Journal of Machine Learning Research 21 (2020) 1-42 Submitted 2/19; Revised 3/20; Published 5/20

Asymptotic Consistency of α-Rényi-Approximate Posteriors

Prateek Jaiswal JAISWALP@PURDUE.EDU
School of Industrial Engineering, Purdue University, West Lafayette, IN 47907, USA

Vinayak Rao VARAO@PURDUE.EDU
Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

Harsha Honnappa HONNAPPA@PURDUE.EDU
School of Industrial Engineering, Purdue University, West Lafayette, IN 47907, USA

Editor: XuanLong Nguyen

Abstract

We study the asymptotic consistency properties of α-Rényi approximate posteriors, a class of variational Bayesian methods that approximate an intractable Bayesian posterior with a member of a tractable family of distributions, the member chosen to minimize the α-Rényi divergence from the true posterior. Unique to our work is that we consider settings with $\alpha > 1$, resulting in approximations that upperbound the log-likelihood, and consequently have wider spread than traditional variational approaches that minimize the Kullback-Liebler (KL) divergence from the posterior. Our primary result identifies sufficient conditions under which consistency holds, centering around the existence of a 'good' sequence of distributions in the approximating family that possesses, among other properties, the right rate of convergence to a limit distribution. We further characterize the good sequence by demonstrating that a sequence of distributions that converges too quickly cannot be a good sequence. We also extend our analysis to the setting where α equals one, corresponding to the minimizer of the reverse KL divergence, and to models with local latent variables. We also illustrate the existence of good sequence with a number of examples. Our results complement a growing body of work focused on the frequentist properties of variational Bayesian methods.

Keywords: α-Rényi divergence, Asymptotic consistency, Bayesian computation, Variational inference

1. Introduction

Bayesian statistics forms a powerful and flexible framework that allows practitioners to bring prior knowledge to statistical problems, and to coherently manage uncertainty resulting from finite and noisy datasets. A Bayesian represents the unknown state of the world with a possibly vector-valued parameter θ, over which they place a prior probability $\pi(\theta)$, representing *a priori* beliefs they might have. θ can include global parameters shared across the entire dataset, as well as local variables specific to each observation. A likelihood $p(\mathbf{X}_n|\theta)$ then specifies a probability distribution over the observed dataset $\mathbf{X}_n$. Given observations $\mathbf{X}_n$, prior beliefs $\pi(\theta)$ are updated to a posterior distribution $\pi(\theta|\mathbf{X}_n)$ calculated through Bayes' rule.

While conceptually straightforward, computing $\pi(\theta|\mathbf{X}_n)$ is intractable for many interesting and practical models, and the field of Bayesian computation is focused on developing scalable and accurate computational techniques to approximate the posterior distribution. Traditionally, much of this has involved Monte Carlo and Markov chain Monte Carlo techniques to construct sampling approximations to the posterior distribution. In recent years, developments from machine learning have sought to leverage tools from optimization to construct tractable posterior approximations. An early and still popular instance of this methodology is *variational Bayes* (VB) (Blei et al., 2017).

At a high level, the idea behind VB is to approximate the intractable posterior $\pi(\theta|\mathbf{X}_n)$ with an element $q(\theta)$ of some simpler class of distributions $\mathcal{Q}$. Examples of $\mathcal{Q}$ include the family of Gaussian distributions, delta functions, or the family of factorized 'mean-field' distributions that discard correlations between components of θ. The variational solution q is the element of $\mathcal{Q}$ that is closest to $\pi(\theta|\mathbf{X}_n)$, where closeness is measured in terms of the Kullback-Leibler (KL) divergence. Thus, q is the solution to:

$$q(\theta) = \operatorname{argmin}_{\tilde{q}\in\mathcal{Q}} \mathrm{KL}(\tilde{q}(\theta) \| \pi(\theta|\mathbf{X}_n)). \tag{1}$$

We term this as the KL-VB method. From the non-negativity of the KL divergence, we can view this as maximizing a lower-bound to the logarithm of the model *evidence*, $\log p(\mathbf{X}_n) = \log\left(\int p(\mathbf{X}_n, \theta)\mathrm{d}\theta\right)$. This lower-bound, called the variational lower-bound or evidence lower bound (ELBO) is defined as

$$\mathrm{ELBO}(\tilde{q}(\theta)) = \log p(\mathbf{X}_n) - \mathrm{KL}(\tilde{q}(\theta) \| p(\theta|\mathbf{X}_n)). \tag{2}$$

Optimizing the two equations above with respect to q does not involve either calculating expectations with respect to the intractable posterior $\pi(\theta|\mathbf{X}_n)$, or evaluating the posterior normalization constant. As a consequence, a number of standard optimization algorithms can be used to select the best approximation $q(\theta)$ to the posterior distribution, examples including expectation-maximization (Neal and Hinton, 1998) and gradient-based (Kingma and Welling, 2014) methods. This has allowed the application of Bayesian methods to increasingly large datasets and high-dimensional settings. Despite their widespread popularity in the machine learning, and more recently, the statistics communities, it is only recently that variational Bayesian methods have been studied theoretically (Alquier and Ridgway, 2020; Chérief-Abdellatif and Alquier, 2018; Wang and Blei, 2018; Yang et al., 2020; Zhang and Gao, 2020).

1.1. Rényi Divergence Minimization

Despite its popularity, variational Bayes has a number of well-documented limitations. An important one is its tendency to produce approximations that underestimate the spread of the posterior distribution (Turner and Sahani, 2011; Li and Turner, 2016): in essence, the variational Bayes solution tends to match closely with the dominant mode of the posterior. This arises from the choice of the divergence measure $\mathrm{KL}(q(\theta)\|\pi(\theta|\mathbf{X}_n)) = \mathbb{E}_q[\log(q(\theta)/\pi(\theta|\mathbf{X}_n))]$, which does not penalize solutions where $q(\theta)$ is small while $\pi(\theta|\mathbf{X}_n)$ is large. While many statistical applications only focus on the mode of the distribution, definite calculations of the variance and higher moments are critical in predictive and decision-making problems.

A natural solution is to consider different divergence measures than those used in variational Bayes. Expectation propagation (EP) (Minka, 2001a) was developed to minimize $\mathbb{E}_p[\log(p/q)]$ instead, though this requires an expectation with respect to the intractable posterior. Consequently, EP can only minimize an approximation of this objective.

More recently, Rényi's α-divergence (Van Erven and Harremos, 2014) has been used as a family of parametrized divergence measures for variational inference (Li and Turner, 2016; Dieng et al., 2017). The α-Rényi divergence is defined as

$$D_\alpha\left(\pi(\theta|\mathbf{X}_n)\|q(\theta)\right) := \frac{1}{\alpha-1}\log\int_\Theta q(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{q(\theta)}\right)^\alpha d\theta.$$

The parameter α spans a number of divergence measures and, in particular, we note that as $\alpha \to 1$ we recover the EP objective $\mathrm{KL}(\pi(\theta|\mathbf{X}_n)\|q(\theta))$, we will call its minimizer 1−Rényi approximate posterior. Settings of $\alpha > 1$ are particularly interesting since, in contrast to VB which lower-bounds the log-likelihood of the data (2), one obtains tractable upper bounds. Precisely, using Jensen's inequality,

$$p(\mathbf{X}_n)^\alpha = \left(\int p(\theta,\mathbf{X}_n)\ \frac{q(\theta)}{q(\theta)}\ d\theta\right)^\alpha \le \mathbb{E}_q\left[\left(\frac{p(\theta,\mathbf{X}_n)}{q(\theta)}\right)^\alpha\right].$$

Applying the logarithm function on either side,

$$\alpha\log p(\mathbf{X}_n) \le \log\mathbb{E}_q\left[\left(\frac{\pi(\theta,\mathbf{X}_n)}{q(\theta)}\right)^\alpha\right] \tag{3}$$

$$= \alpha\log p(\mathbf{X}_n) + \log\mathbb{E}_q\left[\left(\frac{\pi(\theta|\mathbf{X}_n)}{q(\theta)}\right)^\alpha\right] := \mathcal{F}_2(q). \tag{4}$$

Observe that the second term in the expression for $\mathcal{F}_2(q)$ is just $(\alpha-1)D_\alpha(p(\theta|\mathbf{X}_n)\|q(\theta))$. Like with the ELBO lower bound, evaluating this upper bound only involves expectations with respect to $q(\theta)$, and only requires evaluating $p(\theta,\mathbf{X}_n)$, the unnormalized posterior distribution. Optimizing this upper bound over some class of distributions $\mathcal{Q}$, we obtain the α-*Rényi* approximation. As noted before, standard variational Bayes, which optimizes a lower-bound, tends to produce approximating distributions that underestimate the posterior variance, resulting in predictions that are overconfident and ignore high-risk regions in the support of the posterior. We illustrate this in Figure 1 below that reproduces a result from Li and Turner (2016). The true posterior distribution is an anisotropic Gaussian distribution and the variational family consists of isotropic (or mean-field) Gaussian distributions. Standard KL-VB, represented by the curve $\alpha = 0$, clearly fits the mode of the posterior, but completely underestimates the dominant eigen-direction. On the other hand, for large values of α (shown as $\alpha \to +\infty$), the α-Rényi approximate posterior matches the mode and does a better job of capturing the spread of the posterior. The figure also presents results for the $\alpha = 1$ and the $\alpha \to -\infty$ cases. As an aside, we observe that our parametrization of the Rényi divergence is different from Li and Turner (2016), where the upper-bounds considered in Li and Turner (2016) emerge as $\alpha \to -\infty$.

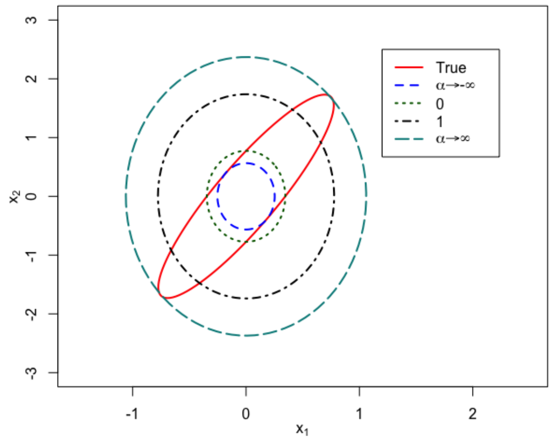

Figure 1: Isotropic variational α-Rényi approximations to an anisotropic Gaussian, for different values of α (see also Li and Turner (2016)).

We note, furthermore, that in tasks such as model selection, the marginal likelihood of the data is of fundamental interest (Grosse et al., 2015), and the α-Rényi upper bound provides an approximation that complements the VB lower bound. Recent developments in stochastic optimization have allowed the α-Rényi objective to be optimized fairly easily; see Li and Turner (2016) and Dieng et al. (2017).

1.2. Large Sample Properties

Despite often state-of-the-art empirical results, variational methods still present a number of unanswered theoretical questions. This is particularly true for α-Rényi divergence minimization which has empirically demonstrated very promising results for a number of applications (Li and Turner, 2016; Dieng et al., 2017). In recent work, Zhang and Gao (2020) have shown conditions under which α-Rényi variational methods are consistent when α is less than one. Their results followed from a proof for the regular Kullback-Leibler variational algorithm, and thus only apply to situations when a *lower-bound* is optimized. As we mentioned before, the setting with α greater than 1 is qualitatively different from both Kullback-Leibler and Rényi divergence with $\alpha < 1$. This setting, which is also of considerable practical interest, is the focus of our paper and we address the question of asymptotic consistency of the approximate posterior distribution obtained by minimizing the Rényi divergence.

Asymptotic consistency (van der Vaart, 1998) is a basic frequentist requirement of any statistical method, guaranteeing that the 'true' parameter is recovered as the number of observations tends to infinity. Table 1 summarizes the current known results on consistency of VI and EP, and highlights the gap that this paper is intended to fill. We note that in this work, we are not analyzing the actual EP algorithm (Wainwright and Jordan, 2008), and are instead looking at the global minimizer of the ideal EP objective.

Methods	Papers
KL-VB	Wang and Blei (2018),Zhang and Gao (2020)
α-Rényi ($\alpha < 1$)	Zhang and Gao (2020)
α-Rényi ($\alpha > 1$)	This paper
1-Rényi ($\alpha \to 1$, global EP)	This paper

Table 1: Known results on the asymptotic consistency of variational methods.

As we will see, filling these gaps will require new developments. This follows from two complicating factors: 1) Rényi divergence with $\alpha > 1$ *upper-bounds* the log-likelihood, and 2) this requires new analytical approaches involving expectations with respect to the intractable $\pi(\theta|\mathbf{X}_n)$. We thus emphasize that the results in our paper are not a consequence of recent analysis in Wang and Blei (2018) and Zhang and Gao (2020) for the KL-VB, and our proofs differ substantially from theirs.

We establish our main result in Theorem 9 under mild regularity conditions. First, in Assumption 1 we assume that the prior distribution places positive mass in the neighborhood of the true parameter θ_0, and that it is uniformly bounded. The former condition is a reasonable assumption to make - clearly, if the prior does not place any mass in the neighborhood of the true parameter (assuming one exists) then neither will the posterior. The uniform boundedness condition on the other hand is attendant to a loss of generality. In particular, we cannot assume certain heavy-tailed priors (such as Pareto) which might be important for some engineering applications. Second, we also make the mild assumption that the likelihood function is locally asymptotically normal (LAN) in Assumption 2. This is a standard assumption that holds for a variety of statistical/stochastic models. However, while the LAN assumption will be critical for establishing the asymptotic consistency results, it is unclear if it is necessary as well. We observe that Wang and Blei (2018) make a similar assumption in analyzing the consistency of KL-VB. We note that any model P_θ that is twice differentiable in the parameter θ satisfies the LAN condition (van der Vaart, 1998). Also critical to the consistency result are the properties of the variational family. Assumption 3 is a mild condition that insists on there existing Dirac delta distributions in an open neighborhood of the true parameter θ_0. This is usually easy to verify: if the variational family consists of Gaussian distributions, for instance, then Dirac delta distributions are present at all points in the parameter space. Next, we assume that the variational family contains 'good sequences' that are constructed so as to converge at the same rate as the true posterior (in sequence with the sample size), with the first moment of an element in the sequence the maximum likelihood estimator of the parameter (at a given sample size). We also require the tails of the good sequence to bound the tails of the true posterior. We provide examples that verify the existence of good sequences in commonly used variational families, such as the mean-field family.

The proof of Theorem 9 is a consequence of a series of auxiliary results. First, in Lemma 6 we characterize α-Rényi minimizers and show that the sequence must have a Dirac delta distribution at the true parameter θ_0 in the large sample limit. Then, in Lemma 7 we argue that any convex combination of a Dirac delta distribution at the true parameter θ_0 with any other distribution can not achieve zero α-Rényi divergence in the limit. Next, we show in Proposition 8 that the α-Rényi divergence between the true posterior and the closest variational approximator is bounded above in the large sample limit. We demonstrate this by showing that a 'good sequence' of distributions (see Assumption 4) has asymptotically bounded α-Rényi divergence, implying that the minimizers do as well. Note that this does not yet prove that the minimizing sequence converges to a Dirac delta distribution at θ_0.

The next stage of the analysis is concerned with demonstrating that the minimizing sequence does indeed converge to a Dirac delta distribution concentrated at the true parameter. We demonstrate this fact as a consequence of Proposition 8, Lemma 6, and Lemma 7. In essence, Theorem 9 shows that, α-Rényi minimizing distributions are arbitrarily close to a good sequence, in the sense of Rényi divergence with the posterior in the large sample limit.

In our next result in Theorem 11, under additional regularity conditions, we further characterize the rate of convergence of the $\alpha-$Rényi minimizers. We demonstrate that the $\alpha-$Rényi minimizing sequence cannot concentrate to a point in the parameter space at a faster rate than the true posterior concentrates at the true parameter θ_0. Consequently, the tail mass in the α-Rényi minimizer could dominate that of the true posterior. This is in contrast with KL-VB, where the evidence lower bound (ELBO) maximizer typically under-estimates the variance of the true posterior.

Here is a brief roadmap of the paper. In Section 2, we formally introduce the α-Rényi methodology, and rigorously state the necessary regularity assumptions. We present our main result in Section 3, presenting only the proofs of the primary results. In Section 4 we also recover the consistency of $1-$Rényi, approximate posteriors, the global minimizer of EP objective as a consequence of the results in Section 3. In Section 5, we generalize the notion of good sequence to the models with local latent parameters and under some additional regularity conditions, prove asymptotic consistency of the α-Rényi approximate posterior over global latent parameters. All proofs of auxiliary and technical results are delayed to the Appendix.

2. Variational Approximation Using $\alpha-$Rényi Divergence

We assume that the data-generating distribution is parametrized by $\theta \in \Theta \subseteq \mathbb{R}^d$, $d \geq 1$ and is absolutely continuous with respect to the Lebesgue measure, so that the likelihood function $p(\cdot|\theta)$ is well-defined. We place a prior $\pi(\theta)$ on the unknown θ, and denote $\pi(\theta|\mathbf{X}_n) \propto p(\theta, \mathbf{X}_n)$ as the posterior distribution, where $\mathbf{X}_n = \{\xi_1, \ldots, \xi_n\}$ are the n independent and identically distributed (i.i.d.) observed samples generated from the 'true' measure P_{θ_0} in the likelihood family. In this paper we will study the $\alpha-$Rényi-approximate posterior q_n^* that minimizes the $\alpha-$Rényi divergence between $\pi(\theta|\mathbf{X}_n)$ and $\tilde{q}(\cdot)$ in some set $\mathcal{Q}$ for a given $\alpha > 1$; that is,

$$q_n^*(\theta) := \mathrm{argmin}_{\tilde{q} \in \mathcal{Q}} \left\{ D_\alpha \left(\pi(\theta|\mathbf{X}_n) \| \tilde{q}(\theta) \right) := \frac{1}{\alpha - 1} \log \int_\Theta \tilde{q}(\theta) \left(\frac{\pi(\theta|\mathbf{X}_n)}{\tilde{q}(\theta)} \right)^\alpha d\theta \right\}. \tag{5}$$

Recall that

Definition 1 (Dominating distribution) *The distribution Q dominates the distribution P ($P \ll Q$), when P is absolutely continuous with respect to Q; that is, $supp(P) \subseteq supp(Q)$.*

Clearly, when $\alpha > 1$, the α−Rényi divergence in (5) is infinite for any distribution $q(\theta) \in \mathcal{Q}$ that does not dominate the true posterior distribution (Van Erven and Harremos, 2014). Intuitively, this is the reason why the α-Rényi approximation can better capture the spread of the posterior distribution.

Our goal is to study the statistical properties of the α−Rényi-approximate posterior as defined in (5). In particular, we show that under certain regularity conditions on the likelihood, the prior, and the variational family the α−Rényi-approximate posterior is consistent or converges weakly to a Dirac delta distribution at the true parameter θ_0 as the number of observations $n \to \infty$.

