(0, I_d)$ and

$$f((m, C), \xi) := \log p_{m+C\xi}(Y_1^n) + \log \frac{\mathrm{d}\Phi_{m,CC^t}}{\pi}(m + C\xi).$$

For algorithmic reasons we constrain the minimization to be over an Euclidean ball that is (8) is transformed into

$$\min_{x\in\mathbb{B}\cap\mathbb{R}^d\times\mathcal{S}_+^d} \mathbb{E}\left[f(x, \xi)\right],$$

where $\mathbb{B} = \left\{x \in \mathbb{R}^{d^2+d}, \|x\|_2 \leq B\right\}$. In addition we can define the corresponding family of Gaussian distribution

$$\mathcal{F}_B^\Phi = \left\{\Phi(d\theta; m, CC^t), \quad (m, C) \in \mathbb{B} \cap \mathbb{R}^d \times \mathcal{S}_+^d\right\}.$$

The objective can now be replaced by a Monte Carlo estimate and we can use stochastic gradient descent as described in Algorithm 1.

Algorithm 1 Stochastic Variational Bayes

Input: x_0, X_1^n, γ_T

For $i \in \{1, \cdots, T\}$,

a. Sample $\xi_t \sim \mathcal{N}(0, I_d)$

b. Update $x_t \leftarrow \mathcal{P}_{\mathbb{B}}\left(x_{t-1} - \gamma_T \nabla f(x_{t-1}, \xi_t)\right)$

End For .

Output: $\bar{x}_T = \frac{1}{T}\sum_{t=1}^T x_t$

Assumption 4.2 *Assume that f, as defined in equation (7), is convex in its first component x and that it has L-Lipschitz gradients.*

Define $\tilde{\pi}_{n,\alpha}^k(\mathrm{d}\theta|X_1^n)$ to be the k-th iterate of Algorithm 1, the Gaussian distribution with parameters $\bar{x}_k = (\bar{m}_k, \bar{C}_k)$.

Theorem 4.2 *Let Assumptions 4.2 and 4.1 be verified, define*

$$r_n = \frac{B_1}{n} \vee \frac{B_2}{n^2} \vee \left\{\frac{d}{n}\left[\frac{1}{2}\log\left(\vartheta^2 n^2 C_{p,d}\right) + \frac{1}{n\vartheta^2}\right] + \frac{\|\theta_0\|^2}{n\vartheta^2} - \frac{d}{2n}\right\}$$

with $C_{p,d}$ as in Theorem 4.1. Let B be such that $B > \|\theta_0\|_2 + \frac{1}{n\sqrt{d}}$ then for $\tilde{\pi}_{n,\alpha}^T(\mathrm{d}\theta|X_1^n)$ obtained by Algorithm 1 with $\gamma_T = \frac{B}{L\sqrt{2T}}$, we get

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}^T(\mathrm{d}\theta|X_1^n)\right] \leq \frac{1+\alpha}{1-\alpha} r_n + \frac{1}{n(1-\alpha)}\sqrt{\frac{2BL}{T}}.$$

Remark 4.1 *On most example the gradient will be written as a sum of at least n components. If each term is Lipschitz with constant L_i then an estimate of the constant will be $L \leq n \max_i L_i$. The last term of the above bound is therefore of the order $\frac{1}{(1-\alpha)}\sqrt{\frac{2B\max L_i}{nT}}$, hence one might want to take $T = O(\sqrt{n})$ to mitigate the impact of the numerical approximation on the rate.*

4.2 Example: logistic regression

We consider the case of a binary regression model. Although estimation of parameters is relatively simple for small datasets [11], it remains challenging when the size of the dictionary is large. Furthermore usual deterministic methods do not come with theoretical guaranties as would a gradient descent algorithm for maximum likelihood.

The logistic regression is not conjugate in the sense that we cannot find an iterative scheme based on a mean field approximation as for the matrix completion example.

Let $X_i = (Y_i, Z_i) \in \{-1, 1\} \times \mathbb{R}^d$ be such that

$$\mathbb{P}\{Y = y | Z = z, \theta\} = \frac{e^{yz\theta}}{1 + e^{yz\theta}},$$

We will consider the case of estimation with a Gaussian prior $\pi(d\theta) = \Phi(d\theta; 0, \vartheta I_d)$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ (other cases are easily incorporated in the theory).

We will prove results in the case of random design where we suppose that the distribution of Z_1^n does not depend on the parameter.

Corollary 4.3 *Let the family of approximation be any $\mathcal{F}$ with $\mathcal{F}_{id}^{\Phi} \subset \mathcal{F}$ as defined above and assume that $K_1 := 2\mathbb{E}\left\|X_1\right\|, K_2 := 4\mathbb{E}\left\|X_1\right\|^2 < \infty$. We put*

$$r_n = \frac{K_1}{n} \vee \frac{K_2}{n^2} \vee \left\{ \frac{d}{n} \left[\frac{1}{2} \log\left(\vartheta^2 n^2 \sqrt{d}\right) + \frac{1}{n\vartheta^2} \right] + \frac{\|\theta_0\|^2}{n\vartheta^2} - \frac{d}{2n} \right\}$$

then for any $\alpha \in (0, 1)$, for any η, ϵ

$$\mathbb{P}\left[\int D_\alpha(P_\theta, P_{\theta_0}) \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n) \leq \frac{(\alpha+1) r_n + \alpha \sqrt{\frac{r_n}{n\eta}} + \frac{\log\left(\frac{1}{\varepsilon}\right)}{n}}{1-\alpha} \right] \geq 1 - \varepsilon - \eta.$$

To apply Theorem 4.2 we need to add some constraint on the covariance matrix. The optimization will be written over $\mathbb{B}_\psi := \mathbb{B} \cap \{C \in \mathcal{S}_+, CC^T \succeq \psi I_{d\times d}\}$ (this is done only to ensure that $\log|\Sigma|$ has Lipschitz gradients).

