The Annals of Statistics
2020, Vol. 48, No. 3, 1475–1497
https://doi.org/10.1214/19-AOS1855
© Institute of Mathematical Statistics, 2020

CONCENTRATION OF TEMPERED POSTERIORS AND OF THEIR VARIATIONAL APPROXIMATIONS

BY PIERRE ALQUIER[1] AND JAMES RIDGWAY[2]

[1]*RIKEN Center for Advanced Intelligence Project, pierre.alquier.stat@gmail.com*
[2]*Capital Fund Management, james.lp.ridgway@gmail.com*

While Bayesian methods are extremely popular in statistics and machine learning, their application to massive data sets is often challenging, when possible at all. 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 $\mathcal{F}$. 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 collaborative filtering, image and video processing or NLP to name a few. However, despite nice results in practice, the theoretical properties of these approximations are not known. We propose a general oracle inequality that relates the quality of the VB approximation to the prior π and to the structure of $\mathcal{F}$. We provide a simple condition that allows to derive rates of convergence from this oracle inequality. We apply our theory to various examples. First, we show that for parametric models with log-Lipschitz likelihood, Gaussian VB leads to efficient algorithms and consistent estimators. We then study a high-dimensional example: matrix completion, and a nonparametric example: density estimation.

1. Introduction.

1.1. *Motivation.* In many applications of Bayesian statistics, the posterior is not tractable. Markov Chain Monte Carlo algorithms (MCMC) were developed to allow the statistician to sample from the posterior distribution even in situations where a closed-form expression is not available. MCMC methods were successfully used in many applications, and are still one of the most valuable tools in the statistician's 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 compute variational approximations of the posterior by using optimization algorithms. 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 [26], NLP on massive datasets [21], video processing [25], classification with Gaussian processes [17], among others. Chapter 10 in [7] is a good introduction to VB and [8] provides an exhaustive survey.

Despite its practical success very little attention has been put toward theoretical guaranties for VB. Asymptotic results in exponential models were provided in [38]. More recently, [39] proposed a very nice asymptotic study of approximations in parametric models. The main problem with these results is that by nature they cannot be applied to high-dimensional or nonparametric models, or to model selection. In the machine learning community, [4] also

Received June 2018; revised April 2019.
MSC2010 subject classifications. Primary 62G15; secondary 62C10, 62C20, 62F25, 62G25.
Key words and phrases. Concentration of the posterior, rate of convergence, variational approximation, PAC-Bayesian bounds.

studied VB approximations. In a distribution-free setting, there is actually no likelihood, but a pseudo-likelihood can be defined through a suitable loss function, and thus it is possible to define a pseudo-posterior. Thanks to PAC-Bayesian inequalities from [11, 12], [4] derived rates of convergence for VB approximation of this pseudo-posterior. However, the tools used in [4] are valid for bounded loss functions, so there is no direct way to adapt this method to study VB approximations 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 [15], an account of recent advances can be found in in [16, 31]. Recently, [6] studied the situation where the likelihood $L(\theta)$ is replaced by $L^{\alpha}(\theta)$ for $0<\alpha<1$, leading to what is usually called a *fractional* or *tempered* posterior. They proved that concentration of the fractional posterior requires actually fewer hypothesis than concentration of the (true) posterior. Extending the technique of [6], we analyze the concentration of VB approximations of (fractional) posteriors. Especially, we derive a condition for the VB approximation to concentrate at the same rate as the fractional posterior.

1.2. *Definitions and notation.* We observe a collection of n i.i.d. random variables $(X_1,\dots,X_n)=X_1^n$ in a measured sample space $(\mathbb{X},\mathcal{X},\mathbb{P})$. Let $\{P_\theta,\theta\in\Theta\}$ be a statistical model (a collection of probability distributions). The objective here is to estimate the distribution of the X_i's. Most 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. Let us first assume that $\mathbb{P}\equiv P_{\theta_0}^{\otimes n}$ (we will explicitly mention when this will no longer be the case).

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 $\mathcal{M}_1^+(E)$ be the set of all probability distributions on a measurable space $(E,\mathcal{E})$. Assume Θ is equipped with some σ-algebra $\mathcal{T}$. Let $\pi\in\mathcal{M}_1^+(\Theta)$ denote the prior. The likelihood and the negative log-likelihood ratio will be denoted respectively[1] by

$$\forall(\theta,\theta')\in\Theta^2,\quad L_n(\theta)=\prod_{i=1}^n p_\theta(X_i)\quad\text{and}\quad r_n(\theta,\theta')=\sum_{i=1}^n\log\frac{p_{\theta'}(X_i)}{p_\theta(X_i)}.$$

DEFINITION 1.1. Let $\alpha\in(0,1)$. Let P and R be two probability measures. Let μ be any measure such that $P\ll\mu$ and $R\ll\mu$, for example, $\mu=P+R$. The α-Rényi divergence and the Kullback–Leibler (KL) divergence between two probability distributions P and R are respectively defined by

$$D_\alpha(P,R)=\frac{1}{\alpha-1}\log\int\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)^{\alpha}\left(\frac{\mathrm{d}R}{\mathrm{d}\mu}\right)^{1-\alpha}\mathrm{d}\mu,$$

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

REMARK 1.1. We remind the reader of a few properties proven in [37]. First, it is obvious that $D_\alpha(P,R)$ does actually not depend on the choice of the reference measure μ. This is sometimes made explicit by the (informal) statement $D_\alpha(P,R)=(1/(\alpha-$

[1] In order to manipulate these quantities, 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.

1)) $\log \int (\mathrm{d}P)^\alpha (\mathrm{d}R)^{1-\alpha}$. The measures P and R are mutually singular if and only if $D_\alpha(P, R) = (\frac{1}{\alpha-1}) \log(0) = +\infty$.

We have $\lim_{\alpha \to 1} D_\alpha(P, R) = \mathcal{K}(P, R)$ which gives ground to the notation $D_1(P, R) = \mathcal{K}(P, R)$. For $\alpha \in (0, 1]$, $(\alpha/2) d_{\mathrm{TV}}^2(P, R) \leq D_\alpha(P, R)$, d_{TV} being the total variation distance – for $\alpha = 1$ this is Pinsker's inequality. The map $\alpha \mapsto D_\alpha(P, R)$ is nondecreasing. Also, the authors of [6] note that the α-Rényi divergences are all equivalent for $0 < \alpha < 1$, through the formula $\frac{\alpha}{\beta} \frac{1-\beta}{1-\alpha} D_\beta \leq D_\alpha \leq D_\beta$ for $\alpha \leq \beta$. Additivity holds: $D_\alpha(P_1 \otimes P_2, R_1 \otimes R_2) = D_\alpha(P_1, R_1) + D_\alpha(P_2, R_2)$, thus $D_\alpha(P^{\otimes n}, R^{\otimes n}) = n D_\alpha(P, R)$; $D_{1/2}(P, R) \geq 2[1 - \exp(-(1/2) D_{1/2}(P, R))] = H^2(P, R)$ the squared Hellinger distance.

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 [6]. The variational approximation $\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n)$ of $\pi_{n,\alpha}(\mathrm{d}\theta | X_1^n)$ is defined as the projection in KL divergence onto a predefined family of distributions $\mathcal{F}$.

DEFINITION 1.2 (Variational Bayes approximation). Let $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$,

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

In Section 2, we state general theorems on the concentration of $\tilde{\pi}_{n,\alpha}(\cdot | X_1^n)$, for example, Theorem 2.4. One of the key assumptions is that the prior gives enough mass to neighborhoods of the true parameter, a condition also required to prove the concentration of the posterior [6, 15, 31]. Here, an additional, but completely natural assumption is required: $\mathcal{F}$ must actually contain distributions concentrated around the true parameter. The choice of $\mathcal{F}$ has thus 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 known results on $\pi_{n,\alpha}$. But this is 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 ensure concentration.

