The Annals of Statistics
2020, Vol. 48, No. 4, 2180–2207
https://doi.org/10.1214/19-AOS1883

CONVERGENCE RATES OF VARIATIONAL POSTERIOR DISTRIBUTIONS

BY FENGSHUO ZHANG* AND CHAO GAO†

*Department of Statistics, University of Chicago, *fengshuo@galton.uchicago.edu; †chaogao@galton.uchicago.edu*

We study convergence rates of variational posterior distributions for nonparametric and high-dimensional inference. We formulate general conditions on prior, likelihood and variational class that characterize the convergence rates. Under similar "prior mass and testing" conditions considered in the literature, the rate is found to be the sum of two terms. The first term stands for the convergence rate of the true posterior distribution, and the second term is contributed by the variational approximation error. For a class of priors that admit the structure of a mixture of product measures, we propose a novel prior mass condition, under which the variational approximation error of the mean-field class is dominated by convergence rate of the true posterior. We demonstrate the applicability of our general results for various models, prior distributions and variational classes by deriving convergence rates of the corresponding variational posteriors.

1. Introduction. Variational Bayes inference is a popular technique to approximate difficult-to-compute probability posterior distributions. Given a posterior distribution $\Pi(\cdot|X^{(n)})$, and a variational family $\mathcal{S}$, variational Bayes inference seeks a $\widehat{Q} \in \mathcal{S}$ that best approximates $\Pi(\cdot|X^{(n)})$ under the Kullback–Leibler divergence. Though it is not exact Bayes inference, the variational class $\mathcal{S}$ often gives computational advantage and leads to algorithms such as coordinate ascent that can be efficiently implemented on large-scale data sets. Researchers in many fields have used variational Bayes inference to solve real problems. Successful examples include statistical genetics [8, 26], natural language processing [7, 22], computer vision [31] and network analysis [4, 38] to name a few. We refer the readers to an excellent recent review [6] on this topic.

The goal of this paper is to study the variational posterior distribution $\widehat{Q}$ from a theoretic perspective. We propose general conditions on the prior, the likelihood and the variational class to characterize the convergence rate of the variational posterior to the true data generating process.

Before discussing our results, we give a brief review on the theory of convergence rates of the posterior distributions in the literature. In order that the posterior distribution concentrates around the true parameter with some rate, the "prior mass and testing" framework requires three conditions on the prior and the likelihood: (a) The prior is required to put a minimal amount of mass in a neighborhood of the true parameter; (b) Restricted to a subset of the parameter space, there exists a testing function that can distinguish the truth from the complement of its neighborhood; (c) The prior is essentially supported on the subset described above. Rigorous statements of these three conditions can be found in seminal papers [17, 18, 30]. Earlier versions of these conditions go back to [2, 3, 23, 29]. We also mention another line of work [10, 19, 35, 40] that established posterior rates of convergence using other approaches.

Received February 2018; revised June 2019.
MSC2020 subject classifications. Primary 62C10; secondary 62F15.
Key words and phrases. Posterior contraction, mean-field variational inference, density estimation, Gaussian sequence model, piecewise constant model, empirical Bayes.

In this paper, we show that under almost the same three conditions, the variational posterior $\widehat{Q}$ also converges to the true parameter, and the rate of convergence is given by

$$\epsilon_n^2 + \frac{1}{n} \inf_{Q \in \mathcal{S}} P_0^{(n)} D\big(Q \| \Pi(\cdot | X^{(n)})\big). \tag{1}$$

The first term ϵ_n^2 is the rate of convergence of the posterior distribution $\Pi(\cdot|X^{(n)})$. The second term is the variational approximation error with respect to the class $\mathcal{S}$ under the data generating process $P_0^{(n)}$. Since we are able to generalize the "prior mass and testing" theory with the same old conditions, many well-studied problems in the literature can now be revisited under our framework of variational Bayes inference with very similar proof techniques. This will be illustrated with several examples considered in the paper.

Remarkably, for a special class of prior distributions and a corresponding variational class, the second term of (1) will be automatically dominated by ϵ_n^2 under a modified "prior mass" condition. We illustrate this result by a prior distribution of product measure

$$d\Pi(\theta) = \prod_j d\Pi_j(\theta_j),$$

and a mean-field variational class

$$\mathcal{S}_{\mathrm{MF}} = \left\{ Q : dQ(\theta) = \prod_j dQ_j(\theta_j) \right\}.$$

As long as there exists a subset $\bigotimes_j \widetilde{\Theta}_j \subset \{\theta : D_\rho(P_0^{(n)} \| P_\theta^{(n)}) \le C_1 n\epsilon_n^2\}$, such that the prior mass condition

$$\Pi\left(\bigotimes_j \widetilde{\Theta}_j \right) \ge \exp(-C_2 n\epsilon_n^2) \tag{2}$$

holds together with the testing conditions, then the variational posterior distribution $\widehat{Q}$ converges to the true parameter with the rate ϵ_n^2. In other words, the variational approximation error term in (1) is dominated under this stronger prior mass condition (2). This is the result of Theorem 2.4. Here, $D_\rho(\cdot\|\cdot)$ stands for a Rényi divergence with some $\rho > 1$. The implication of the condition (2) is important. It says that as long as the prior satisfies a "prior mass" condition that is coherent with the structure of the variational class, the resulted variational approximation error will always be small compared with the statistical error from the true posterior. Therefore, the condition (2) offers a practical guidance on how to choose a good prior for variational Bayes inference. In addition, as a condition only on the prior mass, (2) is usually very easy to check. This mathematical simplicity is not just for independent priors and the mean-field class. In Section 4, a more general condition is proposed that includes the setting of (2) as a special case.

Besides the general formulation of conditions to ensure convergence of the variational posteriors, several interesting aspects of variational Bayes inference are also discussed in the paper. We show that for a general likelihood with a sieve prior, its mean-field variational approximation of the posterior distribution has an interesting relation to an empirical Bayes procedure. We also show that the empirical Bayes procedure is exactly a variational Bayes procedure using a specially designed variational class. This connection between empirical Bayes and variational Bayes is interesting, and may suggest similar theoretical properties of the two.

Finally, we would like to remark that the general rate (1) for variational posteriors is only an upper bound. It is *not* always true that the variational posterior has a slower convergence rate than the true posterior. Sometimes the variational posterior may not be a good approximation to the true posterior, but it can still contract faster to the true parameter if additional regularity is imposed by the variational class $\mathcal{S}$. We construct examples in Section 5.2 to illustrate this point.

Related work. Statistical properties of variational posterior distributions have also been studied in the literature. A recent work by [36] established Bernstein–von Mises type of results for parametric models. We refer the readers to [6, 36] for other related references on theories for parametric variational Bayes inference. For nonparametric and high-dimensional models, recent work by [1, 37] studied variational approximation to tempered posteriors, where the likelihood $dP_\theta^{(n)}/dP_0^{(n)}$ is replaced by $(dP_\theta^{(n)}/dP_0^{(n)})^\alpha$ for some $\alpha \in (0,1)$. Just as the convergence of tempered posteriors [34], the convergence of the variational approximation can also be established under generalizations of the prior mass condition. In addition, the paper [1] also studied convergence rates under model misspecification, and the paper [37] considered a more general setting that can handle latent variables, which is quite useful to analyze mixture models. We would like to point out that these results do not apply to the usual posterior distributions with $\alpha = 1$. After the first version of our paper was posted, similar results on $\alpha = 1$ have also been obtained independently by [25].[1] An early related work on this topic is by [40], where the results cover both posterior distributions and their variational approximations. However, the conditions in [40] are rather abstract and are not easy to check in applications.

Organization. The rest of the paper is organized as follows. In Section 2, we formulate the problem and introduce the general conditions that characterize convergence rates of variational posteriors. This section also includes results for the mean-field variational class, where the variational approximation error can be explicitly analyzed. In Section 3, we apply our general theory to three examples that use three different variational classes. Then, in Section 4, for a general class of prior distributions and a mean-field class under a model selection setting, we propose a new prior mass condition that leads to an automatic control of the variational approximation error. In Section 5, we discuss the relation between variational Bayes and empirical Bayes. We also discuss possible situations where the variational posterior outperforms the true posterior in this section. An extension of the main results under model misspecification is also discussed in Section 5. All the proofs will be given in the Supplementary Materials [39].

Notation. We close this section by introducing notation that will be used later. For $a, b \in \mathbb{R}$, let $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. For a positive real number x, $\lceil x \rceil$ is the smallest integer no smaller than x and $\lfloor x \rfloor$ is the largest integer no larger than x. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if $a_n \le C b_n$ for all n with some constant $C > 0$ that does not depend on n. The relation $a_n \asymp b_n$ holds if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold. For an integer m, $[m]$ denotes the set $\{1, 2, \ldots, m\}$. Given a set S, $|S|$ denotes its cardinality, and $\mathbf{1}_S$ is the associated indicator function. The ℓ_p norm of a vector $v \in \mathbb{R}^m$ with $1 \le m \le \infty$ is defined as $\|v\|_p = (\sum_{j=1}^m |v_j|^p)^{1/p}$ for $1 \le p < \infty$ and $\|v\|_\infty = \sup_{1 \le k \le m} |v_k|$. Moreover, we use $\|v\|$ to denote the ℓ_2 norm $\|v\|_2$ by convention. For any function f, the ℓ_p norm is defined in a similar way, that is, $\|f\|_p = (\int f(x)^p \, dx)^{1/p}$. Specifically, $\|f\|_\infty = \sup_x |f(x)|$. We use $\mathbb{P}$ and $\mathbb{E}$ to denote generic probability and expectation whose distribution is determined from the context. The notation $\mathbb{P}f$ also means expectation of f under $\mathbb{P}$ so that $\mathbb{P}f = \int f \, d\mathbb{P}$. Throughout the paper, C, c and their variants denote generic constants that do not depend on n. Their values may change from line to line.

[1] Some extensions of the results of [25] were later added in the revised version of [37] by the same authors.

2. Main results.

2.1. *Definitions and settings*. We start this section by introducing a class of divergence functions.

DEFINITION 2.1 (Rényi divergence). Let $\rho > 0$ and $\rho \neq 1$. The ρ-Rényi divergence between two probability measures P_1 and P_2 is defined as

$$D_\rho(P_1 \| P_2) = \begin{cases} \dfrac{1}{\rho - 1} \log \displaystyle\int \left(\frac{dP_1}{dP_2}\right)^{\rho-1} dP_1 & \text{if } P_1 \ll P_2, \\ +\infty & \text{otherwise.} \end{cases}$$

The relations between the Rényi divergence and other divergence functions are summarized below:

1. When $\rho \to 1$, the Rényi divergence converges to the Kullback–Leibler divergence, defined as

$$D_1(P_1 \| P_2) = \begin{cases} \displaystyle\int \log\left(\frac{dP_1}{dP_2}\right) dP_1 & \text{if } P_1 \ll P_2, \\ +\infty & \text{otherwise.} \end{cases}$$

From now on, we use $D(P_1 \| P_2)$ without the subscript to denote $D_1(P_1 \| P_2)$.

2. When $\rho = 1/2$, the Rényi divergence is related to the Hellinger distance by

$$D_{1/2}(P_1 \| P_2) = -2 \log\bigl(1 - H(P_1, P_2)^2\bigr),$$

and the Hellinger distance is defined as

$$H(P_1, P_2) = \sqrt{\frac{1}{2} \int (\sqrt{dP_1} - \sqrt{dP_2})^2}.$$

3. When $\rho = 2$, the Rényi divergence is related to the χ^2-divergence by

$$D_2(P_1 \| P_2) = \log\bigl(1 + \chi^2(P_1 \| P_2)\bigr),$$

and the χ^2-divergence is defined as

$$\chi^2(P_1 \| P_2) = \int \frac{(dP_1)^2}{dP_2} - 1.$$

DEFINITION 2.2 (total variation). The total variation distance between two probability measures P_1 and P_2 is defined as

$$TV(P_1, P_2) = \frac{1}{2} \int |dP_1 - dP_2|.$$

The relation among the divergence functions defined above is given by the following proposition (see [32]).

PROPOSITION 2.1. *With the above definitions, the following inequalities hold*:

$$\begin{aligned} TV(P_1, P_2)^2 &\le 2H(P_1, P_2)^2 \le D_{1/2}(P_1 \| P_2) \\ &\le D(P_1 \| P_2) \le D_2(P_1 \| P_2) \le \chi^2(P_1 \| P_2). \end{aligned}$$

Moreover, the Rényi divergence $D_\rho(P_1 \| P_2)$ is a nondecreasing function of ρ.

Now we are ready to introduce the variational posterior distribution. Given a statistical model $P_\theta^{(n)}$ parametrized by θ, and a prior distribution $\theta \sim \Pi$, the posterior distribution is defined by

$$d\Pi\big(\theta|X^{(n)}\big) = \frac{dP_\theta^{(n)}(X^{(n)})\,d\Pi(\theta)}{\int dP_\theta^{(n)}(X^{(n)})\,d\Pi(\theta)}.$$

To address possible computational difficulty of the posterior distribution, variational approximation is a way to find the closest object in a class $\mathcal{S}$ of probability measures to $\Pi(\cdot|X^{(n)})$. The class $\mathcal{S}$ is usually required to be computationally or analytically tractable. The most popular mathematical definition of variational approximation is given through the KL-divergence.