Corollary 4.4 *Let the family of approximation be any $\mathcal{F}$ with $\mathcal{F}_{id}^{\Phi} \subset \mathcal{F}$ and assume that $K_1 := 2\mathbb{E}\left\|X_1\right\|, K_2 := 4\mathbb{E}\left\|X_1\right\|^2 < \infty$, let B be such that $B > \|\theta_0\|_2 + \frac{1}{n\sqrt{d}}$ then for $\pi_{n,\alpha}^T(\mathrm{d}\theta | X_1^n)$ obtained by Algorithm 1 with $\gamma_T = \frac{B}{L\sqrt{2T}}$ and where $\mathbb{B}$ is replaced by $\mathbb{B}_\psi$ for any $\psi \leq \frac{1}{n\sqrt{d}}$, we get*

$$\begin{aligned}
&\mathbb{E}\left[\int D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n}) \tilde{\pi}_{n,\alpha}^T(\mathrm{d}\theta | X_1^n) \right] \\
&\leq \frac{1+\alpha}{1-\alpha} \left(\frac{K_1}{n} \vee \frac{K_2}{n^2} \vee \left\{ \frac{d}{n} \left[\frac{1}{2} \log\left(\vartheta^2 n^2 \sqrt{d}\right) + \frac{1}{n\vartheta^2\sqrt{d}} \right] + \frac{\|\theta_0\|^2}{n\vartheta^2} - \frac{d}{2n} \right\} \right) \\
&\qquad + \frac{1}{(1-\alpha)n} \sqrt{\frac{2BL}{T}}.
\end{aligned}$$

5 Conclusion

Based on PAC-Bayesian inequalities, we introduced a generic way to study the concentration of variational Bayesian approximations. This is a very

general approach that can be applied to many models. We studied applications to matrix completion, and generic Gaussian approximations in regular models. Still, some questions remain open. From a theoretical perspective, the oracle inequality in Theorem 2.7 compares a Renyi divergence to a Kullback-Leibler divergence. It would be very interesting to obtain a result with the Kullback divergence in the left-hand side. This is probably more difficult, if possible at all. We believe that tools from [14] could be of some help, but the assumptions used in this paper might be restrictive. From an applied perspective, one of the major applications of variational approximations is mixture models - and beyond, models with latent variables. We did not find a way to include this case in our analysis, and it would surely be of major interest.

6 Proofs

6.1 Proof of Theorem 2.1

We adapt the proof given in [4]. Fix $\alpha \in (0,1)$. It's immediate to check that

$$\mathbb{E}\left[\exp\left(-\alpha r_n(\theta,\theta_0)\right)\right] = \exp\left[-(1-\alpha)D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n})\right].$$

Thus,

$$\mathbb{E}\left[\exp\left(-\alpha r_n(\theta,\theta_0) + (1-\alpha)D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n})\right)\right] = 1.$$

Integrate with respect to π,

$$\int \mathbb{E}\left[\exp\left(-\alpha r_n(\theta,\theta_0) + (1-\alpha)D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n})\right)\right] \pi(\mathrm{d}\theta) = 1$$

and using Fubini's theorem,

$$\mathbb{E}\left[\int \exp\left(-\alpha r_n(\theta,\theta_0) + (1-\alpha)D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n})\right) \pi(\mathrm{d}\theta)\right] = 1. \tag{9}$$

The key argument here, introduced by [9], is to use Lemma 2.2. Note that almost surely with respect to the sample, we know that

$$h(\theta) := -\alpha r_n(\theta,\theta_0) + (1-\alpha)D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n})$$

satisfies $\int \exp(h)\mathrm{d}\pi < \infty$, otherwise, the expectation in (9) would be infinite. So, the conditions of Lemma 2.2 are satisfied almost surely with respect to the sample, and we obtain:

$$\mathbb{E}\left\{\exp\left[\sup_{\rho\in\mathcal{M}_1^+(\Theta)}\left(\int\left(-\alpha r_n(\theta,\theta_0) + (1-\alpha)D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n})\right)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho,\pi)\right)\right]\right\} = 1.$$

Multiply both sides by ε to get

$$\mathbb{E}\left\{\exp\left[\sup_{\rho\in\mathcal{M}_1^+(\Theta)}\left(\int\left(-\alpha r_n(\theta,\theta_0) + (1-\alpha)D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n})\right)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho,\pi)\right) - \log\left(\frac{1}{\varepsilon}\right)\right]\right\} = \varepsilon.$$

Using Markov's inequality,

$$\mathbb{P}\Bigg[\sup_{\rho\in\mathcal{M}_1^+(\Theta)}\Bigg(\int\left(-\alpha r_n(\theta,\theta_0)+(1-\alpha)D_\alpha(P_\theta^{\otimes n},P_{\theta_0}^{\otimes n})\right)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho,\pi)\Bigg) - \log\left(\frac{1}{\varepsilon}\right)\geq 0\Bigg]\leq\varepsilon.$$

Taking the complementary event,

$$\mathbb{P}\Bigg(\forall\rho\in\mathcal{M}_1^+(\Theta),\quad \int\left(-\alpha r_n(\theta,\theta_0)+(1-\alpha)D_\alpha(P_\theta^{\otimes n},P_{\theta_0}^{\otimes n})\right)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho,\pi) - \log\left(\frac{1}{\varepsilon}\right)\leq 0\Bigg)\geq 1-\varepsilon.$$