In Sections 3, 4 and 5, we apply our general results in various settings. In Section 3, we study the parametric family of Gaussian approximations

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

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 $d \times d$ symmetric positive definite matrices. We show that other approximations are possible, that is, by constraining the variance of the approximation to be a positive diagonal matrix $\Sigma \in \mathrm{Diag}_+^d(\mathbb{R})$. Gaussian approximations have been studied in [29, 35]. We specify those results in the case of a logistic regression in Section 3.2. There 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. In Section 4, we study the case of mean field approximations corresponding to block-independent distributions

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

in the context of matrix completion. While the VB approximation leads to feasible approximation algorithms [26], our theorem shows that $\tilde{\pi}_{n,\alpha}$ concentrates at the minimax-optimal rate. In Section 5, we provide a nonparametric example: density estimation. The more important proofs are gathered in Section 7. The supplementary material [3] contains the remaining proofs and additional comments.

2. Main results.

2.1. *A PAC-Bayesian inequality.* We start with a variant of a result of [6].

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

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

It is tempting to minimize the right-hand side of the inequality in order to ensure a good estimation. The minimizer of the right-hand side can actually be explicitly given. In order to do this, let us recall Donsker and Varadhan's variational inequality (Lemma 1.1.3 in [12]).

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

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

with the convention $\infty - \infty = -\infty$. *Moreover, when h is upper bounded on the support of* π *the supremum with respect to* ρ *in the right-hand side is reached by* π_h *given by* $\mathrm{d}\pi_h/\mathrm{d}\pi(\theta) = \exp(h(\theta))/\int \exp(h)\,\mathrm{d}\pi$.

Using Lemma 2.2 with $h(\theta) = -\alpha r_n(\theta,\theta_0)$ and the definition of $\pi_{n,\alpha}$, we obtain

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

so the minimizer of the right-hand side of Theorem 2.1 is actually $\pi_{n,\alpha}(\mathrm{d}\theta|X_1^n)$.

The statement of Theorem 2.1 for $\rho = \pi_{n,\alpha}(\mathrm{d}\theta|X_1^n)$ is Theorem 3.5 in [6]. The proof of Theorem 2.1 requires a straightforward extension, we provide it in Section 7 for the sake of completeness. Our extension is crucial though as we will have to use it with $\rho = \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)$.

REMARK 2.1. Theorem 2.1 can be used to study other approximations of the posterior. For example, as suggested by one of the referees, we can use it to study distributions centered around the *maximum a posteriori* (MAP) or the maximum likelihood estimate (MLE). For example, Laplace approximations are Gaussian distributions centered at the MLE. However, there are models $(P_\theta, \theta \in \Theta)$ where the MLE and the MAP are not defined, while the posterior and some variational approximations are consistent. Such an example is provided in the supplementary material [3].

2.2. *Concentration of VB approximations.* We specialize the above results to the variational approximation. Elementary calculations show that

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

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

COROLLARY 2.3. *For any $\alpha\in(0,1)$ and $\varepsilon\in(0,1)$, with probability at least $1-\varepsilon$,*

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

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 recover as a special case an upper bound on the risk of the tempered posterior. We are now in position to state our main result.

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

$$\int\mathcal{K}(P_{\theta_0},P_\theta)\rho_n(\mathrm{d}\theta)\le\varepsilon_n,\qquad\int\mathbb{E}\left[\log^2\left(\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}\right)\right]\rho_n(\mathrm{d}\theta)\le\varepsilon_n \tag{2.1}$$

and

$$\mathcal{K}(\rho_n,\pi)\le n\varepsilon_n. \tag{2.2}$$

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

$$\mathbb{P}\left[\int D_\alpha(P_\theta,P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)\le\frac{(\alpha+1)\varepsilon_n+\alpha\sqrt{\frac{\varepsilon_n}{n\eta}}+\frac{\log(\frac{1}{\varepsilon})}{n}}{1-\alpha}\right]\ge1-\varepsilon-\eta.$$

This theorem is a consequence of Corollary 2.3, its proof is provided in Section 7. Let us now discuss the main consequences of this theorem.

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

$$B(r)=\left\{\theta\in\Theta:\mathcal{K}(P_{\theta_0},P_\theta)\le r,\mathbb{E}\left[\log^2\left(\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}\right)\right]\le r\right\}.$$

Then the choice $\rho_n=\pi_{|B(\varepsilon_n)}$, i.e., π restricted to $B(\varepsilon_n)$, ensures immediately (2.1), and (2.2) can be rewritten

$$-\log\pi\big(B(\varepsilon_n)\big)\le n\varepsilon_n.$$

This assumption is standard to study concentration of the posterior; see Theorem 2.1 page 503 in [15] or Section 3.2 in [31]. Our message is that in the studies of concentration of the posterior, the choice $\rho_n=\pi_{|B(\varepsilon_n)}$ was hidden. Other choices might lead to easier calculations in some situations. More importantly, in the relevant case $\mathcal{F}\subsetneq\mathcal{M}_1^+(\Theta)$, $\pi_{|B(\varepsilon_n)}\notin\mathcal{F}$ in general. Thus, $-\log\pi(B(\varepsilon_n))\le n\varepsilon_n$ is no longer sufficient, and (2.1) and (2.2) are natural extensions of this assumption to study VB. They provide an explicit condition on the family $\mathcal{F}$ in order to ensure concentration of the approximation.

Choosing $\eta = \frac{1}{n\varepsilon_n}$ and $\varepsilon = \exp(-n\varepsilon_n)$, we obtain a more readable concentration result. It shows that, as soon as $(1/n) \ll \varepsilon_n \ll 1$, the sequence ε_n gives a concentration rate for VB.

COROLLARY 2.5. *Under the same assumptions as in Theorem 2.4*,

$$\begin{aligned}\mathbb{P}\biggl[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n) \le \frac{2(\alpha+1)}{1-\alpha}\varepsilon_n\biggr] \\ \ge 1 - \frac{1}{n\varepsilon_n} - \exp(-n\varepsilon_n) \\ \ge 1 - \frac{2}{n\varepsilon_n}.\end{aligned}$$

REMARK 2.2. As a special case, when $\alpha = 1/2$, the theorem leads to a concentration result in terms of the more classical Hellinger distance

$$\mathbb{P}\biggl[\int H^2(P_\theta, P_{\theta_0})\tilde{\pi}_{n,1/2}(\mathrm{d}\theta|X_1^n) \le 6\varepsilon_n\biggr] \ge 1 - \frac{2}{n\varepsilon_n}.$$

Also, with a general $\alpha \in (0,1)$, from the properties recalled in Remark 1.1, we have, for $0 < \beta \le \alpha$,

$$\mathbb{P}\biggl[\int D_\beta(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n) \le \frac{2(\alpha+1)}{1-\alpha}\varepsilon_n\biggr] \ge 1 - \frac{2}{n\varepsilon_n},$$

and for $\alpha \le \beta < 1$,

$$\mathbb{P}\biggl[\int D_\beta(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n) \le \frac{2\beta(\alpha+1)}{\alpha(1-\beta)}\varepsilon_n\biggr] \ge 1 - \frac{2}{n\varepsilon_n}.$$

2.3. *A simpler result in expectation.* It is possible to simplify the assumptions at the price of stating a result in expectation instead of concentration.