2.1. Asymptotic Notations

We first define asymptotic notations that frequently appear in our proofs and assumptions. We write $a_n \sim b_n$ when the sequence $\{a_n\}$ can be approximated by a sequence $\{b_n\}$ for large n, so that the ratio $\frac{a_n}{b_n}$ approaches 1 as $n \to \infty$, $a_n = O(b_n)$ as $n \to \infty$, when there exists a positive number M and $n_0 \geq 1$, such that $a_n \leq M b_n \ \forall n \geq n_0$, and $a_n \lesssim b_n$ when the sequence $\{a_n\}$ is bounded above by a sequence $\{b_n\}$ for large n.

2.2. Assumptions and Definitions

First, we assume the following restrictions on permissible priors.

Assumption 1 (Prior Density)

(1) The prior density function $\pi(\theta)$ is continuous with non-zero measure in the neighborhood of the true parameter θ_0, and

(2) there exists a constant $M_p > 0$ such that $\pi(\theta) \leq M_p \ \forall \theta \in \Theta$ and $\mathbb{E}_{\pi(\theta)}[|\theta|] < \infty$.

Assumption 1(1) is typical in Bayesian consistency analyses - quite obviously, if the prior does not place any mass around the true parameter then the (true) posterior will not either. Indeed, it is well known (Schwartz, 1965; Ghosal, 1997) that for any prior that satisfies Assumption 1(1), under very mild assumptions,

$$\pi(U|\mathbf{X}_n) = \int_U \pi(\theta|\mathbf{X}_n) d\theta \to 1 \qquad P_{\theta_0} - a.s. \text{ as } n \to \infty, \tag{6}$$

where P_{θ_0} represents the true data-generating distribution, U is some neighborhood of the true parameter θ_0. Assumption 1(2), on the other hand, is a mild technical condition which is satisfied by a large class of prior distributions, for instance, many of the exponential-family distributions. For simplicity, we write $q_n(\theta) \Rightarrow q(\theta)$ to represent weak convergence of the distributions corresponding to the densities $\{q_n\}$ and q.

We define a generic probabilistic order term, $o_{P_\theta}(1)$ with respect to measure P_θ as follows

Definition 2 *A sequence of random variables $\{\xi_n\}$ is of probabilistic order $o_{P_\theta}(1)$ when*

$$\lim_{n\to\infty} P_\theta(|\xi_n| > \delta) = 0, \text{ for any } \delta > 0 \ .$$

Next, we assume the likelihood function satisfies the following asymptotic normality property (see van der Vaart (1998) as well),

Assumption 2 (Local Asymptotic Normality) *Fix $\theta_0 \in \Theta$. The sequence of log-likelihood functions $\{\log P_n(\theta) = \sum_{i=1}^{n} \log p(x_i|\theta)\}$ satisfies a* local asymptotic normality (LAN) *condition, if there exists a sequence of matrices $\{r_n\}$, a matrix $I(\theta_0)$ and a sequence of random vectors $\{\Delta_{n,\theta_0}\}$ weakly converging to $\mathcal{N}(0, I(\theta_0)^{-1})$ as $n \to \infty$, such that for every compact set $K \subset \mathbb{R}^d$*

$$\sup_{h\in K}\left|\log P_n(\theta_0 + r_n^{-1}h) - \log P_n(\theta_0) - h^T I(\theta_0)\Delta_{n,\theta_0} + \frac{1}{2}h^T I(\theta_0)h\right| \xrightarrow{P_{\theta_0}} 0 \text{ as } n \to \infty \ .$$

The LAN condition is standard, and holds for a wide variety of models. The assumption affords significant flexibility in the analysis by allowing the likelihood to be asymptotically approximated by a scaled Gaussian centered around θ_0 (van der Vaart, 1998). We observe that Wang and Blei (2018) makes a similar assumption in their consistency analysis of the variational lower bound. All statistical models P_θ, which are differentiable in quadratic mean with respect to parameter θ, satisfy the LAN condition with $r_n = \sqrt{n}I$, where I is an identity matrix (van der Vaart, 1998, Chapter-7). Also, all models P_θ which are twice continuously differentiable in θ are also differentiable in quadratic mean and thus satisfy LAN condition, for instance most exponential family models satisfy the LAN condition.

Now, let δ_θ represent the Dirac delta, or singular distribution, concentrated at the parameter θ.

Definition 3 (Degenerate distribution) *A sequence of distributions $\{q_n(\theta)\}$ converges weakly to $\delta_{\theta'}$ that is, $q_n(\theta) \Rightarrow \delta_{\theta'}$ for some $\theta' \in \Theta$, if and only if $\forall \eta > 0$*

$$\lim_{n\to\infty} \int_{\{|\theta-\theta'|>\eta\}} q_n(\theta)d\theta = 0.$$

We use the term 'non-degenerate' for a sequence of distributions that does not converge in distribution to a Dirac delta distribution. We also use the term 'non-singular' to refer to a distribution that does not contain any singular components (i.e., it is absolutely continuous with respect to the Lebesgue measure). If a distribution contains both singularities and absolutely continuous components we term it a 'singular distribution'. More formally,

Definition 4 (Singular distributions) *Let $d(\theta)$ be a distribution with support Θ and for any $i \in \{1, \ldots, K\}$ and $K < \infty$ denote δ_{θ_i}, as the Dirac delta distributions at θ_i for any $\theta_i \in \Theta$, then we define singular distribution $q(\theta)$;*

$$q(\theta) := wd(\theta) + \sum_{i=1}^{K} w^i \delta_{\theta_i},$$

where $w, \{w^i\}_{i=1}^K \in [0, 1)$ and $w + \sum_{i=1}^K w^i = 1$ with at least one of the weights $\{w^i\}_{i=1}^K$ strictly positive.

Finally, we come to the conditions on the variational family $\mathcal{Q}$.

Assumption 3 (Variational Family) *The variational family $\mathcal{Q}$ must contain all Dirac delta distributions in some open neighborhood of $\theta_0 \in \Theta$.*

Since we know that the posterior converges weakly to a Dirac delta distribution function, this assumption is a necessary condition to ensure that the variational approximator exists in the limit. Next, we define the rate of convergence of a sequence of distributions to a Dirac delta distribution as follows.

Definition 5 (Rate of convergence) *A sequence of distributions $\{q_n(\theta)\}$ converges weakly to δ_{θ_1}, $\forall \theta_1 \in \Theta$ at the rate of γ_n if*

(1) the sequence of means $\{\check{\theta}_n := \int \theta q_n(\theta) d\theta\}$ converges to θ_1 as $n \to \infty$, and

(2) the variance of $\{q_n(\theta)\}$ satisfies

$$E_{q_n(\theta)}[|\theta - \check{\theta}_n|^2] = O\left(\frac{1}{\gamma_n^2}\right).$$

A crucial assumption, on which rests the proof of our main result, is the existence of what we call a 'good sequence' in $\mathcal{Q}$.

Assumption 4 (Good sequence) *For any $\bar{M} > 0$, the variational family $\mathcal{Q}$ contains a sequence of distributions $\{\bar{q}_n(\theta)\}$ with the following properties:*

(1) there exists $n_1 \geq 1$ such that $\int_\Theta \theta \bar{q}_n(\theta) d\theta = \hat{\theta}_n$, where $\hat{\theta}_n$ is the maximum likelihood estimate, for each $n \geq n_1$,

(2) there exists $n_{\bar{M}} \geq 1$ such that the rate of convergence is $\gamma_n = \sqrt{n}$, that is $E_{\bar{q}_n(\theta)}[|\theta - \hat{\theta}_n|^2] \leq \frac{\bar{M}}{\gamma_n^2}$ for each $n \geq n_{\bar{M}}$,

(3) there exist a compact ball $K \subset \Theta$ containing the true parameter θ_0 and $n_2 \geq 1$, such that the sequence of Radon-Nikodym derivatives of the posterior density with respect to the sequence $\{\bar{q}_n\}$ exists and is bounded above by a finite positive constant M_r outside of K for all $n \geq n_2$:

$$\frac{\pi(\theta|\mathbf{X}_n)}{\bar{q}_n(\theta)} \leq M_r, \ \forall \theta \in \Theta \backslash K \text{ and } \forall n \geq n_2, \quad P_{\theta_0} - a.s.$$

(4) there exists $n_3 \geq 1$ such that the good sequence $\{\bar{q}_n(\theta)\}$ is log-concave in θ for all $n \geq n_3$.

We term such a sequence of distributions as 'good sequences'.

The first two parts of the assumption hold so long as the variational family $\mathcal{Q}$ contains an open neighborhood of distributions around δ_{θ_0}. The third part essentially requires that for $n \geq n_2$, the tails of $\{\bar{q}_n(\theta)\}$ must decay no faster than the tails of the posterior distribution. Since, the good sequence converges weakly to δ_{θ_0}, this assumption is a mild technical condition. The last assumption implies that the good sequence is, for large sample sizes, a maximum entropy distribution under some deviation constraints on the entropy maximization problem (Grechuk et al., 2009). Note that this does not imply that the good sequence is necessarily Gaussian (which is the maximum entropy distribution specifically under standard deviation constraints).

We note that this assumption is on the family $\mathcal{Q}$, and not on the minimizer of the Rényi divergence. We demonstrate the existence of good sequences for some example models.

Example 1 *Consider a model whose likelihood is an m-dimensional multivariate Gaussian likelihood with unknown mean vector $\boldsymbol{\mu}$ and known covariance matrix $\boldsymbol{\Sigma}$. Using an m-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu_0}$ and covariance matrix $\boldsymbol{\Sigma}$ as conjugate prior, the posterior distribution is*

$$\pi(\boldsymbol{\mu}|\mathbf{X}_n) = \sqrt{\frac{(n+1)^m}{(2\pi)^m det\left(\boldsymbol{\Sigma}\right)}} e^{-\frac{n+1}{2}\left(\boldsymbol{\mu} - \frac{\sum_{i=1}^n X_i + \boldsymbol{\mu_0}}{n+1}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu} - \frac{\sum_{i=1}^n X_i + \boldsymbol{\mu_0}}{n+1}\right)},$$

where exponents 'T' and '-1' denote transpose and inverse. Next, consider the mean-field variational family, that is the product of m 1-dimensional normal distributions. Consider a sequence in the variational family with mean $\{\mu_{q_n}^j, j \in \{1, 2, \ldots, m\}\}$ and variance $\left\{\frac{\sigma_j^2}{\gamma_n^2}, j \in \{1, 2, \ldots, m\}\right\}$:

$$q_n(\boldsymbol{\mu}) = \prod_{j=1}^m \sqrt{\frac{\gamma_n^2}{2\pi\sigma_j^2}} e^{-\frac{\gamma_n^2}{2\sigma_j^2}\left(\mu_j - \mu_{q_n}^j\right)^2} = \sqrt{\frac{\gamma_n^{2m}}{(2\pi)^m det(\mathbf{I}_\sigma)}} e^{-\frac{\gamma_n^2}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_{q_n})^T \mathbf{I}_\sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_{q_n})},$$

where $\boldsymbol{\mu}_{q_n} = \{\mu_{q_n}^1, \mu_{q_n}^2, \ldots, \mu_{q_n}^m\}$ and $\mathbf{I}_\sigma$ is an $m \times m$ diagonal matrix with diagonal elements $\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2\}$. Notice that γ_n is the rate at which the sequence $\{q_n(\boldsymbol{\mu})\}$ converges weakly. It is straightforward to observe that the variational family contains sequences that satisfy properties (1) and (2) in Assumption 4, that is

$$\gamma_n = \sqrt{n} \text{ and } \boldsymbol{\mu_{q_n}} = \frac{\sum_{i=1}^n X_i + \boldsymbol{\mu_0}}{n+1}.$$

For brevity, denote $\tilde{\boldsymbol{\mu}}_n := \boldsymbol{\mu} - \boldsymbol{\mu}_{q_n} = \boldsymbol{\mu} - \frac{\sum_{i=1}^n X_i + \boldsymbol{\mu_0}}{n+1}$. To verify property (3) in Assumption 4 consider the ratio,

$$\frac{\pi(\boldsymbol{\mu}|\mathbf{X}_n)}{q_n(\boldsymbol{\mu})} = \frac{\sqrt{\frac{(n+1)^m}{(2\pi)^m det(\boldsymbol{\Sigma})}} e^{-\frac{n+1}{2}\tilde{\boldsymbol{\mu}}_n^T \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\mu}}_n}}{\sqrt{\frac{\gamma_n^{2m}}{(2\pi)^m det(\mathbf{I}_\sigma)}} e^{-\frac{\gamma_n^2}{2}\tilde{\boldsymbol{\mu}}_n^T \mathbf{I}_\sigma^{-1}\tilde{\boldsymbol{\mu}}_n}}.$$

Using the fact that $\gamma_n^2 = n < n+1$, $\frac{n+1}{\gamma_n^2} = 1 + \frac{1}{n} < 2$, therefore the ratio above can be bounded above by

$$\frac{\pi(\boldsymbol{\mu}|\mathbf{X}_n)}{q_n(\boldsymbol{\mu})} \leq \sqrt{\frac{2^m det(\mathbf{I}_\sigma)}{det\left(\boldsymbol{\Sigma}\right)}} \frac{e^{-\frac{n+1}{2}\tilde{\boldsymbol{\mu}}_n^T \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\mu}}_n}}{e^{-\frac{n+1}{2}\tilde{\boldsymbol{\mu}}_n^T \mathbf{I}_\sigma^{-1}\tilde{\boldsymbol{\mu}}_n}} = \sqrt{\frac{2^m det(\mathbf{I}_\sigma)}{det\left(\boldsymbol{\Sigma}\right)}} e^{-\frac{n+1}{2}\tilde{\boldsymbol{\mu}}_n^T(\boldsymbol{\Sigma}^{-1} - \mathbf{I}_\sigma^{-1})\tilde{\boldsymbol{\mu}}_n}.$$

Observe that if the matrix $\left(\mathbf{\Sigma}^{-1} - \mathbf{I}_\sigma^{-1}\right)$ *is positive definite then the ratio above is bounded by* $\sqrt{\frac{2^m det(\mathbf{I}_\sigma)}{det(\mathbf{\Sigma})}}$ *and if $\mathcal{Q}$ is large enough it will contain distributions that satisfy this condition. To fix the idea, consider the univariate case, where the positive definiteness implies that the variance of the good sequence is greater than the variance of the posterior for all large enough 'n'. That is, the tails of the good sequence decay slower than the tails of the posterior.*

Example 2 *Consider a model whose likelihood is a univariate normal distribution with unknown mean μ and known variance σ. Using a univariate normal distribution with the mean μ_0 and the variance σ as prior, the posterior distribution is*

$$\pi(\mu|\mathbf{X}_n) = \sqrt{\frac{n+1}{2\pi\sigma^2}} e^{-\frac{(n+1)}{2\sigma^2}\left(\mu - \frac{\mu_0 + \sum_{i=1}^n X_i}{n+1}\right)^2}. \tag{7}$$

Next, suppose the variational family $\mathcal{Q}$ is the set of all Laplace distributions. Consider a sequence $\{q_n(\mu)\}$ in $\mathcal{Q}$ with the location and the scale parameter k_n and b_n respectively, that is

$$q_n(\mu) = \frac{1}{2b_n} e^{-\frac{|\mu - k_n|}{b_n}}.$$

To satisfy properties (1) and (2) in Assumption 4, we can choose $k_n = \frac{\mu_0 + \sum_{i=1}^n X_i}{n+1}$ *and* $b_n = \sqrt{\frac{\pi\alpha^{\frac{1}{\alpha-1}}\sigma^2}{2n}}$, $\forall \alpha > 1$. *For brevity denote* $\tilde{\mu}_n = \mu - \frac{\mu_0 + \sum_{i=1}^n X_i}{n+1}$. *To verify property (3) in Assumption 4 consider the ratio,*

$$\frac{\pi(\mu|\mathbf{X}_n)}{q_n(\mu)} = \frac{\sqrt{\frac{n+1}{2\pi\sigma^2}} e^{-\frac{(n+1)}{2\sigma^2}\tilde{\mu}_n^2}}{\frac{1}{2}\sqrt{\frac{2n}{\pi\alpha^{\frac{1}{\alpha-1}}\sigma^2}} e^{-\frac{\sqrt{2n}|\tilde{\mu}_n|}{\sqrt{\pi\alpha^{\frac{1}{\alpha-1}}\sigma^2}}}} \leq \sqrt{\frac{2}{\alpha^{\frac{1}{\alpha-1}}}} \frac{e^{-\frac{(n+1)}{\pi\alpha^{\frac{1}{\alpha-1}}\sigma^2}\tilde{\mu}_n^2}}{e^{-\left|\frac{\sqrt{2(n+1)}|\tilde{\mu}_n|}{\sqrt{\pi\alpha^{\frac{1}{\alpha-1}}\sigma^2}}\right|}} \leq \sqrt{\frac{2}{\alpha^{\frac{1}{\alpha-1}}}} e^{1/2},$$

where the last inequality follows due to the fact that $e^{-(\frac{x^2}{2} - |x|)} < e^{1/2}$.

For the same posterior, we can also choose $\mathcal{Q}$ to be the set of all Logistic distributions. Consider a sequence $\{q_n(\mu)\}$ in this variational family with the mean and the scale parameter m_n and s_n respectively; that is

$$q_n(\mu) = \frac{1}{s_n}\left(e^{\frac{\mu - m_n}{2s_n}} + e^{-\frac{\mu - m_n}{2s_n}}\right)^{-2}.$$

To satisfy properties (1) and (2) in Assumption 4, we can choose $m_n = \frac{\mu_0 + \sum_{i=1}^n X_i}{n+1}$ *and* $s_n = \sqrt{\frac{2\pi\alpha^{\frac{1}{\alpha-1}}\sigma^2}{n+1}}$, $\forall \alpha > 1$. *For brevity denote* $\tilde{\mu}_n = \mu - \frac{\mu_0 + \sum_{i=1}^n X_i}{n+1}$. *To verify property (3) in Assumption 4 observe that,*

$$\frac{\pi(\lambda|\mathbf{X}_n)}{q_n(\lambda)} = \frac{\sqrt{\frac{n+1}{2\pi\sigma^2}} e^{-\frac{(n+1)}{2\sigma^2}\left(\mu - \frac{\mu_0 + \sum_{i=1}^n X_i}{n+1}\right)^2}}{\frac{1}{s_n}\left(e^{\frac{\mu - m_n}{2s_n}} + e^{-\frac{\mu - m_n}{2s_n}}\right)^{-2}} = \frac{1}{\sqrt{\alpha^{\frac{1}{\alpha-1}}}} e^{-\left(\frac{\tilde{\mu}_n}{s_n}\right)^2} \left(e^{\left(\frac{\tilde{\mu}_n}{2s_n}\right)} + e^{-\left(\frac{\tilde{\mu}_n}{2s_n}\right)}\right) \leq \frac{1}{\sqrt{\alpha^{\frac{1}{\alpha-1}}}} 2e^{1/16},$$

where the last inequality follows due to the fact that $e^{-x^2}\left(e^{x/2} + e^{-x/2}\right) < 2e^{1/16}$.

