

α -Variational Inference with Statistical Guarantees

Yun Yang ^{*1}, Debdeep Pati^{†2}, and Anirban Bhattacharya^{‡2}

¹Department of Statistics, Florida State University

²Department of Statistics, Texas A&M University

Abstract

We propose a family of variational approximations to Bayesian posterior distributions, called α -VB, with provable statistical guarantees. The standard variational approximation is a special case of α -VB with $\alpha = 1$. When $\alpha \in (0, 1]$, a novel class of variational inequalities are developed for linking the Bayes risk under the variational approximation to the objective function in the variational optimization problem, implying that maximizing the evidence lower bound in variational inference has the effect of minimizing the Bayes risk within the variational density family. Operating in a frequentist setup, the variational inequalities imply that point estimates constructed from the α -VB procedure converge at an optimal rate to the true parameter in a wide range of problems. We illustrate our general theory with a number of examples, including the mean-field variational approximation to (low)-high-dimensional Bayesian linear regression with spike and slab priors, mixture of Gaussian models, latent Dirichlet allocation, and (mixture of) Gaussian variational approximation in regular parametric models.

KEYWORDS: *Bayes risk; Evidence lower bound; Latent variable models; Rényi divergence; Variational inference.*

1 Introduction and preliminaries

Variational inference [25, 38] is a widely-used tool for approximating complicated probability densities, especially those arising as posterior distributions from complex hierarchical Bayesian models. It provides an alternative strategy to Markov chain Monte Carlo (MCMC,

*yyang@stat.fsu.edu

†debdeep@stat.tamu.edu

‡anirbanb@stat.tamu.edu

[20, 15]) sampling by turning the sampling/inference problem into an optimization problem, where a closest member, relative to the Kullback–Leibler (KL) divergence, in a family of approximate densities is picked out as a proxy to the target density. Variational inference has found its success in a variety of contexts, especially in models involving latent variables, such as Hidden Markov models [30], graphical models [3, 38], mixture models [23, 14, 35], and topic models [9, 11] among others. See the recent review paper [10] by Blei et al. for a comprehensive introduction to variational inference.

The popularity of variational methods can be largely attributed to their computational advantages over MCMC. It has been empirically observed in many applications that variational inference operates orders of magnitude faster than MCMC for achieving the same approximation accuracy. Moreover, compared to MCMC, variational inference tends to be easier to scale to big data due to its inherent optimization nature, and can take advantage of modern optimization techniques such as stochastic optimization [27, 26] and distributed optimization [1]. However, unlike MCMC that is guaranteed to produce (almost) exact samples from the target density for ergodic chains [33], variational inference does not enjoy such general theoretical guarantee.

Several threads of research have been devoted to characterize statistical properties of the variational proxy to the true posterior distribution; refer to Section 5.2 of [10] for a relatively comprehensive survey of the theoretical literature on variational inference. However, almost all these studies are conducted in a case-by-case manner, by either explicitly analyzing the fixed point equation of the variational optimization problem, or directly analyzing the iterative algorithm for solving the optimization problem. In addition, these analyses require certain structural assumptions on the priors such as conjugacy, and is not applicable to broader classes of priors.

This article introduces a novel class of variational approximations and studies their large sample convergence properties in a unified framework. The new variational approximation, termed α -VB, introduces a fixed temperature parameter α inside the usual VB objective function which controls the relative trade-off between model-fit and prior regularization. The usual VB approximation is retained as a special case corresponding to $\alpha = 1$. The α -VB objective function is partly motivated by *fractional posteriors* [39, 4]; specific connections are drawn in §2.1. The general α -VB procedure also inherits all the computational tractability and scalability from the $\alpha = 1$ case, and implementation-wise only requires simple modifications to existing variational algorithms.

For $\alpha \in (0, 1]$, we develop novel variational inequalities for the Bayes risk under the variational solution. These variational inequalities link the Bayes risk with the α -VB objective function, implying that maximizing the evidence lower bound has the effect of minimizing

the Bayes risk within the variational density family. A crucial upshot of this analysis is that point estimates constructed from the variational posterior concentrate at the true parameter at the same rate as those constructed from the actual posterior for a variety of problems. There is now a well-developed literature on the frequentist concentration properties of posterior distributions in nonparametric problems; refer to [34] for a detailed review, and the present paper takes a step towards developing similar general-purpose theoretical guarantees for variational solutions. We applied our theory to a number of examples where VB is commonly used, including mean-field variational approximation to high-dimensional Bayesian linear regression with spike and slab priors, mixtures of Gaussian models, latent Dirichlet allocation, and Gaussian-mixture variational approximation to regular parametric models.

The $\alpha < 1$ case is of particular interest as the major ingredient of the variational inequality involves the prior mass assigned to appropriate Kullback–Leibler neighborhoods of the truth which can be bounded in a straightforward fashion in the aforesaid models and beyond. We mention here a recent preprint by Alquier and Ridgeway [2] where variational approximations to tempered posteriors (without latent variables) are conducted. The α -VB objective function considered here incorporates a much broader class of models involving latent variables, and the corresponding variational inequality recovers the risk bound of [2] when no latent variables are present. The variational inequalities for the $\alpha < 1$ case do not immediately extend to the $\alpha = 1$ case under a simple limiting operation, and require a separate treatment under stronger assumptions. In particular, we make use of additional testability assumptions on the likelihood function detailed in §3.2. Similar assumptions have been used to study concentration of the usual posterior [17].

It is a well-known fact [41, 43] that the covariance matrices from the variational approximations are typically “too small” compared with those for the sampling distribution of the maximum likelihood estimator, which combined with the Bernstein von-Mises theorem [36] implies that the variational approximation may not converge to the true posterior distribution. This fact combined with our result illustrate the landscape of variational approximation—minimizing the KL divergence over the variational family forces the variational distribution to concentrate around the truth at the optimal rate (due to the heavy penalty on the tails in the KL divergence); however, the local shape of the obtained density function around the truth can be far away from that of the true posterior due to mis-match between the distributions in the variational family and the true posterior. Overall, our results reveal that concentration of the posterior measure is not only useful in guaranteeing desirable statistical properties, but also has computational benefits in certifying consistency and concentration of variational approximations.

In the remainder of this section, we introduce key notation used in the paper and provide

necessary background on variational inference.

1.1 Notation

We briefly introduce notation that will be used throughout the paper. Let $h(p||q) = (\int (p^{1/2} - q^{1/2})^2 d\mu)^{1/2}$ and $D(p||q) = \int p \log(p/q) d\mu$ denote the Hellinger distance and Kullback–Leibler divergence, respectively, between two probability density functions p and q relative to a common dominating measure μ . We define an additional discrepancy measure $V(p||q) = \int p \log^2(p/q) d\mu$, which will be referred to as the V -divergence. For a set A , we use the notation I_A to denote its indicator function. For any vector μ and positive semidefinite matrix Σ , we use $\mathcal{N}(\mu, \Sigma)$ to denote the normal distribution with mean μ and covariance matrix Σ , and use $\mathcal{N}(\theta; \mu, \Sigma)$ to denote its pdf at θ .

For any $\alpha \in (0, 1)$, let

$$D_\alpha(p||q) = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} d\mu \quad (1.1)$$

denote the Rényi divergence of order α . Jensen’s inequality implies that $D_\alpha(p||q) \geq 0$ for any $\alpha \in (0, 1)$, and the equality holds if and only if $p = q$. The Hellinger distance can be related with the α -divergence with $\alpha = 1/2$ by $D_{1/2}(p||q) = -2 \log\{1 - (1/2)h^2(p||q)\} \geq h^2(p||q)$ using the inequality $\log(1+t) < t$ for $t > -1$. More details and properties of the α -divergence can be found in [37].