Now, for a given ρ, it might be that

$$\int D_\alpha(P_\theta^{\otimes n},P_{\theta_0}^{\otimes n})\rho(\mathrm{d}\theta)=\infty$$

but then, the previous equation implies that

$$\int r_n(\theta,\theta_0)\rho(\mathrm{d}\theta)+\mathcal{K}(\rho,\pi)=\infty$$

and so the statement of the theorem is trivially satisfied as $\infty\leq\infty$. On the other hand, assuming that

$$\int D_\alpha(P_\theta^{\otimes n},P_{\theta_0}^{\otimes n})\rho(\mathrm{d}\theta)<\infty$$

we can rearrange terms to get

$$\mathbb{P}\Bigg(\forall\rho\in\mathcal{M}_1^+(\Theta),\int\frac{D_\alpha(P_\theta^{\otimes n},P_{\theta_0}^{\otimes n})}{n}\rho(\mathrm{d}\theta) \leq\frac{\alpha}{1-\alpha}\int\frac{r_n(\theta,\theta_0)}{n}\rho(\mathrm{d}\theta)+\frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{1}{\varepsilon}\right)}{n(1-\alpha)}\Bigg)\geq 1-\varepsilon.$$

Finally, using the additive property of the Renyi divergence reminded in Remark 1.1 we have $D_\alpha(P_\theta^{\otimes n},P_{\theta_0}^{\otimes n})=nD_\alpha(P_\theta,P_{\theta_0})$ and we obtain the statement of the theorem:

$$\mathbb{P}\Bigg(\forall\rho\in\mathcal{M}_1^+(\Theta),\int D_\alpha(P_\theta,P_{\theta_0})\rho(\mathrm{d}\theta) \leq\frac{\alpha}{1-\alpha}\int\frac{r_n(\theta,\theta_0)}{n}\rho(\mathrm{d}\theta)+\frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{1}{\varepsilon}\right)}{n(1-\alpha)}\Bigg)\geq 1-\varepsilon.$$

6.2 Proof of Theorem 2.4

Fix $\eta\in(0,1)$ and define

$$\rho^*=\operatorname*{arg\,min}_{\rho\in\mathcal{F}}\left\{\frac{\alpha}{1-\alpha}\int\frac{\mathbb{E}[r_n(\theta,\theta_0)]}{n}\rho(\mathrm{d}\theta)\right.$$

$$+ \frac{\alpha}{n(1-\alpha)} \sqrt{\frac{\mathrm{Var}[\int r_n(\theta,\theta_0)\rho(\mathrm{d}\theta)]}{\eta}} + \frac{\mathcal{K}(\rho,\pi)}{n(1-\alpha)} \Bigg\}.$$

Chebyshev's inequality leads to

$$\mathbb{P}\Bigg\{ \frac{\alpha}{1-\alpha} \int \frac{r_n(\theta,\theta_0)}{n} \rho(\mathrm{d}\theta) \geq \frac{\alpha}{1-\alpha} \int \frac{\mathbb{E}[r_n(\theta,\theta_0)]}{n} \rho(\mathrm{d}\theta) + \frac{\alpha}{n(1-\alpha)} \sqrt{\frac{\mathrm{Var}[\int r_n(\theta,\theta_0)\rho(\mathrm{d}\theta)]}{\eta}} + \frac{\mathcal{K}(\rho,\pi)}{n(1-\alpha)} \Bigg\} \leq \eta$$

and so

$$\mathbb{P}\Bigg\{ \frac{\alpha}{1-\alpha} \int \frac{r_n(\theta,\theta_0)}{n} \rho(\mathrm{d}\theta) \geq \frac{\alpha}{1-\alpha} \int \mathcal{K}(P_{\theta_0}, P_\theta) \rho(\mathrm{d}\theta) + \frac{\alpha}{1-\alpha} \sqrt{\frac{\mathrm{Var}\left[\log \frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}\right]}{n\eta}} + \frac{\mathcal{K}(\rho,\pi)}{n(1-\alpha)} \Bigg\} \leq \eta.$$

Now apply take the union bound of this inequality and of the inequality in Corollary 2.3. We obtain, for any $\alpha \in (0,1)$, for any $\varepsilon \in (0,1)$, with probability at least $1-\varepsilon-\eta$,

$$\begin{aligned}
&\int D_\alpha(P_\theta, P_{\theta_0}) \pi_{n,\alpha}(\mathrm{d}\theta | X_1^n) \\
&\leq \inf_{\rho \in \mathcal{F}} \Bigg\{ \frac{\alpha \int \left[\mathcal{K}(P_{\theta_0}, P_\theta) + \sqrt{\frac{1}{n\eta} \mathbb{E}\left[\log^2 \left(\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} \right) \right]} \right] \rho(\mathrm{d}\theta)}{1-\alpha} \\
&\qquad + \frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{1}{\varepsilon}\right)}{n(1-\alpha)} \Bigg\} \\
&\leq \frac{\alpha \int \left[\mathcal{K}(P_{\theta_0}, P_\theta) + \sqrt{\frac{1}{n\eta} \mathbb{E}\left[\log^2 \left(\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} \right) \right]} \right] \rho_n(\mathrm{d}\theta)}{1-\alpha} \\
&\qquad + \frac{\mathcal{K}(\rho_n,\pi) + \log\left(\frac{1}{\varepsilon}\right)}{n(1-\alpha)} \\
&\leq \frac{\alpha \left(r_n + \sqrt{\frac{r_n}{n\eta}} \right)}{1-\alpha} + \frac{n r_n + \log\left(\frac{1}{\varepsilon}\right)}{n(1-\alpha)}.
\end{aligned}$$

6.3 Proof of Theorem 2.6

The beginning is as for Theorem 2.1. Fix $\alpha \in (0,1)$, then

$$\mathbb{E}\left[\exp\left(-\alpha r_n(\theta,\theta_0) - (1-\alpha) D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n}) \right) \right] = 1.$$