Example 3 *Consider a univariate exponential likelihood model with the unknown rate parameter λ. For some prior distribution $\pi(\lambda)$, the posterior distribution is*

$$\pi(\lambda|\mathbf{X}_n) = \frac{\pi(\lambda)\lambda^n e^{-\lambda\sum_{i=1}^n X_i}}{\int \pi(\lambda)\lambda^n e^{-\lambda\sum_{i=1}^n X_i} d\lambda}.$$

Choose $\mathcal{Q}$ to be the set of Gamma distributions. Consider a sequence $\{q_n(\mu)\}$ in the variational family with the shape and the rate parameter k_n and β_n respectively, that is

$$q_n(\lambda) = \frac{\beta_n^{k_n}}{\Gamma(k_n)}\lambda^{k_n-1}e^{-\lambda\beta_n},$$

where $\Gamma(\cdot)$ is the $\Gamma-$ function. To satisfy properties (1) and (2) in Assumption 4, we can choose $k_n = n+1$ and $\beta_n = \sum_{i=1}^n X_i$. To verify property (3) in Assumption 4 consider the ratio,

$$\frac{\pi(\lambda|\mathbf{X}_n)}{q_n(\lambda)} = \frac{\pi(\lambda)\lambda^n e^{-\lambda\sum_{i=1}^n X_i}}{\frac{\beta_n^{k_n}}{\Gamma(k_n)}\lambda^{k_n-1}e^{-\lambda\beta_n}\int \pi(\lambda)\lambda^n e^{-\lambda\sum_{i=1}^n X_i} d\lambda} = \frac{\pi(\lambda)\Gamma(n+1)}{\left(\sum_{i=1}^n X_i\right)^{n+1}\int \pi(\lambda)\lambda^n e^{-\lambda\sum_{i=1}^n X_i} d\lambda}.$$

Now, observe that $\frac{\left(\sum_{i=1}^n X_i\right)^{n+1}}{\Gamma(n+1)}\lambda^n e^{-\lambda\sum_{i=1}^n X_i}$ is the density of Gamma distribution with the mean $\frac{n+1}{\sum_{i=1}^n X_i}$ and the variance $\frac{1}{n+1}\left(\frac{n+1}{\sum_{i=1}^n X_i}\right)^2$. Since, we assumed in Assumption 1(2) that $\pi(\lambda)$ is bounded from above by M_p, therefore for large n, $\frac{\left(\sum_{i=1}^n X_i\right)^{n+1}}{\Gamma(n+1)}\int \pi(\lambda)\lambda^n e^{-\lambda\sum_{i=1}^n X_i} d\lambda \sim \pi\left(\frac{n+1}{\sum_{i=1}^n X_i}\right)$. Hence, it follows that for large enough n

$$\frac{\pi(\lambda|\mathbf{X}_n)}{q_n(\lambda)} \leq \frac{M_p}{\pi(\lambda_0)},$$

where $\frac{\sum_{i=1}^n X_i}{n+1} \to \frac{1}{\lambda_0}$ as $n \to \infty$.

3. Consistency of $\alpha-$Rényi Approximate Posterior

Recall that the $\alpha-$Rényi-approximate posterior q_n^* is defined as

$$q_n^*(\theta) := \text{argmin}_{\tilde{q}\in\mathcal{Q}}\left\{D_\alpha\left(\pi(\theta|\mathbf{X}_n)\|\tilde{q}(\theta)\right) := \frac{1}{\alpha-1}\log\int_\Theta \tilde{q}(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{\tilde{q}(\theta)}\right)^\alpha d\theta\right\}. \tag{8}$$

We now show that under the assumptions in the previous section, the $\alpha-$Rényi approximators are asymptotically consistent as the sample size increases in the sense that $q_n^* \Rightarrow \delta_{\theta_0}$ in -P_{θ_0} probability as $n \to \infty$. To illustrate the ideas clearly, we present our analysis assuming a univariate parameter space, and that the model P_θ is twice differentiable in parameter θ, and therefore satisfies the LAN condition with $r_n = \sqrt{n}$ (van der Vaart, 1998). The LAN condition together with the existence of a sequence of test functions (van der Vaart, 1998, Theorem 10.1) also implies that the posterior distribution converges weakly to δ_{θ_0} at the rate of $\sqrt{n}$. The analysis can be easily adapted to multivariate parameter spaces.

We will first establish some structural properties of the minimizing sequence of distributions. We show that for any sequence of distributions converging weakly to a non-singular distribution the $\alpha-$Rényi divergence is unbounded in the limit.

Lemma 6 *Under Assumptions 1, 2, 3, and 4, the $\alpha-$Rényi divergence between the true posterior and the sequence $\{q_n(\theta)\} \subset \mathcal{Q}$ can only be finite in the limit if $q_n(\theta)$ converges weakly to a singular distribution $q(\theta)$ with a Dirac delta distribution at the true parameter θ_0.*

The result above implies that the $\alpha-$Rényi approximate posterior must have a Dirac delta distribution component at θ_0 in the limit; that is, it should converge in distribution to δ_{θ_0} or a convex combination of δ_{θ_0} with singular or non-singular distributions as $n \to \infty$. Next, we consider a sequence $\{q'_n(\theta)\} \subset \mathcal{Q}$ that converges weakly to a convex combination of δ_{θ_0} and singular or non-singular distributions $q_i(\theta),\ i \in \{1, 2, \dots\}$ such that for weights $\{w^i \in (0,1) : \sum_{i=1}^{\infty} w^i = 1\}$,

$$q'_n(\theta) \Rightarrow w^j \delta_{\theta_0} + \sum_{i=1, i\neq j}^{\infty} w^i q_i(\theta). \tag{9}$$

In the following result, we show that the $\alpha-$Rényi divergence between the true posterior and the sequence $\{q'_n(\theta)\}$ is bounded below by a positive number.

Lemma 7 *Under Assumption 1, the $\alpha-$Rényi divergence between the true posterior and the sequence $\{q'_n(\theta) \in \mathcal{Q}\}$ is bounded away from zero; that is*

$$\liminf_{n\to\infty} D_\alpha(\pi(\theta|\mathbf{X}_n) \| q'_n(\theta)) \geq \eta > 0 \quad P_{\theta_0} - a.s.$$

We also show in Lemma 22 in the appendix that if in (9) the components $\{q_i(\theta)\ i \in \{1, 2, \dots\}\}$ are singular, then with w^j is the weight of δ_{θ_0}, we have

$$\liminf_{n\to\infty} D_\alpha(\pi(\theta|\mathbf{X}_n) \| q'_n(\theta)) \geq 2(1 - w^j)^2 > 0 \quad P_{\theta_0} - a.s.$$

A consistent sequence asymptotically achieves zero $\alpha-$Rényi divergence. To show its existence, we first provide an asymptotic upper-bound on the minimal $\alpha-$Rényi divergence in the next proposition. This, coupled with the previous two structural results, will allow us to prove the consistency of the minimizing sequence.

Proposition 8 *For a given $\alpha > 1$ and under Assumptions 1, 2, 3, and 4, for any good sequence $\bar{q}_n(\theta)$ there exist $n_0 \geq 1$ and $\bar{M} > 0$ such that for all $n \geq n_0$, the minimal $\alpha-$Rényi divergence satisfies*

$$\min_{q\in\mathcal{Q}} D_\alpha(\pi(\theta|\mathbf{X}_n) \| q(\theta)) \leq D_\alpha(\pi(\theta|\mathbf{X}_n) \| \bar{q}_n(\theta)) \leq B = \frac{1}{2}\log\left(\frac{\bar{e}\bar{M} I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}}\right) + o_{P_{\theta_0}}(1), \tag{10}$$

where $I(\theta_0)$ is defined in Assumption 2 and $\bar{e}$ is the Euler's constant.

Now Proposition 8, Lemma 6, and Lemma 7 allow us to prove our main result that the $\alpha-$Rényi approximate posterior converges weakly to δ_{θ_0}.

Theorem 9 *Under Assumptions 1, 2, 3, and 4, the $\alpha-$Rényi approximate posterior $q_n^*(\theta)$ converges weakly to a Dirac delta distribution at the true parameter θ_0; that is,*

$$q_n^* \Rightarrow \delta_{\theta_0} \quad \textit{in-}P_{\theta_0} \textit{ probability as } n \to \infty.$$

Proof First, we argue that there always exists a sequence $\{\tilde{q}_n(\theta)\} \subset \mathcal{Q}$ such that for every $\eta > 0$

$$\lim_{n\to\infty} P_{\theta_0}\left(D_\alpha(\pi(\theta|\mathbf{X}_n)\|\tilde{q}_n(\theta)) \leq \eta\right) = 1.$$

We demonstrate the existence of $\tilde{q}_n(\theta)$ by construction. Recall from Proposition 8(2) that there exist $0 < \bar{M} < \infty$ and $n_0 \geq 1$, such that for all $n \geq n_0$

$$D_\alpha(\pi(\theta|\mathbf{X}_n)\|\bar{q}_n(\theta)) \leq \frac{1}{2}\log\frac{\bar{e}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}} + o_{P_{\theta_0}}(1),$$

where $\bar{q}_n(\theta)$ is the good sequence as defined in Assumption 4 and $\bar{e}$ is the Euler's constant. Now using the definition of $o_{P_{\theta_0}}(1)$, for every $\eta > 0$, it follows from the inequality above that

$$\lim_{n\to\infty} P_{\theta_0}\left(D_\alpha(\pi(\theta|\mathbf{X}_n)\|\bar{q}_n(\theta)) - \frac{1}{2}\log\frac{\bar{e}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}} > \eta\right) \leq \lim_{n\to\infty} P_{\theta_0}\left(o_{P_{\theta_0}}(1) > \eta\right) = 0. \tag{11}$$

Now a specific good sequence can be chosen by fixing $\bar{M} = \tilde{M} := \frac{\alpha^{\frac{1}{\alpha-1}}}{\bar{e}I(\theta_0)}$, implying that

$$\lim_{n\to\infty} P_{\theta_0}\left(D_\alpha(\pi(\theta|\mathbf{X}_n)\|\tilde{q}_n(\theta)) > \eta\right) = 0. \tag{12}$$

The above result implies that there exist a sequence in family $\mathcal{Q}$ such that $D_\alpha(\pi(\theta|\mathbf{X}_n)\|\tilde{q}_n(\theta)) \to 0$ in P_{θ_0} -probability.
Next, we will show that the minimizing sequence must converge to a Dirac delta distribution in probability. The previous result shows that the minimizing sequence must have zero α-Rényi divergence in the limit. Lemma 6 shows that the minimizing sequence must have a delta at θ_0, since otherwise the α-Rényi divergence is unbounded. Similarly, Lemma 7 shows that it cannot be a mixture of such a delta with other components, since otherwise the α-Rényi divergence is bounded away from zero.
Therefore, it follows that the $\alpha-$Rényi approximate posterior $q_n^*(\theta)$ must converge weakly to a Dirac delta distribution at the true parameter θ_0, in $-P_{\theta_0}$ probability, thereby completing the proof. ■

Note that the choice of $\bar{M}$ in the proof essentially determines the variance of the good sequence. As noted before, the asymptotic log-concavity of the good sequence implies that it is eventually an entropy maximizing sequence of distributions (Grechuk et al., 2009). It does not necessarily follow that the sequence is Gaussian, however. If such a choice can be made (i.e., the variational family contains Gaussian distributions) then the choice of good sequence amounts to matching the entropy of a Gaussian distribution with variance $\frac{\alpha^{\frac{1}{\alpha-1}}}{\bar{e}I(\theta_0)}$.

We further characterize the rate of convergence of the $\alpha-$Rényi approximate posterior under additional regularity conditions. In particular, we establish an upper bound on the rate of convergence of the possible candidate $\alpha-$Rényi approximators when the variational family is sub-Gaussian. Additionally, we require that the posterior distribution satisfies the Bernstein-von Mises Theorem, that is for any compact set K containing θ_0

$$\int_K \pi(\theta|\mathbf{X}_n)d\theta = \int_K \mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})d\theta + o_{P_{\theta_0}}(1). \tag{13}$$

According to Theorem 10.1 in van der Vaart (1998), the Bernstein-von Mises Theorem holds under Assumption 1, 2, and the following additional assumption on the existence of consistent test functions:

Assumption 5 (Consistent Tests) *For every $\epsilon > 0$ there exists a sequence of tests $\phi_n(\mathbf{X}_n)$ such that i) $\lim_{n\to\infty}\mathbb{E}_{P_{\theta_0}}(\phi_n(\mathbf{X}_n)) = 0$, and $\lim_{n\to\infty}\sup_{\|\theta-\theta_0\|\geq\epsilon}\mathbb{E}_{P_{\theta_0}}(1-\phi_n(\mathbf{X}_n)) = 0$.*

A further modeling assumption is to choose a sub-Gaussian variational family $\mathcal{Q}$ that limits the variance. We choose a sub-Gaussian sequence of distributions $\{q_n(\theta)\}\subset\mathcal{Q}$, that is for some positive constant B and any $t\in\mathbb{R}$,

$$\mathbb{E}_{q_n(\theta)}[e^{t\theta}] \leq e^{\tilde{\theta}_n t + \frac{B}{2\gamma_n^2}t^2}, \tag{14}$$

where $\tilde{\theta}_n$ is the mean of $q_n(\theta)$ and γ_n is the rate (see Definition 5) at which $q_n(\theta)$ converges weakly to a Dirac delta distribution as $n\to\infty$.

Lemma 10 *Consider a sequence of sub-Gaussian distributions $\{q_n(\theta)\}\subset\mathcal{Q}$, with parameters B and t, that converges weakly to some Dirac delta distribution faster than the posterior converges weakly to δ_{θ_0} (that is, $\gamma_n > \sqrt{n}$), and suppose the true posterior distribution satisfies the Bernstein-von Mises Theorem* (13). *Then, there exists an $n_0 \geq 1$ such that the $\alpha-$Rényi divergence $D_\alpha(\pi(\theta|\mathbf{X}_n)\|q_n(\theta))$ is infinite for all $n > n_0$.*

We use the above result to show that, when the variational family $\mathcal{Q}$ is sub-Gaussian, then the $\alpha-$Rényi appropriate posterior cannot converge at a rate γ_n faster than $\sqrt{n}$, that is the rate at which the posterior converges weakly to δ_{θ_0}.

Theorem 11 *Under Assumptions 1, 2, 3, 4, and 5, and $\mathcal{Q}$ is a family of sub-Gaussian distribution, then the rate of convergence, γ_n, of $\alpha-$Rényi approximate posterior is bounded above by $\sqrt{n}$, that is $\gamma_n \leq \sqrt{n}$.*

Proof Since we choose the variational family to be sub-Gaussian, the $\alpha-$Rényi approximate posterior must be one of the sequences satisfying (14) and as a consequence of Theorem 9, $\tilde{\theta}_n$ must converge to θ_0 as $n\to\infty$. On the other hand, using Lemma 10, it follows that the rate of convergence γ_n of $\alpha-$Rényi approximate posterior must be bounded above by $\sqrt{n}$, that is $\gamma_n \leq \sqrt{n}$. ■

4. Consistency of α-Rényi Approximate Posterior as $\alpha \to 1$

Our results on the consistency of α-Rényi variational approximators in Section 3 can be a step forward in understanding the consistency of posterior approximations obtained using expectation propogation (EP) (Minka, 2001a,b). Observe that for any $n \geq 1$, as $\alpha \to 1$,

$$D_\alpha\left(\pi(\theta|\mathbf{X}_n)\|\tilde{q}(\theta)\right) \to \mathrm{KL}\left(\pi(\theta|\mathbf{X}_n)\|\tilde{q}(\theta)\right), \tag{15}$$

where the limit is the EP objective using KL divergence. We define the 1-Rényi-approximate posterior s_n^* as the distribution in the variational family $\mathcal{Q}$ that minimizes the KL divergence between $\pi(\theta|\mathbf{X}_n)$ and $\tilde{s}(\theta)$, where $\tilde{s}(\theta)$ is an element of $\mathcal{Q}$:

$$s_n^*(\theta) := \mathrm{argmin}_{\tilde{s}\in\mathcal{Q}} \left\{ \mathrm{KL}\left(\pi(\theta|\mathbf{X}_n)\|\tilde{s}(\theta)\right) := \int_\Theta \pi(\theta|\mathbf{X}_n) \log\left(\frac{\pi(\theta|\mathbf{X}_n)}{\tilde{s}(\theta)}\right) d\theta \right\}. \tag{16}$$

We note that the EP algorithm (Minka, 2001a) is a message-passing algorithm that optimizes an approximations to this objective (Wainwright and Jordan, 2008). Nevertheless, understanding this idealized objective is an important step towards understanding the actual EP algorithm. Furthermore, ideas from Li and Turner (2016) can be used to construct alternate algorithms that directly minimize (16). We thus focus on this objective, and show that under the assumptions in Section 2, the 1-Rényi-approximate posterior is asymptotically consistent as the sample size increases, in the sense that $s_n^* \Rightarrow \delta_{\theta_0}$, in-$P_{\theta_0}$ probability as $n \to \infty$. The proofs in this section are corollaries of the results in the previous section.

Recall that the KL divergence lower-bounds the $\alpha-$Rényi divergence when $\alpha > 1$; that is

$$\mathrm{KL}\left(p(\theta)\|q(\theta)\right) \leq D_\alpha\left(p(\theta)\|q(\theta)\right). \tag{17}$$

This is a direct consequence of Jensen's inequality. Analogous to Proposition 8, we first show that the minimal KL divergence between the true Bayesian posterior and the variational family $\mathcal{Q}$ is asymptotically bounded.

Proposition 12 *For a given $\alpha > 1$, and under Assumptions 1, 2, 3, 4, and for any good sequence $\bar{q}_n(\theta)$ there exist $n_0 \geq 1$ and $\bar{M} > 0$ such that the minimal* KL *divergence satisfies*

$$\min_{\tilde{s}\in\mathcal{Q}} \mathrm{KL}\left(\pi(\theta|\mathbf{X}_n)\|\tilde{s}(\theta)\right) < B = \frac{1}{2}\log\left(\frac{\bar{e}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}}\right) + o_{P_{\theta_0}}(1). \tag{18}$$

where $I(\theta_0)$ is defined in Assumption 2 and $\bar{e}$ is the Euler's constant.

Proof The result follows immediately from Proposition 8 and (17), since for any $\tilde{s}(\theta) \in \mathcal{Q}$ and $\alpha > 1$,

$$\mathrm{KL}\left(\pi(\theta|\mathbf{X}_n)\|\tilde{s}(\theta)\right) \leq D_\alpha\left(\pi(\theta|\mathbf{X}_n)\|\tilde{s}(\theta)\right).$$

■

Next, we demonstrate that any sequence of distributions $\{s_n(\theta)\} \subset \mathcal{Q}$ that converges weakly to a distribution $s(\theta) \in \mathcal{Q}$ with positive probability outside the true parameter θ_0 cannot achieve zero KL divergence in the limit. Observe that this result is weaker than Lemma 6, and does not show that the KL divergence is necessarily infinite in the limit. This loses some structural insight.