1.2 Review of variational inference

Suppose we have observations $Y^n = (Y_1, \dots, Y_n) \in \mathcal{Y}^n$ with n denoting the sample size. Let $\mathbb{P}_\theta^{(n)}$ be the distribution of Y^n given parameter $\theta \in \Theta$ that admits a density $p_\theta^{(n)}$ relative to the Lebesgue measure. We will also interchangeably use $P(Y^n | \theta)$ and $p(Y^n | \theta)$ to denote $\mathbb{P}_\theta^{(n)}$ and its density function (likelihood function) $p_\theta^{(n)}$. Assume additionally that the likelihood $p(Y^n | \theta)$ can be represented as

$$p(Y^n | \theta) = \sum_{s^n} p(Y^n | S^n = s^n, \theta) p(S^n = s^n | \theta),$$

where S^n denotes a collection of latent or unobserved variables; the superscript n signifies the possible dependence of the number of latent variables on n ; for example, when there are observation specific latent variables. In certain situations, the latent variables may be introduced for purely computational reasons to simplify an otherwise intractable likelihood, such as the latent cluster indicators in a mixture model. Alternatively, a complex proba-

bilistic model $p(Y^n | \theta)$ may itself be defined in a hierarchical fashion by first specifying the distribution of the data given latent variables and parameters, and then specifying the latent variable distribution given parameters; examples include the latent Dirichlet allocation and many other prominent Bayesian hierarchical models. For ease of presentation, we have assumed discrete latent variables in the above display and continue to do so subsequently, although our development seamlessly extends to continuous latent variables by replacing sums with integrals; further details are provided in a supplemental document.

Let P_θ denote a prior distribution on θ with density function p_θ , and denote $W^n = (\theta, S^n) \in \mathcal{W}^n$. In a Bayesian framework, all inference is based on the augmented posterior density $p(W^n | Y^n)$ given by

$$p(W^n | Y^n) = p(\theta, S^n | Y^n) \propto p(Y^n | \theta, S^n) p(S^n | \theta) p_\theta(\theta). \quad (1.2)$$

In many cases, $p(W^n | Y^n)$ can be inconvenient for conducting direct analysis due to its intractable normalizing constant and expensive to sample from due to the slow mixing of standard MCMC algorithms. Variational inference aims to bypass these difficulties by turning the inference problem into an optimization problem, which can be solved by using iterative algorithms such as coordinate descent [7] and alternating minimization.

Let Γ denote a pre-specified family of density functions over \mathcal{W}^n that can be either parameterized by some “variational parameters”, or required to satisfy some structural constraints (see below for examples of Γ). The goal of variational inference is to approximate this conditional density $p(W^n | Y^n)$ by finding the closest member of this family in KL divergence to the conditional density $p(W^n | Y^n)$ of interest, that is, computing the minimizer

$$\begin{aligned} \hat{q}_{W^n} &:= \operatorname{argmin}_{q_{W^n} \in \Gamma} D[q_{W^n}(\cdot) || p(\cdot | Y^n)] \\ &= \operatorname{argmin}_{q_{W^n} \in \Gamma} \left\{ - \int_{\mathcal{W}^n} q_{W^n}(w^n) \log \frac{p(w^n | Y^n)}{q_{W^n}(w^n)} dw^n \right\} \\ &= \operatorname{argmin}_{q_{W^n} \in \Gamma} \left\{ - \underbrace{\int_{\mathcal{W}^n} q_{W^n}(w^n) \log \frac{p(Y^n | w^n) p_{W^n}(w^n)}{q_{W^n}(w^n)} dw^n}_{L(q_{W^n})} \right\} \end{aligned} \quad (1.3)$$

where the last step follows by using Bayes’ rule and the fact that the marginal density $p(Y^n)$ does not depend on W^n and q_{W^n} . The function $L(q_{W^n})$ inside the argmin-operator above (without the negative sign) is called the evidence lower bound (ELBO, [10]) since it provides

a lower bound to the log evidence $\log p(Y^n)$,

$$\log p(Y^n) = L(q_{W^n}) + D[q_{W^n}(\cdot) \parallel p(\cdot | Y^n)] \geq L(q_{W^n}), \quad (1.4)$$

where the equality holds if and only if $q_{W^n} = p(\cdot | Y^n)$. The decomposition (1.4) provides an alternative interpretation of variational inference to the original derivation from Jensen’s inequality[25]—minimizing the KL divergence over the variational family Γ is equivalent to maximizing the ELBO over Γ . When Γ is composed of all densities over \mathscr{W}^n , this variational approximation \hat{q}_{W^n} exactly recovers $p(W^n | Y^n)$. In general, the variational family Γ is chosen to balance between the computational tractability and the approximation accuracy. Some common examples of Γ are provided below.

Example: (Exponential variational family) When there is no latent variable and $W^n = \theta \in \Theta$ corresponds to the parameter in the model, a popular choice of the variational family is an exponential family of distributions. Among the exponential variational families, the Gaussian variational family, $q_\theta(\theta; \mu, \Sigma) \equiv \mathcal{N}(\theta; \mu, \Sigma)$ for $\theta \in \mathbb{R}^d$, is the most widely-used owing to the Bernstein von-Mises theorem (Section 10.2 of [36]), stating that for regular parametric models, the posterior distribution converges to a Gaussian limit relative to the total variation metric as the sample size tends to infinity. There are also some recent developments by replacing the single Gaussian with a Gaussian-mixture as the variational family to improve finite-sample approximation [47], which is useful when the posterior distribution is skewed or far away from Gaussian for the given sample size.

Example: (Mean-field variational family) Suppose that W^n can be decomposed into m components (or blocks) as $W^n = (W_1, W_2, \dots, W_m)$ for some $m > 1$, where each component $W_j \in \mathscr{W}_j$ can be multidimensional. The mean-field variational family Γ_{MF} is composed of all density functions over $\mathscr{W}^n = \prod_{j=1}^m \mathscr{W}_j$ that factorizes as

$$q_{W^n}(w^n) = \prod_{j=1}^m q_{W_j}(w_j), \quad w^n = (w_1, \dots, w_m) \in \mathscr{W}^n,$$

where each variational factor q_{W_j} is a density function over \mathscr{W}_j for $j = 1, \dots, m$. A natural mean-field decomposition is to let $q_{W^n}(w^n) = q_\theta(\theta) q_{S^n}(s^n)$, with q_{S^n} often further decomposed as $q_{S^n}(s^n) = \prod_{i=1}^n q_{S_i}(s_i)$.

Note that we have not specified the parametric form of the individual variational factors, which are determined by properties of the model— in some cases, the optimal q_{W_j} is in

the same parametric family as the conditional distribution of W_j given the parameter. The corresponding mean-field variational approximation \widehat{q}_{W^n} , which is necessarily of the form $\prod_{j=1}^m \widehat{q}_{W_j}(w_j)$, can be computed via the coordinate ascent variational inference (CAVI) algorithm [7, 10] which iteratively optimizes each variational factor keeping others fixed at their present value and resembles the EM algorithm in the presence of latent variables.

The mean-field variational family can be further constrained by restricting each factor q_{W_j} to belong to a parametric family, such as the exponential family in the previous example. In particular, it is a common practice to restrict the variational density q_θ of the parameter into a structured family (for example, the mean-field family if θ is multi-dimensional), which will be denoted by Γ_θ in the sequel.