Integrate with respect to π,

$$\int \mathbb{E}\left[\exp\left(-\alpha r_n(\theta,\theta_0) - (1-\alpha) D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n}) \right) \right] \pi(\mathrm{d}\theta) = 1$$

and using Fubini's theorem and Lemma 2.2

$$\mathbb{E}\Bigg\{ \exp\Bigg[\sup_{\rho \in \mathcal{M}_1^+(\Theta)} \Bigg(\int \left(-\alpha r_n(\theta,\theta_0) - (1-\alpha) D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n}) \right) \rho(\mathrm{d}\theta)$$

$$- \mathcal{K}(\rho, \pi)\Big) \Big] \Big\} = 1.$$

This is where things change: we now use Jensen's inequality to obtain

$$\mathbb{E}\Bigg[\sup_{\rho \in \mathcal{M}_1^+(\Theta)} \Bigg(\int \left(-\alpha r_n(\theta, \theta_0) - (1-\alpha) D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n})\right) \rho(\mathrm{d}\theta) \\ - \mathcal{K}(\rho, \pi)\Bigg)\Bigg] = 0$$

and so as a special case

$$\mathbb{E}\Bigg[\int \left(-\alpha r_n(\theta, \theta_0) - (1-\alpha) D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n})\right) \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n) \\ - \mathcal{K}(\tilde{\pi}_{n,\alpha}(\cdot | X_1^n), \pi)\Bigg] = 0.$$

Rearranging terms,

$$\begin{aligned}
&\mathbb{E}\left[\int D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n}) \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n)\right] \\
&\leq \mathbb{E}\left[\frac{\alpha}{1-\alpha} \int r_n(\theta, \theta_0) \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n) + \frac{\mathcal{K}(\tilde{\pi}_{n,\alpha}(\cdot | X_1^n), \pi)}{1-\alpha}\right] \\
&= \mathbb{E}\left\{\inf_{\rho \in \mathcal{F}} \left[\frac{\alpha}{1-\alpha} \int r_n(\theta, \theta_0) \rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho, \pi)}{1-\alpha}\right]\right\} \text{ by dfn.} \\
&\leq \inf_{\rho \in \mathcal{F}} \left\{\mathbb{E}\left[\frac{\alpha}{1-\alpha} \int r_n(\theta, \theta_0) \rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho, \pi)}{1-\alpha}\right]\right\} \\
&= \inf_{\rho \in \mathcal{F}} \left\{\frac{n\alpha}{1-\alpha} \int \mathcal{K}(P_{\theta_0}, P_\theta) \rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho, \pi)}{1-\alpha}\right\}
\end{aligned}$$

and so

$$\begin{aligned}
\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0}) \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n)\right] &= \mathbb{E}\left[\int \frac{D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n})}{n} \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n)\right] \\
&\leq \inf_{\rho \in \mathcal{F}} \left\{\frac{\alpha}{1-\alpha} \int \mathcal{K}(P_{\theta_0}, P_\theta) \rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho, \pi)}{n(1-\alpha)}\right\}.
\end{aligned}$$

6.4 Proof of Theorem 3.1

Fix $B > 0$, $r \geq 1$ and any pair $(\bar{U}, \bar{V}) \in \mathcal{M}_{r,B}$ and define for $\delta \in (0, B)$ that will be chosen later,

$$\rho_n(\mathrm{d}U, \mathrm{d}V, \mathrm{d}\gamma) \propto \mathbf{1}(\|U - \bar{U}\|_\infty \leq \delta, \|U - \bar{U}\|_\infty \leq \delta) \pi(\mathrm{d}U, \mathrm{d}V, \mathrm{d}\gamma).$$

Note that it can be factorized so it belongs to the family $\mathcal{F}$.

We adapt the calculations from [1, 2] but simplify a lot. First, note that

$$\mathcal{K}(P_M, P_{UV^T}) = \frac{1}{mp} \sum_{i=1}^m \sum_{j=1}^p \frac{(M_{i,j} - (UV^T)_{i,j})^2}{2\sigma^2} = \frac{\|M - UV^T\|_F^2}{2\sigma^2 mp}$$

and that for any (U, V) in the support of ρ_n we have

$$\|M - UV^T\|_F = \|M - \bar{U}\bar{V}^T + \bar{U}\bar{V}^T - \bar{U}V^T + \bar{U}V^T - UV^T\|_F$$

$$
\begin{aligned}
&\leq \|M - \bar{U}\bar{V}^T\|_F + \|\bar{U}\bar{V}^T - \bar{U}V^T\|_F + \|\bar{U}V^T - UV^T\|_F \\
&= \|M - \bar{U}\bar{V}^T\|_F + \|\bar{U}(\bar{V}^T - V^T)\|_F + \|(\bar{U} - U)V^T\|_F \\
&\leq \|M - \bar{U}\bar{V}^T\|_F + \|\bar{U}\|_F\|\bar{V} - V\|_F + \|\bar{U} - U\|_F\|V^T\|_F \\
&\leq \|M - \bar{U}\bar{V}^T\|_F + mp\|\bar{U}\|_\infty^{1/2}\|\bar{V} - V\|_\infty^{1/2} + mp\|V\|_\infty^{1/2}\|\bar{U} - U\|_\infty^{1/2} \\
&\leq \|M - \bar{U}\bar{V}^T\|_F + mp(B^{1/2}\delta^{1/2} + (B+\delta)^{1/2}\delta^{1/2}) \\
&\leq \|M - \bar{U}\bar{V}^T\|_F + 2mp\delta^{1/2}(B+\delta)^{1/2} \\
&\leq \|M - \bar{U}\bar{V}^T\|_F + 2^{3/2}mp\delta^{1/2}B^{1/2} \\
&= \|M - \bar{U}\bar{V}^T\|_F + B/n
\end{aligned}
$$

with the choice $\delta = B/[8(nmp)^2]$ which satisfies $0 < \delta < B$.