DEFINITION 2.3 (Variational posterior). Let $\mathcal{S}$ be a family of distributions. The variational approximation of the posterior is defined as

$$\widehat{Q} = \operatorname*{argmin}_{Q\in\mathcal{S}} D\big(Q\|\Pi(\cdot|X^{(n)})\big). \tag{3}$$

Just like the posterior distribution $\Pi(\cdot|X^{(n)})$, the variational posterior $\widehat{Q}$ is a data-dependent measure that summarizes information from both the prior and the data. For a variational set $\mathcal{S}$, the corresponding variational posterior can be regarded as the projection of the true posterior onto $\mathcal{S}$ under KL-divergence. When $\mathcal{S}$ is the set of all distributions, $\widehat{Q}$ turns out to be the true posterior $\Pi(\cdot|X^{(n)})$. The choice of the class $\mathcal{S}$ usually determines the difficulty of the optimization (3). In this paper, our main goal is to study the statistical property of the data-dependent measure $\widehat{Q}$ for a general $\mathcal{S}$.

2.2. *Results for general variational posteriors.* Assume the observation $X^{(n)}$ is generated from a probability measure $P_0^{(n)}$, and $\widehat{Q}$ is the variational posterior distribution driven by $X^{(n)}$. The goal of this paper is to analyze $\widehat{Q}$ from a frequentist perspective. In other words, we study statistical properties of $\widehat{Q}$ under $P_0^{(n)}$. The first theorem gives conditions that guarantee convergence of the variational posterior $\widehat{Q}$.

THEOREM 2.1. *Suppose ϵ_n is a sequence that satisfies $n\epsilon_n^2 \geq 1$. Consider a loss function $L(\cdot,\cdot)$, such that for any two probability measures P_1 and P_2, $L(P_1,P_2)\geq 0$. Let $C, C_1, C_2, C_3 > 0$ be constants such that $C > C_2 + C_3 + 2$. We assume*:

- *For any $\epsilon > \epsilon_n$, there exists a set $\Theta_n(\epsilon)$ and a testing function ϕ_n, such that*

$$P_0^{(n)}\phi_n + \sup_{\substack{\theta\in\Theta_n(\epsilon)\\ L(P_\theta^{(n)},P_0^{(n)})\geq C_1 n\epsilon^2}} P_\theta^{(n)}(1-\phi_n) \leq \exp\big(-Cn\epsilon^2\big). \tag{C1}$$

- *For any $\epsilon > \epsilon_n$, the set $\Theta_n(\epsilon)$ above satisfies*

$$\Pi\big(\Theta_n(\epsilon)^c\big) \leq \exp\big(-Cn\epsilon^2\big). \tag{C2}$$

- *For some constant $\rho > 1$,*

$$\Pi\big(D_\rho\big(P_0^{(n)}\|P_\theta^{(n)}\big) \leq C_3 n\epsilon_n^2\big) \geq \exp\big(-C_2 n\epsilon_n^2\big). \tag{C3}$$

Then for the variational posterior $\widehat{Q}$ defined in (3), we have

$$P_0^{(n)}\widehat{Q}L\big(P_\theta^{(n)}, P_0^{(n)}\big) \leq Mn\big(\epsilon_n^2 + \gamma_n^2\big), \tag{4}$$

for some constant M only depending on C_1, C and ρ, where the quantity γ_n^2 is defined as

$$\gamma_n^2 = \frac{1}{n} \inf_{Q \in \mathcal{S}} P_0^{(n)} D(Q \| \Pi(\cdot | X^{(n)})).$$

Conditions (C1)–(C3) resemble the three conditions of "prior mass and testing" in [17]. Interestingly, Theorem 2.1 shows that with a slight modification, these three conditions also lead to the convergence of the variational posterior. The testing conditions (C1) and (C2) are required to hold for all $\epsilon > \epsilon_n$. In the prior mass condition (C3), the neighborhood of $P_0^{(n)}$ is defined through a Rényi divergence with a $\rho > 1$, compared with the KL-divergence used in [17]. According to Proposition 2.1, $D_\rho(P_1 \| P_2) \geq D(P_1 \| P_2)$ for $\rho > 1$, so the condition (C3) in our paper is slightly stronger than that in [17]. This stronger "prior mass" condition ensures that the loss $L(P_\theta^{(n)}, P_0^{(n)})$ is exponentially integrable under the true posterior $\Pi(\cdot | X^{(n)})$, which is a key step in the proof of Theorem 2.1. In all the examples considered in this paper, we will check (C3) with $D_2(P_0^{(n)} \| P_\theta^{(n)})$, which turns out to be a very convenient choice.

The convergence rate is the sum of two terms, ϵ_n^2 and γ_n^2. The first term ϵ_n^2 is the convergence rate of the true posterior $\Pi(\cdot | X^{(n)})$. The second term γ_n^2 characterizes the approximation error given by the variational set $\mathcal{S}$. A larger $\mathcal{S}$ means more expressive power given by the variational approximation, and thus the rate of γ_n^2 is smaller.

It is worth mentioning that we characterize the convergence of the variational posterior $\widehat{Q}$ through the expected loss $P_0^{(n)} \widehat{Q} L(P_\theta^{(n)}, P_0^{(n)})$. Bounds for this quantity are also obtained by [25] independently with a stronger testing condition on the entire space. We remark that convergence in $P_0^{(n)} \widehat{Q} L(P_\theta^{(n)}, P_0^{(n)})$ automatically implies that the entire variational posterior distribution concentrates in a neighborhood of the true distribution $P_0^{(n)}$ with a radius of the same rate. When the loss function is convex, it also implies the existence of a point estimator that enjoys the same convergence rate. We summarize these results in the next corollary.

COROLLARY 2.1. *Under the same setting of Theorem 2.1, for any diverging sequence $M_n \to \infty$, we have*

$$P_0^{(n)} \widehat{Q}\big(L\big(P_\theta^{(n)}, P_0^{(n)}\big) > M_n n\big(\epsilon_n^2 + \gamma_n^2\big)\big) \to 0.$$

Furthermore, if the loss $L(P_\theta^{(n)}, P_0^{(n)})$ is convex respect to θ, then the variational posterior mean $\widehat{\theta} = \widehat{Q}\theta$ satisfies

$$P_0^{(n)} L\big(P_{\widehat{\theta}}^{(n)}, P_0^{(n)}\big) \leq M n\big(\epsilon_n^2 + \gamma_n^2\big),$$

where M is the same constant in (4).

PROOF. The first result is an application of Markov's inequality

$$P_0^{(n)} \widehat{Q}\big(L\big(P_\theta^{(n)}, P_0^{(n)}\big) > M_n n\big(\epsilon_n^2 + \gamma_n^2\big)\big) \leq \frac{P_0^{(n)} \widehat{Q} L(P_\theta^{(n)}, P_0^{(n)})}{M_n n(\epsilon_n^2 + \gamma_n^2)} \leq \frac{M}{M_n} \to 0.$$

The second result is directly implied by Jensen's inequality that

$$P_0^{(n)} L\big(P_{\widehat{Q}\theta}^{(n)}, P_0^{(n)}\big) \leq P_0^{(n)} \widehat{Q} L\big(P_\theta^{(n)}, P_0^{(n)}\big) \leq M n\big(\epsilon_n^2 + \gamma_n^2\big). \qquad \square$$

To apply Theorem 2.1 to specific problems, we need to analyze the variational approximation error $\gamma_n^2 = \frac{1}{n} \inf_{Q \in \mathcal{S}} P_0^{(n)} D(Q || \Pi(\cdot | X^{(n)}))$ in each individual setting. However, this

task may not be trivial for many problems. Now we borrow a technique in [40] to get a useful upper bound for γ_n^2. For any $Q \in \mathcal{S}$, we have

$$\begin{aligned} n\gamma_n^2 &\le P_0^{(n)} D(Q\|\Pi(\cdot|X^{(n)})) = D(Q\|\Pi) + Q\bigg[\int \log\bigg(\frac{dP_\Pi^{(n)}}{dP_\theta}\bigg) dP_0^{(n)}\bigg] \\ &= D(Q\|\Pi) + Q\big[D\big(P_0^{(n)}\|P_\theta^{(n)}\big) - D\big(P_0^{(n)}\|P_\Pi^{(n)}\big)\big] \\ &\le D(Q\|\Pi) + Q\big[D\big(P_0^{(n)}\|P_\theta^{(n)}\big)\big], \end{aligned}$$

where $P_\Pi^{(n)} = \int P_\theta^{(n)}\, d\Pi(\theta)$. Then we obtain the upper bound

$$\gamma_n^2 \le \inf_{Q\in\mathcal{S}} R(Q),$$

where

$$R(Q) = \frac{1}{n}\big(D(Q\|\Pi) + Q\big[D\big(P_0^{(n)}\|P_\theta^{(n)}\big)\big]\big). \tag{5}$$

Now, it is easy to see that a sufficient condition for the variational posterior to converge at the same rate as the true posterior is

$$\inf_{Q\in\mathcal{S}} R(Q) \lesssim \epsilon_n^2. \tag{C4}$$

We incorporate this condition into the next theorem.

THEOREM 2.2. *Suppose ϵ_n is a sequence that satisfies $n\epsilon_n^2 \ge 1$, for which the conditions* (C1), (C2), (C3), (C4) *hold. Then, for the variational posterior $\widehat{Q}$ that is defined in* (3), *we have*

$$P_0^{(n)} \widehat{Q} L\big(P_\theta^{(n)}, P_0^{(n)}\big) \lesssim n\epsilon_n^2. \tag{6}$$

We would like to remark that the quantity $\inf_{Q\in\mathcal{S}} R(Q)$ is easier to analyze compared with the original definition of γ_n^2. According to its definition given by (5), it is sufficient to find a distribution $Q \in \mathcal{S}$, such that

$$D(Q\|\Pi) \lesssim n\epsilon_n^2 \quad \text{and} \quad Q\big[D\big(P_0^{(n)}\|P_\theta^{(n)}\big)\big] \lesssim n\epsilon_n^2. \tag{7}$$

These are exactly the two conditions formulated by [1] as a natural extension of the prior mass condition. The relation between the prior mass condition and (7) has also been discussed in [37].

One way to construct such a distribution Q that satisfies the above two inequalities is to focus on those whose supports are within the set $\mathcal{C} = \{\theta : D(P_0^{(n)}\|P_\theta^{(n)}) \le Cn\epsilon_n^2\}$ for some constant $C > 0$. We summarize this method into the following theorem.

THEOREM 2.3. *Suppose there exist constants $C_1, C_2 > 0$, such that*

$$\inf_{Q\in\mathcal{S}\cap\mathcal{E}} D(Q\|\Pi) \le C_1 n\epsilon_n^2, \tag{C4*}$$

where $\mathcal{E} = \{Q : \operatorname{supp}(Q) \subset \mathcal{C}\}$ with $\mathcal{C} = \{\theta : D(P_0^{(n)}\|P_\theta^{(n)}) \le C_2 n\epsilon_n^2\}$. Then we have

$$\inf_{Q\in\mathcal{S}} R(Q) \le (C_1 + C_2)\epsilon_n^2.$$

2.3. *Results for mean-field variational posteriors.* A special choice of $\mathcal{S}$ is the mean-field class of distributions. Not only does this class lead to computationally efficient algorithms such as coordinate ascent, but in this section, we will also show that the structure of this class leads to a convenient convergence analysis. We begin with its definition.

DEFINITION 2.4 (Mean-field class). For parameters in a product space that can be written as $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ with some $1 \le m \le \infty$, the mean-field variational family is defined as

$$\mathcal{S}_{\mathrm{MF}} = \left\{ Q : dQ(\boldsymbol{\theta}) = \prod_{j=1}^{m} dQ_j(\theta_j) \right\}.$$

The following theorem can be viewed as an application of Theorem 2.3 to the mean-field class.

THEOREM 2.4. *Suppose there exists a* $\widetilde{Q} \in \mathcal{S}_{\mathrm{MF}}$ *and a subset* $\bigotimes_{j=1}^{m} \widetilde{\Theta}_j$, *such that*

$$\bigotimes_{j=1}^{m} \widetilde{\Theta}_j \subset \left\{ \theta : D\big(P_0^{(n)} \| P_\theta^{(n)}\big) \le C_1 n\epsilon_n^2, \log \frac{d\widetilde{Q}(\theta)}{d\Pi(\theta)} \le C_2 n\epsilon_n^2 \right\} \tag{8}$$

and

$$-\sum_{j=1}^{m} \log \widetilde{Q}_j(\widetilde{\Theta}_j) \le C_3 n\epsilon_n^2, \tag{9}$$

for some constants $C_1, C_2, C_3 > 0$. *Then we have*

$$\inf_{Q \in \mathcal{S}_{\mathrm{MF}}} R(Q) \le (C_1 + C_2 + C_3)\epsilon_n^2.$$

Note that the condition (9) can also be written as

$$\widetilde{Q}\left(\bigotimes_{j=1}^{m} \widetilde{\Theta}_j \right) \ge \exp\big(-C_3 n\epsilon_n^2\big).$$

In other words, Theorem 2.4 gives an interesting "distribution mass" type of characterization for $\inf_{Q \in \mathcal{S}} R(Q)$. Checking (9) is very similar to checking the "prior mass" condition (C3), and is usually not hard in many examples. We only need to make sure that $\widetilde{Q}$ is not too far away from the prior Π in the sense of (8). In fact, if the prior Π belongs to the class $\mathcal{S}_{\mathrm{MF}}$, then one can take $\widetilde{Q} = \Pi$, and the conditions of Theorem 2.4 simply become a "prior mass" condition $\Pi(\bigotimes_{j=1}^{m} \widetilde{\Theta}_j) \ge \exp(-C_3 n\epsilon_n^2)$, with the choice of $\bigotimes_{j=1}^{m} \widetilde{\Theta}_j$ being a subset of the KL-neighborhood $\{\theta : D(P_0^{(n)} \| P_\theta^{(n)}) \le C_1 n\epsilon_n^2\}$. A more general characterization of the variational approximation error under model selection setting through a prior mass condition will be studied in Section 4.