The rest of the paper is organized as follows. In §2, we introduce the α -VB objective function and relate it to usual VB. §3 presents our general theoretical results concerning finite sample risk bounds for the α -VB solution. In §4, we apply the theory to concrete examples. We conclude with a discussion in §5. All proofs and some additional discussions are provided in a separate supplemental document. The supplemental document also contains a detailed simulation study.

2 The α -VB procedure

Before introducing the proposed family of objective functions, we first represent the KL term $D[q_{W^n}(\cdot) \parallel p(\cdot | Y^n)]$ in a more convenient form which provides intuition into how VB works in the presence of latent variables and aids our subsequent theoretical development.

2.1 A further decomposition of the ELBO

To aid our subsequent development, we introduce some additional notation and make some simplifying assumptions. First, we decompose $\theta = (\mu, \pi)$, with $p(Y^n | S^n = s^n, \theta) = p(Y^n | S^n = s^n, \mu)$ and $\pi_{s^n} := p(S^n = s^n | \theta)$. In other words, μ is the parameter characterizing the conditional distribution $P(Y^n | S^n, \mu)$ of the observation Y^n given latent variable S^n , and $\pi = (\pi_{s^n})$ characterizes the distribution $P(S^n | \pi)$ of the latent variables. We shall also assume the mean-field decomposition

$$q_{W^n}(w^n) = q_\theta(\theta) q_{S^n}(s^n) \tag{2.1}$$

throughout, and let $\Gamma = \Gamma_\theta \times \Gamma_{S^n}$ denote the class of such product variational distributions. When necessary subsequently, we shall further assume $q_{S^n}(s^n) = \prod_{i=1}^n q_{S_i}(s_i)$ and $q_\theta(\theta) =$

$q_\mu(\mu) q_\pi(\pi)$, which however is not immediately necessary for this subsection.

The KL divergence $D[q_{W^n}(\cdot) \parallel p(\cdot | Y^n)]$ in (1.3) involves both parameters and latent variables. Separating out the KL divergence for the parameter part leads to the equivalent representation

$$D[q_{W^n}(\cdot) \parallel p(\cdot | Y^n)] = D(q_\theta \parallel p_\theta) - \underbrace{\int_{\Theta} \left[\sum_{s^n} q_{S^n}(s^n) \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{q_{S^n}(s^n)} \right] q_\theta(d\theta)}_{\widehat{\ell}_n(\theta)}. \quad (2.2)$$

Observe that, using concavity of $x \mapsto \log x$ and Jensen's inequality,

$$\log p(Y^n | \theta) = \log \left[\sum_{s^n} q_{S^n}(s^n) \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{q_{S^n}(s^n)} \right] \geq \sum_{s^n} q_{S^n}(s^n) \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{q_{S^n}(s^n)}.$$

The quantity $\widehat{\ell}_n(\theta)$ in (2.2) can therefore be recognized as an approximation (from below) to the log likelihood $\ell_n(\theta) := \log p(Y^n | \theta)$ in terms of the latent variables. Define an average Jensen gap Δ_J due to the variational approximation to the log-likelihood,

$$\Delta_J(q_\theta, q_{S^n}) = \int_{\Theta} [\ell_n(\theta) - \widehat{\ell}_n(\theta)] q_\theta(d\theta) \geq 0.$$

With this, write the KL divergence $D[q_{W^n}(\cdot) \parallel p(\cdot | Y^n)]$ as

$$D[q_{W^n}(\cdot) \parallel p(\cdot | Y^n)] = - \int_{\Theta} \ell_n(\theta) q_\theta(d\theta) + \Delta_J(q_\theta, q_{S^n}) + D(q_\theta \parallel p_\theta), \quad (2.3)$$

which splits as a sum of three terms: an integrated (w.r.t. the variational distribution) negative log-likelihood, the KL divergence between the variational distribution q_θ and the prior p_θ for θ , and the Jensen gap Δ_J due to the latent variables. In particular, the role of the latent variable variational distribution q_{S^n} is conveniently confined to Δ_J .

Another view of the above is an equivalent formulation of the ELBO decomposition (1.4),

$$\log p(Y^n) = L(q_{W^n}) + \Delta_J(q_\theta, q_{S^n}) + D[q_\theta(\theta) \parallel p(\theta | Y^n)]. \quad (2.4)$$

which readily follows since

$$D[q_\theta(\theta) \parallel p(\theta | Y^n)] = - \int_{\Theta} \ell_n(\theta) q_\theta(d\theta) + D(q_\theta \parallel p_\theta).$$

Thus, in latent variable models, maximizing the ELBO $L(q_{W^n})$ is equivalent to minimizing a sum of the Jensen gap Δ_J and the KL divergence between the variational density and the

posterior density of the parameters. When there is no likelihood approximation with latent variables, $\Delta_J = 0$.

2.2 The α -VB objective function

Here and in the rest of the paper, we adopt the frequentist perspective by assuming that there is a true data generating model $\mathbb{P}_{\theta^*}^{(n)}$ that generates the data Y^n , and θ^* will be referred to as the true parameter, or simply truth. Let $\ell_n(\theta, \theta^*) = \ell_n(\theta) - \ell_n(\theta^*)$ be the log-likelihood ratio. Define

$$\Psi_n(q_\theta, q_{S^n}) = - \int_{\Theta} \ell_n(\theta, \theta^*) q_\theta(d\theta) + \Delta_J(q_\theta, q_{S^n}) + D(q_\theta || p_\theta), \quad (2.5)$$

and observe that Ψ_n differs from the KL divergence $D[q_{W^n}(\cdot) || p(\cdot | Y^n)]$ in (2.3) only by $\ell_n(\theta^*)$ which does not involve the variational densities. Hence, minimizing $D[q_{W^n}(\cdot) || p(\cdot | Y^n)]$ is equivalent to minimizing $\Psi_n(q_\theta, q_{S^n})$. We note here that the introduction of the $\ell_n(\theta^*)$ term is to develop theoretical intuition and the actual minimization does not require the knowledge of θ^* .

The objective function Ψ_n in (2.5) elucidates the trade-off between model-fit and fidelity to the prior underlying a variational approximation, which is akin to the classical bias-variance trade-off for shrinkage or penalized estimators. The model-fit term consists of two constituents: the first term is an averaged (with respect to the variational distribution) log-likelihood ratio which tends to get small as the variational distribution q_θ places more mass near the true parameter θ^* , while the second term is the Jensen gap Δ_J due to the variational approximation with the latent variables. On the other hand, the regularization or penalty term $D(q_\theta || p_\theta)$ prevents over-fitting to the data by constricting the KL divergence between the variational solution and the prior.