Then, we derive

$$
\begin{aligned}
\mathcal{K}(\rho_n, \pi) &= \log \frac{1}{\pi\left(\|U - \bar{U}\|_\infty \leq \delta, \|U - \bar{U}\|_\infty \leq \delta\right)} \\
&= \log \frac{1}{\int \pi\left(\|U - \bar{U}\|_\infty \leq \delta, \|U - \bar{U}\|_\infty \leq \delta|\gamma\right)\pi(\mathrm{d}\gamma)} \\
&= \log \frac{1}{\int \pi\left(\|U - \bar{U}\|_\infty \leq \delta|\gamma\right)\pi(\mathrm{d}\gamma)} + \log \frac{1}{\int \pi\left(\|V - \bar{V}\|_\infty \leq \delta|\gamma\right)\pi(\mathrm{d}\gamma)} \\
&= \log \frac{1}{\int_E \pi\left(\|U - \bar{U}\|_\infty \leq \delta|\gamma\right)\pi(\mathrm{d}\gamma)} + \log \frac{1}{\int_E \pi\left(\|V - \bar{V}\|_\infty \leq \delta|\gamma\right)\pi(\mathrm{d}\gamma)}
\end{aligned}
$$

for any event E. We actually take $E = \{\gamma_1, \ldots, \gamma_r \in [B^2, 2B^2], \gamma_{r+1}, \ldots, \gamma_K \in [s, 2s]\}$ and $s \in (0, B^2)$ is to be chosen later.

Then note that

$$
\begin{aligned}
&\pi\left(\|U - \bar{U}\|_\infty \leq \delta|\gamma\right) = \pi\left(\forall i, k|U_{i,k} - \bar{U}_{i,k}| \leq \delta|\gamma\right) \\
&= \prod_{i=1}^m \prod_{k=1}^r \pi\left(|U_{i,k} - \bar{U}_{i,k}| \leq \delta|\gamma_k\right)\pi\left(\max_{i=1}^m \max_{k=1}^K |U_{i,k}| \leq \delta \middle| \gamma_{r+1}, \ldots, \gamma_K\right).
\end{aligned}
$$

First,

$$
\begin{aligned}
\pi\left(\max_{i=1}^m \max_{k=1}^K |U_{i,k}| \leq \delta \middle| \gamma_{r+1}, \ldots, \gamma_K\right) &= 1 - \pi\left(\max_{i=1}^m \max_{k=1}^K |U_{i,k}| > \delta \middle| \gamma_{r+1}, \ldots, \gamma_K\right) \\
&\geq 1 - \pi\left(\sum_{i=1}^m \sum_{k=1}^K |U_{i,k}| > \delta \middle| \gamma_{r+1}, \ldots, \gamma_K\right) \\
&\geq 1 - \frac{\sum_{i=1}^m \sum_{k=1}^K \pi\left(|U_{i,k}||\,\gamma_k\right)}{\delta} \\
&\geq 1 - \frac{mK \max_k \sqrt{\gamma_k}}{\delta} \\
&\geq 1 - \frac{mK\sqrt{2s}}{\delta} \\
&= \frac{1}{2} \text{ with the choice } s = \frac{1}{2}\left(\frac{\delta}{2(m \vee p)K}\right)^2
\end{aligned}
$$

(note that this choice of s satisfies $0 < s < B^2$). Then, for $k \leq r$,

$$
\begin{aligned}
\pi\left(|U_{i,k} - \bar{U}_{i,k}| \leq \delta|\gamma_k\right) &= \frac{1}{\sqrt{2\pi\gamma_k}} \int_{\bar{U}_{i,k}-\delta}^{\bar{U}_{i,k}+\delta} \exp\left(-\frac{x^2}{2\gamma_k}\right)\mathrm{d}x \\
&\geq \frac{2\delta \exp\left(-\frac{(B+\delta)^2}{2\gamma_k}\right)}{\sqrt{2\pi\gamma_k}}
\end{aligned}
$$

$$\geq \frac{\delta \exp\left(-\frac{(B+\delta)^2}{2B^2}\right)}{B\sqrt{\pi}} \text{ as } B^2 \leq \gamma_k \leq 2B^2$$
$$\geq \frac{\delta \exp(-2)}{B\sqrt{\pi}} \text{ as } \delta < 1 \leq B$$

and so

$$\prod_{i=1}^{m}\prod_{k=1}^{r} \pi\left(|U_{i,k} - \bar{U}_{i,k}| \leq \delta | \gamma_k\right) \geq \left(\frac{\delta}{B\sqrt{\pi}}\right)^{mr} \exp(-2mr).$$

Finally

$$\begin{aligned}\int_E \pi\left(\|U - \bar{U}\|_\infty \leq \delta | \gamma\right) \pi(\mathrm{d}\gamma) &\geq \int_E \frac{1}{2}\left(\frac{\delta}{B\sqrt{\pi}}\right)^{mr} \exp(-2mr)\, \pi(\mathrm{d}\gamma) \\ &= \frac{1}{2}\left(\frac{\delta}{B\sqrt{\pi}}\right)^{mr} \exp(-2mr)\, \pi(\gamma \in E)\end{aligned}$$