3. Applications. In this section, we consider several examples to illustrate the theory developed in Section 2.

3.1. *Gaussian sequence model.* Consider observations generated by a Gaussian sequence model,

$$Y_j = \theta_j + \frac{1}{\sqrt{n}} Z_j, \quad Z_j \overset{i.i.d}{\sim} N(0,1),\ j \ge 1. \tag{10}$$

We use the notation $P_\theta^{(n)} = \bigotimes_j N(\theta_j, n^{-1})$ for the distribution above. Our goal is to use variational Bayes methods to estimate the true parameter θ^* that belongs to the following Sobolev ball:

$$\Theta_\alpha(B) = \left\{\theta = (\theta_j)_{j=1}^\infty : \sum_{j=1}^\infty j^{2\alpha}\theta_j^2 \le B^2\right\}. \tag{11}$$

Here, the smoothness $\alpha > 0$ and the radius $B > 0$ are considered as constants throughout the paper. The loss function for this problem is $L(P_\theta^{(n)}, P_{\theta^*}^{(n)}) = n\|\theta - \theta^*\|^2$, which is a natural choice for the Gaussian sequence model.

The prior distribution $\theta \sim \Pi$ is described through the following sampling process:

1. Sample $k \sim \pi$;
2. Conditioning on k, sample $\theta_j \sim f_j$ for all $j \in [k]$, and set $\theta_j = 0$ for all $j > k$.

In other words, the prior on θ is a mixture of product measures,

$$d\Pi(\theta) = \sum_{k=1}^\infty \pi(k) \prod_{j=1}^k f_j(\theta_j) \prod_{j>k} \delta_0(\theta_j)\, d\theta. \tag{12}$$

Priors of similar forms are also considered in [15, 16, 27, 28]. Direct calculation implies that the posterior is also in the form of a mixture of product measures.

Consider the variational posterior $\widehat{Q}$ defined by (3) with $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}$. That is, we seek a data-dependent measure in a more tractable form of a product measure. In most cases, the variational posterior does not have a closed form and needs to be solved by coordinate ascent algorithms [6]. However, for the Gaussian sequence model (10) with the prior distribution (12), one can write down the exact form of the mean-field variational posterior distribution.

THEOREM 3.1. *Consider the variational posterior $\widehat{Q}$ induced by the likelihood* (10), *the prior* (12) *and the mean-field variational set $\mathcal{S}_{\mathrm{MF}}$. The distribution $\widehat{Q}$ is a product measure with the density of each coordinate specified by*

$$q_j = \begin{cases} \widetilde{f}_j & j < \widetilde{k}, \\ \widetilde{p}\delta_0 + (1-\widetilde{p})\widetilde{f}_{\widetilde{k}} & j = \widetilde{k}, \\ \delta_0 & j > \widetilde{k}. \end{cases} \tag{13}$$

where

$$\widetilde{f}_j(\theta_j) \propto f_j(\theta_j)\exp\left(-\frac{n}{2}(\theta_j - Y_j)^2\right),$$

$$\widetilde{p} = \frac{\pi(k-1|Y)}{\pi(k-1|Y) + \pi(k|Y)}$$

and

$$\widetilde{k} = \operatorname*{argmax}_k \big(\pi(k-1|Y) + \pi(k|Y)\big). \tag{14}$$

The number $\pi(k|Y)$ is the posterior probability of the model dimension, and according to Bayes formula, it is

$$\pi(k|Y) \propto \pi(k) \prod_{j\le k} \int f_j(\theta_j)\exp\left(-\frac{n(\theta_j - Y_j)^2}{2}\right) d\theta_j \prod_{j>k} \exp\left(-\frac{nY_j^2}{2}\right).$$

In other words, the mean-field variational posterior $\widehat{Q}$ is nearly equivalent to a thresholding rule. It estimates all θ_j^* by 0 after $\widetilde{k}$ and applies the usual posterior distribution for each coordinate before $\widetilde{k}$. A mixed strategy is applied to the $\widetilde{k}$th coordinate. The effective model dimension $\widetilde{k}$ is found in a data-driven way through (14).

Next, we will show that even though the posterior itself is not a product measure, using $\widehat{Q}$ from the mean-field class still gives us a rate-optimal contraction result. The conditions on the prior distributions are summarized below.

- There exist some constants $C_1, C_2 > 0$ such that

$$\sum_{j=k}^{\infty} \pi(j) \le C_1 \exp(-C_2 k) \quad \text{for all } k. \tag{15}$$

- There exist some constants $C_3, C_4 > 0$ such that for $k_0 = \lceil (\frac{n}{\log n})^{\frac{1}{2\alpha+1}} \rceil$,

$$\pi(k_0) \ge C_3 \exp(-C_4 k_0 \log k_0). \tag{16}$$

- For the k_0 defined above, there exist some constants $c_0 \in \mathbb{R}$ and $c_1 > 0$ such that

$$-\log f_j(x) \le c_0 + c_1 j^{2\alpha+1} x^2 \quad \text{for all } j \le k_0 \text{ and } x \in \mathbb{R}. \tag{17}$$

These three conditions on Π include a large class of prior distributions. We remark that even though (17) involves α, it does not mean that one needs to know α when defining the prior Π. For example, the choice that $\pi(k) \propto e^{-\tau k}$ and f_j being $N(0, \sigma^2)$ for some constants $\tau, \sigma^2 > 0$ easily satisfies all the three conditions (15)–(17).

Conditions (15)–(17) will be used to derive the four conditions in Theorem 2.2. To be specific, (C1) and (C2) are consequences of (15) (see Lemma B.7 in the Supplementary Materials [39]), and (C3) and (C4) can be derived from (16) and (17) (see Lemma B.8 in the Supplementary Materials [39]). Then, by Theorem 2.2, we obtain the following result.

THEOREM 3.2. *Consider the prior* Π *that satisfies* (15)–(17). *Then, for any* $\theta^* \in \Theta_\alpha(B)$, *we have*

$$P_{\theta^*}^{(n)} \widehat{Q} \|\theta - \theta^*\|^2 \lesssim n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^{\frac{2\alpha}{2\alpha+1}},$$

where $\widehat{Q}$ *is the variational posterior defined by* (3) *with* $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}$.

It is well known that the minimax rate of estimating θ^* in $\Theta_\alpha(B)$ is $n^{-\frac{2\alpha}{2\alpha+1}}$ [20]. Using a mean-field variational posterior, we achieve the minimax rate up to a logarithmic factor. In fact, the following proposition demonstrates that this rate cannot be improved for a very general class of priors.

PROPOSITION 3.1. *Consider the prior* Π *specified in* (12). *Assume that* $\max_j \|f_j\|_\infty \le a$ *and* $\pi(k)$ *is nonincreasing over* k. *Then we have*

$$\sup_{\theta^* \in \Theta_\alpha(B)} P_{\theta^*}^{(n)} \widehat{Q} \|\theta - \theta^*\|^2 \gtrsim n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^{\frac{2\alpha}{2\alpha+1}},$$

where $\widehat{Q}$ *is the variational posterior defined by* (3) *with* $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}$.

On the other hand, the extra logarithmic factor can actually be removed by a rescaling of the prior. Details of this improvement are given in Appendix A.1 of the Supplementary Materials [39].

3.2. *Infinite dimensional exponential families.* In this section, we study another interesting variational family. The Gaussian mean-field family is defined as

$$\mathcal{S}_{\mathrm{G}} = \left\{ Q = \bigotimes_j N(\mu_j, \sigma_j^2) : \mu_j \in \mathbb{R}, \sigma_j^2 \geq 0 \right\}. \tag{18}$$

This class offers better interpretability of the results because every distribution in $\mathcal{S}_G$ is fully determined by a sequence of mean and variance parameters. Note that we allow σ_j^2 to be zero and $N(\mu_j, 0)$ is understood as the delta measure δ_{μ_j} on μ_j.

The application of $\mathcal{S}_G$ is illustrated by an infinite dimensional exponential family model. We define the probability measure P_θ by

$$\frac{dP_\theta}{d\ell} = \exp\left(\sum_{j=0}^{\infty} \theta_j \phi_j - c(\theta) \right), \tag{19}$$

where ℓ denotes the Lebesgue measure on $[0, 1]$, ϕ_j is the jth Fourier basis function of $L^2[0, 1]$, and $c(\theta)$ is given by

$$c(\theta) = \log \int_0^1 \exp\left(\sum_{j=0}^{\infty} \theta_j \phi_j(x) \right) dx.$$

Since $\phi_0(x) = 1$ and θ_0 can take arbitrary values without changing P_θ, we simply set $\theta_0 = 0$. In other words, P_θ is fully parameterized by $\theta = (\theta_1, \theta_2, \ldots)$. Given i.i.d. observations from $P_{\theta^*}^n$, our goal is to estimate $P_{\theta *}$, where θ^* is assumed to belong to the Sobolev ball $\Theta_\alpha(B)$ defined in (11). The loss function is chosen as n times the squared Hellinger distance $L(P_\theta^n, P_{\theta^*}^n) = nH^2(P_\theta, P_{\theta *})$.

We consider a prior distribution Π that is similar to the one used in Section 3.1. Its sampling process is described as follows:

1. Sample $k \sim \pi$;
2. Conditioning on k, sample $\theta_j \sim f_j$ for all $j \in [k]$, and set $\theta_j = 0$ for all $j > k$.

We impose the following conditions on the prior Π:

- There exist some constants $C_1, C_2 > 0$ such that

$$\sum_{j=k}^{\infty} \pi(j) \leq C_1 \exp(-C_2 k \log k) \quad \text{for all } k. \tag{20}$$

- There exist some constants $C_3, C_4 > 0$ such that for $k_0 = \lceil (\frac{n}{\log n})^{\frac{1}{2\alpha+1}} \rceil$

$$\pi(k_0) \geq C_3 \exp(-C_4 k_0 \log k_0). \tag{21}$$

- There exist some constants $c_0 \in \mathbb{R}$ and $c_1, \beta > 0$ such that

$$-\log f_j(x) \geq c_0 + c_1 |x|^\beta, \tag{22}$$

  for all $x \in \mathbb{R}$ and $j \in [k_0]$ with k_0 defined above.
- For the k_0 defined above, there exist some constants $c_0' \in \mathbb{R}$ and $c_1' > 0$ such that

$$-\log f_j(x) \leq c_0' + c_1' j^{2\alpha+1} x^2 \quad \text{for all } j \leq k_0 \text{ and } x \in \mathbb{R}. \tag{23}$$

The conditions (20)–(23) are satisfied by a large class of prior distributions. For example, one can choose $k \sim \text{Poisson}(\tau)$ and f_j being the density of $N(0, \sigma^2)$ for some constants $\tau, \sigma^2 > 0$, and then the four conditions are easily satisfied.

THEOREM 3.3. *Consider the prior* Π *that satisfies* (20)–(23). *Then, for any* $\theta^* \in \Theta_\alpha(B)$ *with some* $\alpha > 1/2$, *we have*

$$P_{\theta^*}^n \widehat{Q} H^2(P_\theta, P_{\theta^*}) \lesssim n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^{\frac{2\alpha}{2\alpha+1}},$$

where $\widehat{Q}$ *is the variational posterior defined by* (3) *with* $\mathcal{S} = \mathcal{S}_{\mathrm{G}}$.

The theorem shows that the Gaussian mean-field variational posterior is able to achieve the minimax rate $n^{-\frac{2\alpha}{2\alpha+1}}$ up to a logarithmic factor. We remark that the same result also holds for the mean-field variational posterior defined with $\mathcal{S}_{\mathrm{MF}}$. This is because $\mathcal{S}_{\mathrm{G}} \subset \mathcal{S}_{\mathrm{MF}}$, and thus $\inf_{Q \in \mathcal{S}_{\mathrm{MF}}} R(Q) \le \inf_{Q \in \mathcal{S}_{\mathrm{G}}} R(Q)$. Compared with the class $\mathcal{S}_{\mathrm{MF}}$, the objective function using the parametric family $\mathcal{S}_{\mathrm{G}}$ can be optimized by algorithms such as stochastic gradient descent over the parameters (μ_j, σ_j^2). The objective function can be greatly simplified according to the general mean-field solution given in Theorem 5.1.