Lemma 13 *There exists an $\eta > 0$ in the extended real line such that the* KL *divergence between the true posterior and sequence $\{s_n(\theta)\}$ is bounded away from zero; that is,*

$$\liminf_{n\to\infty} \mathrm{KL}(\pi(\theta|\mathbf{X}_n)\|s_n(\theta)) \geq \eta > 0 \quad P_{\theta_0} - a.s.$$

Now using Proposition 12 and Lemma 13 we show that the 1-Rényi-approximate posterior converges weakly to the δ_{θ_0}.

Theorem 14 *Under Assumptions 1, 2, 3, and 4, the 1-Rényi-approximate posterior $s^*_n(\theta)$ satisfies*

$$s^*_n \Rightarrow \delta_{\theta_0} \qquad \text{in-}P_{\theta_0} \text{ probability as } n \to \infty.$$

Proof Recall (12) from the proof of Theorem 9 that there exists a good sequence $\tilde{q}_n(\theta)$, such that

$$D_\alpha(\pi(\theta|\mathbf{X}_n)\|\tilde{q}_n(\theta)) \to 0 \text{ in-}P_{\theta_0} \text{ probability as } n \to \infty.$$

Since the KL divergence is always non-negative, using (17) it follows that

$$\mathrm{KL}(\pi(\theta|\mathbf{X}_n)\|\tilde{q}_n(\theta)) \to 0 \text{ in-}P_{\theta_0} \text{ probability as } n \to \infty.$$

Consequently, the sequence of 1-Rényi-approximate posteriors must also achieve zero KL divergence from the true posterior in the large sample limit with high probability. Finally, as demonstrated in Lemma 13, any other sequence of distribution that converges weakly to a distribution, that has positive probability at any point other that θ_0 cannot achieve zero KL divergence. Therefore, it follows that the 1-Rényi-approximate posterior $s^*_n(\theta)$ must converge weakly to a Dirac delta distribution at the true parameter θ_0, in-P_{θ_0} probability as $n \to \infty$, thereby completing the proof. ■

5. Models with Local Latent Parameters

We generalize the model we have worked with so far to include a collection of n independent local latent variables $z_{1:n} := \{z_1, z_2, \ldots, z_n\} \in \mathcal{Z}^n$, one for each observation ξ_i. We assume these are distributed as $\pi(z_i|\theta)$ for each i, with the observations distributed as $p(\xi_i|z_i, \theta)$. Recall that θ is the global latent variable with prior distribution $\pi(\theta)$. Denote by z_0 and θ_0 the true local and global latent parameters respectively. For brevity we denote the model P_{θ_0,z_0} as P_0. The posterior distribution over θ and $z_{1:n}$ is defined as

$$\pi(\theta, z_{1:n}|\mathbf{X}_n) := \frac{\pi(\theta)\prod_{i=1}^n \pi(z_i|\theta)p(\xi_i|z_i,\theta)}{\int\int \pi(\theta)\prod_{i=1}^n \pi(z_i|\theta)p(\xi_i|z_i,\theta)d\theta dz_{1:n}}.$$

We denote the denominator above as $P(\mathbf{X}_n)$, the model *evidence*, and the numerator as $p(\theta, \mathbf{X}_n, z_{1:n})$. Since computing $P(\mathbf{X}_n)$ is difficult, an approximate posterior can be obtained by minimizing the following objective over an appropriately chosen variational family $\mathcal{Q}$:

$$D_\alpha\left(\pi(\theta, z_{1:n}|\mathbf{X}_n)\|q(\theta, z_{1:n})\right) := \frac{1}{\alpha-1}\log\int_{\Theta\times\mathcal{Z}^n} q(\theta, z_{1:n})\left(\frac{\pi(\theta, z_{1:n}|\mathbf{X}_n)}{q(\theta, z_{1:n})}\right)^\alpha d\theta dz_{1:n}, \text{ where } \alpha > 1.$$

This objective can be derived as an upper-bound to the model evidence similar to (4). It is common to assume that the variational family $\mathcal{Q}$ factorizes into components $\mathcal{Q}^n$ (over local variables) and $\bar{\mathcal{Q}}$ (over θ). Define the Rényi approximate posterior over the global parameter θ as

$$q_n^*(\theta) := \text{argmin}_{q(\theta)\in\bar{\mathcal{Q}}} \min_{q(z_{1:n})\in\mathcal{Q}^n} \log \int_{\Theta\times\mathcal{Z}^n} q(\theta)q(z_{1:n}) \left(\frac{p(\theta, z_{1:n}, \mathbf{X}_n)}{q(\theta)q(z_{1:n})}\right)^{\alpha} d\theta dz_{1:n}. \tag{19}$$

In this section, we aim to show that $q_n^*(\theta)$ converges weakly to the Dirac delta distribution at θ_0. To show this we require some additional assumptions. First, define the profile likelihood at $\theta = \theta_0 + n^{-1/2}h_n$ for any bounded and stochastic $h_n = O_{P_0}(1)$ as $p(\mathbf{X}_n|\theta_0 + n^{-1/2}h_n, z_{1:n}^p)$, where $z_{1:n}^p = \text{argmax}_{z_{1:n}} p(\mathbf{X}_n|\theta_0 + n^{-1/2}h_n, z_{1:n})$ is the maximum profile likelihood estimate of $z_{1:n}$ at $\theta = \theta_0 + n^{-1/2}h_n$. Denote $d_H(z_{1:n}, z_{1:n}^p) := H(P_{\theta_0,z_{1:n}}, P_{\theta_0,z_{1:n}^p})$ as the Helinger distance between models $P_{\theta_0,z_{1:n}}$ and $P_{\theta_0,z_{1:n}^p}$ Furthermore, for any $\rho > 0$ and for all bounded and stochastic $h_n = O_{P_0}(1)$, define $D(\theta_0 + n^{-1/2}h_n, \rho) = \{z_{1:n} : d_H(z_{1:n}, z_{1:n}^p) < \rho\}$ as the Hellinger ball of radius ρ around $z_{1:n}^p$.

Next we impose regularity conditions on the conditioned posterior $p(z_{1:n}|\mathbf{X}_n, \theta_0)$. The assumption below follows Wang and Blei (2018, Proposition 10), and is motivated by Bickel and Kleijn (2012, Theorem 4.2).

Assumption 6 (Conditioned latent posterior) *The conditioned latent posterior $p(z_{1:n}|\mathbf{X}_n, \theta_0)$ satisfies*

1. *The conditioned latent posterior is consistent under $n^{-1/2}$-perturbation at some rate ρ_n with $\rho_n \downarrow 0$ and $n\rho_n^2 \to \infty$, that is, for all bounded, stochastic $h_n = O_{P_0}(1)$, $p(z_{1:n}|\mathbf{X}_n, \theta_0)$ converges as*

$$\int_{D^c(\theta_0+n^{-1/2}h_n,\rho_n)} p(z_{1:n}|\mathbf{X}_n, \theta = \theta_0 + n^{-1/2}h_n)dz_{1:n} = o_{P_0}(1).$$

2. *The sequence $\{\rho_n\}$ as defined above should also satisfy the following conditions for all bounded and stochastic $h_n = O_{P_0}(1)$:*

$$(i) \sup_{z_{1:n}\in\{z_{1:n}:d_H(z_{1:n},z_{1:n}^p)<\rho_n\}} \mathbb{E}_{P_{\theta_0,z_{1:n}}}\left[\frac{p(\mathbf{X}_n|z_{1:n}, \theta_0 + n^{-1/2}h_n)}{p(\mathbf{X}_n|z_{1:n}, \theta_0)}\right] = O(1), \quad (ii)\ d_H(z_0, z_{1:n}^p) = o(\rho_n).$$

The first condition ensures that conditioned latent posterior converges slower than the true posterior and the second condition is an additional regularity condition on the expected likelihood ratio. Bickel and Kleijn (2012, Lemma 4.3) identifies mild differentiablity conditions on the likelihood ratio that imply condition 2(i) above. Also, Theorem 3.1 in Bickel and Kleijn (2012) provide the regularity conditions under which the the conditioned latent posterior satisfies the first condition above.

The next assumption, adapted from Bickel and Kleijn (2012), is an extension of LAN condition in Assumption 2 to models with both global and local latent parameters.

Assumption 7 (Stochastic LAN (s-LAN)) *Fix $\theta_0 \in \Theta$ and recall that $z^p_{1:n}$ is the profile likelihood maximizer. The sequence of log-likelihood functions $\{P^n_{\theta_0, z^p_{1:n}} := p(\mathbf{X}_n|\theta_0, z^p_{1:n})\}$ satisfies* stochastic local asymptotic normality (s-LAN) *condition if there exists a matrix $I(\theta_0, z_0)$ and a sequence of random vectors $\{\Delta_{n,(\theta_0,z_0)}\} \in L_2(P^n_{\theta_0,z_{1:n}})$ such that for every bounded and stochastic sequence $\{h_n\}$, that is $h_n = O_{P_0}(1)$, we have*

$$\log \frac{P^n_{\theta_0+n^{-1/2}h_n, z^p_{1:n}}}{P^n_{\theta_0, z^p_{1:n}}} = h_n^T I(\theta_0, z_0)\Delta_{n,(\theta_0,z_0)} - \frac{1}{2} h_n^T I(\theta_0, z_0) h_n + o_{P_0}(1),$$

where $P_0 = P_{\theta_0, z_0}$.

Stochastic LAN is slightly stronger than the usual LAN property. In most of the examples, the ordinary LAN property often extends to stochastic LAN without significant difficulties (Bickel and Kleijn, 2012). Also, Theorem 1 in Murphy and van der Vaart (2000) identifies conditions under which the above LAN assumption is satisfied by models with both global and local latent variables. It must be noted that if $\hat{\theta}_n$ is an asymptotically efficient estimator of θ_0, then according to Lemma 25.25 in van der Vaart (1998) $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \Delta_{n,(\theta_0,z_0)} + o_{P_0}(1)$.

Next we state a modified version of Assumption 4(3) for the models that contain local latent variables:

Assumption 8 (Good Sequence-Local) *For any $\bar{M} > 0$, the variational family $\bar{\mathcal{Q}}$ contains a sequence of distributions $\{\bar{q}_n(\theta)\}$ with the following properties:*

(1) there exists $n_1 \geq 1$ such that $\int_\Theta \theta \bar{q}_n(\theta) d\theta = \hat{\theta}_n$, where $\hat{\theta}_n$ is the maximum likelihood estimate, for each $n \geq n_1$,

(2) there exists $n_{\bar{M}} \geq 1$ such that the rate of convergence is $\gamma_n = \sqrt{n}$, that is $E_{\bar{q}_n(\theta)}[|\theta - \hat{\theta}_n|^2] \leq \frac{\bar{M}}{\gamma_n^2}$ for each $n \geq n_{\bar{M}}$,

(3) there exist a compact ball $K \subset \Theta$ containing the true parameter θ_0 and $n_2 \geq 1$, such that the sequence of Radon-Nikodym derivatives of the Bayes posterior density with respect to the sequence $\{\bar{q}_n\}$ exists and is bounded above by a finite positive constant M_r outside of K for all $n \geq n_2$; that is,

$$\frac{\pi(\theta|\mathbf{X}_n, z^0_{1:n})}{\bar{q}_n(\theta)} \leq M_r, \ \forall \theta \in \Theta \backslash K \text{ and } \forall n \geq n_2, \quad P_{\theta_0} - a.s,$$

where $z^0_{1:n}$ is the first n components of the true local latent parameter z_0.

(4) there exists $n_3 \geq 1$ such that the good sequence $\{\bar{q}_n(\theta)\}$ is log-concave in θ for all $n \geq n_3$.

Example 4 (Bayesian mixture model) *Consider a mixture of uncorrelated L univariate Gaussians, each with mean $\mu_i, i \in \{1, 2, \ldots, L\}$ and unit variance. Each observation X_i*

is assumed to be generated using the following model:

$$\mu_l \sim \pi, \forall l \in \{1,2,\dots,L\}$$
$$z_i \sim Categorical\left(\frac{1}{L},\frac{1}{L},\dots,\frac{1}{L}\right), \forall i \in \{1,2,\dots,n\}$$
$$X_i \sim \mathcal{N}(z_i^T\boldsymbol{\mu},1)\forall i \in \{1,2,\dots,n\}$$

Notice that $\boldsymbol{\mu}$ is the global and $z_{1:n}$ are the local latent parameters. Now observe that

$$\begin{aligned}
\pi(\boldsymbol{\mu}|\mathbf{X}_n, z^0_{1:n}) = \frac{\prod_{l=1}^L \pi(\mu_l)\prod_{i=1}^n p(z_i^0, X_i|\boldsymbol{\mu})}{\int \prod_{l=1}^L \pi(\mu_l)\prod_{i=1}^n p(z_i^0, X_i|\boldsymbol{\mu})d\boldsymbol{\mu}} &= \frac{\prod_{l=1}^L \pi(\mu_l)\prod_{i=1}^n p(X_i|\boldsymbol{\mu}, z_i^0)}{\int \prod_{l=1}^L \pi(\mu_l)\prod_{i=1}^n p(X_i|\boldsymbol{\mu}, z_i^0)d\boldsymbol{\mu}} \\
&= \frac{\prod_{l=1}^L \pi(\mu_l)\prod_{i=1}^n \mathcal{N}(X_i|\boldsymbol{\mu}^T z_i^0, 1)}{\int \prod_{l=1}^L \pi(\mu_l)\prod_{i=1}^n \mathcal{N}(X_i|\boldsymbol{\mu}^T z_i^0, 1)d\boldsymbol{\mu}} \qquad (20) \\
&= \frac{\prod_{l=1}^L \left[\pi(\mu_l)\prod_{j=1}^{n_l} \mathcal{N}(X_j^l|\mu_l, 1)\right]}{\int \prod_{l=1}^L \pi(\mu_l)\prod_{j=1}^{n_l} \mathcal{N}(X_j^l|\mu_l, 1)d\boldsymbol{\mu}}, \qquad (21)
\end{aligned}$$

where X_j^l is the j^{th} observation in the l^{th} cluster and $n_l = \sum_{i=1}^n z^0_{i,l}$ is the total number of observations in the l^{th} cluster. In practice, $\pi(\mu_l) = \mathcal{N}(\mu_l|m,\sigma^2)\}$ is assumed to be a conjugate Gaussian with known mean m and variance σ^2. In this case, the distribution in (21) can be computed analytically, that is

$$\pi(\mu|\mathbf{X}_n, z^0_{1:n}) = \frac{\prod_{l=1}^L \pi(\mu_l)\prod_{i=1}^n p(z_i^0, X_i|\mu)}{\int \prod_{l=1}^L \pi(\mu_l)\prod_{i=1}^n p(z_i^0, X_i|\mu)d\mu} = \prod_{l=1}^L \mathcal{N}\left(\mu_l \middle| \frac{1}{\frac{1}{\sigma^2}+n_l}\left(\frac{m}{\sigma^2}+\sum_{j=1}^{n_l} X_j^l\right), \left(\frac{1}{\sigma^2}+n_l\right)^{-1}\right).$$

In practice $\bar{\mathcal{Q}}$ is chosen to be a mean-field approximate family, viz. a product of L univariate Gaussians. Now consider the following sequence of distributions in $\bar{\mathcal{Q}}$

$$q_n(\mu) = \prod_{l=1}^L \mathcal{N}\left(\mu_l|m_{n,l},\sigma^2_{n,l}\right).$$

Choosing $m_{n,l} = \frac{1}{\frac{1}{\sigma^2}+n_l}\left(\frac{m}{\sigma^2}+\sum_{j=1}^{n_l} X_j^l\right)$ and $\sigma^2_{n,l} = \left(\frac{1}{\sigma^2}+n_l\right)^{-1}$, the ratio $\frac{\pi(\mu|\mathbf{X}_n, z^0_{1:n})}{\bar{q}_n(\theta)}$ is bounded by 1.

The s-LAN assumption for finite mixtures model follows from the finiteness of the support of local latent variables (Murphy and van der Vaart, 1996, 2000).

In the next result we show that a consistent sequence asymptotically achieves zero $\alpha-$Rényi divergence. To show its existence, we first provide an asymptotic upper-bound on the minimum of the LHS in (25) in the next proposition. This will allow us to prove the consistency of the minimizing sequence.

Proposition 15 *For a given $\alpha > 1$ and under Assumptions 1, 3 (for $\bar{\mathcal{Q}}$), 6, 7, 8, and for any good sequence there exist $n_0 \geq 1$ and $\bar{M} > 0$ such that for all $n \geq n_0$, the minimal*

$\alpha-$Rényi divergence satisfies

$$\min_{q\in\bar{\mathcal{Q}}}\min_{q(z_{1:n})\in\mathcal{Q}^n} D_\alpha(\pi(\theta,z_{1:n}|\mathbf{X}_n)\|q(\theta)q(z_{1:n})) \le \min_{q(z_{1:n})\in\mathcal{Q}^n} D_\alpha(\pi(\theta,z_{1:n}|\mathbf{X}_n)\|\bar{q}_n(\theta)q(z_{1:n}))$$
$$\le B = \frac{1}{2}\log\left(\frac{\bar{e}\bar{M}I(\theta_0,z_0)}{\alpha^{\frac{1}{\alpha-1}}}\right) + o_{P_0}(1) \qquad (22)$$

where $\bar{e}$ is the Euler's constant and $I(\theta_0,z_0)$ is as defined in Assumption 7.

Since the term on the RHS above in (22) is non-negative for all $n\ge n_0$, implying that $\bar{M}\ge\frac{\alpha^{\frac{1}{\alpha-1}}}{\bar{e}I(\theta_0,z_0)}$ for all $n\ge n_0$. Therefore, a specific good sequence can be chosen by fixing $\tilde{M}=\frac{\alpha^{\frac{1}{\alpha-1}}}{\bar{e}I(\theta_0,z_0)}$, implying that $\limsup_{n\to\infty}\min_{q(z_{1:n})\in\mathcal{Q}^n} D_\alpha(\pi(\theta,z_{1:n}|\mathbf{X}_n)\|\tilde{q}_n(\theta)q(z_{1:n})) = 0\ \forall n\ge n_0$. Now analogous to the parametric case we are only left to show that the global Rényi approximator necessarily converges to a Dirac delta distribution concentrated at the true global parameter θ_0 to achieve zero Rényi divergence.