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

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

Assume that $\varepsilon_n > 0$ is such that there is distribution $\rho_n \in \mathcal{F}$ such that

$$\int \mathcal{K}(P_{\theta_0}, P_\theta)\rho_n(\mathrm{d}\theta) \le \varepsilon_n \quad \textit{and} \quad \mathcal{K}(\rho_n, \pi) \le n\varepsilon_n.$$

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

$$\mathbb{E}\biggl[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)\biggr] \le \frac{1+\alpha}{1-\alpha}\varepsilon_n.$$

2.4. *Extension of the result in expectation to the misspecified case.* In this section, we do not assume any longer that the true distribution is in $\{P_\theta, \theta \in \Theta\}$. In order not to change all the notation we define an extended parameter set $\Theta \cup \{\theta_0\}$ where $\theta_0 \notin \Theta$ and define P_{θ_0} as the true distribution. Theorem 2.6 can be applied to this setting, and we obtain

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

Now, rewriting, for $\theta^* \in \Theta$,

$$\mathcal{K}(P_{\theta_0}, P_\theta) = \mathcal{K}(P_{\theta_0}, P_{\theta^*}) + \mathbb{E}\biggl[\log\frac{p_{\theta^*}(X_i)}{p_\theta(X_i)}\biggr],$$

we obtain the following result.

THEOREM 2.7. *Assume that, for $\theta^* = \arg\min_{\theta\in\Theta}\mathcal{K}(P_{\theta_0}, P_\theta)$, there is $\varepsilon_n > 0$ and $\rho_n \in \mathcal{F}$ with*

$$\int \mathbb{E}_{\theta_0}\biggl[\log\frac{p_{\theta^*}(X_i)}{p_\theta(X_i)}\biggr]\rho_n(\mathrm{d}\theta) \leq \varepsilon_n \quad \textit{and} \quad \mathcal{K}(\rho_n,\pi) \leq n\varepsilon_n,$$

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

$$\mathbb{E}\biggl[\int D_\alpha(P_\theta, P_{\theta_0})\tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)\biggr] \leq \frac{\alpha}{1-\alpha}\min_{\theta\in\Theta}\mathcal{K}(P_{\theta_0}, P_\theta) + \frac{1+\alpha}{1-\alpha}\varepsilon_n.$$

In the well-specified case, $\theta^* = \theta_0$ and we recover Theorem 2.6. Otherwise, this result takes the form of an oracle inequality. It is not a sharp oracle inequality as that the risk measure used in the left-hand side and the right-hand side are not the same, but remains informative when $\mathcal{K}(P_{\theta_0}, P_{\theta^*})$ is small. For example, in Section 5 below, we provide a nonparametric example where $\mathcal{K}(P_{\theta_0}, P_{\theta^*})$ and Theorem 2.7 leads to the minimax rate of convergence.

3. Gaussian variational Bayes. In this section, we consider $\Theta \subset \mathbb{R}^d$ and the class of Gaussian approximations

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

thus the algorithm will consist in projecting onto the set of Gaussian distributions. Depending on the hypotheses made on the covariance matrix, we can build different approximations. For instance, define

$$\mathcal{F}_{\text{diag}}^\Phi := \{\Phi(d\theta; m, \Sigma), m \in \mathbb{R}^d, \Sigma \in \mathrm{Diag}_+^d(\mathbb{R})\},$$
$$\mathcal{F}_{\text{id}}^\Phi := \{\Phi(d\theta; m, \sigma^2 I_d), m \in \mathbb{R}^d, \sigma^2 \in \mathbb{R}_{+\star}\}.$$

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

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 α-Rényi 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 $\frac{1}{n}\mathcal{K}(\rho,\pi) \leq \varepsilon_n$; many other distribution could be used, providing different rates.

In the rest of the section, we assume that the density is log Lipschitz.

ASSUMPTION 3.1. There is a measurable real function $M(\cdot)$ such that

$$\left|\log p_\theta(X_1) - \log p_{\theta'}(X_1)\right| \le M(X_1)\|\theta - \theta'\|_2.$$

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

An example is logistic regression; see Section 3.2 below.

THEOREM 3.1. *Let the approximation family be $\mathcal{F}$ with $\mathcal{F}_{\mathrm{id}}^{\Phi} \subset \mathcal{F}$ as defined above and that the model satisfies Assumption 3.1. We put*

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

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

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

3.1. *Stochastic variational Bayes.* In many cases, the model is not conjugate, that is, the VB objective does not have a closed-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 [35].

We may write our variational bound as the following minimization problem:

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

or after dropping the constants,

$$\begin{aligned}\min_{m\in\mathbb{R}^d, \Sigma\in\mathcal{S}_+^d} &\left\{-\alpha \int \log p_\theta(Y^n)\Phi(d\theta; m, \Sigma)\right. \\ &\left. + \int \log \frac{d\Phi(\theta; m, \Sigma)}{d\pi(\theta)}\Phi(d\theta; m, \Sigma)\right\}.\end{aligned} \tag{3.1}$$

In [35], 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 notation 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 (3.1), where $\xi \sim \mathcal{N}(0, I_d)$ and

$$f\left((m, C), \xi\right) := -\alpha \log p_{m+C\xi}\left(Y_1^n\right) + \log \frac{\mathrm{d}\Phi_{m,CC^t}}{\mathrm{d}\pi}(m + C\xi). \tag{3.2}$$

In order to be able to state nonasymptotic results on the stochastic gradient algorithms, we restrict the parameter space to an Euclidean ball, that is, (3.1) is transformed into

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

where $\mathbb{B} = \{x \in \mathbb{R}^{d^2+d}, \|x\|_2 \le B\}$ for some $B > 0$. We will then let $\mathcal{P}_{\mathbb{B}}$ denote the orthogonal projection onto $\mathbb{B}$. In addition, we can define the corresponding family of Gaussian distributions