3.3. *Piecewise constant model.* The previous two sections consider examples of the mean-field variational set and its variant. In this section, we use another example to illustrate a situation where the mean-field variational set only gives a trivial rate. On the other hand, we show that alternative variational classes with appropriate dependence structures are able to achieve the optimal rate.

We consider the following piecewise constant model:

$$X_i = \theta_i + \sigma Z_i, \quad i \in [n], \tag{24}$$

where $Z_i \sim N(0,1)$ independently for all $i \in [n]$. We assume $n \ge 2$ throughout the section. The true parameter θ^* is assumed to belong to the class $\Theta_{k^*}(B) = \{\theta \in \Theta_{k^*} : \|\theta\|_\infty \le B\}$, where for a general $k \in [n]$,

$$\begin{aligned} \Theta_k = \big\{ &\theta \in \mathbb{R}^n : \text{ there exist } \{a_j\}_{j=0}^k \text{ and } \{\mu_j\}_{j=1}^k \text{ such that} \\ &0 = a_0 \le a_1 \le \cdots \le a_k = n, \text{ and } \theta_i = \mu_j \text{ for all } i \in (a_{j-1} : a_j] \big\}. \end{aligned} \tag{25}$$

Here, for any two integers $a < b$, we use $(a : b]$ to denote all integers from $a+1$ to b. We assume both $B > 0$ and $\sigma^2 > 0$ are constants throughout this section. A vector $\theta^* \in \Theta_{k^*}(B)$ is a piecewise constant signal with at most k^* pieces. We use $P_\theta^{(n)}$ to denote the probability distribution of $N(\theta, \sigma^2 I_n)$ in this section.

The piecewise constant model is widely studied in the literature of change-point analysis. Recently, the minimax rate of the class Θ_{k^*} is derived by [14]. When $2 < k^* \le n^{1-\delta}$ for some constant $\delta \in (0,1)$, the minimax rate is $\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta_{k^*}} \mathbb{E}_{\theta^*}^{(n)} \|\widehat{\theta} - \theta^*\|^2 \asymp k^* \log n$. With an extra constraint on the infinity norm, the minimax rate for $\Theta_{k^*}(B)$ is still $k^* \log n$, with a slight modification of the proof in [14]. Since $D_\rho(P_\theta^{(n)}, P_{\theta'}^{(n)}) = \frac{\rho}{2\sigma^2}\|\theta - \theta'\|^2$ in this case, it is natural to choose the loss function as $L(P_\theta^{(n)}, P_{\theta^*}^{(n)}) = \|\theta - \theta^*\|^2$.

We put a prior distribution Π on the parameter θ. Consider Π that has the following sampling process:

1. Sample $w \sim \mathrm{Beta}(\alpha_0, \beta_0)$;
2. Conditioning on w, sample $z_i \sim \mathrm{Bernoulli}(w)$ for $i = 2, 3, \ldots, n$;
3. Conditioning on $(z_2, \ldots, z_n)$, sample $\theta_1 \sim g$, and then for $i = 2, 3, \ldots, n$, sample θ_i according to $\theta_i \sim g$ if $z_i = 1$ and $\theta_i = \theta_{i-1}$ if $z_i = 0$.

We first consider variational inference via the mean-field class, defined as

$$\mathcal{S}_{\mathrm{MF}} = \left\{ Q : dQ(\theta) = \prod_{i=1}^n dQ_i(\theta_i) \right\}.$$

We also define $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}^{\mathrm{joint}}$ on the joint distribution of (w, z, θ) by

$$\mathcal{S}_{\mathrm{MF}}^{\mathrm{joint}} = \Bigg\{ Q : dQ(w,z,\theta) = dQ^{(w)}(w)\, dQ^{(z)}(z)\, dQ^{(\theta)}(\theta), \\ dQ^{(z)}(z) = \prod_{i=2}^{n} dQ_i^{(z)}(z_i),\ Q^{(\theta)} \in \mathcal{S}_{\mathrm{MF}} \Bigg\}.$$

The variational posteriors $\widehat{Q}_{\mathrm{MF}}$ and $\widehat{Q}_{\mathrm{MF}}^{\mathrm{joint}}$ are given by (3) with variational classes defined above, respectively.[2] Interestingly, for the piecewise constant model, both $\widehat{Q}_{\mathrm{MF}}$ and $\widehat{Q}_{\mathrm{MF}}^{\mathrm{joint}}$ give a trivial rate.

THEOREM 3.4. *For the prior* Π *specified above with any* g *absolutely continuous with respect to the Lebesgue measure, we have*

$$\sup_{\theta^* \in \Theta_{k^*}(B)} P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MF}} \|\theta - \theta^*\|^2 = \sup_{\theta^* \in \Theta_{k^*}(B)} P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MF}}^{\mathrm{joint}} \|\theta - \theta^*\|^2 \gtrsim n,$$

for any $k^* \in [n]$, *where* $\widehat{Q}_{\mathrm{MF}}$ *and* $\widehat{Q}_{\mathrm{MF}}^{\mathrm{joint}}$ *are the variational posteriors defined by* (3) *with* $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}$ *and* $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}^{\mathrm{joint}}$, *respectively.*

The result of Theorem 3.4 shows that the mean-field variational posteriors $\widehat{Q}_{\mathrm{MF}}$ and $\widehat{Q}_{\mathrm{MF}}^{\mathrm{joint}}$ are unable to achieve a better rate than simply estimating θ^* by the naive estimator $\widehat{\theta} = X$. The proof, given in Appendix B.5 in the Supplementary Materials [39], reveals the reason of this phenomenon. Since the independence structure of the two classes fails to capture the underlying dependence structure of the parameter space $\Theta_{k^*}(B)$, the variational posterior distributions are equivalent to the posterior distribution induced by the prior $\Pi = \bigotimes_{i=1}^{n} g$ and, therefore, the condition (C4) is violated. Note that this is the first negative result in the literature on the statistical convergence of the mean-field approximation.

In order to achieve the minimax rate of the space $\Theta_{k^*}(B)$, it is necessary to introduce some dependence structure in the variational class. One of the simplest classes of dependent distributions is the class of first-order Markov chains, defined by

$$\mathcal{S}_{\mathrm{MC}} = \Bigg\{ Q : dQ(\theta) = dQ_1(\theta_1) \prod_{i=2}^{n} dQ_i(\theta_i | \theta_{i-1}) \Bigg\}.$$

The class $\mathcal{S}_{\mathrm{MC}}$ introduces a natural dependence structure for the piecewise constant model, and it is compatible with the prior distribution Π, because conditioning on the change point pattern z, the prior distribution of $\theta|z$ belongs to the class $\mathcal{S}_{\mathrm{MC}}$. We also introduce a similar variational class on the joint distribution of (w, z, θ), defined by

$$\mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}} = \Bigg\{ Q : dQ(w,z,\theta) = dQ^{(w)}(w)\, dQ^{(z)}(z)\, dQ^{(\theta)}(\theta), \\ dQ^{(z)}(z) = \prod_{i=2}^{n} dQ_i^{(z)}(z_i),\ Q^{(\theta)} \in \mathcal{S}_{\mathrm{MC}} \Bigg\}.$$

Besides the distribution of θ restricted to $\mathcal{S}_{\mathrm{MC}}$, the distributions of w and z are both in the mean-field classes.

In order to derive the rates for the variational posterior distributions induced by $\mathcal{S}_{\mathrm{MC}}$ and $\mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$, we impose the following conditions on the prior distribution Π:

[2] To be rigorous, the posterior distribution $\Pi(\cdot|X^{(n)})$ used in $D(Q\|\Pi(\cdot|X^{(n)}))$ are the marginal posterior of θ and the joint posterior of (w, z, θ), respectively.

- There exist some constants $C_2 > C_1 > 1$ such that

$$(n+\alpha_0)n^{C_1} \le \beta_0 \le \alpha_0 n^{C_2} - n. \tag{26}$$

- There exists a constant $c > 0$ such that

$$g(x) \ge c \quad \text{for all } |x| \le B+1. \tag{27}$$

According to Theorem 2.2, we get the following result.

THEOREM 3.5. *Consider a prior distribution* Π *that satisfies* (26) *and* (27). *Then, for any* $\theta^* \in \Theta_{k^*}(B)$, *we have*

$$P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MC}} \|\theta - \theta^*\|^2 \lesssim k^* \log n,$$
$$P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}} \|\theta - \theta^*\|^2 \lesssim k^* \log n,$$

where $\widehat{Q}_{\mathrm{MC}}$ *and* $\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}$ *are the variational posterior distributions defined by* (3) *with* $\mathcal{S} = \mathcal{S}_{\mathrm{MC}}$ *and* $\mathcal{S} = \mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$, *respectively.*

Theorem 3.5 shows that both $\widehat{Q}_{\mathrm{MC}}$ and $\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}$ are able to achieve the minimax rate of the problem. This example illustrates the importance of the choice of the variational class. According to Theorem 2.1, the rate of a variational posterior is upper bounded by ϵ_n^2, the rate of the true posterior, plus γ_n^2, the variational approximation error. The choice of $\mathcal{S}_{\mathrm{MF}}$ for the piecewise constant model leads to a very large γ_n^2, and thus a trivial rate in Theorem 3.4. On the other hand, the variational approximation errors given by the classes $\mathcal{S}_{\mathrm{MC}}$ and $\mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$ are small, which are dominated by the minimax rate.

Though the statistical properties of the two classes $\mathcal{S}_{\mathrm{MC}}$ and $\mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$ are both satisfactory, the class $\mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$ enjoys a computational advantage, and the solution $\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}$ can be computed exactly via dynamic programming. In order to characterize the solution $\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}$, we consider the following discrete optimization problem:

$$\max_{1\le k\le n}\Bigg\{\max_{0=a_0<a_1<\cdots<a_k=n}\sum_{j=1}^{k}\log\int g(\theta)\exp\Bigg(-\frac{1}{2}\sum_{i\in(a_{j-1}:a_j]}(X_i-\theta)^2\Bigg)d\theta + \log\big(\Gamma(k-1+\alpha_0)\Gamma(n-k+\beta_0)\big)\Bigg\}. \tag{28}$$

The solution of (28) is denoted as the sequence $0 = \widehat{a}_0 < \widehat{a}_1 < \cdots < \widehat{a}_{\widehat{k}} = n$. We remark that under the condition (26), the penalty term of (28) comes from the fact that

$$-\log\frac{\Gamma(k-1+\alpha_0)\Gamma(n-k+\beta_0)\Gamma(\alpha_0+\beta_0)}{\Gamma(n-1+\alpha_0+\beta_0)\Gamma(\alpha_0)\Gamma(\beta_0)} \asymp k\log n,$$

which coincides with the minimax rate.

THEOREM 3.6. *Let the maximizer of* (28) *be* $(\widehat{a}_0, \widehat{a}_1, \ldots, \widehat{a}_{\widehat{k}})$. *For* $d\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}(w,z,\theta) = d\widehat{Q}^{(w)}(w)\,d\widehat{Q}^{(z)}(z)\,d\widehat{Q}^{(\theta)}(\theta)$, *the distributions* $\widehat{Q}^{(w)}$, $\widehat{Q}^{(z)}$ *and* $\widehat{Q}^{(\theta)}$ *are specified as follows*:

1. *Under* $\widehat{Q}^{(z)}$, $z_{\widehat{a}_j+1} = 1$ *for* $j = 1, \ldots, \widehat{k}-1$, *and* $z_i = 0$ *elsewhere with probability* 1.
2. *We have* $\widehat{Q}^{(w)} = \mathrm{Beta}(\widehat{k}+\alpha_0-1, n-\widehat{k}+\beta_0)$.

Algorithm 1: Computation of (28)

Input : The data $X_1, \dots, X_n$.

Output: The set of knots $A_{\widehat{k},n} = \{\widehat{a}_1, \cdots, \widehat{a}_{\widehat{k}-1}\}$.

For j in $1:n$, set $A_{1,j} = \varnothing$, and compute
$B_{1,j} = S_{(0:j]}$.

For k in $2:n$
For j in $k:n$, compute
$B_{k,j} = \max_{k-1\le m\le j-1}\{B_{k-1,m} + S_{(m:j]}\}$,
$a_{k,j} = \operatorname{argmax}_{k-1\le m\le j-1}\{B_{k-1,m} + S_{(m:j]}\}$,
$A_{k,j} = A_{k-1,a_{k,j}} \cup \{a_{k,j}\}$.

Compute
$\widehat{k} = \operatorname{argmax}_{1\le k\le n}\{B_{k,n} + \log(\Gamma(k-1+\alpha_0)\Gamma(n-k+\beta_0))\}$.