and it remains to lower bound

$$\begin{aligned}\pi(\gamma \in E) &= \left(\prod_{k=1}^{r} \pi(1 \leq \gamma_k \leq 2)\right)\left(\prod_{k=r+1}^{K} \pi(0 \leq \gamma_k \leq s)\right) \\ &= \left(\frac{b^a}{\Gamma(a)}\right)^K \left[\int_{B^2}^{2B^2} x^{-a-1} \exp\left(-\frac{b}{x}\right) \mathrm{d}x\right]^r \left[\int_s^{2s} x^{-a-1} \exp\left(-\frac{b}{x}\right) \mathrm{d}x\right]^{K-r} \\ &\geq \left(\frac{b^a}{\Gamma(a)}\right)^K \left[B^2 (2B^2)^{-a-1} \exp\left(-\frac{b}{B^2}\right)\right]^r \left[s(2s)^{-a-1} \exp\left(-\frac{b}{s}\right)\right]^{K-r} \\ &= \left(\frac{b^a}{2^{a+1}\Gamma(a)}\right)^K \exp\left[-\frac{b}{B^2} r - \frac{b}{s}(K-r)\right] (B^2)^{-(a+1)r} s^{-a(K-r)} \\ &\geq \left(\frac{b^a}{(B^2)^a 2^{a+1}\Gamma(a)}\right)^K \exp\left[-\frac{Kb}{s}\right]\end{aligned}$$

as $s < B^2$. So, finally,

$$\mathcal{K}(\rho_n, \pi) \leq r(m+p) \log\left(\frac{B\sqrt{\pi}\exp(2)}{\delta}\right) + K\left[\log\left(\frac{2^{a+1}\Gamma(a)(B^2)^a}{b^a}\right) + \frac{b}{s}\right] + 2\log(2).$$

The choice $b = s$ leads to

$$\begin{aligned}\mathcal{K}(\rho_n, \pi) &\leq r(m+p) \log\left(\frac{B\sqrt{\pi}\exp(2)}{\delta}\right) + K \log\left(\frac{\mathrm{e}2^{a+1}\Gamma(a)(B^2)^a}{s^a}\right) + 2\log(2) \\ &\leq r(m+p) \log\left(8\sqrt{\pi}\exp(2)(nmp)^2\right) + 4aK \log(nmp) \\ &\quad + K \log(\mathrm{e}2^{10a+1}\Gamma(a)) + 2\log(2)\end{aligned}$$

where we replaced δ and s by their respective value. In order to keep the expressions as simple as possible we can use $K \leq m \vee p \leq m+p \leq r(m+p)$ and $2 \leq m+p \leq r(m+p)$ to get

$$\mathcal{K}(\rho_n, \pi) \leq 2(1+2a)r(m+p)\left[\log(nmp) + \underbrace{\log(8\sqrt{\pi}\Gamma(a)2^{10a+1}) + 3}_{=:\mathcal{C}(a)}\right].$$

We are now in position to apply Theorem 2.6. Then

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0}) \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n)\right]$$

$$\leq \frac{\alpha}{1-\alpha}\int \mathcal{K}(P_{\theta_0}, P_\theta)\rho_n(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho_n,\pi)}{n(1-\alpha)}$$
$$\leq \frac{\alpha}{1-\alpha}\frac{\left[\|M-\bar{U}\bar{V}^T\|_F + \frac{\sqrt{B}}{n}\right]^2}{2\sigma^2 mp} + \frac{2(1+\alpha)(1+2a)r(m+p)\left[\log(nmp)+\mathcal{C}(a,B)\right]}{n(1-\alpha)}.$$

6.5 Proof of Theorem 4.1

We start by defining a sequence $\rho_n(d\theta) := \Phi(d\theta; \theta_0, \sigma_n^2 I) \in \mathcal{F}_\Phi^{id}$ indexed by a positive scalar σ_n^2 to be later defined. As before by proving the result on the smallest family of distribution, it will remain true on larger ones using the fact that $\min_{\mathcal{F}^{id}} \leq \min_{\mathcal{F}^{diag}} \leq \min_{\mathcal{F}^{full}}$. Under Assumption 4.1 we can check the hypotheses on the KL between the likelihood terms as required in Theorem 2.4. We have

$$\mathcal{K}(P_{\theta_0}, P_\theta) = \mathbb{E}\left[\log p_{\theta_0}(X) - \log p_\theta(X)\right] \leq \mathbb{E}\left[M(X)\right]\|\theta-\theta_0\|_2^{2p}$$

and

$$\mathbb{E}\left[\log^2 \frac{p_{\theta_0}}{p_\theta}(X)\right] = \mathbb{E}\left[\left(\log p_{\theta_0}(X) - \log p_\theta(X)\right)^2\right] \leq \mathbb{E}\left[M(X)^2\right]\|\theta-\theta_0\|_2^{4p}$$

When integrating with respect to ρ_n two cases must be studied. We start by the case when $p = \frac{1}{2}$ direct calculations yield

$$\int \mathcal{K}(P_{\theta_0}, P_\theta)\rho_n(d\theta) \leq B_1\sigma_n\sqrt{d}$$
$$\int \mathbb{E}\left[\log^2 \frac{p_{\theta_0}}{p_\theta}(X)\right]\rho_n(d\theta) \leq B_2\sigma_n^2 d,$$

In the case where $p \in \mathbb{N}^\star$ we can write using the second part of Assumption 4.1 and the expression of the moments of Chi-square distribution

$$\int \mathcal{K}(P_{\theta_0}, P_\theta)\rho_n(d\theta) \leq \mathbb{E}\left[M(X)\right]\int \|\theta-\theta_0\|_2^{2p}\rho_n(d\theta)$$
$$= \sigma_n^{2p} B_1 2^p \frac{\Gamma(p+\frac{1}{2}d)}{\Gamma(\frac{1}{2}d)}$$

and

$$\int \mathbb{E}\left[\log^2 \frac{p_{\theta_0}}{p_\theta}(X)\right]\rho_n(d\theta) \leq \mathbb{E}\left[M^2(X)\right]\int \|\theta-\theta_0\|_2^{4p}\rho_n(d\theta)$$
$$= \sigma_n^{4p} B_2 2^{2p} \frac{\Gamma(2p+\frac{1}{2}d)}{\Gamma(\frac{1}{2}d)}.$$

In the following define $C_{p,d} = \begin{cases} \sqrt{d} & \text{if } p = \frac{1}{2} \\ 2^{2p}\frac{\Gamma(2p+\frac{1}{2}d)}{\Gamma(\frac{1}{2}d)} & \text{otherwise} \end{cases}$.