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

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: $x_0$, $X_1^n$, $\gamma_T$
For $t \in \{1, \dots, T\}$,
    a. Sample $\xi_t \sim \mathcal{N}(0, I_d)$
    b. Update $x_t \leftarrow \mathcal{P}_{\mathbb{B}}(x_{t-1} - \gamma_T \nabla f(x_{t-1}, \xi_t))$
End For .
Output: $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$
```

ASSUMPTION 3.2. Assume that f, as defined in (3.2), 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 kth iterate of Algorithm 1, the Gaussian distribution with parameters $\bar{x}_k = (\bar{m}_k, \bar{C}_k)$.

THEOREM 3.2. *Let Assumptions 3.2 and 3.1 be verified, and define ε_n as in Theorem 3.1. Let B be such that $B > \|\theta_0\|_2 + 1/n\sqrt{d}$. 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*

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

REMARK 3.1. In most examples, the gradient is a sum of at least n components. If each term is Lipschitz with constant L_i, an estimate of the constant will be $L \leq n \max_i L_i$. The additional term of the bound is therefore of the order $(2B \max L_i/(nT))^{1/2}/(1-\alpha)$, hence a good choice is $T = O(\sqrt{n})$ to mitigate the impact of the numerical approximation on the rate.

3.2. *Example*: *Logistic regression.* We consider the case of a binary regression model. Although estimation of parameters is relatively simple for small datasets [13], it remains challenging when the size of the dictionary is large. Furthermore, usual deterministic methods do not come with theoretical guarantees as would a gradient descent algorithm for maximum likelihood. Note that the logistic regression is not conjugate in the sense that we cannot find an iterative scheme based on a mean field approximation, as will be done for the matrix completion example in Section 4.

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

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

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

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

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

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

To apply Theorem 3.2, we need to add some constraint on the covariance matrix. The optimization will be written over $\mathbb{B}_\psi := \mathbb{B} \cap \{C \in \mathbb{R}^{d\times d}, CC^t \succeq \psi I_{d\times d}\}$ (this is done only to ensure that $\log|\Sigma|$ has Lipschitz gradients).

COROLLARY 3.4. *Let the family of approximation be any* $\mathcal{F}$ *with* $\mathcal{F}_{\mathrm{id}}^{\Phi} \subset \mathcal{F}$ *and assume that* $K_1 := 2\mathbb{E}\|X_1\|$, $K_2 := 4\mathbb{E}\|X_1\|^2 < \infty$, *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*

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

Note that the only assumption on the distribution of X_1 is that $K_2 < \infty$. Still, it is interesting to compute K_1 and K_2 on some examples. For example, when X_1 is uniform on the unit sphere, $K_1 \leq 2$ and $K_2 \leq 4$. When $X_1 \sim \mathcal{N}(0, s^2 I_d)$, then $K_2 = 4s^2 d$ and $K_1 \leq 2\sqrt{s^2 d}$. In both cases, the terms in K_1 and K_2 do not deteriorate the parametric rate of convergence d/n. Furthermore, the Lipschitz constant can be bounded explicitly under additional assumptions on the design matrix (e.g., bounded singular value) and leads to $L = \mathcal{O}(nd + \frac{d}{\psi})$. Hence taking $\psi = 1/(n\sqrt{d})$, one would get a bound in $\mathcal{O}(\sqrt{\frac{d^{3/2}}{nT}} + (d/n)\log nd)$. We can take T of the order $\frac{n}{d^{1/4}}$ in order not to deteriorate the rate.

4. Application to matrix completion.

4.1. *Context.* Challenging applications such as collaborative filtering made matrix completion one of the most important machine learning problems in the past few years. Let us describe briefly the model: in this case, our parameter θ is a matrix $M \in \mathbb{R}^{m\times p}$, with $m, p \geq 1$. For clarity, we will denote by M_0 the true matrix θ_0 and use M as a notation for a generic parameter instead of θ. Under P_M, the observations are random entries of this matrix with possible noise

$$Y_k = M_{i_k, j_k} + \xi_k \quad \text{for } 1 \leq k \leq n,$$

where the (i_k, j_k) are i.i.d. $\mathcal{U}(\{1, \ldots, m\} \times \{1, \ldots, p\})$. For the sake of simplicit, we will assume that the ξ_k are i.i.d. $\mathcal{N}(0, \sigma^2)$, and that σ^2 is known, so we only have to estimate M. Note that for $\alpha \leq 1$, $\mathcal{D}_\alpha(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)) = \alpha(\mu_1 - \mu_2)^2/(2\sigma^2)$; see (10) page 3800 in [37]. Thus, for $0 < \alpha < 1$,

$$\mathcal{D}_\alpha(P_M, P_N) = \frac{1}{\alpha - 1}\log\left[\frac{1}{mp}\sum_{i=1}^{m}\sum_{j=1}^{p}\exp\left(\frac{\alpha(\alpha-1)(M_{i,j} - N_{i,j})^2}{2\sigma^2}\right)\right],$$

which depends only on α, σ_2 and the matrices M and N so we will use the notation $d_{\alpha,\sigma}(M,N) = \mathcal{D}_\alpha(P_M, P_N)$. In the case $\alpha = 1$,

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

where $\|\cdot\|_F$ denotes the Frobenius norm. In the noiseless case $\sigma^2 = 0$, [10] proved that it is possible to recover exactly M under the assumption that its rank is small enough. Various extensions to noisy settings, approximately low-rank matrices, or other loss functions can be found in [2, 9, 22, 23]. The main message of these 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 [1, 24, 32, 42] using factorized Gaussian priors. Convergence of the posterior mean was proven in [27] for a bounded prior, excluding the Gaussian prior used in practice. Similarly, [34] proves concentration of a truncated version of the posterior. For very large datasets, the MCMC algorithm proposed in [32] is too slow, a VB approximation was proposed in [26] with very good results on the Netflix dataset. This approximation was re-used and extended by many authors including [1, 5, 14, 28, 30]. But the consistency of the Bayesian estimator with Gaussian priors and of its variational approximations are opened questions.

First, we will recall the Gaussian prior [32] and the VB approximation [26]. We will then prove the concentration of the VB approximation, and as a consequence, the concentration of the tempered posterior.

4.2. *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 $\mathrm{rank}(M) \leq K$, 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, $1/\gamma_k \sim \Gamma(a,b)$. These choices ensure conjugacy: put $\gamma = (\gamma_1,\dots,\gamma_K)$, it is then possible to compute the conditional posteriors of $U|V,\gamma$, of $V|U,\gamma$ and $\gamma|U,V$. This allows to use the Gibbs sampler [32]. For large datasets, [26] proposed mean-field VB with $\mathcal{F}$ given by

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

The minimization of the VB program is shown in many cited papers; see [1] and all the references therein. Shortly: 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 $\mathbf{m}_{i,\cdot}$, some $p \times K$ matrix $\mathbf{n}$ whose rows are denoted by $\mathbf{n}_j$, and some vector $\beta = (\beta_1,\dots,\beta_K)$. The parameters are updated iteratively through the formulae:

1. moments of U:

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

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

2. moments of V:

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

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

3. parameter of γ:

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

(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 $\mathcal{W}_j$).

4.3. *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, satisfying the following constraints: $\bar{U}_{i,\ell} = 0$ for $i > r$, $\|\bar{U}\|_\infty := \max_{i,\ell}|\bar{U}_{i,\ell}| \leq B$, and similarly $\bar{V}_{j,\ell} = 0$ for $j > r$ and $\|\bar{V}\|_\infty \leq B$.

THEOREM 4.1. *Fix a as any constant. There is a small enough $b > 0$ such that*

$$\mathbb{E}\left[\int d_{\alpha,\sigma}(M, M_0)\tilde{\pi}_{n,\alpha}(\mathrm{d}M|X_1^n)\right]$$
$$\leq \inf_{1\leq r\leq K}\inf_{(\bar{U},\bar{V})\in\mathcal{M}(r,B)}\left\{\frac{\alpha}{1-\alpha}\frac{[\|M_0 - \bar{U}\bar{V}^t\|_F + \frac{\sqrt{B}}{n}]^2}{2\sigma^2 mp} + \frac{2(1+\alpha)(1+2a)r(m+p)[\log(nmp) + \mathcal{C}(a)]}{n(1-\alpha)}\right\},$$

where the constant $\mathcal{C}(a) = \log(8\sqrt{\pi}\Gamma(a)2^{10a+1}) + 3$. In particular, the result holds for the choice $b = B^2/\{512(nmp)^4[(m \vee p)K]^2\}$.

In practice, it is important that b is small to ensure a good approximation of low-rank matrices [1]. We do not claim that $b = B^2/\{512(nmp)^4[(m \vee p)K]^2\}$ is the optimal value, [1] recommends cross-validation to tune b.

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

$$\mathbb{E}\left[\int d_{\alpha,\sigma}(M, M_0)\tilde{\pi}_{n,\alpha}(\mathrm{d}M|X_1^n)\right] \leq \frac{2(1+\alpha)(1+2a)r(m+p)[\log(nmp) + \mathcal{C}(a) + \frac{\alpha B}{2\sigma^2 mp}]}{n(1-\alpha)}. \tag{4.1}$$

This result is the first consistency result for the VB approximation with Gaussian priors, that is, used in practice. Still, it is stated for a "weak" distance criterion $d_{\alpha,\sigma}(M, M_0)$. Under additional assumptions, it is actually possible to relate this criterion to the standard Frobenius norm. Assume that there is a known C such that $\max_{i,j}|(M_0)_{i,j}| \leq C$. This assumption is satisfied in many applications like collaborative filtering: in the Netflix data, the entries are between 1 and 5. Then it is natural to project any estimator to the set of matrices with bounded entries. Precisely, define for any M the matrix $\mathrm{clip}_C(M)$ its (i, j)th entry: $\min(\max(M_{i,j}, -C), C)$. A simple study of $d_{\alpha,\sigma}$, detailed in the proofs section, leads to the following result.

COROLLARY 4.2. *Under the assumptions of Theorem 4.1, and when in addition* $\max_{i,j}|(M_0)_{i,j}| \leq C$, *then*

$$\mathbb{E}\left[\int \|\mathrm{clip}_C(M) - M_0\|_F^2 \tilde{\pi}_{n,\alpha}(\mathrm{d}M|X_1^n)\right] \\ \leq \frac{8C^2(1+\alpha)(1+2a)r(m+p)[\log(nmp) + \mathcal{C}(a) + \frac{\alpha B}{2\sigma^2 mp}]}{n[1-\exp(2C^2\alpha(\alpha-1)/\sigma^2)]}.$$

Note that once the Gaussian approximation of the posterior is known, it is easy to sample from it and to clip the samples to approximate the posterior mean of $\mathrm{clip}(M)$. So under the boundedness assumption, we have a bound based on the Frobenius norm for an effective procedure based on VB. It is known that for the squared Frobenius norm, the rate $r(m+p)/n$ is minimax optimal—maybe up to log terms [23].

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 4.1.

THEOREM 4.3. *Assume* $M = \bar{U}\bar{V}^t$ *for* $(\bar{U}, \bar{V}) \in \mathcal{M}(r, B)$ *and take* b *as in Theorem 4.1. Then*

$$\mathbb{P}\left[\int d_{\alpha,\sigma}(M, M_0)\tilde{\pi}_{n,\alpha}(\mathrm{d}M|X_1^n) \leq \frac{2(\alpha+1)}{1-\alpha}\varepsilon_n\right] \geq 1 - \frac{2}{n\varepsilon_n},$$

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

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

5. Nonparametric regression estimation.

In this section, we provide a nonparametric example. Thus, the parameter will actually be a function f. We assume that $X_1 = (W_1, Y_1), \ldots, X_n = (W_n, Y_n)$ are i.i.d. from a distribution P_{f_0}, and the model (P_f) is given by $W_i \sim \mathcal{U}([-1, 1])$ and

$$Y_i = f(W_i) + \xi_i,$$

where $\xi_i \sim \mathcal{N}(0, 1)$. We will provide a prior and a mean-field approximation of the posterior. We will show that we estimate the functions f belonging to a Sobolev ellipsoid $\mathcal{W}(r, C^2)$ at the minimax rate of convergence, up to log terms (the definitions of the ellipsoids will be reminded below). The reader might think that this example is not the most striking application of VB. On the other hand, it is an illustration of the generality of our method. We will estimate f using projections on the Fourier basis and the choice of the number of coefficients will be done by model selection. It appears that in this case, model selection can be seen as a variational approximation where the constraint on the posterior is to give all its mass to only one model. This leads to adaptation of the estimator, in the sense that it is not required to know r nor C to compute the estimator.

5.1. *Construction of the prior.*

First, we recall the definition of the trigonometric basis $(\varphi_k)_{k=1}^{\infty}$:

$$\varphi_1(t) = 1, \qquad \varphi_{2k}(t) = \cos(\pi k t), \qquad \varphi_{2k+1}(t) = \sin(\pi k t), \quad k = 1, 2, \ldots.$$

We now define a prior distribution π by describing how to draw from π: we first draw K from a geometric distribution, $\pi(K = k) = 2^{-k}$. We then draw $\beta_1, \ldots, \beta_K$ i.i.d. from a $\mathcal{N}(0, 1)$

distribution. We finally put

$$f(x) = \sum_{k=1}^{K} \beta_k \varphi_k(x).$$

Note that when $f_0(\cdot) = \sum_{k=1}^{\infty} \beta_k^0 \varphi_k(\cdot)$ and all the β_k^0's are nonzero, such a function is never "produced" by the prior. Still, we will see that the prior gives enough mass to functions in the neighborhood of f_0, ensuring consistency.

5.2. *Construction of the variational approximation.* Note that the support of $\pi(\cdot|K)$ has dimension K, but the support of π is infinite-dimensional. Thus, we can expect the support of the tempered posterior $\pi_{n,\alpha}$ to be also infinite-dimensional, and $\pi_{n,\alpha}$ to be intractable. We define a variational approximation that will fix these problems.

First, for $K \geq 1$ define $\mathcal{F}_K$ as the set of probability measures $\rho_{\mathbf{m},s^2}$ where $\mathbf{m} = (m_1, \ldots, m_K)$ on functions $f(\cdot) = \sum_{k=1}^{K} \beta_k \varphi_k(\cdot)$ such that under $\rho_{\mathbf{m},s^2}$, the β_k's are independent and $\beta_k \sim \mathcal{N}(m_k, s^2)$. We put $\mathcal{F} = \bigcup_{k=1}^{\infty} \mathcal{F}_k$. Note that the choice of a constant variance s^2 was motivated by the fact that the estimator of β_k studied, for example, in [36], $\hat{\beta}_k = (1/n)\sum_{i=1}^{n} Y_i \varphi_k(X_i)$, satisfies $\hat{\beta}_k \sim \mathcal{N}(\beta_k, \sigma^2/n)$. Then

$$\begin{aligned}
\tilde{\pi}_{n,\alpha} &= \operatorname*{arg\,min}_{\rho \in \mathcal{F}} \left\{ \alpha \int r_n(f, f_0)\rho(\mathrm{d}f) + \mathcal{K}(\rho, \pi) \right\} \\