In this article, we study a wider class of variational objective functions $\Psi_{n,\alpha}$ indexed by a scalar parameter $\alpha \in (0, 1]$ which encompass the usual VB,

$$\Psi_{n,\alpha}(q_\theta, q_{S^n}) = - \underbrace{\int_{\Theta} \ell_n(\theta, \theta^*) q_\theta(d\theta)}_{\text{model fit}} + \underbrace{\alpha^{-1} D(q_\theta || p_\theta)}_{\text{regularization}}, \quad (2.6)$$

and define the α -VB solution as

$$(\hat{q}_{\theta,\alpha}, \hat{q}_{S^n,\alpha}) = \operatorname{argmin}_{(q_\theta, q_{S^n}) \in \Gamma} \Psi_{n,\alpha}(q_\theta, q_{S^n}). \quad (2.7)$$

Observe that the α -VB criterion $\Psi_{n,\alpha}$ differs from Ψ_n only in the regularization term, where

the inverse temperature parameter α controls the amount of regularization, with smaller α implying a stronger penalty. When $\alpha = 1$, $\Psi_{n,\alpha}$ reduces to the usual variational objective function Ψ_n in (2.5), and we shall denote the solution of (2.7) by \widehat{q}_θ and \widehat{q}_{S^n} as before. As we shall see in the sequel, the introduction of the temperature parameter α substantially simplifies the theoretical analysis and allows one to certify (near-)minimax optimality of the α -VB solution for $\alpha < 1$ under only a prior mass condition, whereas analysis of the usual VB solution ($\alpha = 1$) requires more intricate testing arguments.

The α -VB solution can also be interpreted as the minimizer of a certain divergence function between the product variational distribution $q_\theta(\theta) \times q_{S^n}(s^n)$ and the joint α -fractional posterior distribution [4] of (θ, S^n) ,

$$P_\alpha(\theta \in B, s^n | Y^n) = \frac{\int_B [p(Y^n | \mu, s^n) \pi_{s^n}]^\alpha p_\theta(\theta) d\theta}{\int_\Theta \sum_{s^n} [p(Y^n | \mu, s^n) \pi_{s^n}]^\alpha p_\theta(\theta) d\theta}, \quad (2.8)$$

which is obtained by raising the joint likelihood of (θ, s^n) to the fractional power α , and combining with the prior p_θ using Bayes' rule. We shall use $p_\alpha(\cdot | Y^n)$ to denote the fractional posterior density. The fractional posterior is a specific example of a Gibbs posterior [24] and shares a nice coherence property with the usual posterior when viewed as a mechanism for updating beliefs [8].

Proposition 2.1 (Connection with fractional posteriors). *The α -VB solution $(\widehat{q}_{\theta,\alpha}, \widehat{q}_{S^n,\alpha})$ satisfy,*

$$(\widehat{q}_{\theta,\alpha}, \widehat{q}_{S^n,\alpha}) = \operatorname{argmin}_{(q_\theta, q_{S^n}) \in \Gamma} \left[D[q_{W^n}(\cdot) || p_\alpha(\cdot | Y^n)] + (1 - \alpha)\mathcal{H}(q_{S^n}) \right],$$

where $\mathcal{H}(q_{S^n}) = -\sum_{s^n} q_{S^n}(s^n) \log q_{S^n}(s^n)$ is the entropy of q_{S^n} , and $p_\alpha(\cdot | Y^n)$ is the joint α -fractional posterior density of $w^n = (\theta, s^n)$.

The entropy term $\mathcal{H}(q_{S^n})$ encourages the latent-variable variational density q_{S^n} to be concentrated to the uniform distribution, in addition to minimizing the KL divergence between $q_{W^n}(\cdot)$ and $p_\alpha(\cdot | Y^n)$. In particular, if there are no latent variables, the entropy term disappears and the objective function reduces to a KL divergence between q_θ and $p_\alpha(\theta | Y^n)$.

We conclude this section by remarking that the additive decomposition of the model-fit term in (2.6) provides a peak into why mean-field approximations work for latent variable models, since the roles of the variational density q_{S^n} for the latent variables and q_θ for the model parameters are de-coupled. Roughly speaking, a good choice of q_{S^n} should aim to make the Jensen gap Δ_J small, while the choice of q_θ should balance the integrated log-likelihood ratio and the penalty term. This point is crucial for the theoretical analysis.

3 Variational risk bounds for α -VB

In this section, we investigate concentration properties of the α -VB posterior under a frequentist framework assuming the existence of a true data generating parameter θ^* . We first focus on the $\alpha < 1$ case, and then separately consider the $\alpha = 1$ case. The main take-away message from our theoretical results below is that under fairly general conditions, the α -VB procedure concentrates at the true parameter at the same rate as the actual posterior, and as a result, point estimates obtained from the α -VB can provide rate-optimal frequentist estimators. These results thus compliment the empirical success of VB in a wide variety of models.

We present our results in the form of Bayes risk bounds for the variational distribution. Specifically, for a suitable loss function $r(\theta, \theta^*)$, we aim to obtain a high-probability (under the data generating distribution $\mathbb{P}_{\theta^*}^{(n)}$) to the variational risk

$$\int r(\theta, \theta^*) \widehat{q}_{\theta, \alpha}(d\theta). \quad (3.1)$$

In particular, if $r(\cdot, \cdot)$ is convex in its first argument, then the above risk bound immediately translates into a risk bound for the α -VB point estimate $\widehat{\theta}_{\text{VB}, \alpha} = \int \theta \widehat{q}_{\theta, \alpha}(d\theta)$ using Jensen's inequality:

$$r(\widehat{\theta}_{\text{VB}, \alpha}, \theta^*) \leq \int r(\theta, \theta^*) \widehat{q}_{\theta, \alpha}(d\theta).$$

Specifically, our goal will be to establish general conditions under which $\widehat{\theta}_{\text{VB}, \alpha}$ concentrates around θ^* at the minimax rate for the particular problem.

3.1 Risk bounds for the $\alpha < 1$ case:

We use the shorthand

$$\frac{1}{n} D_{\alpha}^{(n)}(\theta, \theta^*) := \frac{1}{n} D_{\alpha}[p_{\theta}^{(n)} || p_{\theta^*}^{(n)}]$$

to denote the averaged α -divergence between $\mathbb{P}_{\theta}^{(n)}$ and $\mathbb{P}_{\theta^*}^{(n)}$. We adopt the theoretical framework of [4] to use this divergence as our loss function $r(\theta, \theta^*)$ for measuring the closeness between any $\theta \in \Theta$ and the truth θ^* . Note that in case of i.i.d. observations, this averaged divergence $n^{-1} D_{\alpha}^{(n)}(\theta, \theta^*)$ simplifies to $D_{\alpha}[p_{\theta} || p_{\theta^*}]$, which is stronger than the squared Hellinger distance $h^2[p_{\theta} || p_{\theta^*}]$ between p_{θ} and p_{θ^*} for any fixed $\alpha \in [1/2, 1)$.

Our first main result provides a general finite-sample upper bound to the variational Bayes risk (3.1) for the above choice of $r(\theta, \theta^*)$.

Theorem 3.1 (Variational risk bound). *Recall the α -VB objective function $\Psi_{n,\alpha}(q_\theta, q_{S^n})$ from (2.6). For any $\zeta \in (0, 1)$, it holds with $\mathbb{P}_{\theta^*}^n$ probability at least $(1 - \zeta)$ that for any probability measure $q_\theta \in \Gamma_\theta$ with $q_\theta \ll p_\theta$ and any probability measure $q_{S^n} \in \Gamma_{S^n}$ on S^n ,*

$$\int \frac{1}{n} D_\alpha^{(n)}(\theta, \theta^*) \widehat{q}_{\theta,\alpha}(\theta) d\theta \leq \frac{\alpha}{n(1-\alpha)} \Psi_{n,\alpha}(q_\theta, q_{S^n}) + \frac{1}{n(1-\alpha)} \log(1/\zeta).$$

Here and elsewhere, the probability statement is uniform over all $(q_\theta, q_{S^n}) \in \Gamma$. Theorem 3.1 links the variational Bayes risk for the α -divergence to the objective function $\Psi_{n,\alpha}$ in (2.6). As a consequence, minimizing $\Psi_{n,\alpha}$ in (2.6) has the same effect as minimizing the variational Bayes risk. To apply Theorem 3.1 to various problems, we now discuss strategies to further analyze and simplify $\Psi_{n,\alpha}$ under appropriate structural constraints of Γ_θ and Γ_{S^n} . To that end, we make some simplifying assumptions.