To apply Theorem 2.4 it remains to compute the KL between the approximation of the pseudo-posterior and the prior,

$$\frac{1}{n}\mathcal{K}(\rho_n,\pi) = \frac{d}{n}\left[\frac{1}{2}\log\left(\frac{\vartheta^2}{\sigma^2}\right) + \frac{\sigma^2}{\vartheta^2}\right] + \frac{\|\theta_0\|^2}{n\vartheta^2} - \frac{d}{2n}.$$

To obtain an estimate of the rate r_n of Theorem 2.4 we put together those bounds. Choosing $\sigma_n^{2p} = \frac{1}{nC_{p,d}}$ we can apply it with

$$r_n = \frac{B_1}{n} \vee \frac{B_2}{n^2} \vee \left\{\frac{d}{n}\left[\frac{1}{2}\log\left(\vartheta^2 n^2 C_{p,d}\right) + \frac{1}{n\vartheta^2}\right] + \frac{\|\theta_0\|^2}{n\vartheta^2} - \frac{d}{2n}\right\}.$$

6.6 Proof of Theorem 4.2

From the proof of Theorem 2.6 we get

$$\begin{aligned}
&\mathbb{E}\left[\int D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n})\tilde{\pi}_{n,\alpha}^k(\mathrm{d}\theta|X_1^n)\right] \\
&\quad\leq \mathbb{E}\left[\frac{\alpha}{1-\alpha}\int r_n(\theta,\theta_0)\tilde{\pi}_{n,\alpha}^k(\mathrm{d}\theta|X_1^n) + \frac{\mathcal{K}(\tilde{\pi}_{n,\alpha}^k(\cdot|X_1^n),\pi)}{1-\alpha}\right] \\
&\quad= \mathbb{E}\left\{\inf_{\rho\in\mathcal{F}_B^\Phi}\left[\frac{\alpha}{1-\alpha}\int r_n(\theta,\theta_0)\rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho,\pi)}{1-\alpha}\right]\right\} \\
&\qquad+\left\{\mathbb{E}\left[\frac{\alpha}{1-\alpha}\int r_n(\theta,\theta_0)\tilde{\pi}_{n,\alpha}^k(\mathrm{d}\theta|X_1^n) + \frac{\mathcal{K}(\tilde{\pi}_{n,\alpha}^k(\cdot|X_1^n),\pi)}{1-\alpha}\right]\right. \\
&\qquad\left.-\mathbb{E}\left\{\inf_{\rho\in\mathcal{F}_B^\Phi}\left[\frac{\alpha}{1-\alpha}\int r_n(\theta,\theta_0)\rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho,\pi)}{1-\alpha}\right]\right\}\right\}.
\end{aligned}$$

By definition of f we get

$$\begin{aligned}
&\mathbb{E}\left[\int D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n})\tilde{\pi}_{n,\alpha}^k(\mathrm{d}\theta|X_1^n)\right] \\
&\quad= \mathbb{E}\left\{\inf_{\rho\in\mathcal{F}_B^\Phi}\left[\frac{\alpha}{1-\alpha}\int r_n(\theta,\theta_0)\rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho,\pi)}{1-\alpha}\right]\right\} + \frac{1}{1-\alpha}\mathbb{E}\left\{\mathbb{E}f(\bar{x}_k,\xi) - \inf_{u\in\mathbb{B}}\mathbb{E}f(u,\xi)\right\} \\
&\quad\leq \inf_{\rho\in\mathcal{F}_B^\Phi}\left\{\mathbb{E}\left[\frac{\alpha}{1-\alpha}\int r_n(\theta,\theta_0)\rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho,\pi)}{1-\alpha}\right]\right\} + \frac{1}{1-\alpha}\mathbb{E}\left\{\mathbb{E}f(\bar{x}_k,\xi) - \inf_{u\in\mathbb{B}}\mathbb{E}f(u,\xi)\right\} \\
&\quad= \inf_{\rho\in\mathcal{F}_B^\Phi}\left\{\frac{n\alpha}{1-\alpha}\int \mathcal{K}(P_{\theta_0},P_\theta)\rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho,\pi)}{1-\alpha}\right\} + \frac{1}{1-\alpha}\mathbb{E}\left\{\mathbb{E}f(\bar{x}_k,\xi) - \inf_{u\in\mathbb{B}}\mathbb{E}f(u,\xi)\right\}.
\end{aligned}$$

Following the rest of the proof of 2.6 we get

$$\begin{aligned}
&\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}^k(\mathrm{d}\theta|X_1^n)\right] \\
&\quad\leq \inf_{\rho\in\mathcal{F}}\left\{\frac{\alpha}{1-\alpha}\int \mathcal{K}(P_{\theta_0},P_\theta)\rho(\mathrm{d}\theta) + \frac{\mathcal{K}(\rho,\pi)}{n(1-\alpha)}\right\} + \frac{1}{n(1-\alpha)}\mathbb{E}\left\{\mathbb{E}f(\bar{x}_k,\xi) - \inf_{u\in\mathbb{B}}\mathbb{E}f(u,\xi)\right\}.
\end{aligned}$$

To bound the first term of the right hand-side we use Assumption 4.1 and the proof of Theorem 4.1. In particular notice that $\Phi(d\theta;\theta_0,\frac{1}{n\sqrt{d}}I_d)\in\mathcal{F}_B^\Phi$, we get straight away

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}^k(\mathrm{d}\theta|X_1^n)\right] \quad\leq \frac{1+\alpha}{1-\alpha}r_n + \frac{1}{n(1-\alpha)}\mathbb{E}\left\{\mathbb{E}f(\bar{x}_k,\xi) - \inf_{u\in\mathbb{B}}\mathbb{E}f(u,\xi)\right\}$$

We now study the term inside the brackets on the right hand-side.