Now notice that for any $n\ge 1$,

$$\begin{aligned}
&\min_{q(z_{1:n})\in\mathcal{Q}^n}\log\int_\Theta q(\theta)\left(\frac{\pi(\theta)}{q(\theta)}\right)^\alpha\int_{\mathcal{Z}^n} q(z_{1:n})\left(\frac{p(z_{1:n},\mathbf{X}_n|\theta)}{q(z_{1:n})}\right)^\alpha dz_{1:n}d\theta\\
&\ge\log\int_\Theta q(\theta)\left(\frac{\pi(\theta)}{q(\theta)}\right)^\alpha\min_{q(z_{1:n})\in\mathcal{Q}^n}\int_{\mathcal{Z}^n} q(z_{1:n})\left(\frac{p(z_{1:n},\mathbf{X}_n|\theta)}{q(z_{1:n})}\right)^\alpha dz_{1:n}d\theta\\
&=\log\int_\Theta q(\theta)\left(\frac{\pi(\theta)M(\mathbf{X}_n|\theta)}{q(\theta)}\right)^\alpha d\theta, \qquad (23)
\end{aligned}$$

where $M(\mathbf{X}_n|\theta)$ is the variational likelihood define as

$$M(\mathbf{X}_n|\theta) := \left[\min_{q(z_{1:n})\in\mathcal{Q}^n}\int_{\mathcal{Z}^n} q(z_{1:n})\left(\frac{p(z_{1:n},\mathbf{X}_n|\theta)}{q(z_{1:n})}\right)^\alpha dz_{1:n}\right]^{1/\alpha}. \qquad (24)$$

Observe that subtracting the $\log P(\mathbf{X}_n)^\alpha$ from either side of (23) yields:

$$\min_{q(z_{1:n})\in\mathcal{Q}^n} D_\alpha(\pi(\theta,z_{1:n}|\mathbf{X}_n)\|q(\theta)q(z_{1:n})) \ge D_\alpha(\pi^*(\theta|\mathbf{X}_n)\|q(\theta)), \qquad (25)$$

where the ideal posterior $\pi^*(\theta|\mathbf{X}_n)$ is defined as

$$\pi^*(\theta|\mathbf{X}_n) := \frac{\pi(\theta)M(\mathbf{X}_n|\theta)}{\int\pi(\theta)M(\mathbf{X}_n|\theta)d\theta}. \qquad (26)$$

In the subsequent lemma we show that under certain regularity conditions $M(\mathbf{X}_n|\theta)$ satisfies the LAN condition with the similar expansion as of the true likelihood model for a given local latent parameter z_0. The proof parallels that of Wang and Blei (2018, Proposition 10).

Lemma 16 *Fix $\theta \in \Theta$. Under Assumptions 6 and 7, the sequence of variational log-likelihood functions $\{M_n(\theta) := \log M(\mathbf{X}_n|\theta)$ satisfies* s-LAN *condition, that is there exists a matrix $I(\theta_0, z_0)$ and a sequence of random vectors $\{\Delta_{n,(\theta_0,z_0)}\}$ as defined in Assumption 7, such that for every bounded and stochastic sequence $\{h_n\}$, that is $h_n = O_{P_0}(1)$, we have*

$$\log \frac{M_n(\theta_0 + n^{-1/2}h_n)}{M_n(\theta_0)} = h_n^T I(\theta_0, z_0)\Delta_{n,(\theta_0,z_0)} - \frac{1}{2}h_n^T I(\theta_0, z_0)h_n + o_{P_0}(1).$$

Next, we will show that the minimizing sequence must converge to a Dirac delta distribution at θ_0 using the results in Proposition 15 and Lemma 16.

Theorem 17 *For a given $\alpha > 1$ and under Assumptions 1, 3 (for $\bar{\mathcal{Q}}$) , 6, and 8, the $\alpha-$Rényi approximate posterior $q_n^*(\theta)$ over global latent parameters θ as defined in* (19) *converges weakly to a Dirac delta distribution at the true parameter θ_0; that is,*

$$q_n^*(\theta) \Rightarrow \delta_{\theta_0} \quad \textit{in } P_0 - \textit{probability as } n \to \infty.$$

Proof Using the result in Proposition 15 and following similar steps as used in Theorem 9, we can show that the minimizing sequence must have zero α-Rényi divergence in the limit with high probability. Recall the inequality in (25)

$$\min_{q(z_{1:n})\in\mathcal{Q}^n} D_\alpha(\pi(\theta, z_{1:n}|\mathbf{X}_n)\|q(\theta)q(z_{1:n})) \geq D_\alpha(\pi^*(\theta|\mathbf{X}_n)\|q(\theta)). \tag{27}$$

Also note that $q_n^*(\theta)$ is the minimizer of the LHS in the equation above. Since the variational likelihood satisfies the LAN condition due to Lemma 16, under the consistent testability assumption, the ideal posterior $\pi^*(\theta|\mathbf{X}_n)$ also degenerates to a Dirac delta distribution at the true parameter θ_0 (Kleijn and van der Vaart, 2012).

Now recall Lemma 6 and 7. Following the arguments in Lemma 6, and using the inequality in (27) we can argue that any sequence of distributions in $\bar{\mathcal{Q}}$ that minimizes the LHS in (27) must converge weakly to a Dirac delta distribution at the true parameter θ_0 in the large sample limit, since otherwise the objective in the LHS of (27) is unbounded. In addition, using Lemma 7 and the inequality in (27) we can also show that any sequence of distribution in $\bar{\mathcal{Q}}$ that converges weakly to a convex combination of a Dirac delta distribution at θ_0 with any other distribution can not achieve zero $\alpha-$Rényi divergence in the limit. This completes the proof. ∎

Acknowledgments

Acknowledgments This research is supported by the National Science Foundation (NSF) through awards DMS/1812197 and IIS/1816499, and the Purdue Research Foundation (PRF). We thank the Action Editor and anonymous reviewers for their helpful comments and suggestions.

Appendix A. Proofs

A.1. Proofs in Section 3

We begin with the following well known result.

Lemma 18 *[Laplace Approximation] Consider an integral of the form*

$$I = \int_a^b h(y) e^{-ng(y)} dy,$$

where $g(y)$ is a smooth function which has a local minimum at $y^* \in (a, b)$ and $h(y)$ is a smooth function. Then

$$I \sim h(y^*) e^{-ng(y^*)} \sqrt{\frac{2\pi}{ng''(y^*)}} \text{ as } n \to \infty.$$

Proof Readers are directed to Wong (1989, Chapter-2) for the proof. ■

Now we prove a technical lemma that bounds the differential entropy of the good sequence.

Lemma 19 *For a good sequence $\bar{q}_n(\theta)$, there exist an $n_M \geq 1$ and $\bar{M} > 0$, such that for all $n \geq n_M$*

$$-\int \bar{q}_n(\mu) \log \bar{q}_n(\mu) \leq \frac{1}{2} \log\left(2\pi \bar{e} \frac{\bar{M}}{n}\right),$$

where $\bar{e}$ is the Euler's constant.

Proof Recall from Assumption 4 that the $\bar{q}_n(\theta)$ converges weakly to δ_{θ_0} at the rate of $\sqrt{n}$. It follows from the Definition 5 for rate of convergence that,

$$E_{\bar{q}_n(\theta)}[|\theta - \hat{\theta}_n|^2] = O\left(\frac{1}{n}\right).$$

There exist an $n_M \geq 1$ and $\bar{M} > 0$, such that for all $n \geq n_M$

$$\mathbb{E}_{\bar{q}_n(\theta)}[(\theta - \hat{\theta}_n)^2] \leq \frac{\bar{M}}{n}.$$

Using the fact that, the differential entropy of random variable with a given variance is bounded by the differential entropy of the Gausian distribution of the same variance (Cover, 2006, Theorem 9.6.5)), it follows that the differential entropy of $\bar{q}_n(\mu)$ is bounded by $\frac{1}{2}\log(2\pi\bar{e}\frac{\bar{M}}{n})$, where $\bar{e}$ is the Euler's constant. ■

Next, we prove the following result on the prior distributions. This result will be useful in proving Lemma 21 and 6.

Lemma 20 *Given a prior distribution $\pi(\theta)$ with $\mathbb{E}_{\pi(\theta)}[|\theta|] < \infty$, for any $\beta > 0$, there exists a sequence of compact sets $\{K_n\} \subset \Theta$ such that*

$$\int_{\Theta \backslash K_n} \pi(\gamma) d\gamma = O(n^{-\beta}).$$

Proof Fix $\theta_1 \in \Theta$. Define a sequence of compact sets

$$K_n = \{\theta \in \Theta : |\theta - \theta_1| \leq n^{\beta}\} \forall \beta > 0.$$

Clearly, as n increases K_n approaches Θ. Now, using Markov's inequality followed by the triangule inequality,

$$\begin{aligned} \int_{\Theta \backslash K_n} \pi(\gamma) d\gamma = \int_{\{\gamma \in \Theta : |\gamma - \theta_1| > n^{\beta}\}} \pi(\gamma) d\gamma &\leq n^{-\beta} \mathbb{E}_{\pi(\theta)}[|\gamma - \theta_1|] \\ &\leq n^{-\beta} \left(\mathbb{E}_{\pi(\theta)}[|\gamma|] + |\theta_1| \right). \end{aligned} \tag{28}$$

Since, $\mathbb{E}_{\pi(\gamma)}[|\gamma|] < \infty$, it follows that $\forall \beta > 0$, $\int_{\Theta \backslash K_n} \pi(\gamma) d\gamma = O(n^{-\beta})$. ■

The next result approximates the normalizing sequence of the posterior distribution using the lemma above and the LAN condition.

Lemma 21 *There exists a sequence of compact balls $\{K_n \subset \Theta\}$, such that $\theta_0 \in K_n$ and under Assumptions 1 and 2, the normalizing sequence of the posterior distribution*

$$\begin{aligned} &\int_{\Theta} \prod_{i=1}^{n} \frac{p(X_i|\gamma)}{p(X_i|\theta_0)} \pi(\gamma) d\gamma \\ &= \sqrt{\frac{2\pi}{nI(\theta_0)}} e^{\left(\frac{1}{2} nI(\theta_0)((\hat{\theta}_n - \theta_0)^2)\right)} \left(e^{o_{P_{\theta_0}}(1)} \int_{K_n} \pi(\gamma) \mathcal{N}(\gamma; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\gamma + o(1) \right). \end{aligned} \tag{29}$$

Proof Let $\{K_n \subset \Theta\}$ be a sequence of compact balls such that $\theta_0 \in K_n$, where θ_0 is any point in Θ where prior distribution $\pi(\theta)$ places positive density. Using Lemma 20, we can always find a sequence of sets $\{K_n\}$ for a prior distribution, such that $\theta_0 \in K_n$ and for any positive constant $\beta > \frac{3}{2}$,

$$\int_{\Theta \backslash K_n} \pi(\gamma) d\gamma = O(n^{-\beta}). \tag{30}$$

Observe that

$$\int_{\Theta} \prod_{i=1}^{n} \frac{p(X_i|\gamma)}{p(X_i|\theta_0)} \pi(\gamma) d\gamma = \left(\int_{K_n} \prod_{i=1}^{n} \frac{p(X_i|\gamma)}{p(X_i|\theta_0)} \pi(\gamma) d\gamma + \int_{\Theta \backslash K_n} \pi(\gamma) \prod_{i=1}^{n} \frac{p(X_i|\gamma)}{p(X_i|\theta_0)} d\gamma \right). \tag{31}$$

Consider the first term in (31); following similar steps as in (49) and (50) and using Assumption 2, we have

$$\begin{aligned}
\int_{K_n} \prod_{i=1}^{n} & \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma \\
&= e^{o_{P_{\theta_0}}(1)} \exp\left(\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n-\theta_0)^2\right)\right)\int_{K_n}\pi(\gamma)\exp\left(-\frac{1}{2}nI(\theta_0)\left((\gamma-\hat{\theta}_n)^2\right)\right)d\gamma \\
&= e^{o_{P_{\theta_0}}(1)} \exp\left(\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n-\theta_0)^2\right)\right)\sqrt{\frac{2\pi}{nI(\theta_0)}}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma, \quad (32)
\end{aligned}$$

where the last equality follows from the definition of Gaussian density, $\mathcal{N}(\cdot;\hat{\theta}_n,(nI(\theta_0))^{-1})$.

Substituting (32) into (31), we obtain

$$\begin{aligned}
\int_{\Theta} \prod_{i=1}^{n} & \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma \\
&= \exp\left(\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n-\theta_0)^2\right)\right)\sqrt{\frac{2\pi}{nI(\theta_0)}}\Bigg(e^{o_{P_{\theta_0}}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma \\
&\quad + \exp\left(-\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n-\theta_0)^2\right)\right)\sqrt{\frac{nI(\theta_0)}{2\pi}}\int_{\Theta\backslash K_n}\pi(\gamma)\prod_{i=1}^{n}\frac{p(X_i|\gamma)}{p(X_i|\theta_0)}d\gamma\Bigg). \quad (33)
\end{aligned}$$

Next, using the Markov's inequality and then Fubini's Theorem, for arbitrary $\delta > 0$, we have

$$\begin{aligned}
P_{\theta_0}\left(\sqrt{\frac{nI(\theta_0)}{2\pi}}\int_{\Theta\backslash K_n}\prod_{i=1}^{n}\frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma > \delta\right) &\leq \sqrt{\frac{nI(\theta_0)}{\delta^2 2\pi}}\mathbb{E}_{P_{\theta_0}}\left[\int_{\Theta\backslash K_n}\prod_{i=1}^{n}\frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma\right] \\
&= \sqrt{\frac{nI(\theta_0)}{\delta^2 2\pi}}\int_{\Theta\backslash K_n}\mathbb{E}_{P_{\theta_0}}\left[\prod_{i=1}^{n}\frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\right]\pi(\gamma)d\gamma \\
&= \sqrt{\frac{nI(\theta_0)}{\delta^2 2\pi}}\int_{\Theta\backslash K_n}\pi(\gamma)d\gamma, \quad (34)
\end{aligned}$$

since $\mathbb{E}_{P_{\theta_0}}\left[\prod_{i=1}^{n}\frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\right] = 1$.

Hence, using (30) for $\beta > 3/2$, it is straightforward to observe that

$$P_{\theta_0}\left(\sqrt{\frac{nI(\theta_0)}{2\pi}}\int_{\Theta\backslash K_n}\prod_{i=1}^{n}\frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma > \delta\right) \leq \sqrt{\frac{I(\theta_0)}{\delta^2 2\pi}}\frac{1}{n^{\beta-1/2}}.$$

Since the upper bound above is summable, using First Borel-Cantelli Theorem it follows that

$$\sqrt{\frac{nI(\theta_0)}{2\pi}}\int_{\Theta\backslash K_n}\prod_{i=1}^{n}\frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma = o(1)\ P_{\theta_0} - \text{a.s.}. \quad (35)$$

Since, $\exp\left(-\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n-\theta_0)^2\right)\right) \le 1$, it follows from substituting (35) into (33) that

$$\begin{aligned}
&\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma \\
&\quad= \exp\left(\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n-\theta_0)^2\right)\right)\sqrt{\frac{2\pi}{nI(\theta_0)}}\left(e^{o_{P_{\theta_0}}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma + o(1)\right).
\end{aligned}$$

■

Next we prove Lemma 6, showing that the α–Rényi divergence between the posterior and any non-degenerate distribution diverges in the large sample limit.

Proof of Lemma 6 Let $K_n \subset \Theta$ be a sequence of compact sets such that $\theta_0 \in K_n$, where θ_0 is any point in Θ where prior distribution $\pi(\theta)$ places positive density. Using Lemma 20, we can always find a sequence of sets $\{K_n\}$ for a prior distribution, such that $\theta_0 \in K_n$ and for any positive constant $\beta > \frac{1}{2}$,

$$\int_{\Theta\backslash K_n}\pi(\gamma)d\gamma = O(n^{-\beta}). \tag{36}$$

Now, observe that

$$\begin{aligned}
&\frac{\alpha-1}{\alpha}D_\alpha(\pi(\theta|\mathbf{X}_n)\|q_n(\theta)) \\
&\quad= \frac{1}{\alpha}\log\left(\int_{K_n}q_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{q_n(\theta)}\right)^\alpha d\theta + \int_{\Theta\backslash K_n}q_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{q_n(\theta)}\right)^\alpha d\theta\right) \\
&\quad\ge \frac{1}{\alpha}\log\left(\int_{K_n}q_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{q_n(\theta)}\right)^\alpha d\theta\right),
\end{aligned} \tag{37}$$

where the last inequality follows from the fact that the integrand is always positive.

Next, we approximate the ratio in the integrand on the right hand side of the above equation using the LAN condition in Assumption 2. Let $\Delta_{n,\theta_0} := \sqrt{n}(\hat{\theta}_n-\theta_0)$, such that $\hat{\theta}_n \to \theta_0$, $P_{\theta_0}-a.s.$ and Δ_{n,θ_0} converges in distribution to $\mathcal{N}(0,I(\theta_0)^{-1})$. Re-parameterizing the expression with $\theta = \theta_0 + n^{-1/2}h$, we have

$$\begin{aligned}
\int_{K_n}q_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{q_n(\theta)}\right)^\alpha d\theta &= n^{-1/2}\int_{K_n}q_n(\theta_0+n^{-1/2}h)\left(\frac{\pi(\theta_0+n^{-1/2}h)\prod_{i=1}^n\frac{p(X_i|(\theta_0+n^{-1/2}h))}{p(X_i|\theta_0)}}{q_n(\theta_0+n^{-1/2}h)\int_\Theta\prod_{i=1}^n\frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma}\right)^\alpha dh \\
&= n^{-1/2}\int_{K_n}q_n(\theta_0+n^{-1/2}h)\left(\frac{\pi(\theta_0+n^{-1/2}h)\prod_{i=1}^n\frac{p(X_i|(\theta_0+n^{-1/2}h))}{p(X_i|\theta_0)}}{q_n(\theta_0+n^{-1/2}h)\int_\Theta\prod_{i=1}^n\frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma}\right)^\alpha dh \qquad (38) \\
&= n^{-1/2}\int_{K_n}q_n(\theta_0+n^{-1/2}h)\left(\pi(\theta_0+n^{-1/2}h)\frac{\exp(hI(\theta_0)\Delta_{n,\theta_0}-\frac{1}{2}h^2I(\theta_0)+o_{P_{\theta_0}}(1))}{q_n(\theta_0+n^{-1/2}h)\int_\Theta\prod_{i=1}^n\frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma}\right)^\alpha dh. \qquad (39)
\end{aligned}$$

Resubstituting $h = \sqrt{n}(\theta - \theta_0)$ in the expression above and reverting to the previous parametrization,

$$= \int_{K_n} q_n(\theta) \left(\pi(\theta) \frac{\exp\left(\sqrt{n}(\theta-\theta_0) I(\theta_0)\Delta_{n,\theta_0} - \frac{1}{2}n(\theta-\theta_0)^2 I(\theta_0) + o_{P_{\theta_0}}(1)\right)}{q_n(\theta)\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma} \right)^\alpha d\theta$$
$$= \int_{K_n} q_n(\theta) \left(\pi(\theta) \frac{e^{o_{P_{\theta_0}}(1)} \exp\left(-\frac{1}{2}nI(\theta_0)\left((\theta-\theta_0)^2 - 2(\theta-\theta_0)(\hat{\theta}_n - \theta_0)\right)\right)}{q_n(\theta)\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma} \right)^\alpha d\theta.$$

Now completing the square by dividing and multiplying the numerator by $\exp\left(\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n - \theta_0)^2\right)\right)$ we obtain

$$= \int_{K_n} q_n(\theta) \left(\pi(\theta) \frac{e^{o_{P_{\theta_0}}(1)} \exp\left(\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n - \theta_0)^2\right)\right)\exp\left(-\frac{1}{2}nI(\theta_0)\left((\theta-\hat{\theta}_n)^2\right)\right)}{q_n(\theta)\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma} \right)^\alpha d\theta$$
$$= \int_{K_n} q_n(\theta) \left(\pi(\theta) \frac{e^{o_{P_{\theta_0}}(1)} \exp\left(\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n - \theta_0)^2\right)\right)\sqrt{\frac{2\pi}{nI(\theta_0)}}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})}{q_n(\theta)\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma} \right)^\alpha d\theta, \tag{40}$$

where, in the last equality we used the definition of Gaussian density, $\mathcal{N}(\cdot;\hat{\theta}_n,(nI(\theta_0))^{-1})$.