3. *We have* $d\widehat{Q}^{(\theta)}(\theta) = d\widehat{Q}_1^{(\theta)}(\theta_1)\prod_{i=2}^n d\widehat{Q}_i^{(\theta)}(\theta_i|\theta_{i-1})$, *where*

$$\begin{cases} d\widehat{Q}_1^{(\theta)}(\theta_1) \propto g(\theta_1)\exp\Big(-\dfrac{1}{2}\displaystyle\sum_{i\in(\widehat{a}_0:\widehat{a}_1]}(X_i-\theta_1)^2\Big)\,d\theta_1, \\ d\widehat{Q}_i^{(\theta)}(\theta_i|\theta_{i-1}) \propto g(\theta_i)\exp\Big(-\dfrac{1}{2}\displaystyle\sum_{l\in(\widehat{a}_{j-1}:\widehat{a}_j]}(X_l-\theta_i)^2\Big)\,d\theta_i, \quad i=\widehat{a}_{j-1}+1,\ j>1, \\ d\widehat{Q}_i^{(\theta)}(\theta_i|\theta_{i-1}) = \delta_{\theta_{i-1}}(\theta_i)\,d\theta_i \qquad \text{otherwise.} \end{cases}$$

By Theorem 3.6, in order to get $\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}$, it is sufficient to solve (28). This can be done through a dynamic programming given in Algorithm 1. To simplify the notation, we define

$$S_{(a:b]} = \log\int g(\theta)\exp\Big(-\frac{1}{2}\sum_{i\in(a:b]}(X_i-\theta)^2\Big)\,d\theta, \tag{29}$$

for any integers $0 \le a < b \le n$.

We note that the computational cost of the dynamic programming above is $O(n^3)$ (see [13]), and for any integers $0 \le a < b \le n$, (29) has a closed form as long as we use a conjugate $g(\cdot)$.

4. Variational Bayes with model selection.

4.1. *General settings.* In this section, we consider a general form of probability models

$$\mathcal{M} = \{P_{k,\theta^{(k)}}^{(n)} : k \in \mathcal{K}, \theta^{(k)} \in \Theta^{(k)}\}.$$

Here, the probability $P_{k,\theta^{(k)}}^{(n)}$ is determined by an index k and a parameter $\theta^{(k)}$. We assume that the set $\mathcal{K}$ is either countable or finite. For a given k, the probability $P_{k,\theta^{(k)}}^{(n)}$ is parametrized by a $\theta^{(k)}$ in a parameter space $\Theta^{(k)}$ that is indexed by this k. Without loss of generality, we assume that the parameter $\theta^{(k)}$ can be written in a blockwise structure

$$\theta^{(k)} = (\theta_1^{(k)}, \dots, \theta_{m_k}^{(k)}).$$

Note that the dimension of $\theta^{(k)}$ may vary with k.

The model $\mathcal{M}$ is very natural for many applications. One can think of k as a model dimension index, which determines the complexity of the parameter space $\Theta^{(k)}$. A leading example is the mixture density model, where k stands for the number of components.

To model the hierarchical structure of $(k, \theta^{(k)})$, one naturally uses a hierarchical prior distribution, which is specified through the following sampling process:

1. First, sample $k \sim \pi$ from $\mathcal{K}$;
2. Conditioning on k, sample $\theta^{(k)}$ from the probability measure $\Pi^{(k)}$, and $\Pi^{(k)}$ has a product structure

$$d\Pi^{(k)}(\theta^{(k)}) = \prod_{j=1}^{m_k} d\Pi_j^{(k)}(\theta_j^{(k)}). \tag{30}$$

For variational inference, we consider a mean-field class that naturally takes advantage of the structure of the prior distribution. For a given $k \in \mathcal{K}$, the corresponding mean-field class is defined as

$$\mathcal{S}_{\mathrm{MF}}^{(k)} = \left\{ Q^{(k)} : dQ^{(k)}(\theta^{(k)}) = \prod_{j=1}^{m_k} dQ_j^{(k)}(\theta_j^{(k)}) \right\}. \tag{31}$$

In order to select the best model from the data, we consider optimizing the evidence lower bound (ELBO). With the notation $p(X^{(n)}|\theta^{(k)})$ standing for the joint likelihood function, the marginal likelihood given a model $k \in \mathcal{K}$ is defined by

$$p(X^{(n)}|k) = \int p(X^{(n)}|\theta^{(k)})\, d\Pi^{(k)}(\theta^{(k)}). \tag{32}$$

Then a straightforward model selection procedure is to maximize $\log(p(X^{(n)}|k)\pi(k))$ over $k \in \mathcal{K}$. In order to overcome the intractability of the integral (32), we instead optimize a lower bound, which is given by

$$\begin{aligned} &\log\bigl(p(X^{(n)}|k)\pi(k)\bigr) \\ &\quad \geq \int \log p(X^{(n)}|\theta^{(k)})\, dQ^{(k)}(\theta^{(k)}) - D(Q^{(k)} \| \Pi^{(k)}) + \log \pi(k), \end{aligned} \tag{33}$$

which can be derived by a direct application of Jensen's inequality. Denote the right-hand side of (33) by $F(Q^{(k)}, k)$, and we will solve the following optimization problem:

$$\max_{k \in \mathcal{K}} \max_{Q^{(k)} \in \mathcal{S}_{\mathrm{MF}}^{(k)}} F(Q^{(k)}, k). \tag{34}$$

Finally, the solution to (34) leads to the variational posterior distribution $\widehat{Q} = \widehat{Q}^{(\widehat{k})}$ that we use in a model selection context. A similar variational approximation to the tempered posterior in the model selection setting was studied by [12].

4.2. *Convergence rates.* Assume the observation $X^{(n)}$ is generated from a probability measure $P_0^{(n)}$, and $\widehat{Q} = \widehat{Q}^{(\widehat{k})}$ is the variational posterior that is a solution to (34). For the general settings described above, we show that the variational approximation error can be automatically controlled by a prior mass condition. Let Π be the prior distribution on $P_{k,\theta^{(k)}}$ induced by the sampling process of $(k, \theta^{(k)})$.

THEOREM 4.1. *Suppose ϵ_n is a sequence that satisfies $n\epsilon_n^2 \geq 1$. Let $\rho > 1$ be a constant and $C_2, C_3 > 0$ be constants. We assume that there exists a $k_0 \in \mathcal{K}$ and a subset $\Theta^{(k_0)} = \bigotimes_{j=1}^{m_{k_0}} \Theta_j^{(k_0)} \subset \{\theta^{(k_0)} : D_\rho(P_0^{(n)} \| P_{k_0,\theta^{(k_0)}}^{(n)}) \leq C_3 n\epsilon_n^2\}$, such that*

$$-\log \pi(k_0) - \sum_{j=1}^{m_{k_0}} \log \Pi_j^{(k_0)}(\Theta_j^{(k_0)}) \leq C_2 n\epsilon_n^2, \tag{C3*}$$

where $\pi(k_0)$ *and* $\Pi_j^{(k_0)}$ *are defined in the prior sampling procedure. Moreover, assume that the conditions* (C1) *and* (C2) *hold for all* $\epsilon > \epsilon_n$ *with respect to prior procedure* Π *and some constant* $C > C_2 + C_3 + 2$. *Then for the variational posterior* $\widehat{Q}^{(\widehat{k})}$ *defined as the solution of* (34), *we have*

$$P_0^{(n)} \widehat{Q}^{(\widehat{k})} L\big(P^{(n)}_{\widehat{k},\theta^{(\widehat{k})}}, P_0^{(n)}\big) \lesssim n\epsilon_n^2. \tag{35}$$

Theorem 4.1 characterizes the convergence rate of mean-field variational posterior with model selection using the conditions (C1), (C2) and (C3*). Given the structure of the prior distribution, an equivalent way of writing (C3*) is

$$\Pi\big(\{P_{k,\theta^{(k)}} : k = k_0, \theta^{(k_0)} \in \Theta^{(k_0)}\}\big) \geq \exp\big(-C_2 n\epsilon_n^2\big),$$

for the factorized structure of $\Theta^{(k_0)}$. Therefore, our three conditions (C1), (C2) and (C3*) still fall into the "prior mass and testing" framework, and directly correspond to the three conditions in [17] for convergence rates of the true posterior.

An interesting special case is when the set $\mathcal{K}$ is a singleton. Then, for a product prior measure and the mean-field variational class, the condition (C3*) is reduced to (2) discussed in Section 1.

4.3. *Density estimation via location-scale mixtures.* In this section, we consider the location-scale mixture model as an application of the theory. The location-scale mixture density is defined as

$$p\big(x|k,\theta^{(k)}\big) = \sum_{j=1}^{k} w_j \psi_\sigma(x - \mu_j), \tag{36}$$

where $k \in \mathbb{N}_+$, $\theta^{(k)} = (\mu, w, \sigma)$ with $\sigma > 0$, $\mu = (\mu_1, \ldots, \mu_k) \in \mathbb{R}^k$, $w = (w_1, \ldots, w_k) \in \Delta_k = \{w \in \mathbb{R}^k : w_j \geq 0 \text{ for } 1 \leq j \leq k \text{ and } \sum_{j=1}^k w_j = 1\}$ and

$$\psi_\sigma(x) = \frac{1}{2\sigma\Gamma(1+\frac{1}{p})} \exp\big(-(|x|/\sigma)^p\big), \tag{37}$$

for some positive even integer p. The kernel $\psi_\sigma(\cdot)$ has a prespecified form, for example, Gaussian density when $p = 2$, while the parameters k and $\theta^{(k)} = (w, \mu, \sigma)$ are to be learned from the data.

The location-scale mixture model (36) can be written as a special example of the general probability models introduced in Section 4.1. In this case, the countable set $\mathcal{K}$ is the positive integer set $\mathbb{N}_+$. The parameter space indexed by k is defined as

$$\begin{aligned}\Theta^{(k)} = \big\{&\theta^{(k)} = (\mu, w, \sigma) : \mu = (\mu_1, \ldots, \mu_k) \in \mathbb{R}^k,\\ &w = (w_1, \ldots, w_k) \in \Delta_k, \sigma \in \mathbb{R}_+\big\}.\end{aligned} \tag{38}$$

Given i.i.d. observations $X_1, \ldots, X_n$ sampled from some density function f_0, our goal is to estimate the density f_0 through the location-scale mixture model (36). We denote the probability distribution of the mixture density $p(x|k,\theta^{(k)})$ as $P_{k,\theta^{(k)}}$ and a probability distribution with a general density f as P_f. In the paper [21], a Bayesian procedure is proposed and a nearly minimax optimal convergence rate is derived for the true posterior distribution. We will follow the same setting in [21], but analyze the variational posterior.

We first specify the prior distribution Π through the following sampling process:

1. Sample the number of mixtures $k \sim \pi$;

2. Conditioning on k, sample the location parameters $\mu_1, \ldots, \mu_k$ independently from p_μ, sample the weights $w = (w_1, \ldots, w_k)$ from $p_w^{(k)}$, and then sample the precision parameter $\tau = \sigma^{-2}$ from p_τ.

In order to optimize (34) in the variational Bayes framework, we specify the blockwise structure (31) in this case as

$$\mathcal{S}_{\mathrm{MF}}^{(k)} = \left\{ Q^{(k)} : dQ^{(k)}\big(\theta^{(k)}\big) = dQ_\sigma(\sigma)\, dQ_w^{(k)}(w) \prod_{j=1}^{k} dQ_{\mu_j}(\mu_j) \right\}. \tag{39}$$

Note that we do not factorize $dQ_w^{(k)}(w)$ because of the constraint $\sum_{j=1}^k w_j = 1$. The variational posterior distribution is defined as $\widehat{Q} = \widehat{Q}^{(\widehat{k})}$ that solves (34). The loss function here is chosen as n times squared Hellinger distance, that is, $L(P_f^n, P_{f_0}^n) = nH^2(P_f, P_{f_0})$.

In order that $\widehat{Q}$ enjoys a good convergence rate, we need conditions on the prior distribution and the true density function f_0. We first list the conditions on the prior:

1. There exist constants $C_1, C_2 > 0$, such that

$$\sum_{m=k}^{\infty} \pi(m) \le C_1 \exp(-C_2 k \log k), \tag{40}$$

for all $m > 0$. There exist constants $t, C_3, C_4 > 0$, such that

$$\pi(k_0) \ge C_3 \exp(-C_4 k_0 \log k_0), \tag{41}$$

for all $n^{\frac{1}{2\alpha+1}} \le k_0 \le n^{\frac{1}{2\alpha+1}+t}$.

2. There exist constants $c_1, c_2, c_3 > 0$, such that

$$\int_{-\infty}^{-x_0} p_\mu(x)\, dx + \int_{x_0}^{\infty} p_\mu(x)\, dx \le c_1 \exp\big(-c_2 x_0^{c_3}\big), \tag{42}$$

for all $x_0 > 0$ and constants c_4, c_5, c_6, such that

$$p_\mu(x) \ge c_4 \exp\big(-c_5 |x|^{c_6}\big), \tag{43}$$

for all x.

3. There exist constants $t, d_1, d_2, d_3 > 0$, such that

$$\int_{w \in \Delta_{k_0}(w_0, \epsilon)} p_w^{(k_0)}(x)\, dx \ge d_1 \exp\left(-d_2 k_0 (\log k_0)^{d_3} \log\left(\frac{1}{\epsilon}\right)\right), \tag{44}$$

for all $w_0 \in \Delta_{k_0}$ and $n^{\frac{1}{2\alpha+1}} \le k_0 \le n^{\frac{1}{2\alpha+1}+t}$, where $\Delta_{k_0}(w_0, \epsilon) = \{w \in \Delta_{k_0} : \|w - w_0\|_1 \le \epsilon\}$.