First, we assume a further mean-field decomposition $q_{S^n}(s^n) = \prod_{i=1}^n q_{S_i}(s_i)$ for the latent variables S^n , where each factor q_{S_i} is restriction-free. Second, the inconsistency of the mean-field approximation for state-space models proved in [40] indicates that this mean-field approximation for the latent variables may not generally work for non-independent observations with non-independent latent variables. For this reason, we assume that the observation latent variable pair (S_i, Y_i) are mutually independent across $i = 1, 2, \dots, n$. In fact, we assume that (S_i, Y_i) are i.i.d. copies of (S, Y) whose density function is given by $p(S, Y | \mu, \pi) = p(Y | S, \mu) p(S | \pi)$. Following earlier notation, let $\pi_S := p(S | \pi)$ denote the probability mass function of the i.i.d. discrete latent variables $\{S_i\}$, with the parameter $\pi = (\pi_1, \pi_2, \dots, \pi_K)$ residing in the K -dim simplex $\mathcal{S}_K = \{\pi \in [0, 1]^K : \sum_k \pi_k = 1\}$. Finally, we assume the variational family Γ_θ of the parameter decomposes into $\Gamma_\mu \otimes \mathcal{S}_K$, where Γ_μ denotes variational family for parameter μ .

Let $p(Y | \theta) = \sum_{s=1}^K \pi_s p(Y | \theta, S = s)$ denote the marginal probability density function of the i.i.d. observations $\{Y_i\}$. The i.i.d. assumption implies a simplified structure of various quantities encountered before, e.g. $\pi_{S^n} = \prod_{i=1}^n \pi_{s_i}$, $p(Y^n | \mu, S^n) = \prod_{i=1}^n \pi_{s_i} p(Y_i | \mu, S_i)$ and $p(Y^n | \theta) = \prod_{i=1}^n p(Y_i | \theta)$. Moreover, under these assumptions, $n^{-1} D_\alpha^{(n)}(\theta, \theta^*) = D_\alpha[p(\cdot | \theta) || p(\cdot | \theta^*)]$.

As discussed in the previous subsection, the decoupling of the roles of q_θ and q_{S^n} in the model fit term aid bounding $\Psi_{n,\alpha}$. Specifically, we first choose a \tilde{q}_{S^n} which controls the Jensen gap Δ_J , and then make a choice of q_θ which controls $\Psi_{n,\alpha}(q_\theta, \tilde{q}_{S^n})$. The choice of q_θ requires a delicate balance between placing enough mass near θ^* and controlling the KL divergence from the prior.

For a fixed q_θ , if we choose q_{S^n} to be the full conditional distribution of S^n given θ , i.e.,

$$q_{S^n}(s^n | \theta) = \prod_{i=1}^n q_{S_i}(s_i | \theta) = \prod_{i=1}^n \frac{\pi_{s_i} p(Y_i | \mu, s_i)}{p(Y_i | \theta)}, \quad s^n \in \{1, 2, \dots, K\}^n,$$

then the normalizing constant of $q_{S_i}(\cdot | \theta)$ is $\sum_{s_i} \pi_{s_i} p(Y_i | \mu, s_i) = p(Y_i | \theta)$, and as a result, the Jensen gap $\Delta_J = 0$. The mean-field approximation precludes us from choosing q_{S^n} dependent on θ , and hence the Jensen gap cannot be made exactly zero in general. However, this naturally suggests replacing θ by θ^* in the above display and choosing $\tilde{q}_{S_i} \propto \pi_{s_i}^* p(Y_i | \mu^*, s_i)$. This leads us to the following corollary.

Corollary 3.2 (i.i.d. observations). *It holds with $\mathbb{P}_{\theta^*}^n$ probability at least $(1 - \zeta)$ that for any probability measure $q_\theta \in \Gamma_\theta$ with $q_\theta \ll p_\theta$*

$$\begin{aligned} & \int \left\{ D_\alpha [p(\cdot | \theta) || p(\cdot | \theta^*)] \right\} \hat{q}_{\theta, \alpha}(\theta) d\theta \leq \frac{\alpha}{n(1 - \alpha)} \Psi_{n, \alpha}(q_\theta, \tilde{q}_{S^n}) + \frac{1}{n(1 - \alpha)} \log(1/\zeta), \\ & = \frac{\alpha}{n(1 - \alpha)} \left[- \int_{\Theta} \sum_{i=1}^n \sum_{s_i} \tilde{q}_{S_i}(s_i) \log \frac{p(Y_i | \mu, s_i) \pi_{s_i}}{p(Y_i | \mu^*, s_i) \pi_{s_i}^*} q_\theta(d\theta) + \frac{D(q_\theta || p_\theta)}{\alpha} + \frac{\log(1/\zeta)}{\alpha} \right], \end{aligned} \quad (3.2)$$

where \tilde{q}_{S^n} is the probability distribution over S^n defined as

$$\tilde{q}_{S^n}(s^n) = \prod_{i=1}^n \tilde{q}_{S_i}(s_i) = \prod_{i=1}^n \frac{\pi_{s_i}^* p(Y_i | \mu^*, s_i)}{p(Y_i | \theta^*)}, \quad s^n \in \{1, 2, \dots, K\}^n. \quad (3.3)$$

The second line of (3.2) follows from the first since

$$\Delta_J(q_\theta, \tilde{q}_{S^n}) = - \int_{\Theta} \sum_{i=1}^n \sum_{s_i} \tilde{q}_{S_i}(s_i) \log \frac{p(Y_i | \mu, s_i) \pi_{s_i}}{p(Y_i | \mu^*, s_i) \pi_{s_i}^*} q_\theta(d\theta) + \int \ell_n(\theta, \theta^*) q_\theta(d\theta).$$

After choosing \tilde{q}_{S^n} as (3.3) in Corollary 3.2, we can make the first term in the r.h.s. of (3.2) small by choosing the variational factor q_θ of θ concentrated around θ^* . In the rest of this subsection, we will apply Corollary 3.2 to derive more concrete variational Bayes risk bounds under some further simplifying assumptions.

As a first application, assume there is no latent variable in the model, that is, $W^n = \theta = \mu$. As discussed before, the α -VB solution in this case coincides with the nearest KL point to the α -fractional posterior of the parameter. A reviewer pointed out a recent preprint by Alquier and Ridgeway [2] where they exploit risk bounds for fractional posteriors developed in [4] to analyze tempered posteriors and their variational approximations, which coincides with the α -VB solution when $W^n = \theta$. The following Theorem 3.3 arrives at a similar conclusion

to Corollary 2.3 of [2]. We reiterate here that our main motivation is models with latent variables not considered in [2], and Theorem 3.3 follows as a corollary of our general result in Theorem 3.1.