Next, we approximate the integral in the denominator of (50). Using Lemma 21, it follows that there exist a sequence of compact balls $\{K_n \subset \Theta\}$, such that $\theta_0 \in K_n$ and

$$\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma$$
$$= \sqrt{\frac{2\pi}{nI(\theta_0)}} e^{\left(\frac{1}{2}nI(\theta_0)((\hat{\theta}_n-\theta_0)^2)\right)} \left(e^{o_{P_{\theta_0}}(1)} \int_{K_n} \pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma + o(1)\right). \tag{41}$$

Substituting (41) into (40) and simplifying, we obtain

$$\int_{K_n} q_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{q_n(\theta)}\right)^\alpha d\theta$$
$$= \int_{K_n} q_n(\theta)^{1-\alpha} \left(\frac{e^{o_{P_{\theta_0}}(1)}\pi(\theta)\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})}{\left(e^{o_{P_{\theta_0}}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma + o(1)\right)} \right)^\alpha d\theta. \tag{42}$$

Observe that

$$\left(\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})\right)^\alpha = \left(\sqrt{\frac{nI(\theta_0)}{2\pi}}\right)^\alpha \left(\sqrt{\frac{2\pi}{n\alpha I(\theta_0)}}\right)\mathcal{N}(\theta;\hat{\theta}_n,(n\alpha I(\theta_0))^{-1}).$$

Substituting this into the right hand side of (42)

$$
\begin{aligned}
&\frac{1}{\alpha}\log\int_{K_n} q_n(\theta)^{1-\alpha}\left(\frac{\pi(\theta)\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})}{\left(e^{o_{P_{\theta_0}}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma+o(1)\right)}\right)^{\alpha}d\theta\\
&=-\log\left(e^{o_{P_{\theta_0}}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma+o(1)\right)+\frac{\alpha-1}{2\alpha}\log n-\frac{\log\alpha}{2\alpha}\\
&\quad+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\pi}+\frac{1}{\alpha}\log\int_{K_n} q_n(\theta)^{1-\alpha}\pi(\theta)^{\alpha}\mathcal{N}(\theta;\hat{\theta}_n,(n\alpha I(\theta_0))^{-1})d\theta. \qquad (43)
\end{aligned}
$$

From the Laplace approximation (Lemma 18) and the continuity of the logarithm, we have

$$
-\log\left(e^{o_{P_{\theta_0}}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma+o(1)\right)\sim-\log\left(e^{o_{P_{\theta_0}}(1)}\pi(\hat{\theta}_n)\right).
$$

Next, using the Laplace approximation on the last term in (43)

$$
\frac{1}{\alpha}\log\int_{K_n} q_n(\theta)^{1-\alpha}\pi(\theta)^{\alpha}\mathcal{N}(\theta;\hat{\theta}_n,(n\alpha I(\theta_0))^{-1})d\theta\sim\frac{\alpha-1}{\alpha}\log\frac{1}{q_n(\hat{\theta}_n)}+\log\pi(\hat{\theta}_n).
$$

Substituting the above two approximations into (43), we have

$$
\begin{aligned}
&\frac{1}{\alpha}\log\int_{K_n} q_n(\theta)^{1-\alpha}\left(\frac{\pi(\theta)\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})}{\left(e^{o_{P_{\theta_0}}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma+o(1)\right)}\right)^{\alpha}d\theta\\
&\sim-\log\left(e^{o_{P_{\theta_0}}(1)}\pi(\hat{\theta}_n)\right)-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\pi}\\
&\quad+\frac{\alpha-1}{2\alpha}\log n-\frac{\alpha-1}{\alpha}\log q_n(\hat{\theta}_n)+\log\pi(\hat{\theta}_n)\\
&\sim-\log\left(\pi(\hat{\theta}_n)\right)-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\pi}+\frac{\alpha-1}{2\alpha}\log n-\frac{\alpha-1}{\alpha}\log q(\hat{\theta}_n)+\log\pi(\hat{\theta}_n)+o_{P_{\theta_0}}(1)\\
&=-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\pi}+\frac{\alpha-1}{2\alpha}\log n-\frac{\alpha-1}{\alpha}\log q(\hat{\theta}_n)+o_{P_{\theta_0}}(1), \qquad (44)
\end{aligned}
$$

where the penultimate approximation follows from the fact that

$$
q_n(\hat{\theta}_n)\sim q(\hat{\theta}_n).
$$

Note that $\hat{\theta}_n \to \theta_0,\ P_{\theta_0} - a.s.$ Therefore, if $q(\theta_0) = 0$, then the right hand side in (44) will diverge as $n \to \infty$ because $\frac{\alpha-1}{2\alpha}\log n$ also diverges as $n \to \infty$. Also observe that, for any $q(\theta)$ that places finite mass on θ_0, the α–Rényi divergence diverges as $n \to \infty$. Hence, α–Rényi approximate posterior must converge weakly to a distribution that has a Dirac delta distribution at the true parameter θ_0. ∎

Next, we show that the α–Rényi divergence between the true posterior and the sequence $\{q_n'(\theta)\} \in \mathcal{Q}$ as defined in (9) is bounded below by a positive number.

Proof of Lemma 7 Van Erven and Harremos (2014, Theorem 19) shows that for any $\alpha > 0$, the $\alpha-$Rényi divergence $D_\alpha(p(\theta)\|q(\theta))$ is a lower semi-continuous function of the pair $(p(\theta), q(\theta))$ in the weak topology on the space of probability measures. Recall from (6) that the true posterior distribution $\pi(\theta|\mathbf{X}_n)$ converges weakly to δ_{θ_0} $P_{\theta_0} - a.s.$ Using this fact it follows that

$$\liminf_{n\to\infty} D_\alpha(\pi(\theta|\mathbf{X}_n)\|q'_n(\theta)) \geq D_\alpha\left(\delta_{\theta_0}\middle\| w^j\delta_{\theta_0} + \sum_{i=1,i\neq j}^{\infty} w^i q_i(\theta)\right) \quad P_{\theta_0} - a.s.$$

Next, using Pinsker's inequality (Cover, 2006) for $\alpha > 1$, we have

$$\begin{aligned} D_\alpha\left(\delta_{\theta_0}\middle\| w^j\delta_{\theta_0} + \sum_{i=1,i\neq j}^{\infty} w^i q_i(\theta)\right) &\geq \frac{1}{2}\left(\int_\Theta \left|\delta_{\theta_0} - w^j\delta_{\theta_0} - \sum_{i=1,i\neq j}^{\infty} w^i q_i(\theta)\right| d\theta\right)^2 \\ &= \frac{1}{2}\left(\int_\Theta \left|(1-w^j)\delta_{\theta_0} - \sum_{i=1,i\neq j}^{\infty} w^i q_i(\theta)\right| d\theta\right)^2. \end{aligned}$$

Now dividing the integral over ball of radius ϵ centered at θ_0, $B(\theta_0,\epsilon)$ and its complement, we obtain

$$\begin{aligned} &\liminf_{n\to\infty} D_\alpha(\pi(\theta|\mathbf{X}_n)\|q'_n(\theta)) \\ &\quad\geq \frac{1}{2}\left(\int_{B(\theta_0,\epsilon)} \left|(1-w^j)\delta_{\theta_0} - \sum_{i=1,i\neq j}^{\infty} w^i q_i(\theta)\right| d\theta + \int_{B(\theta_0,\epsilon)^C} \left|(1-w^j)\delta_{\theta_0} - \sum_{i=1,i\neq j}^{\infty} w^i q_i(\theta)\right| d\theta\right)^2 \\ &\quad\geq \frac{1}{2}\left(\int_{B(\theta_0,\epsilon)^C} \left|(1-w^j)\delta_{\theta_0} - \sum_{i=1,i\neq j}^{\infty} w^i q_i(\theta)\right| d\theta\right)^2 \\ &\quad= \frac{1}{2}\left(\int_{B(\theta_0,\epsilon)^C} \left|-\sum_{i=1,i\neq j}^{\infty} w^i q_i(\theta)\right| d\theta\right)^2 \quad P_{\theta_0} - a.s. \end{aligned} \tag{45}$$

Since, $w^i \in (0,1)$, observe that for any $\epsilon > 0$, there exists $\eta(\epsilon) > 0$, such that

$$\frac{1}{2}\left(\int_{B(\theta_0,\epsilon)^C} \left|-\sum_{i=1,i\neq j}^{\infty} w^i q_i(\theta)\right| d\theta\right)^2 \geq \eta(\epsilon).$$

Therefore, it follows that

$$\liminf_{n\to\infty}\ D_\alpha(\pi(\theta|\mathbf{X}_n)\|q'_n(\theta)) \geq \eta(\epsilon) > 0 \quad P_{\theta_0} - a.s.$$

■

In the following result, we show that if $q_i(\theta), i \in \{1,2,\ldots\}$ in the definition of $\{q'_n(\theta)\}$ in (9) are Dirac delta distributions then

$$\liminf_{n\to\infty} D_\alpha(\pi(\theta|\mathbf{X}_n)\|q'_n(\theta)) \geq 2(1-w^j)^2 > 0 \quad P_{\theta_0} - a.s,$$

where w^j is the weight of δ_{θ_0}. Consider a sequence $\{q_n(\theta)\}$, that converges weakly to a convex combination of $\delta_{\theta_i}, i \in \{1,2,\ldots\}$ such that for weights $\{w^i \in (0,1) : \sum_{i=1}^{\infty} w^i = 1\}$,

$$q_n(\theta) \Rightarrow \sum_{i=1}^{\infty} w^i \delta_{\theta_i}, \tag{46}$$

where for any $j \in \{1,2,\ldots\}$, $\theta_j = \theta_0$ and for all $i \in \{1,2,\ldots\} \backslash \{j\}$, $\theta_j \neq \theta_0$.

Lemma 22 *The $\alpha-$Rényi divergence between the true posterior and sequence $\{q_n(\theta)\}$ is bounded below by a positive number $2(1-w^j)^2$; that is,*

$$\liminf_{n\to\infty} D_\alpha(\pi(\theta|\mathbf{X}_n)\|q_n(\theta)) \geq 2(1-w^j)^2 > 0 \quad P_{\theta_0} - a.s,$$

where w^j is the weight of δ_{θ_0} in the definition of sequence $\{q_n(\theta)\}$.

Proof Van Erven and Harremos (2014, Theorem 19) shows that for any $\alpha > 0$, the $\alpha-$Rényi divergence $D_\alpha(p(\theta)\|q(\theta))$ is a lower semi-continuous function of the pair $(p(\theta), q(\theta))$ in the weak topology on the space of probability measures. Recall from (6) that the true posterior distribution $\pi(\theta|\mathbf{X}_n)$ converges weakly to δ_{θ_0}, $P_{\theta_0} - a.s.$ Using this fact it follows that

$$\liminf_{n\to\infty} D_\alpha(\pi(\theta|\mathbf{X}_n)\|q_n(\theta)) \geq D_\alpha\left(\delta_{\theta_0}\middle\| \sum_{i=1}^{\infty} w_i\delta_{\theta_i}\right) \quad P_{\theta_0} - a.s.$$

Next, using Pinsker's inequality (Cover, 2006) for $\alpha > 1$, we have

$$\begin{aligned} D_\alpha\left(\delta_{\theta_0}\middle\| \sum_{i=1}^{\infty} w^i\delta_{\theta_i}\right) &\geq \frac{1}{2}\left(\int_\Theta \left|\delta_{\theta_0} - \sum_{i=1}^{\infty} w^i\delta_{\theta_i}\right| d\theta\right)^2 \\ &= \frac{1}{2}\left(\int_\Theta \left|(1-w^j)\delta_{\theta_0} - \sum_{i=1, i\neq j}^{\infty} w^i\delta_{\theta_i}\right| d\theta\right)^2 \\ &= \frac{1}{2}\left(\int_{B(\theta_0,\epsilon)} (1-w^j)|\delta_{\theta_0}| d\theta + \sum_{i=1,i\neq j}^{\infty} w^i \int_{B(\theta_i,\epsilon)} |-\delta_{\theta_i}| d\theta\right)^2 \\ &= \frac{1}{2}\left((1-w^j) + \sum_{i=1,i\neq j}^{\infty} w^i\right)^2 = 2(1-w^j)^2, \end{aligned} \tag{47}$$

where $B(\theta_i, \epsilon)$ is the ball of radius ϵ centered at θ_i. Note that, there always exist an $\epsilon > 0$, such that $\bigcap_{i=1}^{\infty} B(\theta_i, \epsilon) = \phi$. Since, by the definition of sequence $\{q_n(\theta)\}$, $w^j \in (0,1)$, therefore $2(1-w^j)^2 > 0$ and the lemma follows. ■

Now we show that any sequence of distributions $\{s_n(\theta)\} \subset \mathcal{Q}$ that converges weakly to a distribution $s(\theta) \in \mathcal{Q}$, that has positive density at any point other than the true parameter θ_0, cannot achieve zero KL divergence in the limit.

Proof of Proposition 8 Observe that for any good sequence $\{\bar{q}_n(\theta)\}$

$$\min_{q\in\mathcal{Q}} D_\alpha(\pi(\theta|\mathbf{X}_n)\|q(\theta)) \leq D_\alpha(\pi(\theta|\mathbf{X}_n)\|\bar{q}_n(\theta)).$$

Therefore, for the second part, it suffices to show that

$$D_\alpha(\pi(\theta|\mathbf{X}_n)\|\bar{q}_n(\theta)) < B + o_{P_{\theta_0}}(1).$$

The subsequent arguments in the proof are for any $n \geq \max(n_1, n_2, n_3, n_M)$, where n_1, n_2, and n_3 are defined in Assumption 4. First observe that, for any compact ball K containing the true parameter θ_0,

$$\begin{aligned}
&\frac{\alpha-1}{\alpha} D_\alpha(\pi(\theta|\mathbf{X}_n)\|\bar{q}_n(\theta)) \\
&\quad = \frac{1}{\alpha}\log\left(\int_K \bar{q}_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{\bar{q}_n(\theta)}\right)^\alpha d\theta + \int_{\Theta\backslash K}\bar{q}_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{\bar{q}_n(\theta)}\right)^\alpha d\theta\right).
\end{aligned} \tag{48}$$

First, we approximate the first integral on the right hand side using the LAN condition in Assumption 2. Let $\Delta_{n,\theta_0} \coloneqq \sqrt{n}(\hat{\theta}_n - \theta_0)$, where $\hat{\theta}_n \to \theta_0$, $P_{\theta_0} - a.s.$ and Δ_{n,θ_0} converges in distribution to $\mathcal{N}(0, I(\theta_0)^{-1})$. Reparameterizing the expression with $\theta = \theta_0 + n^{-1/2}h$, we have

$$\begin{aligned}
&\int_K \bar{q}_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{\bar{q}_n(\theta)}\right)^\alpha d\theta = n^{-1/2}\int_K \bar{q}_n(\theta_0+n^{-1/2}h)\left(\frac{\pi(\theta_0+n^{-1/2}h)\prod_{i=1}^n \frac{p(X_i|(\theta_0+n^{-1/2}h))}{p(X_i|\theta_0)}}{\bar{q}_n(\theta_0+n^{-1/2}h)\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma}\right)^\alpha dh \\
&= n^{-1/2}\int_K \bar{q}_n(\theta_0+n^{-1/2}h)\left(\frac{\pi(\theta_0+n^{-1/2}h)\prod_{i=1}^n \frac{p(X_i|(\theta_0+n^{-1/2}h))}{p(X_i|\theta_0)}}{\bar{q}_n(\theta_0+n^{-1/2}h)\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma}\right)^\alpha dh \\
&= n^{-1/2}\int_K \bar{q}_n(\theta_0+n^{-1/2}h)\left(\pi(\theta_0+n^{-1/2}h)\frac{\exp(hI(\theta_0)\Delta_{n,\theta_0} - \frac{1}{2}h^2 I(\theta_0) + o_{P_{\theta_0}}(1))}{\bar{q}_n(\theta_0+n^{-1/2}h)\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma}\right)^\alpha dh.
\end{aligned} \tag{49}$$

Resubstituting $h = \sqrt{n}(\theta - \theta_0)$ in the expression above and reverting to the previous parametrization,

$$\begin{aligned}
&= \int_K \bar{q}_n(\theta)\left(\pi(\theta)\frac{\exp\left(\sqrt{n}(\theta-\theta_0)I(\theta_0)\Delta_{n,\theta_0} - \frac{1}{2}n(\theta-\theta_0)^2 I(\theta_0) + o_{P_{\theta_0}}(1)\right)}{\bar{q}_n(\theta)\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma}\right)^\alpha d\theta \\
&= \int_K \bar{q}_n(\theta)\left(\pi(\theta)\frac{e^{o_{P_{\theta_0}}(1)}\exp\left(-\frac{1}{2}nI(\theta_0)\left((\theta-\theta_0)^2 - 2(\theta-\theta_0)(\hat{\theta}_n-\theta_0)\right)\right)}{\bar{q}_n(\theta)\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma}\right)^\alpha d\theta.
\end{aligned}$$

Completing the square by dividing and multiplying the numerator by $\exp\left(\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n-\theta_0)^2\right)\right)$

$$\begin{aligned}
&= \int_K \bar{q}_n(\theta)\left(\pi(\theta)\frac{e^{o_{P_{\theta_0}}(1)}\exp\left(\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n-\theta_0)^2\right)\right)\exp\left(-\frac{1}{2}nI(\theta_0)\left((\theta-\hat{\theta}_n)^2\right)\right)}{\bar{q}_n(\theta)\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma}\right)^\alpha d\theta \\
&= \int_K \bar{q}_n(\theta)\left(\pi(\theta)\frac{e^{o_{P_{\theta_0}}(1)}\exp\left(\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n-\theta_0)^2\right)\right)\sqrt{\frac{2\pi}{nI(\theta_0)}}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})}{\bar{q}_n(\theta)\int_\Theta \prod_{i=1}^n \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma}\right)^\alpha d\theta,
\end{aligned} \tag{50}$$

where, in the last equality we used the definition of Gaussian density, $\mathcal{N}(\cdot;\hat{\theta}_n,(nI(\theta_0))^{-1})$.