&= \operatorname*{arg\,min}_{K \geq 1} \operatorname*{arg\,min}_{\mathbf{m},s^2} \left\{ \alpha \int \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{K} \beta_k \varphi_k(W_i) \right)^2 \Phi(\mathrm{d}\beta; \mathbf{m}, s^2 I) \right. \\
&\qquad \left. + \sum_{k=1}^{K} \frac{1}{2} \left[\log\left(\frac{1}{s^2} \right) + s^2 + m_k^2 - 1 \right] + K \log(2) \right\} \\
&= \operatorname*{arg\,min}_{K \geq 1} \operatorname*{arg\,min}_{\mathbf{m},s^2} \left\{ \frac{\alpha}{2} \sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{K} m_k \varphi_k(W_i) \right)^2 + \frac{s^2 \alpha}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \varphi_k^2(W_i) \right. \\
&\qquad \left. + \sum_{k=1}^{K} \frac{1}{2} \left[\log\left(\frac{1}{s^2} \right) + s^2 + m_k^2 - 1 \right] + K \log(2) \right\}
\end{aligned}$$

(e.g., the approximated posterior mean is simply a ridge regression estimator).

5.3. *Nonparametric rates of convergence.* We remind the definition of the Sobolev ellipsoid given (see, e.g., Chapter 1 in [36]) for $C > 0$ and $r \geq 2$:

$$\mathcal{W}(r, C^2) = \left\{ f \in L_2([-1, 1]) : f = \sum_{k=1}^{\infty} \beta_k \varphi_k \text{ and } \sum_{k=1}^{\infty} k^{2r} \beta_k^2 \leq C^2 \right\}.$$

THEOREM 5.1. *Fix* $\alpha \in (0, 1)$. *Assume that there is an* $r \in [2, \infty[$ *and a* $C > 0$ *such that* $f_0 \in \mathcal{W}(r, C^2)$. *Then*

$$\mathbb{E}\left[\int D_\alpha(P_f, P_{f_0}) \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta | X_1^n) \right] = \mathcal{O}\left(\left(\frac{\log(n)}{n} \right)^{\frac{2r}{2r+1}} \right).$$

The proof is in Section 7. Note that on the contrary to previous sections, we only provide an asymptotic statement here. However, from the proof of Theorem 5.1, it is clear that it is possible to provide a nonasymptotic statement as well (with cumbersome constants).

Here again, note that the distance criterion used in the left-hand side is not standard. We actually have

$$D_\alpha(P_f, P_{f_0}) = \frac{1}{\alpha-1}\log\left[\frac{1}{2}\int_{-1}^{1}\exp\left(\frac{\alpha(\alpha-1)(f(x)-f_0(x))^2}{2}\right)\mathrm{d}x\right].$$

However, when $f_0 \in \mathcal{W}_{r,C^2}$, f is bounded by a constant that depends on r and C^2. If we moreover assume that f_0 is bounded by a known constant c_0, we can as in Section 4 define a clip operator: $\mathrm{clip}_{c_0}(f)(x) = \min(\max(-c_0, f(x)), c_0)$ and obtain

$$\mathbb{E}\left[\int \|\mathrm{clip}_{c_0}(f) - f_0\|_2^2 \tilde{\pi}_{n,\alpha}(\mathrm{d}\theta|X_1^n)\right] = \mathcal{O}\left(\left(\frac{\log(n)}{n}\right)^{\frac{2r}{2r+1}}\right).$$

The rate $1/n^{2r/(2r+1)}$ is known to be minimax optimal on $\mathcal{W}(r, C^2)$ for the squared $\|\cdot\|_2$-norm [36]. The additional log term is sometimes referred to as "the price to pay for adaptation." In the case of the $\|\cdot\|_2$-norm, this is misleading as it is actually possible to build an adaptive estimator that reaches the minimax rate without the additional log, but up to our knowledge this is not possible with a fully Bayesian estimator.

6. Conclusion.

Based on PAC-Bayesian inequalities, we introduced a generic method 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 logistic regression, matrix completion and density estimation. Still, some questions remain open. From a theoretical perspective, the oracle inequality in Theorem 2.7 compares a Rényi divergence to a Kullback–Leibler divergence. It would be very interesting to obtain a result with the Kullback divergence on the left-hand side. This is probably more difficult, if possible at all. We believe that tools from [20] could be of some help, but some work is needed to make explicit assumptions of this paper in our context.

Also, since the first version of this work was submitted, extensions were proven by other authors: [40] extended our results to models with hidden variables, such as mixture models, and [41] proved results in the case $\alpha = 1$ and study many nonparametric examples. Note that while $\alpha = 1$ remains the most popular choice in practice, these results require much stronger assumptions and cannot in general be extended to the misspecified case [19].

An important open issue is the choice of the parameter α. It is clear that our results are not helpful to solve this issue. Some previous work proposes to use cross-validation [4], but this is computationally expensive. Moreover, no theoretical guarantees are known in this case. In the misspecified case, [18] proposed an online adaptive tuning of this parameter. However, it is not clear if this method could work in our context. This should be the object of a future work.

Finally, it would be nice to get rid of the extra log in the rates. Catoni's localization technique [12] is a nice tool to remove extra log factors in PAC-Bayesian bounds, but its adaptation to our setting is not direct. It could be the object of future works.

7. Proofs.

7.1. *Proof of Theorem* 2.1.

We adapt the proof given in [6]. Fix $\alpha \in (0, 1)$, and $\theta \in \Theta$. It is immediate to check that

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

Note that it might be that $D_\alpha(P_\theta^{\otimes n}, P_{\theta_0}^{\otimes n}) = nD_\alpha(P_\theta, P_{\theta_0}) = +\infty$. Rewrite $\pi(\mathrm{d}\theta) = \pi(\mathrm{d}\theta)\mathbf{1}_{\{D_\alpha(P_\theta, P_{\theta_0})<+\infty\}} + \pi(\mathrm{d}\theta)\mathbf{1}_{\{D_\alpha(P_\theta, P_{\theta_0})=+\infty\}} = \pi_1(\mathrm{d}\theta) + \pi_2(\mathrm{d}\theta)$. First, when $\pi = \pi_2$,

π-almost surely, P_θ is singular to P_{θ_0}. But then, π-almost surely, $D_\alpha(P_\theta, P_{\theta_0}) = +\infty$ and $r_n(\theta, \theta_0) = +\infty$. This also holds ρ-almost surely for any $\rho \ll \pi$, and thus the statement of the theorem is trivial: $\mathbb{P}(+\infty \leq +\infty) \geq 1 - \varepsilon$.

Assume now that $\pi \neq \pi_2$. This allows to define the renormalization $\tilde{\pi}(\cdot) = \pi_1(\cdot)/\pi(D_\alpha(P_\theta, P_{\theta_0}) < +\infty)$, that is, a probability measure. On the support of $\tilde{\pi}(\cdot)$,

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

Integrate with respect to $\tilde{\pi}$,

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

and using Fubini's theorem,

$$\mathbb{E}\bigg[\int \exp\big(-\alpha r_n(\theta, \theta_0) + (1-\alpha) n D_\alpha(P_\theta, P_{\theta_0})\big)\tilde{\pi}(\mathrm{d}\theta)\bigg] = 1. \tag{7.1}$$

The key argument here, introduced by [11], is to use Lemma 2.2. Note that almost surely with respect to the sample, we know that

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

satisfies $\int \exp(h)\,\mathrm{d}\tilde{\pi} < \infty$, otherwise, the expectation in (7.1) would be infinite. So, the conditions of Lemma 2.2 are satisfied almost surely with respect to the sample, and we obtain

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

Multiply both sides by ε to get

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

Using Markov's inequality,

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

Taking the complementary event,

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