First notice that the sequence $(x_t)_{t\geq 0}$ in Algorithm 1 is equivalent to that of an online gradient descent on the sequence $\{f(x,\xi_t)\}_t$. Hence under Assumption 1 we can apply Corollary 2.7 of [25] with $\gamma_T = \frac{B}{L\sqrt{2T}}$ to get the following bound on the regret for any $u\in\mathbb{B}$

$$\sum_{t=1}^T f(x_t,\xi_t) - \sum_{t=1}^T f(u,\xi_t) \leq \sqrt{2BLT}.$$

Divide by T, take expectation with respect to $(\xi_t)_t$

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}f(x_t,\xi_t) - \mathbb{E}f(u,\xi) \leq \sqrt{\frac{2BL}{T}}.$$

Notice that x_t belongs to the σ-algebra generated by $(x_1, \cdots, x_{t-1})$. By a multiple use of the tower property we get,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}f(x_t,\xi) - \mathbb{E}f(u,\xi) \leq \sqrt{\frac{2BL}{T}}$$

$$\mathbb{E}f(\bar{x},\xi) - \mathbb{E}f(u,\xi) \leq \sqrt{\frac{2BL}{T}}, \text{ by Jensen and the convexity of } f.$$

Putting everything together concludes the proof.

6.7 Proof of Corollary 4.3

Direct calculation shows that the log-likelihood is $2\|X\|$-Lipschitz hence satisfying Assumption 4.1. We conclude using Theorem 4.1 and the assumption on the design matrix.

6.8 Proof of Corollary 4.4

Start by noticing that we can take f as

$$f((m,C),\xi) := \alpha \log p_{m+C\xi}(\tilde{x}) + \mathcal{K}(\rho,\pi),$$

where $\rho(.) = \Phi(.; m, CC^t)$ the likelihood part is convex with Lipschitz gradient as a composition of a convex and gradient Lipschitz function with a affine map. The KL part can be written as $\mathcal{K}(\rho,\pi) = \frac{\|m\|^2}{2\vartheta} + \left(\frac{1}{2\vartheta}\text{trace}(CC^t) - \log|C|\right)$ which is convex for positive semi-definite C. We need to check that the gradients of the objectives are also Lipschitz, the only problematic term is $\log\det(C)$. Denote (λ_i) the eigen values of $\Sigma = CC^T$

$$\begin{aligned}
\Sigma \succeq \psi I_{d\times d} \Rightarrow \forall i \in \{1,\cdots,d\}, \quad & \frac{1}{\lambda_i} \leq \frac{1}{\psi} \\
\Rightarrow \text{trace}(\Sigma^{-1}) &\leq \frac{d}{\psi} \\
\Rightarrow \text{trace}^{\frac{1}{2}}\left(C^{-1}C^{-T}\otimes C^{-1}C^{-T}\right) &\leq \frac{d}{\psi} \\
\Rightarrow \text{trace}^{\frac{1}{2}}\left((C^{-1}\otimes C^{-T})(C^{-1}\otimes C^{-T})\right) &\leq \frac{d}{\psi} \\
\Rightarrow \|\nabla^2_C \log\det C\|_2 &\leq \frac{d}{\psi}.
\end{aligned}$$

To apply Theorem 4.2 we also need to check that the new constraint contains the Gaussian distribution used in the proof. This is the case as long as $\psi \leq \sigma^2 = \frac{1}{n\sqrt{d}}$.

References

[1] P. Alquier, V. Cottet, N. Chopin, and J. Rousseau. Bayesian matrix completion: prior specification. *arXiv preprint arXiv:1406.1440*, 2014.

[2] P. Alquier and B. Guedj. An oracle inequality for quasi-bayesian nonnegative matrix factorization. *Mathematical Methods of Statistics*, 26(1):55–67, 2017.

[3] P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of gibbs posteriors. *JMLR*, 17(239):1–41, 2016.

[4] A. Bhattacharya, D. Pati, and Y. Yang. Bayesian fractional posteriors. *arXiv preprint arXiv:1611.01125*, 2016.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *arXiv:1601.00670*, 2017.

[7] E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[8] E. J. Candès and T. Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010.

[9] O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics. Springer, 2004.

[10] 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.

[11] N. Chopin and J. Ridgway. Leave pima indians alone: binary regression as a benchmark for bayesian computation. *Statistical Science*, 32(1):64–87, 2017.

[12] S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000.

[13] M. N. Gibbs and D. J. C. MacKay. Variational gaussian process classifiers. *IEEE Transactions on Neural Networks*, 11(6):1458–1464, 2000.

[14] Peter D. Grünwald and Nishant A. Mehta. Fast rates with unbounded losses. *arXiv preprint arXiv:1605.00252*, 2016.

[15] 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.

[16] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[17] 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.

[18] 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.

[19] Xin Li and Yunfei Zheng. Patch-based video processing: A variational bayesian approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(1):27–40, 2009.

[20] Y. J. Lim and Y. W. Teh. Variational Bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop*, volume 7, pages 15–21, 2007.

[21] T. T. Mai and P. Alquier. A bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 9(1):823–841, 2015.

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

[23] J. Rousseau. On the frequentist properties of bayesian nonparametric methods. *Annual Review of Statistics and Its Application*, 3:211–231, 2016.

[24] 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.

[25] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[26] T. Suzuki. Convergence rate of bayesian tensor estimator and its minimax optimality. In *ICML*, pages 1273–1282, 2015.

[27] M. Titsias and M. Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1971–1979, 2014.

[28] T. Van Erven and P. Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[29] Bo Wang and DM Titterington. Convergence and asymptotic normality of variational bayesian approximations for exponential family models with missing values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 577–584. AUAI Press, 2004.

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