Next, we approximate the integral in the denominator of (50). Using Lemma 21 (in the appendix) it follows that, there exist a sequence of compact balls $\{K_n \subset \Theta\}$, such that $\theta_0 \in K_n$ and

$$\int_\Theta \prod_{i=1}^{n} \frac{p(X_i|\gamma)}{p(X_i|\theta_0)}\pi(\gamma)d\gamma = \sqrt{\frac{2\pi}{nI(\theta_0)}} e^{\left(\frac{1}{2}nI(\theta_0)((\hat{\theta}_n-\theta_0)^2)\right)} \left(e^{o_{P_{\theta_0}}(1)} \int_{K_n} \pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma + o(1)\right). \tag{51}$$

Substituting (51) into (50), we obtain

$$\int_K \bar{q}_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{\bar{q}_n(\theta)}\right)^\alpha d\theta = \int_K \bar{q}_n(\theta)^{1-\alpha}\left(\frac{e^{o_{P_{\theta_0}}(1)}\pi(\theta)\mathcal{N}(\theta;\hat{\theta}_n,\frac{1}{nI(\theta_0)})}{\left(e^{o_{P_{\theta_0}}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,\frac{1}{nI(\theta_0)})d\gamma + o(1)\right)}\right)^\alpha d\theta. \tag{52}$$

Now, recall the definition of compact ball K, n_1 and n_2 from Assumption 4 and fix $n \geq n'_0$, where $n'_0 = \max(n_1, n_2)$. Note that n_2 is chosen, such that for all $n \geq n_2$, the bound in Assumption 4(3) holds on the set $\Theta\backslash K$. Next, consider the second term inside the logarithm function on the right hand side of (48). Using Assumption 4(3), we obtain

$$\int_{\Theta\backslash K} \bar{q}_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{\bar{q}_n(\theta)}\right)^\alpha d\theta \leq M_r^\alpha \int_{\Theta\backslash K} \bar{q}_n(\theta)d\theta \quad P_{\theta_0}-a.s. \tag{53}$$

Recall that the good sequence $\{\bar{q}_n(\cdot)\}$ exists $P_{\theta_0}-a.s$ with mean $\hat{\theta}_n$, for all $n \geq n_1$ and therefore it converges weakly to δ_{θ_0} (Assumption 4(2)). Combined with the fact that compact set K contains the true parameter θ_0, it follows that the second term in (48) is of $o(1)$, $P_{\theta_0}-a.s.$ Therefore, the second term inside the logarithm function on the right hand side of (48) is $o(1)$:

$$\int_{\Theta\backslash K} \bar{q}_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{\bar{q}_n(\theta)}\right)^\alpha d\theta = o(1) \ P_{\theta_0} - \text{a.s.} \tag{54}$$

Substituting (52) and (54) into (48), we have

$$\begin{aligned}
&\frac{\alpha-1}{\alpha}D_\alpha(\pi(\theta|\mathbf{X}_n)\|\bar{q}_n(\theta)) \\
&= \frac{1}{\alpha}\log\left(\int_K \bar{q}_n(\theta)^{1-\alpha}\left(\frac{e^{o_{P_{\theta_0}}(1)}\pi(\theta)\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})}{\left(e^{o_{P_{\theta_0}}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma + o(1)\right)}\right)^\alpha d\theta + o(1)\right) \\
&= \frac{1}{\alpha}\log\left(e^{o_{P_{\theta_0}}(1)}\int_K \bar{q}_n(\theta)^{1-\alpha}\left(\frac{\pi(\theta)\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})}{\left(e^{o_{P_{\theta_0}}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma + o(1)\right)}\right)^\alpha d\theta + o(1)\right).
\end{aligned} \tag{$\star\star$}$$

Now observe that,

$$
\begin{aligned}
(\star\star) &\sim \frac{1}{\alpha}\log\left(\int_K \bar{q}_n(\theta)^{1-\alpha}\left(\frac{\pi(\theta)\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})}{\left(e^{o_{P_{\theta_0}}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma+o(1)\right)}\right)^{\alpha} d\theta\right)\\
&= \frac{1}{\alpha}\log\left(\int_K \bar{q}_n(\theta)^{1-\alpha}\pi(\theta)^{\alpha}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})^{\alpha}d\theta\right)\\
&\quad -\log\left(e^{o_{P_{\theta_0}}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma+o(1)\right)\\
&\sim \frac{1}{\alpha}\log\left(\int_K \bar{q}_n(\theta)^{1-\alpha}\pi(\theta)^{\alpha}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})^{\alpha}d\theta\right)\\
&\quad -\log\left(\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma\right)+o_{P_{\theta_0}}(1).
\end{aligned}
\tag{55}
$$

Note that $\left(\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})\right)^{\alpha} = \left(\sqrt{\frac{nI(\theta_0)}{2\pi}}\right)^{\alpha}\left(\sqrt{\frac{2\pi}{n\alpha I(\theta_0)}}\right)\mathcal{N}(\theta;\hat{\theta}_n,(n\alpha I(\theta_0))^{-1})$.

Substituting this into (55), for large enough n, we have

$$
\begin{aligned}
&\frac{\alpha-1}{\alpha}D_{\alpha}(\pi(\theta|\mathbf{X}_n)\|\bar{q}_n(\theta))\\
&\quad\sim\frac{\alpha-1}{2\alpha}\log n-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\pi}+\frac{1}{\alpha}\log\int_K \bar{q}_n(\theta)^{1-\alpha}\pi(\theta)^{\alpha}\mathcal{N}(\theta;\hat{\theta}_n,(n\alpha I(\theta_0))^{-1})d\theta\\
&\quad\quad-\log\left(\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma\right).
\end{aligned}
\tag{56}
$$

From the Laplace approximation (Lemma 18) and the continuity of the logarithm, we have

$$
\frac{1}{\alpha}\log\int_K \bar{q}_n(\theta)^{1-\alpha}\pi(\theta)^{\alpha}\mathcal{N}(\theta;\hat{\theta}_n,(n\alpha I(\theta_0))^{-1})d\theta \sim \frac{1-\alpha}{\alpha}\log\bar{q}_n(\hat{\theta}_n)+\log\pi(\hat{\theta}_n).
$$

Next, using the Laplace approximation (Lemma 18) on the last term in (56) yields

$$
-\log\left(\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma\right) \sim -\log\left(\pi(\hat{\theta}_n)\right).
$$

Substituting the above two approximations into (56), for large enough n, we obtain

$$
\begin{aligned}
&\frac{\alpha-1}{\alpha}D_{\alpha}(\pi(\theta|\mathbf{X}_n)\|\bar{q}_n(\theta))\\
&\quad\sim\frac{1-\alpha}{\alpha}\log\bar{q}_n(\hat{\theta}_n)+\log\pi(\hat{\theta}_n)-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\pi}+\frac{\alpha-1}{2\alpha}\log n-\log\pi(\hat{\theta}_n)+o_{P_{\theta_0}}(1)\\
&\quad=\frac{1-\alpha}{\alpha}\log\bar{q}_n(\hat{\theta}_n)-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\pi}+\frac{\alpha-1}{2\alpha}\log n+o_{P_{\theta_0}}(1).
\end{aligned}
\tag{57}
$$

Now, recall Assumption 4(4) which, combined with the monotonicity of logarithm function, implies that $\log \bar{q}_n(\cdot)$ is concave for all $n \geq n_3$. Using Jensen's inequality,

$$\log \bar{q}_n(\hat{\theta}_n) = \log \bar{q}_n\left(\int \theta \bar{q}_n(\theta) d\theta\right) \geq \int \bar{q}_n(\theta) \log \bar{q}_n(\theta) d\theta.$$

Since $\alpha > 1$,

$$\frac{1-\alpha}{\alpha} \log \bar{q}_n(\hat{\theta}_n) \leq -\frac{\alpha-1}{\alpha} \int \bar{q}_n(\theta) \log \bar{q}_n(\theta) d\theta.$$

Using Lemma 19, there exists $n_M \geq 1$ and $0 < \bar{M} < \infty$, such that for all $n \geq n_M$

$$-\frac{\alpha-1}{\alpha} \int \bar{q}_n(\theta) \log \bar{q}_n(\theta) d\theta \leq \frac{\alpha-1}{2\alpha} \log\left(2\pi\bar{e}\frac{\bar{M}}{n}\right) = \frac{\alpha-1}{2\alpha} \log(2\pi\bar{e}\bar{M}) - \frac{\alpha-1}{2\alpha} \log n, \tag{58}$$

where $\bar{e}$ is the Euler's constant. Substituting (58) into the right hand side of (57), we have for all $n \geq n_0$, where $n_0 = \max(n_0', n_3, n_M)$,

$$\begin{aligned}
&\frac{1-\alpha}{\alpha} \log \bar{q}_n(\hat{\theta}_n) - \frac{\log\alpha}{2\alpha} + \frac{\alpha-1}{2\alpha} \log \frac{I(\theta_0)}{2\pi} + \frac{\alpha-1}{2\alpha} \log n. \\
&\leq \frac{\alpha-1}{2\alpha} \log(2\pi\bar{e}\bar{M}) - \frac{\alpha-1}{2\alpha} \log n - \frac{\log\alpha}{2\alpha} + \frac{\alpha-1}{2\alpha} \log \frac{I(\theta_0)}{2\pi} + \frac{\alpha-1}{2\alpha} \log n \\
&= \frac{\alpha-1}{2\alpha} \log(2\pi\bar{e}\bar{M}) - \frac{\log\alpha}{2\alpha} + \frac{\alpha-1}{2\alpha} \log \frac{I(\theta_0)}{2\pi} \\
&= \frac{\alpha-1}{\alpha} \frac{1}{2} \log \frac{\bar{e}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}}.
\end{aligned} \tag{59}$$

Observe that the left hand side in (57) is always non-negative, implying the right hand side must be too for large n. Therefore, the following inequality must hold for all $n \geq n_0$:

$$\frac{\bar{e}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}} \geq 1.$$

Consequently, substituting (59) into (57), we have

$$D_\alpha(\pi(\theta|\mathbf{X}_n) \| \bar{q}_n(\theta)) \leq \frac{1}{2} \log \frac{\bar{e}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}} + o_{P_{\theta_0}}(1) \ \forall n \geq n_0, \tag{60}$$

and the result follows.

■

We next state an important inequality, that is a direct consequence of Hölder's inequality. We use the following result in the proof of Lemma 10.

Lemma 23 *For any set $K \subset \Theta$ and $\alpha > 1$ and any sequence of distributions $\{q_n(\theta)\} \subset \mathcal{Q}$, the following inequality holds true*

$$\int_\Theta q_n(\theta) \left(\frac{\pi(\theta|\mathbf{X}_n)}{q_n(\theta)}\right)^\alpha d\theta \geq \frac{\left(\int_K \pi(\theta|\mathbf{X}_n) d\theta\right)^\alpha}{\left(\int_K q_n(\theta) d\theta\right)^{\alpha-1}}. \tag{61}$$

Proof Fix a set $K \subset \Theta$. Since $\alpha > 1$, using Hölder's inequality for $f(\theta) = \frac{\pi(\theta|\mathbf{X}_n)}{q_n(\theta)^{1-\frac{1}{\alpha}}}$ and $g(\theta) = q_n(\theta)^{1-\frac{1}{\alpha}}$,

$$\begin{aligned}\int_K \pi(\theta|\mathbf{X}_n)d\theta &= \int_K f(\theta)g(\theta)d\theta \\ &\leq \left(\int_K \frac{\pi(\theta|\mathbf{X}_n)^\alpha}{q_n(\theta)^{\alpha-1}}d\theta\right)^{\frac{1}{\alpha}} \left(\int_K q_n(\theta)d\theta\right)^{1-\frac{1}{\alpha}}.\end{aligned}$$

It is straightforward to observe from the above equation that,

$$\int_K \frac{\pi(\theta|\mathbf{X}_n)^\alpha}{q_n(\theta)^{\alpha-1}}d\theta \geq \frac{\left(\int_K \pi(\theta|\mathbf{X}_n)d\theta\right)^\alpha}{\left(\int_K q_n(\theta)d\theta\right)^{\alpha-1}}.$$

Also note that, for any set K, the following inequality holds true,

$$\int_\Theta q_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{q_n(\theta)}\right)^\alpha d\theta \geq \int_K \frac{\pi(\theta|\mathbf{X}_n)^\alpha}{q_n(\theta)^{\alpha-1}}d\theta \geq \frac{\left(\int_K \pi(\theta|\mathbf{X}_n)d\theta\right)^\alpha}{\left(\int_K q_n(\theta)d\theta\right)^{\alpha-1}}, \tag{62}$$

and the result follows immediately. ∎

Proof of Lemma 10 First, we fix $n \geq 1$ and let M_r be a sequence such that $M_r \to \infty$ as $r \to \infty$. Recall that $\hat{\theta}_n$ is the maximum likelihood estimate and denote $\tilde{\theta}_n = \mathbb{E}_{q_n(\theta)}[\theta]$. Define a set

$$K_r := \{\theta \in \Theta : |\theta - \hat{\theta}_n| > M_r\} \bigcup \{\theta \in \Theta : |\theta - \tilde{\theta}_n| > M_r\}.$$

Now, using Lemma 23 with $K = K_r$, we have

$$\int_\Theta q_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n)}{q_n(\theta)}\right)^\alpha d\theta \geq \frac{\left(\int_{K_r} \pi(\theta|\mathbf{X}_n)d\theta\right)^\alpha}{\left(\int_{K_r} q_n(\theta)d\theta\right)^{\alpha-1}}. \tag{63}$$

Note that the left hand side in the above equation does not depend on r and when $r \to \infty$ both the numerator and denominator on the right hand side converges to zero individually. For the ratio to diverge, however, we require the denominator to converge much faster than the numerator. To be more precise, observe that for a given n, since $\alpha - 1 < \alpha$ the tails of $q_n(\theta)$ must decay significantly faster than the tails of the true posterior for the right hand side in (63) to diverge as $r \to \infty$.

We next show that there exists an $n_0 \geq 1$ such that for all $n \geq n_0$, the right hand side in (63) diverges as $r \to \infty$. Since the posterior distribution satisfies the Bernstein-von Mises Theorem (van der Vaart, 1998), we have

$$\int_{K_r} \pi(\theta|\mathbf{X}_n)d\theta = \int_{K_r} \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1})d\theta + o_{P_{\theta_0}}(1).$$

Observe that the numerator on the right hand side of (63) satisfies,

$$\begin{aligned}\left(\int_{K_r}\pi(\theta|\mathbf{X}_n)d\theta\right)^\alpha &= \left(\int_{K_r}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})d\theta + o_{P_{\theta_0}}(1)\right)^\alpha \\ &\geq \left(\int_{\{|\theta-\hat{\theta}_n|>M_r\}}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})d\theta + o_{P_{\theta_0}}(1)\right)^\alpha \\ &= \left(\int_{\{\theta-\hat{\theta}_n>M_r\}}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})d\theta + \int_{\{\theta-\hat{\theta}_n\leq -M_r\}}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})d\theta + o_{P_{\theta_0}}(1)\right)^\alpha \\ &\geq \left(\int_{\{\theta-\hat{\theta}_n>M_r\}}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})d\theta + o_{P_{\theta_0}}(1)\right)^\alpha. \end{aligned} \tag{64}$$

Now, using the lower bound on the Gaussian tail distributions from Feller (1968)

$$\begin{aligned}\left(\int_{K_r}\pi(\theta|\mathbf{X}_n)d\theta\right)^\alpha &= \left(\int_{K_r}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})d\theta + o_{P_{\theta_0}}(1)\right)^\alpha \\ &\geq \left(\frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sqrt{nI(\theta_0)}M_r} - \frac{1}{(\sqrt{nI(\theta_0)}M_r)^3}\right)e^{-\frac{nI(\theta_0)}{2}M_r^2} + o_{P_{\theta_0}}(1)\right)^\alpha \\ &\sim \left(\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{nI(\theta_0)}M_r}e^{-\frac{nI(\theta_0)}{2}M_r^2} + o_{P_{\theta_0}}(1)\right)^\alpha, \end{aligned} \tag{65}$$

where the last approximation follows from the fact that, for large r,

$$\left(\frac{1}{\sqrt{nI(\theta_0)}M_r} - \frac{1}{(\sqrt{nI(\theta_0)}M_r)^3}\right) \sim \frac{1}{\sqrt{nI(\theta_0)}M_r}.$$

Next, consider the denominator on the right hand side of (63). Using the union bound

$$\left(\int_{K_r} q_n(\theta)d\theta\right)^{\alpha-1} \leq \left(\int_{\{|\theta-\tilde{\theta}_n|>M_r\}} q_n(\theta)d\theta + \int_{\{|\theta-\hat{\theta}_n|>M_r\}} q_n(\theta)d\theta\right)^{\alpha-1}. \tag{66}$$

Since, $\tilde{\theta}_n$ and $\hat{\theta}_n$ are finite for all $n \geq 1$, there exists an $\epsilon > 0$ such that for large n, $|\tilde{\theta}_n - \hat{\theta}_n| \leq \epsilon$. Applying the triangle inequality,

$$|\theta - \hat{\theta}_n| \leq |\theta - \tilde{\theta}_n| + |\tilde{\theta}_n - \hat{\theta}_n| \leq |\theta - \tilde{\theta}_n| + \epsilon.$$

Therefore, $\{|\theta - \hat{\theta}_n| > M_r\} \subseteq \{|\theta - \tilde{\theta}_n| > M_r - \epsilon\}$ and it follows from (66) that

$$\left(\int_{K_r} q_n(\theta)d\theta\right)^{\alpha-1} \leq \left(\int_{\{|\theta-\tilde{\theta}_n|>M_r\}} q_n(\theta)d\theta + \int_{\{|\theta-\tilde{\theta}_n|>M_r-\epsilon\}} q_n(\theta)d\theta\right)^{\alpha-1}.$$

Next, using the sub-Gaussian tail distribution bound from (Boucheron et al., 2013, Theorem 2.1),

$$\left(\int_{\{|\theta-\tilde{\theta}_n|>M_r\}} q_n(\theta)d\theta + \int_{\{|\theta-\tilde{\theta}_n|>M_r-\epsilon\}} q_n(\theta)d\theta\right)^{\alpha-1} \leq \left(2e^{-\frac{\gamma_n^2 M_r^2}{2B}} + 2e^{-\frac{\gamma_n^2(M_r-\epsilon)^2}{2B}}\right)^{\alpha-1}. \tag{67}$$

For large r, $M_r \sim M_r - \epsilon$, and it follows that

$$\left(\int_{\{|\theta - \tilde{\theta}_n| > M_r\}} q_n(\theta) d\theta + \int_{\{|\theta - \tilde{\theta}_n| > M_r - \epsilon\}} q_n(\theta) d\theta \right)^{\alpha-1} \lesssim \left(4e^{-\frac{\gamma_n^2 M_r^2}{2B}} \right)^{\alpha-1}. \tag{68}$$

Substituting (65) and (68) into (63), we obtain

$$\int_\Theta q_n(\theta) \left(\frac{\pi(\theta|\mathbf{X}_n)}{q_n(\theta)} \right)^\alpha d\theta \gtrsim \left(\frac{\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{nI(\theta_0)} M_r} e^{-\frac{nI(\theta_0)}{2} M_r^2} + o_{P_{\theta_0}}(1)}{\left(4e^{-\frac{\gamma_n^2 M_r^2}{2B}} \right)^{\frac{\alpha-1}{\alpha}}} \right)^\alpha,$$

for large r. Observe that

$$\frac{\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{nI(\theta_0)} M_r} e^{-\frac{nI(\theta_0)}{2} M_r^2}}{\left(4e^{-\frac{\gamma_n^2 M_r^2}{2B}} \right)^{\frac{\alpha-1}{\alpha}}} = \frac{1}{4^{\frac{\alpha-1}{\alpha}} \sqrt{2\pi}} \frac{1}{M_r} \left(\frac{1}{\sqrt{nI(\theta_0)}} e^{M_r^2 \left(\frac{\alpha-1}{\alpha} \frac{\gamma_n^2}{2B} - \frac{nI(\theta_0)}{2} \right)} \right). \tag{69}$$

Since $\gamma_n^2 > n$, choosing $n_0 = \min\left\{ n : \left(\frac{\alpha-1}{\alpha} \frac{\gamma_n^2}{2B} - \frac{nI(\theta_0)}{2} \right) > 0 \right\}$ implies that for all $n \geq n_0$, as $r \to \infty$, the left hand side in (69) diverges and the result follows.