Now, for a given ρ, it might be that $\int n D_\alpha(P_\theta, P_{\theta_0})\rho(\mathrm{d}\theta) = \infty$ but then, the previous equation implies that $\int r_n(\theta, \theta_0)\rho(\mathrm{d}\theta) + \mathcal{K}(\rho, \tilde{\pi}) = \infty$ and so the statement of the theorem is trivially satisfied as $\infty \leq \infty$. On the other hand, assuming that $\int nD_\alpha(P_\theta, P_{\theta_0})\rho(\mathrm{d}\theta) < \infty$, we rearrange terms to get

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

Now, we decompose $\rho = \rho_1 + \rho_2$ as we decomposed π. First, when $\rho \neq \rho_1$, we have: $\rho(D_\alpha(P_\theta, P_{\theta_0}) = +\infty) > 0$. But then this means that $\rho(r_n(\theta,\theta_0) = +\infty) > 0$, and once again, the statement of the theorem is trivial: $\mathbb{P}(+\infty \leq +\infty) \geq 1-\varepsilon$. So we can assume that $\rho = \rho_1$. But then $\mathcal{K}(\rho,\tilde{\pi}) = \mathcal{K}(\rho,\pi) + \log \pi(D_\alpha(P_\theta, P_{\theta_0}) < +\infty) \leq \mathcal{K}(\rho,\pi)$, thus the statement of the theorem also holds. This completes the proof.

7.2. *Proof of Theorem* 2.4. Fix $\eta \in (0,1)$ and define

$$\rho^* = \operatorname*{arg\,min}_{\rho\in\mathcal{F}}\biggl\{\frac{\alpha}{1-\alpha}\int \frac{\mathbb{E}[r_n(\theta,\theta_0)]}{n}\rho(\mathrm{d}\theta) \\ + \frac{\alpha}{n(1-\alpha)}\sqrt{\frac{\mathrm{Var}[\int r_n(\theta,\theta_0)\rho(\mathrm{d}\theta)]}{\eta}} + \frac{\mathcal{K}(\rho,\pi)}{n(1-\alpha)}\biggr\}.$$

Chebyshev's inequality leads to

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

and so

$$\mathbb{P}\biggl\{\frac{\alpha}{1-\alpha}\int \frac{r_n(\theta,\theta_0)}{n}\rho^*(\mathrm{d}\theta) \geq \frac{\alpha}{1-\alpha}\int \mathcal{K}(P_{\theta_0}, P_\theta)\rho^*(\mathrm{d}\theta) \\ + \frac{\alpha}{1-\alpha}\sqrt{\frac{\int \mathrm{Var}[\log\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}]\rho^*(\mathrm{d}\theta)}{n\eta}} + \frac{\mathcal{K}(\rho^*,\pi)}{n(1-\alpha)}\biggr\} \leq \eta. \tag{7.2}$$

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 $1-\varepsilon-\eta$,

$$\int D_\alpha(P_\theta, P_{\theta_0})\pi_{n,\alpha}(\mathrm{d}\theta|X_1^n) \\ \leq \inf_{\rho\in\mathcal{F}}\biggl\{\frac{\alpha\int r_n(\theta,\theta_0)\rho(\mathrm{d}\theta) + \mathcal{K}(\rho,\pi) + \log(\frac{1}{\varepsilon})}{1-\alpha}\biggr\} \quad \text{by Cor. 2.3} \\ \leq \frac{\alpha\int r_n(\theta,\theta_0)\rho^*(\mathrm{d}\theta) + \mathcal{K}(\rho^*,\pi) + \log(\frac{1}{\varepsilon})}{1-\alpha} \\ \leq \frac{\alpha\int[\mathcal{K}(P_{\theta_0}, P_\theta) + \sqrt{\frac{1}{n\eta}\mathrm{Var}[\int r_n(\theta,\theta_0)\rho^*(\mathrm{d}\theta)]}]}{1-\alpha} \\ + \frac{\mathcal{K}(\rho^*,\pi) + \log(\frac{1}{\varepsilon})}{n(1-\alpha)} \quad \text{by (7.2)}$$

$$
\begin{aligned}
&= \inf_{\rho\in\mathcal{F}}\left\{\frac{\alpha\int[\mathcal{K}(P_{\theta_0},P_\theta)+\sqrt{\frac{1}{n\eta}\mathrm{Var}[\int r_n(\theta,\theta_0)]}]\rho(\mathrm{d}\theta)}{1-\alpha}\right.\\
&\quad\left.+\frac{\mathcal{K}(\rho,\pi)+\log(\frac{1}{\varepsilon})}{n(1-\alpha)}\right\}\quad\text{by definition of } \rho^*\\
&\leq \inf_{\rho\in\mathcal{F}}\left\{\frac{\alpha\int[\mathcal{K}(P_{\theta_0},P_\theta)+\sqrt{\frac{1}{n\eta}\mathbb{E}[\log^2(\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)})]}]\rho(\mathrm{d}\theta)}{1-\alpha}\right.\\
&\quad\left.+\frac{\mathcal{K}(\rho,\pi)+\log(\frac{1}{\varepsilon})}{n(1-\alpha)}\right\}\\
&\leq \frac{\alpha\int[\mathcal{K}(P_{\theta_0},P_\theta)+\sqrt{\frac{1}{n\eta}\mathbb{E}[\log^2(\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)})]}]\rho_n(\mathrm{d}\theta)}{1-\alpha}\\
&\quad+\frac{\mathcal{K}(\rho_n,\pi)+\log(\frac{1}{\varepsilon})}{n(1-\alpha)}\\
&\leq \frac{\alpha(\varepsilon_n+\sqrt{\frac{\varepsilon_n}{n\eta}})}{1-\alpha}+\frac{n\varepsilon_n+\log(\frac{1}{\varepsilon})}{n(1-\alpha)},
\end{aligned}
$$

where in the last step we use the assumptions on ρ_n.

7.3. *Proof of Theorem* 2.6. The beginning is as for Theorem 2.1. Fix $\alpha\in(0,1)$, then

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

Integrate with respect to π,

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

and using Fubini's theorem and Lemma 2.2,

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

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

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

and so as a special case

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

Rearranging terms,

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

and so

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

7.4. *Proof of Theorem* 3.1. We start by defining a sequence $\rho_n(d\theta) := \Phi(d\theta;\theta_0,\sigma_n^2 I) \in \mathcal{F}_\Phi^{\mathrm{id}}$ 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 $\min_{\mathcal{F}^{\mathrm{id}}} \leq \min_{\mathcal{F}^{\mathrm{diag}}} \leq \min_{\mathcal{F}^{\mathrm{full}}}$. Under Assumption 3.1, we can check the hypotheses on the KL between the likelihood terms as required in Theorem 2.4. We have

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

and

$$\mathbb{E}\Big[\log^2\frac{p_{\theta_0}}{p_\theta}(X)\Big] = \mathbb{E}\big[\big(\log p_{\theta_0}(X) - \log p_\theta(X)\big)^2\big] \leq \mathbb{E}\big[M(X)^2\big]\|\theta-\theta_0\|_2.$$

When integrating with respect to ρ_n, we have

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

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

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

To obtain an estimate of the rate ε_n of Theor,em 2.4 we put together those bounds. Choosing $\sigma_n^2 = \frac{1}{n\sqrt{d}}$, we can apply it with

$$\varepsilon_n = \frac{B_1}{n} \vee \frac{B_2}{n^2} \vee \Big\{\frac{d}{n}\Big[\frac{1}{2}\log(\vartheta^2 n^2\sqrt{d}) + \frac{1}{n\vartheta^2}\Big] + \frac{\|\theta_0\|^2}{n\vartheta^2} - \frac{d}{2n}\Big\}.$$

7.5. *Proof of Theorem* 3.2. From the proof of Theorem 2.6, we get

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

By definition of f, we get

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

Following the rest of the proof of 2.6, we get

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

To bound the first term of the right hand-side, we use Assumption 3.1 and the proof of Theorem 3.1. 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

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

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 [33] 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 $(x_1,\dots,x_{t-1})$. By a multiple use of the tower property, we get

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

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

Putting everything together concludes the proof.

7.6. *Proof of Corollary* 3.3. Direct calculation shows that the log likelihood is $2\|X\|$-Lipschitz, hence satisfying Assumption 3.1. We conclude using Theorem 3.1 and the assumption on the design matrix.

7.7. *Proof of Corollary* 3.4. Start by noticing that we can take f as

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

where $\rho(\cdot) = \Phi(\cdot; m, CC^t)$ the likelihood part is convex with Lipschitz gradient as a composition of a convex and gradient Lipschitz function with a affine map. The Lipschitz constant for this term is bounded by $\sum_{i=1}^{n} \|x_i x_i^t\|$. The KL part can be written as $\mathcal{K}(\rho,\pi) = \frac{\|m\|^2}{2\vartheta} + (\frac{1}{2\vartheta}\text{trace}(CC^t) - \log|C|)$, which is convex for positive semidefinite 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 eigenvalues of $\Sigma = CC^t$,