Theorem 3.3 (No latent variable). *It holds with $\mathbb{P}_{\theta^*}^n$ probability at least $(1 - \zeta)$ that for any probability measure $q_\theta \in \Gamma_\theta$ with $q_\theta \ll p_\theta$*

$$\begin{aligned} & \int \left\{ D_\alpha [p(\cdot | \theta) || p(\cdot | \theta^*)] \right\} \widehat{q}_{\theta, \alpha}(\theta) d\theta \\ &= \frac{\alpha}{n(1 - \alpha)} \left[- \int_{\Theta} \log \frac{p(Y^n | \theta)}{p(Y^n | \theta^*)} q_\theta(\theta) d\theta + \frac{D(q_\theta || p_\theta)}{\alpha} + \frac{\log(1/\zeta)}{\alpha} \right], \end{aligned} \quad (3.4)$$

We will illustrate some particular choices of q_θ for typical variational families Γ_Θ in the examples in Section 4.

As a second application, we consider a special case when Γ_θ is restriction-free, which is an ideal example for conveying the general idea of how to choose q_θ to control the upper bound in (3.2). To that end, define two KL neighborhoods around (π^*, μ^*) with radius $(\varepsilon_\pi, \varepsilon_\mu)$ as

$$\begin{aligned} \mathcal{B}_n(\pi^*, \varepsilon_\pi) &= \left\{ D(\pi^* || \pi) \leq \varepsilon_\pi^2, \quad V(\pi^* || \pi) \leq \varepsilon_\pi^2 \right\}, \\ \mathcal{B}_n(\mu^*, \varepsilon_\mu) &= \left\{ \sup_s D[p(\cdot | \mu^*, s) || p(\cdot | \mu, s)] \leq \varepsilon_\mu^2, \quad \sup_s V[p(\cdot | \mu^*, s) || p(\cdot | \mu, s)] \leq \varepsilon_\mu^2 \right\}, \end{aligned} \quad (3.5)$$

where we used the shorthand $D(\pi^* || \pi) = \sum_s \pi_s^* \log(\pi_s^*/\pi_s)$ to denote the KL divergence between categorical distributions with parameters $\pi^* \in \mathcal{S}_K$ and $\pi \in \mathcal{S}_K$ in the K -dim simplex \mathcal{S}_K . By choosing q_θ as the restriction of p_θ into $\mathcal{B}_n(\pi^*, \varepsilon_\pi) \times \mathcal{B}_n(\mu^*, \varepsilon_\mu)$, we obtain the following theorem. Here, we make the assumption of independent priors on μ and π , i.e., $p_\theta = p_\mu \otimes p_\pi$, to simplify the presentation.

Theorem 3.4 (Parameter restriction-free). *For any fixed $(\varepsilon_\pi, \varepsilon_\mu) \in (0, 1)^2$ and $D > 1$, with $\mathbb{P}_{\theta^*}^{(n)}$ probability at least $1 - 5/\{(D - 1)^2 n (\varepsilon_\pi^2 + \varepsilon_\mu^2)\}$, it holds that*

$$\begin{aligned} & \int \left\{ D_\alpha [p(\cdot | \theta) || p(\cdot | \theta^*)] \right\} \widehat{q}_{\theta, \alpha}(d\theta) \leq \frac{D \alpha}{1 - \alpha} (\varepsilon_\pi^2 + \varepsilon_\mu^2) + \\ & \left\{ - \frac{1}{n(1 - \alpha)} \log P_\pi[\mathcal{B}_n(\pi^*, \varepsilon_\pi)] \right\} + \left\{ - \frac{1}{n(1 - \alpha)} \log P_\mu[\mathcal{B}_n(\mu^*, \varepsilon_\mu)] \right\}. \end{aligned} \quad (3.6)$$

Although the results in this section assume discrete latent variables, similar results can be seamlessly obtained for continuous latent variables; see the supplemental document for more details. We will apply this theorem for analysing mean-field approximations for the Gaussian mixture model and the latent Dirichlet allocation in Section 4.

Observe that the variational risk bound in Theorem 3.4 depends only on prior mass assigned to appropriate KL neighborhoods of the truth. This renders an application of Theorem 3.4 to various practical problems particularly straightforward. As we shall see in the next subsection, the $\alpha = 1$ case, i.e. the regular VB, requires more stringent conditions involving the existence of exponentially consistent tests to separate points in the parameter space. The testing condition is even necessary for the actual posterior to contract; see, e.g., [4], and hence one cannot avoid the testing assumption for its usual variational approximation. Nevertheless, we show below that once the existence of such tests can be verified, the regular VB approximation can also be shown to contract optimally.

3.2 Risk bounds for the $\alpha = 1$ case

We consider any loss function $r(\theta, \theta^*)$ satisfying the following assumption.

Assumption T (Statistical identifiability): For some $\varepsilon_n > 0$ and any $\varepsilon \geq \varepsilon_n$, there exists a sieve set $\mathcal{F}_{n,\varepsilon} \subset \Theta$ and a test function $\phi_{n,\varepsilon} : \mathcal{Y}^n \rightarrow [0, 1]$ such that

$$P_\theta(\mathcal{F}_{n,\varepsilon}^c) \leq e^{-cn\varepsilon^2}, \quad (3.7)$$

$$\mathbb{E}_{\theta^*}[\phi_{n,\varepsilon}] \leq e^{-cn\varepsilon_n^2}, \quad (3.8)$$

$$\mathbb{E}_\theta[1 - \phi_{n,\varepsilon}] \leq e^{-cnr(\theta, \theta^*)}, \quad \forall \theta \in \mathcal{F}_{n,\varepsilon} \text{ satisfies } r(\theta, \theta^*) \geq \varepsilon^2. \quad (3.9)$$

Roughly speaking, the sieve set $\mathcal{F}_{n,\varepsilon}$ can be viewed as the effective support of the prior distribution at sample size n , and ε_n the contraction rate of the usual posterior distribution. The first condition (3.7) allows us to focus attention to this important region in the parameter space that is not too large, but still possesses most of the prior mass. The last two conditions (3.8) and (3.9) ensure the statistical identifiability of the parameter under the loss $r(\cdot, \cdot)$ through the existence of a test function $\phi_{n,\varepsilon}$, and require a sufficiently fast decay of the Type I/II error. In the case when Θ is compact and $r(\theta, \theta') = h^2(\theta \parallel \theta')$ is the squared Hellinger distance between p_θ and $p_{\theta'}$, such a test $\phi_{n,\varepsilon}$ always exists [17]. A similar set of assumptions are used for showing the concentration of the usual posterior (for example, see [18] and [17]), with the existence of such sieve sets and test functions verified for numerous model-prior combinations. The only difference in our case is that Assumption T requires the existence of the pair $(\mathcal{F}_{n,\varepsilon}, \phi_{n,\varepsilon})$ for all $\varepsilon \geq \varepsilon_n$, not just at $\varepsilon = \varepsilon_n$. However, this extra requirement appears mild since in most cases a construction of $(\mathcal{F}_{n,\varepsilon}, \phi_{n,\varepsilon})$ at $\varepsilon = \varepsilon_n$ naturally extends to any $\varepsilon \geq \varepsilon_n$.

Our main result for the usual VB ($\alpha = 1$) provides a finite-sample upper bound to the

variational Bayes risk for any loss function $r(\theta, \theta^*)$ satisfying Assumption T. Here, we use Q_θ to denote the probability distribution associated with any member q_θ in the variational density family Γ .