■

A.2. Proofs in Section 4

Proof of Lemma 13 Posner (1975, Theorem 1) shows that, the KL divergence $\mathrm{KL}(p(\theta) \| s(\theta))$ is a lower semi-continuous function of the pair $(p(\theta), s(\theta))$ in the weak topology on the space of probability measures. Recall from (6) that the true posterior distribution $\pi(\theta|\mathbf{X}_n)$ converges weakly to δ_{θ_0}, $P_{\theta_0} - a.s.$ Using this fact it follows that

$$\liminf_{n \to \infty} \mathrm{KL}(\pi(\theta|\mathbf{X}_n) \| s_n(\theta)) \geq \mathrm{KL}\left(\delta_{\theta_0} \| s(\theta) \right) \quad P_{\theta_0} - a.s.$$

Next, using Pinsker's inequality Cover (2006) for $\alpha > 1$, we have

$$\mathrm{KL}\left(\delta_{\theta_0} \| s(\theta) \right) \geq \frac{1}{2} \left(\int_\Theta |\delta_{\theta_0} - s(\theta)| \, d\theta \right)^2.$$

Now, fixing $\epsilon > 0$ such that $s(\theta)$ has positive density in the complement of the ball of radius ϵ centered at θ_0, $B(\theta_0, \epsilon)^C$, we have

$$\begin{aligned} \liminf_{n \to \infty} \mathrm{KL}(\pi(\theta|\mathbf{X}_n) \| s_n(\theta)) &\geq \frac{1}{2} \left(\int_{B(\theta_0,\epsilon)} |\delta_{\theta_0} - s(\theta)| \, d\theta + \int_{B(\theta_0,\epsilon)^C} |\delta_{\theta_0} - s(\theta)| \, d\theta \right)^2 \\ &\geq \frac{1}{2} \left(\int_{B(\theta_0,\epsilon)^C} |\delta_{\theta_0} - s(\theta)| \, d\theta \right)^2 \\ &= \frac{1}{2} \left(\int_{B(\theta_0,\epsilon)^C} |-s(\theta)| \, d\theta \right)^2 \quad P_{\theta_0} - a.s. \end{aligned} \tag{70}$$

Since $s(\theta)$ has positive density in the set $B(\theta_0,\epsilon)^C$, there exists $\eta(\epsilon) > 0$, such that

$$\frac{1}{2}\left(\int_{B(\theta_0,\epsilon)^C} |-s(\theta)|\, d\theta\right)^2 \geq \eta(\epsilon),$$

completing the proof. ∎

A.3. Proofs in Section 5

Proof of Lemma 16 We prove the assertion of the Lemma for the class of local latent parameters z_i that have discrete and finite support. First observe that for $\alpha > 1$, using Jensen's inequality

$$M(\mathbf{X}_n|\theta)^\alpha = \min_{q(z_{1:n})\in\mathcal{Q}^n} \int_{\mathcal{Z}^n} q(z_{1:n}) \left(\frac{p(z_{1:n},\mathbf{X}_n|\theta)}{q(z_{1:n})}\right)^\alpha dz_{1:n} \geq \left[\int_{\mathcal{Z}^n} p(z_{1:n},\mathbf{X}_n|\theta)dz_{1:n}\right]^\alpha. \quad (71)$$

Now since family $\mathcal{Q}^n$ contains point masses, we choose a member of family $\mathcal{Q}^n$ which is a joint distribution of point masses at $z^p_{1:n} := \{z^p_1, z^p_2, \ldots, z^p_n\}$ to obtain

$$M(\mathbf{X}_n|\theta)^\alpha = \min_{q(z_{1:n})\in\mathcal{Q}^n} \int_{\mathcal{Z}^n} q(z_{1:n}) \left(\frac{p(z_{1:n},\mathbf{X}_n|\theta)}{q(z_{1:n})}\right)^\alpha dz_{1:n} \leq \left[p(z^p_{1:n},\mathbf{X}_n|\theta)\right]^\alpha, \quad (72)$$

where $z^p_{1:n}$ is as defined in Assumption 6.

Since, $f(x) = x^\alpha$ is increasing for $\alpha > 1$ and $x > 0$, it follows from (71), (72), and monotonicity of the logarithm function that

$$\log \int_{\mathcal{Z}^n} p(z_{1:n},\mathbf{X}_n|\theta)dz_{1:n} \leq \log M(\mathbf{X}_n|\theta) \leq \log p(z^p_{1:n},\mathbf{X}_n|\theta). \quad (73)$$

Now using Assumption 6 (1) and (2(ii)), that is $d_H(z_0, z^p_{1:n}) = o(\rho_n)$, it follows that at some rate ρ_n with $\rho_n \downarrow 0$ and $n\rho_n^2 \to \infty$; that is for all bounded, stochastic $h_n = O_{P_0}(1)$,

$$\begin{aligned}
&\int_{\{z_{1:n}:d_H(z_{1:n},z_0)\geq\rho_n\}} p(z_{1:n}|\mathbf{X}_n, \theta = \theta_0 + n^{-1/2}h_n)dz_{1:n} \\
&\qquad \leq \int_{\{z_{1:n}:d_H(z_{1:n},z^p_{1:n})+d_H(z_0,z^p_{1:n})\geq\rho_n\}} p(z_{1:n}|\mathbf{X}_n, \theta = \theta_0 + n^{-1/2}h_n)dz_{1:n} \\
&\qquad \leq \int_{\{z_{1:n}:d_H(z_{1:n},z^p_{1:n})\geq\rho_n(1-\epsilon)\}} p(z_{1:n}|\mathbf{X}_n, \theta = \theta_0 + n^{-1/2}h_n)dz_{1:n} = o_{P_0}(1),
\end{aligned}$$

where the first inequality follows from using the fact that $d_H(z_{1:n}, z_0) \leq d_H(z_{1:n}, z^p_{1:n}) + d_H(z_0, z^p_{1:n})$, the second inequality uses the fact that $d_H(z_0, z^p_{1:n}) = o(\rho_n)$, that is for some $\epsilon \in (0,1)$, $d_H(z_0, z^p_{1:n}) < \epsilon\rho_n$ for sufficiently large n, and the last inequality is due to Assumption 6 (1).

Therefore, it can be observed from the above result that the conditioned latent posterior $p(z_{1:n}|\mathbf{X}_n, \theta_0)$ concentrates at z_0. Consequently, when the local latent parameters are discrete it follows that

$$\log \int_{\mathcal{Z}^n} p(z_{1:n},\mathbf{X}_n|\theta_0)dz_{1:n} = \log \int_{\mathcal{Z}^n} \frac{p(z_{1:n}|\mathbf{X}_n,\theta_0)}{p(z_{1:n}|\mathbf{X}_n,\theta_0)} p(z_{1:n},\mathbf{X}_n|\theta_0)dz_{1:n} = \log p(z_0,\mathbf{X}_n|\theta_0) + o_{P_0}(1).$$

Now it follows that

$$\log M(\mathbf{X}_n|\theta_0) = \log p(z_0, \mathbf{X}_n|\theta_0) + o_{P_0}(1) = \log \int_{\mathcal{Z}^n} p(z_{1:n}, \mathbf{X}_n|\theta_0) dz_{1:n} + o_{P_0}(1). \tag{74}$$

Subtracting $\log M(\mathbf{X}_n|\theta_0)$ from (73) and using (74) yields

$$\log \frac{\int_{\mathcal{Z}^n} p(z_{1:n}, \mathbf{X}_n|\theta) dz_{1:n}}{\int_{\mathcal{Z}^n} p(z_{1:n}, \mathbf{X}_n|\theta_0) dz_{1:n}} + o_{P_0}(1) \leq \log \frac{M(\mathbf{X}_n|\theta)}{M(\mathbf{X}_n|\theta_0)} \leq \log \frac{p(z_0, \mathbf{X}_n|\theta)}{p(z_0, \mathbf{X}_n|\theta_0)} + o_{P_0}(1). \tag{75}$$

Now, substituting $\theta = \theta_0 + n^{-1/2} h_n$ for all bounded and stochastic $h_n = O_{P_0}(1)$, and using the result in Bickel and Kleijn (2012, Theorem 4.2) under the conditions in Assumption 6 the RHS and LHS above have the same LAN expansion and the result follows. Notice that, by definition, the s-LAN condition in Assumption 2 is also true at $z_{1:n} = z^p_{1:n}$. Assumption 6 (2(ii)) implies $d_H(z_0, z^p_{1:n}) = o(\rho_n)$ with $\rho_n \downarrow 0$ and $n\rho_n^2 \to \infty$, so that

$$\log \frac{P^n_{\theta_0+n^{-1/2}h_n, z^p_{1:n}}}{P^n_{\theta_0, z^p_{1:n}}} = \log \frac{P^n_{\theta_0+n^{-1/2}h_n, z_0}}{P^n_{\theta_0, z_0}} + o(1).$$

Therefore, $\log \frac{p(z_0, \mathbf{X}_n|\theta_0+n^{-1/2}h_n)}{p(z_0, \mathbf{X}_n|\theta_0)} = \log \frac{p(\mathbf{X}_n|z_0, \theta_0+n^{-1/2}h_n)}{p(\mathbf{X}_n|z_0, \theta_0)} + \log \frac{p(z_0|\theta_0+n^{-1/2}h_n)}{p(z_0|\theta_0)} = \log \frac{P^n_{\theta_0+n^{-1/2}h_n, z_0}}{P^n_{\theta_0, z_0}} +$ $o(1)$ also have the same expansion as given in the s-LAN condition in Assumption 2. ■

Proof of Proposition 15 Observe that for any good sequence $\{\bar{q}_n(\theta)\}$ and $q(z_{1:n})$ as point masses (discrete distribution) at the truth $z^0_{1:n} := \{z^0_1, z^0_2, \ldots, z^0_n\}$, we have

$$\begin{aligned}
&\min_{q \in \mathcal{Q}} \min_{q(z_{1:n}) \in \mathcal{Q}^n} D_\alpha(\pi(\theta, z_{1:n}|\mathbf{X}_n) \| q(\theta) q(z_{1:n})) \\
&\qquad = \min_{q(\theta) \in \bar{\mathcal{Q}}, q(z_{1:n}) \in \mathcal{Q}^n} \frac{1}{\alpha - 1} \log \int_{\Theta \times \mathcal{Z}^n} q(\theta) q(z_{1:n}) \left(\frac{p(\theta, z_{1:n}, \mathbf{X}_n)}{p(\mathbf{X}_n) q(\theta) q(z_{1:n})} \right)^\alpha d\theta dz_{1:n} \\
&\qquad \leq \frac{1}{\alpha - 1} \log \int_\Theta \bar{q}_n(\theta) \left(\frac{p(\theta, z^0_{1:n}, \mathbf{X}_n)}{p(\mathbf{X}_n) \bar{q}_n(\theta)} \right)^\alpha d\theta \\
&\qquad \leq \frac{1}{\alpha - 1} \log \int_\Theta \bar{q}_n(\theta) \left(\frac{\pi(\theta, z^0_{1:n}|\mathbf{X}_n)}{\bar{q}_n(\theta)} \right)^\alpha d\theta.
\end{aligned} \tag{76}$$

Also note that, using the definition of $\pi(\theta, z^0_{1:n}|\mathbf{X}_n)$, we have

$$\pi(\theta, z^0_{1:n}|\mathbf{X}_n) = \frac{\pi(\theta)\pi(z^0_{1:n}|\theta) p(\mathbf{X}_n|\theta, z^0_{1:n})}{\int_{\Theta \times \mathcal{Z}^n} \pi(\theta)\pi(z_{1:n}|\theta) p(\mathbf{X}_n|\theta, z_{1:n}) d\theta dz_{1:n}} \leq \frac{\pi(\theta)\pi(z^0_{1:n}|\theta) p(\mathbf{X}_n|\theta, z^0_{1:n})}{\int_\Theta \pi(\theta)\pi(z^0_{1:n}|\theta) p(\mathbf{X}_n|\theta, z^0_{1:n}) d\theta}, \tag{77}$$

where the second inequality follows from the fact that $z_{1:n}$ is a discrete random variable. Therefore substituting (77) into (76) yields

$$\begin{aligned}\min_{q\in\mathcal{Q}}\min_{q(z_{1:n})\in\mathcal{Q}^n} D_\alpha(\pi(\theta,z_{1:n}|\mathbf{X}_n)\|q(\theta)q(z_{1:n})) &\leq \frac{1}{\alpha-1}\log\int_\Theta \bar{q}_n(\theta)\left(\frac{\pi(\theta)p(\mathbf{X}_n,z^0_{1:n}|\theta)}{\bar{q}_n(\theta)\int_\Theta\pi(\theta)p(\mathbf{X}_n,z^0_{1:n}|\theta)d\theta}\right)^\alpha d\theta \\ &= \frac{1}{\alpha-1}\log\int_\Theta \bar{q}_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n,z^0_{1:n})}{\bar{q}_n(\theta)}\right)^\alpha d\theta \\ &=: D_\alpha(\pi(\theta|\mathbf{X}_n,z^0_{1:n})\|\bar{q}_n(\theta)). \end{aligned} \tag{78}$$

Therefore, for the second part, it suffices to show that

$$D_\alpha(\pi(\theta|\mathbf{X}_n,z^0_{1:n})\|\bar{q}_n(\theta)) < B + o_{P_0}(1).$$

The subsequent arguments in the proof are for any $n \geq \max(n_1,n_2,n_3,n_M)$, where n_1,n_2, and n_3 are defined in Assumption 4. First observe that, for any compact ball K containing the true parameter θ_0,

$$\begin{aligned}&\frac{\alpha-1}{\alpha}D_\alpha(\pi(\theta|\mathbf{X}_n,z^0_{1:n})\|\bar{q}_n(\theta)) \\ &\quad= \frac{1}{\alpha}\log\left(\int_K \bar{q}_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n,z^0_{1:n})}{\bar{q}_n(\theta)}\right)^\alpha d\theta + \int_{\Theta\setminus K}\bar{q}_n(\theta)\left(\frac{\pi(\theta|\mathbf{X}_n,z^0_{1:n})}{\bar{q}_n(\theta)}\right)^\alpha d\theta\right). \end{aligned} \tag{79}$$

First, we approximate the first integral on the right hand side using the LAN condition in Assumption 2. Let $\Delta_{n,(\theta_0,z_0)} := \sqrt{n}(\hat{\theta}_n-\theta_0)$, where $\hat{\theta}_n \to \theta_0$, $P_0-a.s.$ and $\Delta_{n,(\theta_0,z_0)}$ converges in distribution to $\mathcal{N}(0,I(\theta_0,z_0)^{-1})$ (van der Vaart, 1998, Lemma 25.23 and 25.25). Now the proof follows similar steps as used in the proof of Proposition 8. ■

References

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *Annals of Statistics*, 48(3):1475–1497, 2020.

Peter J. Bickel and Bas J.K. Kleijn. The semiparametric Bernstein-von Mises theorem. *The Annals of Statistics*, 40(1):206–237, 2012.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12 (2):2995–3035, 2018.

Thomas M. Cover. *Elements of Information Theory.* Wiley-Interscience, Hoboken, N.J., 2nd ed.. edition, 2006. ISBN 0471241954.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David M. Blei. Variational inference via χ-upper bound minimization. In *Advances in Neural Information Processing Systems*, pages 2732–2741, 2017.

William Feller. *An Introduction to Probability Theory and its Applications.* Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 1968. ISBN 9780471257080.

Subhashis Ghosal. A review of consistency and convergence of posterior distribution. In *Varanashi Symposium in Bayesian Inference, Banaras Hindu University*, 1997.

Bogdan Grechuk, Anton Molyboha, and Michael Zabarankin. Maximum entropy principle with general deviation measures. *Mathematics of Operations Research*, 34(2):445–467, 2009.

Roger B. Grosse, Zoubin Ghahramani, and Ryan P. Adams. Sandwiching the marginal likelihood using bidirectional Monte Carlo. *Stat*, 1050:8, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Bas J.K. Kleijn and Aad W. van der Vaart. The Bernstein-von-Mises Theorem under Misspecification. *Electronic Journal of Statistics*, 6(0):354–381, 2012.

Yingzhen Li and Richard E. Turner. Rényi Divergence Variational Inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

Thomas P. Minka. Expectation Propagation for Approximate Bayesian Inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369, 2001a.

Thomas P. Minka. *A Family of Algorithms for Approximate Bayesian Inference.* PhD thesis, Massachusetts Institute of Technology, 2001b.

Susan A. Murphy and Aad W. van der Vaart. Likelihood inference in the errors-in-variables model. *Journal of Multivariate Analysis*, 59(1):81–108, 1996.

Susan A. Murphy and Aad W. van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000.

Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

Edward Posner. Random coding strategies for minimum entropy. *Information Theory, IEEE Transactions on*, 21(4):388–391, 1975. ISSN 0018-9448.

Lorraine Schwartz. On Bayes procedures. *Probability Theory and Related Fields*, 4(1):10–26, 1965.

Richard E. Turner and Maneesh Sahani. Two problems with variational expectation maximisation for time-series models. In D. Barber, T. Cemgil, and S. Chiappa, editors, *Bayesian Time Series Models*, chapter 5, pages 109–130. Cambridge University Press, 2011.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Tim Van Erven and Peter Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Yixin Wang and David M. Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2018.

Roderick Wong. II - Classical Procedures. In Roderick Wong, editor, *Asymptotic Approximations of Integrals*, pages 55 – 146. Academic Press, 1989. ISBN 978-0-12-762535-5.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *Annals of Statistics*, 48(2):886–905, 2020.

Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *Annals of Statistics*, 2020. (To appear).