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

To apply Theorem 3.2, we also need to check that the new constraint contains the Gaussian distribution used in the proof. This is the case as long as $\psi \leq \sigma^2 = \frac{1}{n\sqrt{d}}$.

Acknowledgments. We would like to thank Badr-Eddine Chérief-Abdellatif, as well as the Associate Editor and the anonymous referees, for their helpful comments and suggestions on the paper.

The first author was supported by 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).

SUPPLEMENTARY MATERIAL

Supplementary Material to "Concentration of tempered posteriors and of their variational approximations" (DOI: 10.1214/19-AOS1855SUPP; .pdf). The supplementary material contains: the toy example mentioned in Remark 2.1 above; the proofs of Theorems 4.1 and 5.1 and of Corollary 4.2.

REFERENCES

[1] ALQUIER, P., COTTET, V., CHOPIN, N. and ROUSSEAU, J. (2014). Bayesian matrix completion: Prior specification. Preprint. Available at arXiv:1406.1440.

[2] ALQUIER, P., COTTET, V. and LECUÉ, G. (2019). Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *Ann. Statist.* **47** 2117–2144. MR3953446 https://doi.org/10.1214/18-AOS1742

[3] ALQUIER, P. and RIDGWAY, J. (2020). Supplement to "Concentration of tempered posteriors and of their variational approximations." https://doi.org/10.1214/19-AOS1855SUPP.

[4] ALQUIER, P., RIDGWAY, J. and CHOPIN, N. (2016). On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.* **17** Paper No. 239, 41. MR3595173

[5] BABACAN, S. D., LUESSI, M. and MOLINA, R. and KATSAGGELOS, A. K. (2011). Low-rank matrix completion by variational sparse Bayesian learning. In *IEEE International Conference on Audio, Speech and Signal Processing* 2188–2191. Prague.

[6] BHATTACHARYA, A., PATI, D. and YANG, Y. (2019). Bayesian fractional posteriors. *Ann. Statist.* **47** 39–66. MR3909926 https://doi.org/10.1214/18-AOS1712

[7] BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. Springer, New York. MR2247587 https://doi.org/10.1007/978-0-387-45528-0

[8] BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

[9] CANDÈS, E. J. and PLAN, Y. (2010). Matrix completion with noise. *Proc. IEEE* **98** 925–936.

[10] CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56** 2053–2080. MR2723472 https://doi.org/10.1109/TIT.2010.2044061

[11] CATONI, O. (2004). *Statistical Learning Theory and Stochastic Optimization*. *Lecture Notes in Math.* **1851**. Springer, Berlin. MR2163920 https://doi.org/10.1007/b99352

[12] CATONI, O. (2007). *Pac-Bayesian Supervised Classification*: *The Thermodynamics of Statistical Learning*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **56**. IMS, Beachwood, OH. MR2483528

[13] CHOPIN, N. and RIDGWAY, J. (2017). Leave Pima Indians alone: Binary regression as a benchmark for Bayesian computation. *Statist. Sci.* **32** 64–87. MR3634307 https://doi.org/10.1214/16-STS581

[14] COTTET, V. and ALQUIER, P. (2018). 1-bit matrix completion: PAC—Bayesian analysis of a variational approximation. *Mach. Learn.* **107** 579–603. MR3761297 https://doi.org/10.1007/s10994-017-5667-z

[15] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

[16] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

[17] GIBBS, M. N. and MACKAY, D. J. C. (2000). Variational Gaussian process classifiers. *IEEE Trans. Neural Netw. Learn. Syst.* **11** 1458–1464.

[18] GRÜNWALD, P. (2012). The safe Bayesian: Learning the learning rate via the mixability gap. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **7568** 169–183. Springer, Heidelberg. MR3042889 https://doi.org/10.1007/978-3-642-34106-9_16

[19] GRÜNWALD, P. and VAN OMMEN, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **12** 1069–1103. MR3724979 https://doi.org/10.1214/17-BA1085

[20] GRÜNWALD, P. D. and MEHTA, N. A. (2016). Fast rates with unbounded losses. Preprint. Available at arXiv:1605.00252.

[21] HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.* **14** 1303–1347. MR3081926

[22] KLOPP, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20** 282–303. MR3160583 https://doi.org/10.3150/12-BEJ486

[23] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869 https://doi.org/10.1214/11-AOS894

[24] LAWRENCE, N. D. and URTASUN, R. (2009). Non-linear matrix factorization with Gaussian processes. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 601–608. ACM, New York.

[25] LI, X. and ZHENG, Y. (2009). Patch-based video processing: A variational Bayesian approach. *IEEE Transactions on Circuits and Systems for Video Technology* **19** 27–40.

[26] LIM, Y. J. and TEH, Y. W. (2007). Variational Bayesian approach to movie rating prediction. *In Proceedings of KDD Cup and Workshop* **7** 15–21.

[27] MAI, T. T. and ALQUIER, P. (2015). A Bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electron. J. Stat.* **9** 823–841. MR3331862 https://doi.org/10.1214/15-EJS1020

[28] MARSDEN, A. and BACALLADO, S. (2017). Sequential matrix completion. Preprint. Available at arXiv:1710.08045.

[29] OPPER, M. and ARCHAMBEAU, C. (2009). The variational Gaussian approximation revisited. *Neural Comput.* **21** 786–792. MR2478318 https://doi.org/10.1162/neco.2008.08-07-592

[30] PAISLEY, J. and CARIN, L. (2010). A nonparametric Bayesian model for kernel matrix completion. In *Proceedings of ICASSP* 2010, Dallas, USA.

[31] ROUSSEAU, J. (2016). On the frequentist properties of Bayesian nonparametric methods. *Annual Review of Statistics and Its Application* **3** 211–231.

[32] SALAKHUTDINOV, R. and MNIH, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the* 25*th International Conference on Machine Learning* 880–887. ACM, New York.

[33] SHALEV-SHWARTZ, S. (2012). Online learning and online convex optimization. *Found. Trends Mach. Learn.* **4** 107–194.

[34] SUZUKI, T. (2015). Convergence rate of Bayesian tensor estimator and its minimax optimality. In *ICML* 1273–1282.

[35] TITSIAS, M. and LÁZARO-GREDILLA, M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *Proceedings of the* 31*st International Conference on Machine Learning* (*ICML*-14) 1971–1979.

[36] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

[37] VAN ERVEN, T. and HARREMOËS, P. (2014). Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. Inform. Theory* **60** 3797–3820. MR3225930 https://doi.org/10.1109/TIT.2014.2320500

[38] WANG, B. and TITTERINGTON, D. M. (2004). Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In *Proceedings of the* 20*th Conference on Uncertainty in Artificial Intelligence* 577–584. AUAI Press.

[39] WANG, Y. and BLEI, D. M. Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc.* To appear. https://doi.org/10.1080/01621459.2018.1473776.

[40] YANG, Y., PATI, D. and BHATTACHARYA, A. In α-variational inference with statistical guarantees. *Ann. Statist.* To appear.

[41] ZHANG, F. and GAO, C. (2018). Convergence rates of variational posterior distributions. Preprint. Available at arXiv:1712.02519.

[42] ZHOU, M., WANG, C., CHEN, M., PAISLEY, J., DUNSON, D. and CARIN, L. (2010). Nonparametric Bayesian matrix completion. In *Proc. IEEE SAM*.