Theorem 3.5. *Under Assumption T, for any $\varepsilon \geq \varepsilon_n$, we have that with $\mathbb{P}_{\theta^*}^{(n)}$ probability at least $1 - 2e^{-c n \varepsilon_n^2/2}$, it holds that for any probability measure $q_\theta \in \Gamma_\theta$ with $q_\theta \ll p_\theta$ and any probability measure $q_{S^n} \in \Gamma_{S^n}$ on S^n that*

$$\begin{aligned} & \frac{1}{n} \left[\widehat{Q}_\theta(\mathcal{F}_{n,\varepsilon}^c) \log \frac{\widehat{Q}_\theta(\mathcal{F}_{n,\varepsilon}^c)}{P_\theta(\mathcal{F}_{n,\varepsilon}^c)} + (1 - \widehat{Q}_\theta(\mathcal{F}_{n,\varepsilon}^c)) \log \frac{1 - \widehat{Q}_\theta(\mathcal{F}_{n,\varepsilon}^c)}{1 - P_\theta(\mathcal{F}_{n,\varepsilon}^c)} \right] \\ & + c \int_{\theta \in \mathcal{F}_{n,\varepsilon}, r(\theta, \theta^*) \geq \varepsilon^2} r(\theta, \theta^*) \widehat{q}_\theta(\theta) d\theta \leq \frac{1}{n} \Psi_n(q_\theta, q_{S^n}) + \frac{c \varepsilon_n^2}{2} + \frac{\log 2}{n}. \end{aligned} \quad (3.10)$$

The first term on the l.h.s. of inequality (3.10) relates the variational complementary probability $\widehat{Q}_\theta(\mathcal{F}_{n,\varepsilon}^c)$ to the prior complementary probability $P_\theta(\mathcal{F}_{n,\varepsilon}^c)$. As a consequence, an upper bound of this term controls the remainder variational probability mass outside the sieve $\mathcal{F}_{n,\varepsilon}$. The second term $\int_{\theta \in \mathcal{F}_{n,\varepsilon}, r(\theta, \theta^*) \geq \varepsilon^2} r(\theta, \theta^*) \widehat{q}_\theta(\theta) d\theta$ in (3.10) is the variational Bayes risk over to the intersection between $\mathcal{F}_{n,\varepsilon}$ and the set of all θ such that the loss $r(\theta, \theta^*)$ is at least ε^2 .

In [32], we proved a risk bound for the $\alpha = 1$ case under the much stronger assumption of a compact parameter space and the existence of a global test ϕ_n with type-I and II error rates bounded above by $e^{-C n \varepsilon_n^2}$. Under those assumptions, the result in [32] can be recovered from our more general result in Theorem 3.5 by setting $\mathcal{F}_{n,\varepsilon} = \Theta$, and $\phi_{n,\varepsilon} = \phi_n$; the global test, for all ε . Such stronger assumptions usually hold when the parameter space Θ is a compact subset of the Euclidean space — however, in other cases such as unbounded parameter spaces or infinite dimensional functional spaces, such a global test function ϕ_n may not exist, signifying the necessity of Theorem 3.5. We also point out the preprint [46], which appeared while this manuscript was in revision, where they consider the usual VB and their analysis is based on a direct application of the variational lemma in the proof of Theorem 3.1 in the supplementary document. However, their results require a stronger prior concentration condition and their analysis does not involve latent variable models.

Similar to the development for $\alpha < 1$, we can further simplify Ψ_n by introducing more assumptions. Due to the space constraint, we only provide a counterpart of Theorem 3.4 under the assumptions made therein. Recall the definition of two KL neighbourhoods $\mathcal{B}_n(\pi^*, \varepsilon)$ and $\mathcal{B}_n(\mu^*, \varepsilon)$ defined in (3.5).

Assumption P (Prior concentration): There exists some constant $C > 0$ such that

$$P_\theta(\mathcal{B}_n(\pi^*, \varepsilon_n)) \geq \exp(-C n \varepsilon_n^2) \quad \text{and} \quad P_\theta(\mathcal{B}_n(\mu^*, \varepsilon_n)) \geq \exp(-C n \varepsilon_n^2).$$

Under Assumptions T and P, Theorem 3.5 leads to a high probability bound on the over variational Bayes risk for loss $r(\theta, \theta^*)$, as summarized in the following Theorem.

Theorem 3.6 (Parameter restriction-free). *Under Assumptions T and P, it holds with $\mathbb{P}_{\theta^*}^{(n)}$ probability at least $1 - 3/\{(D - 1)^2 n \varepsilon_n^2\}$ that for any $\varepsilon \in [\varepsilon_n, e^{c' n \varepsilon_n^2}]$ (for some constant $c' > 0$),*

$$\widehat{Q}_\theta(r(\theta, \theta^*) \geq \varepsilon^2) \leq C_1 \frac{\varepsilon_n^2}{\varepsilon^2}.$$

In particular, this implies for any $R < e^{2c' n \varepsilon_n^2}$,

$$\int_{\theta: r(\theta, \theta^*) \leq R} r(\theta, \theta^*) \widehat{q}_\theta(\theta) d\theta \leq C_3 \varepsilon_n^2 (1 + \log(R/\varepsilon_n)).$$

In particular, if the sequence $\{\varepsilon_n : n \geq 1\}$ satisfies $n \varepsilon_n^2 \rightarrow \infty$, then selecting $\varepsilon = M_n \varepsilon_n$ for $M_n \rightarrow \infty$ ($M_n \leq \varepsilon_n^{-1}$) leads to the asymptotic variational posterior concentration:

$$\widehat{Q}_\theta(r(\theta, \theta^*) \leq M_n^2 \varepsilon^2) \rightarrow 1 \quad \text{in probability, as } n \rightarrow \infty.$$

The extra truncation $r(\theta, \theta^*) \leq R$ in the variational risk bound in the theorem is due to the quadratic decay of our upper bound to $\widehat{Q}_\theta(r(\theta, \theta^*) \geq \varepsilon^2)$. Since the risk upper bound only has a logarithmic dependence on the truncation level R , one can simply set it at a fixed large number to ensure an order $\mathcal{O}(\varepsilon_n^2)$ risk bound in practice. In fact, this truncation can be eliminated under a stronger assumption (as in [32]) that there is a global test function $\phi_n : \mathcal{Y}^n \rightarrow [0, 1]$, such that the type I error bound (3.8) holds, and the following type II error bound holds for all $\theta \in \Theta$ satisfying $r(\theta, \theta^*) \geq \varepsilon_n^2$,

$$\mathbb{E}_\theta[1 - \phi_n] \leq e^{-c n r(\theta, \theta^*)}.$$

This can be seen from Theorem 3.5 by setting $\mathcal{F}_n = \Theta$ and $\varepsilon = \varepsilon_n$ in inequality (3.10), which

implies

$$\begin{aligned} c \int_{\Theta} r(\theta, \theta^*) \widehat{q}_{\theta}(\theta) d\theta &\leq c \varepsilon_n^2 + c \int_{r(\theta, \theta^*) \geq \varepsilon_n^2} r(\theta, \theta^*) \widehat{q}_{\theta}(\theta) d\theta \\ &\leq \frac{1}{n} \Psi_n(q_{\theta}, q_{S^n}) + \frac{3c \varepsilon_n^2}{2} + \frac{\log 2}{n}. \end{aligned}$$

3.3 α -VB using stronger divergences

In this subsection, we consider an extension of our theoretical development for α -VB where the KL divergence in the objective function is replaced by a stronger divergence $\bar{D}[p \parallel q] \geq D[p \parallel q]$, for example, χ^2 divergence and Rényi divergence [29], and the corresponding variational approximation

$$\bar{q}_{W^n} := \operatorname{argmin}_{q_{W^n} \in \Gamma} \bar{D}[q_{W^n}(\cdot) \parallel p(\cdot | Y^n)].$$