4. There exist constants $b_0, b_1, b_2, b_3 > 0$, such that

$$\|p_\tau\|_\infty < b_0, \qquad \int_{\tau_0}^{\infty} p_\tau(x)\, dx \le b_1 \exp\big(-b_2 |\tau_0|^{b_3}\big), \tag{45}$$

for all $\tau_0 > 0$. There exist constants $b_4, b_5 > 0$ and a constant $b_6 \in (0, 1]$ that satisfy

$$p_\tau(x) \ge b_4 \exp\big(-b_5 |x|^{b_6}\big), \tag{46}$$

for all $x > 0$.

The conditions on the prior distribution are quite general. For example, one can choose $k \sim \mathrm{Poisson}(\xi_0)$, $\mu_j \sim N(0, \sigma_0^2)$, $w \sim \mathrm{Dir}(\alpha_0, \alpha_0, \ldots, \alpha_0)$ and $\tau \sim \Gamma(a_0, b_0)$ for some positive constants $\xi_0, \sigma_0, \alpha_0, a_0, b_0$. Then the conditions above are all satisfied.

Next, we list the conditions on the true density function f_0:

B1 (Smoothness) The logarithmic density function $\log f_0$ is assumed to be locally α-Hölder smooth. In other words, for the derivative $l_j(x) = \frac{d^j}{dx^j} \log f_0(x)$, there exists a polynomial $L(\cdot)$ and a constant $\gamma > 0$ such that

$$|l_{\lfloor\alpha\rfloor}(x) - l_{\lfloor\alpha\rfloor}(y)| \le L(x)|x-y|^{\alpha-\lfloor\alpha\rfloor}, \tag{47}$$

for all x, y that satisfies $|x - y| \le \gamma$. Here, the degree and the coefficients of the polynomial $L(\cdot)$ are all assumed to be constants. Moreover, the derivative $l_j(x)$ satisfies the bound $\int |l_j(x)|^{\frac{2\alpha+\epsilon}{j}} f_0(x)\,dx < s_{\max}$ for all $j = 1, \dots, \lfloor\alpha\rfloor$ with some constants $\epsilon, s_{\max} > 0$.

B2 (Tail) There exist positive constants T, ξ_1, ξ_2, ξ_3 such that

$$f_0(x) \le \xi_1 e^{-\xi_2|x|^{\xi_3}}, \tag{48}$$

for all $|x| \ge T$.

B3 (Monotonicity) There exist constants $x_m < x_M$ such that f_0 is nondecreasing on $(-\infty, x_m)$ and is nonincreasing on (x_M, ∞). Without loss of generality, we assume $f_0(x_m) = f_0(x_M) = c$ and $f_0(x) \ge c$ for all $x_m < x < x_M$ with some constant $c > 0$.

These conditions are exactly the same as in [21] and similar conditions are also considered in [24]. The conditions allow a well-behaved approximation to the true density by a location-scale mixture. There are many density functions that satisfy the conditions (B1)–(B3), for which we refer to [21].

The convergence rate of the variational posterior is given by the following theorem.

THEOREM 4.2. *Consider i.i.d. observations generated by $P_{f_0}^n$, and the density function f_0 satisfies conditions (B1)–(B3). For the prior that satisfies (40)–(46), we have*

$$P_{f_0}^n \widehat{Q} H^2(P_{\widehat{k},\theta^{(\widehat{k})}}, P_{f_0}) \lesssim n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^{\frac{2\alpha r}{2\alpha+1}},$$

where $\widehat{Q} = \widehat{Q}^{(\widehat{k})}$ is the solution of (34), and $r = \frac{p}{\min\{p,\xi_3\}} + \max\{d_3 + 1, \frac{c_6}{\min\{p,\xi_3\}}\}$, with p, ξ_3, c_6, d_3 defined in (37), (48), (43) and (44), respectively.

The proof of Theorem 4.2 largely follows the arguments in [21] that are used to establish the corresponding result for the true posterior distribution, thanks to the fact that Theorem 4.1 requires three very similar "prior mass and testing" conditions to that of [17]. The only difference is that function approximations via location-scale mixtures need to be analyzed under a stronger divergence $D_\rho(\cdot\|\cdot)$ for some $\rho > 1$. For this reason, the proof of Theorem 4.2 relies on the construction of a surrogate density function $\widetilde{f}_0$. We first apply Theorem 4.1 and establish a convergence rate under $\widetilde{f}_0$. Then the conclusion is transferred to f_0 with a change-of-measure argument. Details of the proof are given in Appendix B.6 in the Supplementary Materials [39].

4.4. *Dealing with latent variables.* For the mixture model considered in Section 4.3, we discuss a variation of the variational Bayes approach (34) by including latent variables. This facilitates computation and leads to a simple coordinate ascent algorithm that has closed-form updates. In the setting of mixture model, our approach is adaptive to the unknown number of components, and can be regarded as an extension of [25, 37] for variational inference with latent variables.

Since $p(X^{(n)}|k, \theta^{(k)}) = \prod_{i=1}^n \sum_{j=1}^k w_j \psi_\sigma(X_i - \mu_j)$ with $\theta^{(k)} = (\mu, w, \sigma)$, we can write

$$p(X^{(n)}|\theta^{(k)}) = \sum_{z^{(k)} \in [k]^n} p(X^{(n)}|z^{(k)}, \theta^{(k)}) w^{(k)}(z^{(k)}),$$

where $p(X^{(n)}|z^{(k)},\theta^{(k)}) = \prod_{i=1}^{n}\prod_{j=1}^{k}\psi_\sigma(X_i-\mu_j)^{\mathbf{1}\{z_i^{(k)}=j\}}$, and the probability of $z_i^{(k)}=j$ is w_j under $w^{(k)}(\cdot)$. We use the notation $\bar{\Pi}^{(k)}$ for the joint distribution of $(z^{(k)},\theta^{(k)})$, and then the marginal likelihood (32) can be written as

$$p\big(X^{(n)}|k\big) = \int p\big(X^{(n)}|z^{(k)},\theta^{(k)}\big)\,d\bar{\Pi}^{(k)}\big(z^{(k)},\theta^{(k)}\big).$$

Similar to (33), the evidence lower bound with the latent variables is given by

$$\begin{aligned}(49)\qquad &\log\big(p\big(X^{(n)}|k\big)\pi(k)\big)\\ &\quad\geq \int \log p\big(X^{(n)}|z^{(k)},\theta^{(k)}\big)\,d\bar{Q}^{(k)}\big(z^{(k)},\theta^{(k)}\big) - D\big(\bar{Q}^{(k)}\|\bar{\Pi}^{(k)}\big) + \log\pi(k).\end{aligned}$$

The right-hand side of (49) is shorthanded by $\bar{F}(\bar{Q}^{(k)},k)$. Define

$$\bar{\mathcal{S}}_{\mathrm{MF}}^{(k)} = \left\{\bar{Q}^{(k)} : d\bar{Q}^{(k)}\big(z^{(k)},\theta^{(k)}\big) = \prod_{i=1}^{n} dQ_z^{(k)}(z_i)\,dQ_\sigma(\sigma)\,dQ_w^{(k)}(w)\prod_{j=1}^{k} dQ_{\mu_j}(\mu_j)\right\}.$$

Then we solve the following optimization problem:

$$(50)\qquad \max_{k}\max_{\bar{Q}^{(k)}\in\bar{\mathcal{S}}_{\mathrm{MF}}^{(k)}} \bar{F}\big(\bar{Q}^{(k)},k\big).$$

The solution to (50) leads to the variational posterior distribution $\widehat{Q} = \widehat{Q}_{\mathrm{latent}}^{(\widehat{k})}$. It is worth noting that even though $\widehat{Q}$ is a joint distribution of (z,μ,w,σ), the posterior inference only relies on the marginal of (μ,w,σ), since the parametrization of the density $f(\cdot)$ in (36) does not depend on the latent variables. The existence of the latent variables only facilitates computation.

THEOREM 4.3. *Consider i.i.d. observations generated by $P_{f_0}^n$, and the density function f_0 satisfies conditions (B1)–(B3). For the prior that satisfies (40)–(46), we have*

$$P_{f_0}^n\widehat{Q}H^2\big(P_{\widehat{k},\theta^{(\widehat{k})}},P_{f_0}\big) \lesssim n^{-\frac{2\alpha}{2\alpha+1}}(\log n)^{\frac{2\alpha r}{2\alpha+1}},$$

where $\widehat{Q} = \widehat{Q}_{\mathrm{latent}}^{(\widehat{k})}$ is the solution to (50), and $r = \frac{p}{\min\{p,\xi_3\}} + \max\{d_3+1, \frac{c_6}{\min\{p,\xi_3\}}\}$, with p,ξ_3,c_6,d_3 defined in (37), (48), (43) and (44), respectively.

Theorem 4.3 shows that the variational posterior with latent variables achieves the same contraction rate as in Theorem 4.2. In fact, the two variational lower bounds (33) and (49) satisfy the following relation:

$$\log\big(p\big(X^{(n)}|k\big)\pi(k)\big) \geq \max_{Q^{(k)}\in\mathcal{S}_{\mathrm{MF}}^{(k)}} F\big(Q^{(k)},k\big) \geq \max_{\bar{Q}^{(k)}\in\bar{\mathcal{S}}_{\mathrm{MF}}^{(k)}} \bar{F}\big(\bar{Q}^{(k)},k\big),$$

which implies that the introduction of latent variables makes the variational approximation looser. On the other hand, Theorem 4.3 shows that the worse variational approximation does not compromise the statistical convergence rate. Moreover, with the help of latent variables, $\widehat{Q}_{\mathrm{latent}}^{(\widehat{k})}$ can be computed via standard variational inference algorithms. Details of the computational issues are given in Appendix A.2 of the Supplementary Materials [39].

5. Discussion.

5.1. *Variational Bayes and empirical Bayes.* In this section, we discuss an intriguing relation between variational Bayes and empirical Bayes in the context of sieve priors. We consider a nonparametric model $P_\theta^{(n)}$ with an infinite dimensional parameter $\theta = (\theta_j) \in \bigotimes_{j=1}^\infty \Theta_j \subset \mathbb{R}^\infty$. This includes the Gaussian sequence model and the infinite dimensional exponential family discussed in Section 3, as well as nonparametric regression and spectral density estimation. For each dimension, we assume $\Theta_j = \Theta_{j1} \cup \Theta_{j2}$ and $\Theta_{j1} \cap \Theta_{j2} = \varnothing$. Then a sieve prior $\theta \sim \Pi$ is specified by the following sampling process:

1. Sample $k \sim \pi$;
2. Conditioning on k, sample $\theta_j \sim f_{j1}$ for all $j \in [k]$, and sample $\theta_j \sim f_{j2}$ for all $j > k$.

We assume that the densities f_{j1} and f_{j2} satisfy $\int_{\Theta_{j1}} f_{j1} = 1$ and $\int_{\Theta_{j2}} f_{j2} = 1$. A leading example of the sieve prior is case of $\Theta_{j1} = \mathbb{R}\backslash\{0\}$ and $\Theta_{j2} = \{0\}$, as is used in Section 3.1 and Section 3.2.

An empirical Bayes procedure maximizes $e^{m_k(X^{(n)})}\pi(k)$,[3] where

$$m_k\big(X^{(n)}\big) = \log \int p\big(X^{(n)}|\theta\big) \prod_{j\le k} f_{j1}(\theta_j) \prod_{j>k} f_{j2}(\theta_j)\, d\theta$$

is the logarithm of marginal likelihood. With the maximizer $\widehat{k}$, the empirical Bayes posterior is defined as

$$d\widehat{Q}_{\mathrm{EB}}(\theta) \propto p\big(X^{(n)}|\theta\big) \prod_{j\le \widehat{k}} f_{j1}(\theta_j) \prod_{j>\widehat{k}} f_{j2}(\theta_j)\, d\theta. \tag{51}$$

Compared with a hierarchical Bayes approach, the empirical Bayes procedure does not need to evaluate the posterior distribution of k, and thus in many cases is easier to implement.

We also study mean-field approximation of the posterior distribution. In order to characterize its form, we need a few definitions. For any $g = (g_j)_{j=1}^\infty$, define

$$m_k\big(X^{(n)}; g\big) = \int \prod_{j=1}^\infty g_j(\theta_j) \log p\big(X^{(n)}|\theta\big)\, d\theta - \sum_{j\le k} D(g_j \| f_{j1}) - \sum_{j>k} D(g_j \| f_{j2}).$$

By Jensen's inequality, we observe that

$$m_k\big(X^{(n)}\big) \ge m_k\big(X^{(n)}, g\big), \tag{52}$$

for any g. We also define the density classes $\mathcal{G}_{j1} = \{g \ge 0 : \int g = \int_{\Theta_{j1}} g = 1\}$ and $\mathcal{G}_{j2} = \{g \ge 0 : \int g = \int_{\Theta_{j2}} g = 1\}$. The next theorem gives the exact form of the mean-field variational posterior.

THEOREM 5.1. *Consider the variational posterior $\widehat{Q}_{\mathrm{VB}}$ induced by the sieve prior and the mean-field variational set $\mathcal{S}_{\mathrm{MF}}$. The distribution $\widehat{Q}_{\mathrm{VB}}$ is a product measure with the density of each coordinate specified by*

$$q_j = \begin{cases} \widetilde{g}_{j1}^{(\widetilde{k})}, & j < \widetilde{k}, \\ (1-\widetilde{p})\widetilde{g}_{j1}^{(\widetilde{k})} + \widetilde{p}\widetilde{g}_{j2}^{(\widetilde{k})}, & j = \widetilde{k}, \\ \widetilde{g}_{j2}^{(\widetilde{k})}, & j > \widetilde{k}, \end{cases}$$

[3]The canonical form of empirical Bayes has a flat prior on k.

where for each given k, $(\widetilde{g}_{j1}^{(k)})_{j=1}^{k}$ *and* $(\widetilde{g}_{j2}^{(k)})_{j=k}^{\infty}$ *maximize the following objective function*:

$$\pi(k-1)e^{m_{k-1}(X^{(n)},(g_{j1})_{j=1}^{k-1}\cup(g_{j2})_{j=k}^{\infty})}+\pi(k)e^{m_k(X^{(n)},(g_{j1})_{j=1}^{k}\cup(g_{j2})_{j=k+1}^{\infty})}, \tag{53}$$

under the constraints that $g_{j1}\in\mathcal{G}_{j1}$ *and* $g_{j2}\in\mathcal{G}_{j2}$ *for all* j, $\widetilde{k}$ *maximizes*

$$\pi(k-1)e^{m_{k-1}(X^{(n)},(\widetilde{g}_{j1}^{(k)})_{j=1}^{k-1}\cup(\widetilde{g}_{j2}^{(k)})_{j=k}^{\infty})}+\pi(k)e^{m_k(X^{(n)},(\widetilde{g}_{j1}^{(k)})_{j=1}^{k}\cup(\widetilde{g}_{j2}^{(k)})_{j=k+1}^{\infty})}, \tag{54}$$

and finally,

$$\widetilde{p}=\frac{\pi(\widetilde{k}-1)e^{m_{\widetilde{k}-1}(X^{(n)},(\widetilde{g}_{j1}^{(\widetilde{k})})_{j=1}^{\widetilde{k}-1}\cup(\widetilde{g}_{j2}^{(\widetilde{k})})_{j=\widetilde{k}}^{\infty})}}{\pi(\widetilde{k}-1)e^{m_{\widetilde{k}-1}(X^{(n)},(\widetilde{g}_{j1}^{(\widetilde{k})})_{j=1}^{\widetilde{k}-1}\cup(\widetilde{g}_{j2}^{(\widetilde{k})})_{j=\widetilde{k}}^{\infty})}+\pi(\widetilde{k})e^{m_{\widetilde{k}}(X^{(n)},(\widetilde{g}_{j1}^{(\widetilde{k})})_{j=1}^{\widetilde{k}}\cup(\widetilde{g}_{j2}^{(\widetilde{k})})_{j=\widetilde{k}+1}^{\infty})}}.$$

The result of Theorem 5.1 also applies to the class $\mathcal{S}_{\mathrm{G}}$ discussed in Section 3.2 with $\mathcal{G}_{j1}$ replaced by the Gaussian class. We note that Theorem 5.1 can be viewed as an extension of Theorem 3.1. In fact, if the likelihood function can be factorized over each coordinate of θ, the form of $\widehat{Q}_{\mathrm{VB}}$ can be greatly simplified.

COROLLARY 5.1. *Under the same setting of Theorem* 5.1, *if we further assume that* $p(X^{(n)}|\theta)=\prod_{j=1}^{\infty}p(X_j^{(n)}|\theta_j)$, *then we will have*

$$\begin{aligned}\widetilde{g}_{j1}^{(\widetilde{k})}(\theta_j)&\propto f_{j1}(\theta_j)p\big(X_j^{(n)}|\theta_j\big)\mathbf{1}_{\{\theta_j\in\Theta_{j1}\}},\\ \widetilde{g}_{j2}^{(\widetilde{k})}(\theta_j)&\propto f_{j2}(\theta_j)p\big(X_j^{(n)}|\theta_j\big)\mathbf{1}_{\{\theta_j\in\Theta_{j2}\}},\\ \widetilde{k}&=\operatorname*{argmax}_k\big(\pi\big(k-1|X^{(n)}\big)+\pi\big(k|X^{(n)}\big)\big),\end{aligned} \tag{55}$$

and

$$\widetilde{p}=\frac{\pi(k-1|X^{(n)})}{\pi(k-1|X^{(n)})+\pi(k|X^{(n)})},$$

where

$$\pi\big(k|X^{(n)}\big)\propto\pi(k)\prod_{j=1}^{k}\int_{\Theta_{j1}}f_{j1}(\theta_j)p\big(X^{(n)}\big)\,d\theta_j\prod_{j=k+1}^{\infty}\int_{\Theta_{j2}}f_{j2}(\theta_j)p\big(X^{(n)}|\theta_j\big)\,d\theta_j.$$

In light of Theorem 5.1, we can compare the variational Bayes approach and the empirical Bayes approach, especially the definitions of $\widetilde{k}$ and $\widehat{k}$. The empirical Bayes chooses the best model by maximizing $e^{m_k(X^{(n)})}\pi(k)$, or equivalently $\pi(k|X^{(n)})$, while the variational Bayes maximizes (54). There are two major differences. The first difference is that empirical Bayes uses the exact marginal likelihood function $m_k(X^{(n)})$ and variational Bayes uses a mean-field approximation of $m_k(X^{(n)})$. We remark that in the case of likelihood that can be factorized, the mean-field approximation is exact, which leads to (55). The second difference is that empirical Bayes maximizes the posterior probability of the kth model, but the variational Bayes maximizes the sum of the posterior probabilities (or their mean-field approximations) of the $(k-1)$th and the kth models.

Despite the two differences, the empirical Bayes approach and the variational Bayes approach have a lot in common. Both are random probability distributions that summarize the information in data and prior. Both select a submodel according to very similar criteria. To close this section, we show that with a special variational class, the induced variational posterior is exactly the empirical Bayes posterior.

THEOREM 5.2. *Define the following set*:

$$\mathcal{S}_{\mathrm{EB}} = \left\{ Q : Q\left(\left(\bigotimes_{j \le k} \Theta_{j1} \right) \otimes \left(\bigotimes_{j > k} \Theta_{j2} \right) \right) = 1 \text{ for some integer } k \right\}.$$

Then the empirical Bayes posterior $\widehat{Q}_{\mathrm{EB}}$ *defined by* (51) *is the variational posterior induced by the sieve prior and the variational class* $\mathcal{S}_{\mathrm{EB}}$.

The result of Theorem 5.2 shows that for sieve priors, one can view the empirical Bayes approach as a variational Bayes approach, which suggests that it may be possible to unify the theoretical analysis in this paper and the analysis of empirical Bayes procedures in [28].

5.2. *Variational approximation as regularization*. According to Theorem 2.1, the convergence rate of the posterior is determined by the sum of ϵ_n^2, the rate of the true posterior, and γ_n^2, the variational approximation error. Since $\epsilon_n^2 + \gamma_n^2 \ge \epsilon_n^2$, it seems that the convergence rate of variational posterior is always no faster than that of the true posterior. However, Theorem 2.1 just gives an upper bound. In this section, we give two examples, and we show that it is possible for a variational posterior to have a faster convergence rate than that of the true posterior.

Example 1. We consider the setting of Gaussian sequence model (10). The true signal θ^* that generates the data is assumed to belong to the Sobolev ball $\Theta_\alpha(B)$. The prior distribution is specified as

$$\theta \sim d\Pi = \prod_{j \le n} dN(0, j^{-2\beta-1}) \prod_{j>n} \delta_0.$$

Note that a similar Gaussian process prior is well studied in the literature [9, 33]. We force all the coordinates after n to be zero, so that the variational approximation through Kullback–Leibler divergence will not explode. For the specified prior, the posterior contraction rate is $n^{-\frac{2(\alpha\wedge\beta)}{2\beta+1}}$, and when $\beta = \alpha$, the optimal minimax rate $n^{-\frac{2\alpha}{2\alpha+1}}$ is achieved.

Consider the following variational class:

$$\mathcal{S}_{[k]} = \left\{ Q : dQ = \prod_{j \le k} dQ_j \prod_{j=k+1}^{n} dN(0, e^{-jn}) \prod_{j>n} \delta_0 \right\},$$

for a given integer k. It is easy to see that the variational posterior $\widehat{Q}_{[k]}$ defined by (3) with $\mathcal{S} = \mathcal{S}_{[k]}$ can be written as

$$d\widehat{Q}_{[k]} = \prod_{j \le k} dN\left(\frac{n}{n + j^{2\beta+1}} Y_j, \frac{1}{n + j^{2\beta+1}} \right) \prod_{j=k+1}^{n} dN(0, e^{-jn}) \prod_{j>n} \delta_0.$$

In other words, the class $\mathcal{S}_{[k]}$ does not put any constraint on the first k coordinates and shrink all the coordinates after k to zero. Ideally, one would like to use δ_0 for the coordinates after k. However, that would lead to $D(Q\|\Pi(\cdot|Y)) = \infty$ for all $Q \in \mathcal{S}_{[k]}$ given that the support of δ_0 is a singleton. That is why we use $N(0, e^{-jn})$ instead. The rate of $\widehat{Q}_{[k]}$ for each k is given by the following theorem.

THEOREM 5.3. *For the variational posterior* $\widehat{Q}_{[k]}$, *we have*

$$\sup_{\theta^* \in \Theta_\alpha(B)} \mathbb{P}_{\theta^*}^{(n)} \widehat{Q}_{[k]} \|\theta - \theta^*\|^2 \asymp \begin{cases} \dfrac{k}{n} + k^{-2\alpha}, & k \le n^{\frac{1}{2\beta+1}}, \\ n^{-\frac{2(\alpha\wedge\beta)}{2\beta+1}}, & k > n^{\frac{1}{2\beta+1}}, \end{cases}$$

where $\widehat{Q}_{[k]}$ *is the variational posterior defined by* (3) *with* $\mathcal{S} = \mathcal{S}_{[k]}$.

Note that Theorem 5.3 gives both upper and lower bounds for $\widehat{Q}_{[k]}$. This makes the comparison between variational posterior and true posterior possible. Observe that when $k = \infty$, we have $\widehat{Q}_{[\infty]} = \Pi(\cdot|Y)$, and the result is reduced to the posterior contraction rate $n^{-\frac{2(\alpha\wedge\beta)}{2\beta+1}}$ in [9].

Depending on the values of α, β and k, the rate for $\widehat{Q}_{[k]}$ can be better than that of the true posterior. For example, when $\beta < \alpha$, the choice $k = n^{\frac{1}{2\alpha+1}}$ leads to the minimax rate $n^{-\frac{2\alpha}{2\alpha+1}}$, which is always faster than $n^{-\frac{2(\alpha\wedge\beta)}{2\beta+1}}$. This is because for a $\beta < \alpha$, the true posterior distribution undersmooths the data, but the variational class $\mathcal{S}_{[k]}$ with $k = n^{\frac{1}{2\alpha+1}}$ helps to reduce the extra variance resulted from undersmoothing by thresholding all the coordinates after k. On the other hand, when $\beta \geq \alpha$, an improvement through the variational class $\mathcal{S}_{[k]}$ is not possible. In this case, the true posterior has already overly smoothed the data, and the information loss cannot be recovered by the variational class.

Example 2. Consider the problem of sparse linear regression $y \sim N(X\beta^*, I_n)$, where X is a design matrix of size $n \times p$ and β^* belongs to the sparse set $\mathcal{B}(s) = \{\beta \in \mathbb{R}^p : \sum_{j=1}^p \mathbf{1}_{\{\beta_j \neq 0\}} \leq s\}$ for some $s \in [p]$. The prior distribution on β is specified by the Laplace density

$$\frac{d\Pi(\beta)}{d\beta} = \prod_{j=1}^p \left(\frac{\lambda}{2} e^{-\lambda|\beta_j|}\right).$$

Though the posterior distribution has a close connection to LASSO, it is proved in [11] that the posterior distribution cannot adapt to the sparsity of β^*. In particular, the common choice of λ in the theoretical analysis of LASSO only leads to a dense posterior.

In fact, it is known in the literature (e.g., [5]) that the LASSO, which is the posterior mode, achieves a nearly optimal rate over the class $\mathcal{B}(s)$. We show that the posterior mode can be well approximated by applying a simple variational class. Consider the variational class

$$\mathcal{S}_{\tau^2} = \{N(\beta, \tau^2 I_p) : \beta \in \mathbb{R}^p\}.$$

Define $\widehat{Q}_{\tau^2}$ to be the minimizer of $\min_{Q \in \mathcal{S}_{\tau^2}} D(Q\|\Pi(\cdot|y))$.

THEOREM 5.4. *For any $\lambda > 0$ and $\tau > 0$, we have $\widehat{Q}_{\tau^2} = N(\widehat{\beta}, \tau^2 I_p)$, where*

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} \left\{\frac{1}{2}\|y - X\beta\|^2 + \lambda \sum_{j=1}^p \tau h(\beta_j/\tau)\right\}. \tag{56}$$

The function h is defined by $h(x) = 2\phi(x) + x(\Phi(x) - \Phi(-x))$ with $\Phi(x) = \mathbb{P}(N(0,1) \leq x)$ and $\phi(x) = \Phi'(x)$.

Theorem 5.4 shows that the variational approximation is characterized by the penalized least-squares estimator (56). Observe that h is a convex function, and it satisfies $\sup_{x \in \mathbb{R}} |\tau h(x/\tau) - |x|| = \tau\sqrt{\frac{2}{\pi}}$ (see Figure 1), and thus $\widehat{\beta}$ will get arbitrarily close to the LASSO estimator as $\tau \to 0$. Therefore, even though the posterior does not have a good frequentist property, its variational approximation can recover a sparse signal.

By the fact that $\widehat{Q}_{\tau^2} = N(\widehat{\beta}, \tau^2 I_p)$, we have

$$\widehat{Q}_{\tau^2}\|\beta - \beta^*\|^2 = \|\widehat{\beta} - \beta^*\|^2 + p\tau^2. \tag{57}$$

Hence, a risk bound for the penalized least-squares estimator (56) directly leads to the convergence of the variational posterior. To present a bound for $\|\widehat{\beta} - \beta^*\|^2$, we need to introduce

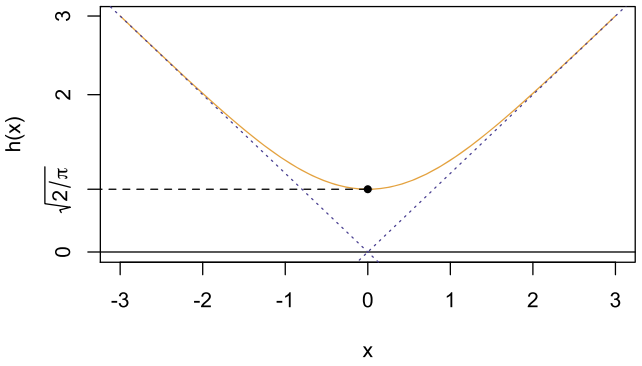

FIG. 1. *The functions $h(x)$ (orange) and $|x|$ (blue).*

some new notation. Let $S = \{j \in [p] : \beta_j^* \neq 0\}$ be the support of β^*. Define the restricted eigenvalue by

$$\kappa = \inf_{\{\Delta \neq 0 : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}} \frac{\frac{1}{\sqrt{n}}\|X\Delta\|}{\|\Delta\|}, \tag{58}$$

where $\|\Delta_S\|_1 = \sum_{j\in S}|\Delta_j|$ and $\|\Delta_{S^c}\|_1$ is defined similarly. The same quantity (58) also appears in the risk bound of LASSO [5].

THEOREM 5.5. *Assume $\|X_{*j}\|/\sqrt{n} \leq L$ for all $j \in [p]$ and $\kappa \leq L$ with some constant $L > 0$. Choose $\lambda = C\sqrt{n \log p}$ and $\tau = O(\frac{1}{np})$ for some sufficiently large constant $C > 0$. The solution to (56) satisfies*

$$\|\widehat{\beta} - \beta^*\|^2 \lesssim \frac{s \log p}{n\kappa^4},$$

with probability at least $1 - p^{-C'}$ uniformly over $\|\beta^*\|_0 \leq s$ for some constant $C' > 0$. As a consequence of (57), we also have

$$\widehat{Q}_{\tau^2}\|\beta - \beta^*\|^2 \lesssim \frac{s \log p}{n\kappa^4},$$

with probability at least $1 - p^{-C'}$.

We note that $\frac{s \log p}{n\kappa^4}$ is the same rate of convergence of LASSO [5]. With τ chosen as small as $O(\frac{1}{np})$, the statistical property of the variational posterior is very similar to that of the LASSO, and thus improves the original dense posterior distribution that is not suitable for sparse recovery.

5.3. *Model misspecification.* In this section, we present an extension of Theorem 2.1 in the context of model misspecification. We consider a data generating process $X^{(n)} \sim P_*^{(n)}$ that may not satisfies the conditions (C1)–(C3). The following theorem shows that the convergence rate of the variational posterior will then have an extra term that characterizes the deviation of $P_*^{(n)}$ to the model specified by the likelihood.

THEOREM 5.6. *Suppose ϵ_n is a sequence that satisfies $n\epsilon_n^2 \geq 1$. Assume that the conditions (C1)–(C3) hold with $P_0^{(n)}$ replaced by $P_{\theta_0}^{(n)}$. Then for the variational posterior $\widehat{Q}$ defined in (3), we have*

$$P_*^{(n)}\widehat{Q}L\big(P_\theta^{(n)}, P_{\theta_0}^{(n)}\big) \leq M\big(n\big(\epsilon_n^2 + \gamma_n^2\big) + D_2\big(P_*^{(n)} \| P_{\theta_0}^{(n)}\big)\big), \tag{59}$$

for some constant M only depending on C_1, C and ρ in (C1)–(C3), where the quantity γ_n^2 is defined as

$$\gamma_n^2 = \frac{1}{n} \inf_{Q \in \mathcal{S}} P_*^{(n)} D\big(Q \| \Pi(\cdot | X^{(n)})\big).$$

We note that here γ_n^2 is defined with respect to $P_*^{(n)}$ instead of $P_0^{(n)}$ in Theorem 2.1. Theorem 2.1 can be viewed as a special case of Theorem 5.6 with $P_0^{(n)} = P_*^{(n)} = P_{\theta_0}^{(n)}$. The extra term in the convergence rate that characterizes model misspecification is given by $D_2(P_*^{(n)} \| P_{\theta_0}^{(n)})$. In fact, it can be replaced by any ρ-Rényi divergence with $\rho > 1$.

Convergence rates of variational approximation to tempered posterior distributions under model misspecification have been studied by [1] (See their Theorem 2.7). Our results complement theirs by considering variational approximation to the ordinary posterior.

The next theorem gives sufficient conditions so that the variational approximation error γ_n^2 is dominated by the sum of the other two terms in (59). It can be viewed as an extension of Theorem 2.3.

THEOREM 5.7. *Suppose there are constants $C_1, C_2 > 0$, such that*

$$\text{(C4**)} \qquad \inf_{Q \in \mathcal{S} \cap \mathcal{E}} D(Q \| \Pi) \le C_1 \big(n\epsilon_n^2 + D_2\big(P_*^{(n)} \| P_{\theta_0}^{(n)}\big)\big),$$

where $\mathcal{E} = \{Q : \operatorname{supp}(Q) \subset \mathcal{C}\}$ with

$$\mathcal{C} = \big\{\theta : D\big(P_*^{(n)} \| P_\theta^{(n)}\big) \le C_2 \big(n\epsilon_n^2 + D_2\big(P_*^{(n)} \| P_{\theta_0}^{(n)}\big)\big)\big\}.$$

Then we have

$$n\gamma_n^2 \le (C_1 + C_2)\big(n\epsilon_n^2 + D_2\big(P_*^{(n)} \| P_{\theta_0}^{(n)}\big)\big).$$

To end this section, we apply Theorem 5.6 and Theorem 5.7 to the piecewise constant model discussed in Section 3.3 and derive oracle inequalities for the variational posterior distributions.

THEOREM 5.8. *Consider a prior distribution Π that satisfies (26) and (27). Then, for any $\theta^* \in \mathbb{R}^n$, we have*

$$P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MC}} \|\theta - \theta^*\|^2 \lesssim \min_{1 \le k \le n} \Big\{ \inf_{\theta_0 \in \Theta_k(B)} \|\theta^* - \theta_0\|^2 + k \log n \Big\},$$

$$P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}} \|\theta - \theta^*\|^2 \lesssim \min_{1 \le k \le n} \Big\{ \inf_{\theta_0 \in \Theta_k(B)} \|\theta^* - \theta_0\|^2 + k \log n \Big\},$$

where the definitions of $\widehat{Q}_{\mathrm{MC}}$ and $\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}$ are given in Theorem 3.5.

Acknowledgments. The authors are grateful to an Associate Editor and two referees who give very insightful feedbacks that lead to the improvement of the paper.

The second author was supported in part by NSF Grant DMS-1712957 and NSF Career Award DMS-1847590.

SUPPLEMENTARY MATERIAL

Supplement to "Convergence rates of variational posterior distributions" (DOI: 10.1214/19-AOS1883SUPP; .pdf). The supplement [39] presents additional theoretical results, additional proofs of main results and proofs of auxiliary results.

REFERENCES

[1] ALQUIER, P. and RIDGWAY, J. (2017). Concentration of tempered posteriors and of their variational approximations. Preprint. Available at arXiv:1706.09293.

[2] BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718 https://doi.org/10.1214/aos/1018031206

[3] BARRON, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Department of Statistics, University of Illinois.

[4] BICKEL, P., CHOI, D., CHANG, X. and ZHANG, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41** 1922–1943. MR3127853 https://doi.org/10.1214/13-AOS1124

[5] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[6] 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

[7] BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

[8] CARBONETTO, P. and STEPHENS, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.* **7** 73–107. MR2896713 https://doi.org/10.1214/12-BA703

[9] CASTILLO, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2** 1281–1299. MR2471287 https://doi.org/10.1214/08-EJS273

[10] CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.* **42** 2058–2091. MR3262477 https://doi.org/10.1214/14-AOS1253

[11] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874 https://doi.org/10.1214/15-AOS1334

[12] CHÉRIEF-ABDELLATIF, B.-E. and ALQUIER, P. (2018). Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electron. J. Stat.* **12** 2995–3035. MR3855643 https://doi.org/10.1214/18-EJS1475

[13] FRIEDRICH, F., KEMPE, A., LIEBSCHER, V. and WINKLER, G. (2008). Complexity penalized M-estimation: Fast computation. *J. Comput. Graph. Statist.* **17** 201–224. MR2424802 https://doi.org/10.1198/106186008X285591

[14] GAO, C., HAN, F. and ZHANG, C.-H. (2017). Minimax risk bounds for piecewise constant models. Preprint. Available at arXiv:1705.06386.

[15] GAO, C., VAN DER VAART, A. W. and ZHOU, H. H. (2015). A general framework for bayes structured linear models. Preprint. Available at arXiv:1506.02174.

[16] GAO, C. and ZHOU, H. H. (2016). Rate exact Bayesian adaptation with modified block priors. *Ann. Statist.* **44** 318–345. MR3449770 https://doi.org/10.1214/15-AOS1368

[17] 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

[18] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274 https://doi.org/10.1214/009053606000001172

[19] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. MR3396985 https://doi.org/10.1214/15-AOS1341

[20] JOHNSTONE, I. M. (2011). Gaussian estimation: Sequence and wavelet models. Manuscript.

[21] KRUIJER, W., ROUSSEAU, J. and VAN DER VAART, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** 1225–1257. MR2735885 https://doi.org/10.1214/10-EJS584

[22] LAFFERTY, J. D. and BLEI, D. M. (2006). Correlated topic models. In *Advances in Neural Information Processing Systems* 147–154.

[23] LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381

[24] MAUGIS-RABUSSEAU, C. and MICHEL, B. (2013). Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM Probab. Stat.* **17** 698–724. MR3126158 https://doi.org/10.1051/ps/2012018

[25] PATI, D., BHATTACHARYA, A. and YANG, Y. (2018). On statistical optimality of variational Bayes. In *International Conference on Artificial Intelligence and Statistics* 1579–1588.

[26] RAJ, A., STEPHENS, M. and PRITCHARD, J. K. (2014). faststructure: Variational inference of population structure in large snp data sets. *Genetics* **197** 573–589.

[27] RIVOIRARD, V. and ROUSSEAU, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal*. **7** 311–333. MR2934953 https://doi.org/10.1214/12-BA710

[28] ROUSSEAU, J. and SZABO, B. (2017). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *Ann. Statist*. **45** 833–865. MR3650402 https://doi.org/10.1214/16-AOS1469

[29] SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26. MR0184378 https://doi.org/10.1007/BF00535479

[30] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist*. **29** 687–714. MR1865337 https://doi.org/10.1214/aos/1009210686

[31] SUDDERTH, E. B. and JORDAN, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman–Yor processes. In *Advances in Neural Information Processing Systems* 1585–1592.

[32] 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

[33] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist*. **36** 1435–1463. MR2418663 https://doi.org/10.1214/009053607000000613

[34] WALKER, S. and HJORT, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **63** 811–821. MR1872068 https://doi.org/10.1111/1467-9868.00314

[35] WALKER, S. G., LIJOI, A. and PRÜNSTER, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist*. **35** 738–746. MR2336866 https://doi.org/10.1214/009053606000001361

[36] WANG, Y. and BLEI, D. M. (2019). Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc*. **114** 1147–1161. MR4011769 https://doi.org/10.1080/01621459.2018.1473776

[37] YANG, Y., PATI, D. and BHATTACHARYA, A. (2017). α-variational inference with statistical guarantees. Preprint. Available at arXiv:1710.03266.

[38] ZHANG, A. Y. and ZHOU, H. H. (2017). Theoretical and computational guarantees of mean field variational inference for community detection. Preprint. Available at arXiv:1710.11268.

[39] ZHANG, F. and GAO, C. (2020). Supplement to "Convergence rates of variational posterior distributions." https://doi.org/10.1214/19-AOS1883SUPP.

[40] ZHANG, T. (2006). From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Ann. Statist*. **34** 2180–2210. MR2291497 https://doi.org/10.1214/009053606000000704