As another example, in some applications of variational inference [47], the minimization of the KL divergence over the variational density q_{W^n} to the conditional density $p(W^n | Y^n)$ may not admit a closed-form updating formula, and some surrogate ELBO $\bar{L}(q_{W^n})$ as a lower bound to the ELBO $L(q_{W^n})$ is employed. Under the perspective of ELBO decomposition (1.4), this replacement is equivalent to using a stronger metric

$$\bar{D}[q_{W^n} \parallel p(\cdot | Y^n)] := \log p(Y^n) - \bar{L}(q_{W^n}) \geq D[q_{W^n} \parallel p(\cdot | Y^n)].$$

The following theorem provides a variational Bayes risk upper bound to \bar{q}_{θ} . To simplify the presentation, the theorem is stated for the $\alpha < 1$ case, although extension to $\alpha = 1$ is straightforward. Define the equivalent objective function

$$\bar{\Psi}_{\alpha}(q_{\theta}, q_{S^n}) = \Psi_{n,\alpha}(q_{\theta}, q_{S^n}) + \left(\bar{D}[q_{W^n} \parallel p(\cdot | Y^n)] - D[q_{W^n} \parallel p(\cdot | Y^n)] \right),$$

and the corresponding α -VB solution $\bar{q}_{\theta,\alpha} = \operatorname{argmin}_{q_{W^n} \in \Gamma} \bar{\Psi}_{\alpha}(q_{\theta}, q_{S^n})$. When \bar{D} is the KL divergence D , objective function $\bar{\Psi}_{\alpha}$ reduces to the $\Psi_{n,\alpha}$ in (2.6).

Theorem 3.7. *For any $\zeta \in (0, 1)$, it holds with $\mathbb{P}_{\theta^*}^n$ probability at least $(1 - \zeta)$ that for any probability measure $q_{\theta} \in \Gamma_{\theta}$ with $q_{\theta} \ll p_{\theta}$ and any probability measure $q_{S^n} \in \Gamma_{S^n}$ on S^n ,*

$$\int \frac{1}{n} D_{\alpha}^{(n)}(\theta, \theta^*) \bar{q}_{\theta,\alpha}(d\theta) \leq \frac{\alpha}{n(1-\alpha)} \bar{\Psi}_{\alpha}(q_{\theta}, q_{S^n}) + \frac{1}{n(1-\alpha)} \log(1/\zeta).$$

This theorem provides a simple replacement rule for α -VB—if the α -VB objective func-

tion $\Psi_{n,\alpha}$ is replaced with an upper bound $\bar{\Psi}_\alpha$, then a variational Bayes risk bound obtained by replacing $\Psi_{n,\alpha}$ with the upper bound $\bar{\Psi}_\alpha$ holds. We will apply this replacement rule to obtain a minimax variational risk bound for the mixture of Gaussian approximation in Section 4.

4 Applications

In this section, we apply our theory in Section 3 to concrete examples: mean field approximation to (low) high-dimensional Bayesian linear regression, (mixture of) Gaussian approximation to regular parametric models, mean field approximation to Gaussian mixture models, and mean field approximation to latent Dirichlet allocation. To simplify the presentation, all results are stated for α -VB with $\alpha < 1$ and the α subscript in $\hat{q}_{\theta,\alpha}$ is dropped. Extensions to the $\alpha = 1$ case are discussed in the supplement.

Example: (Mean field approximation to low-dimensional Bayesian linear regression) Consider the following Bayesian linear model

$$Y^n = X\beta + w, \quad w \sim \mathcal{N}(0, \sigma^2 I_n), \quad (4.1)$$

where $Y^n \in \mathbb{R}^n$ is the n -dim response vector, $X \in \mathbb{R}^{n \times d}$ the design matrix, $\beta \in \mathbb{R}^d$ the unknown regression coefficient vector of interest, and σ the noise level. In this example, we consider the low-dimensional regime where $d \ll n$, and focus on independent prior $p_\beta \otimes p_\sigma$ for parameter pair $\theta = (\beta, \sigma)$ for technical convenience (the result also applies to non-independent priors).

We apply the mean-field approximation by using the following variational family

$$q(\beta, \sigma) = q_\beta(\beta) q_\sigma(\sigma)$$

to approximate the joint α -fractional posterior distribution of $\theta = (\beta, \sigma)$ with $\hat{q}_\theta = \hat{q}_\beta \otimes \hat{q}_\sigma$. This falls into our framework when there is no latent variable and $W^n = \theta$. Computationally, a normal prior for θ and an inverse gamma prior for σ^2 are attractive since they are “conjugate” priors — the resulting variational densities \hat{q}_β and \hat{q}_σ still fall into the same parametric families. An application of Theorem 3.3 leads to the following result.

Corollary 4.1. *Assume that the prior density is continuous, and thick around the truth $\theta^* = (\beta^*, \sigma^*)$, that is, $p_\theta(\theta^*) > 0$ and $p_\sigma(\sigma^*) > 0$. If $d/n \rightarrow 0$ as $n \rightarrow \infty$, then with*

probability tending to one as $n \rightarrow \infty$,

$$\left\{ \int h^2 [p(\cdot | \theta) || p(\cdot | \theta^*)] \hat{q}_\theta(\theta) d\theta \right\}^{1/2} \lesssim \sqrt{\frac{d}{n \min\{\alpha, 1 - \alpha\}} \log(dn)}.$$

The convergence rate of $\mathcal{O}(\sqrt{n^{-1} d \log(dn)})$ under the Hellinger distance implies that the α -VB estimator $\hat{\beta}_{\text{VB}, \alpha} = \int \beta \hat{q}_\beta(\beta) d\beta$ converges towards β^* relative to the ℓ_2 norm at rate $\sqrt{n^{-1} d \log(dn)}$ (under the condition that $n^{-1} X^T X$ has minimal eigenvalue bounded away from zero), which is the minimax rate up to logarithm factors. A similar $n^{-1/2}$ convergence rate has been obtained in [45] by directly analyzing the stationary point of an alternating minimization algorithm. However, their analysis requires the closed-form updating formula based on a conjugate normal prior for β and an inverse gamma prior for σ^2 , and may not be applicable to other priors. On the other hand, Corollary 4.1 only requires the minimal conditions of prior thickness and continuity.

Example: (Mean field approximation to high-dimensional Bayesian linear regression with spike-and-slab priors) In this example we continue to consider the Bayesian linear model (4.1), but we are interested in the high-dimensional regime where $d \gg n$. Following standard practice to make sparsity assumptions in the $d \gg n$ regime, let $s \ll n$ denote the sparsity level, i.e., the number of non-zero coefficients, of the true regression parameter β^* .

We consider the popularly used spike-and-slab priors [16] on β . Following [16], we introduce a latent indicator variable $z_j = I(\beta_j \neq 0)$ for each β_j to indicate whether the j th covariate X_j is included in the model, and call $z = (z_1, \dots, z_d) \in \{0, 1\}^d$ the latent indicator vector. We use the notation β_z to denote the vector of nonzero components of β selected by z , that is $\beta_z = (\beta_j : z_j = 1)$. Consider the following sparsity inducing hierarchical prior $p_{\beta, z}$ over (β, z) :

$$\begin{aligned} z_j &\stackrel{iid}{\sim} \frac{1}{d} \delta_1 + \left(1 - \frac{1}{d}\right) \delta_0, \quad j = 1, \dots, d, \\ \beta_z | z &\sim p_{\beta | z}, \quad \text{and} \quad \sigma \sim p_\sigma, \end{aligned} \tag{4.2}$$

where the prior probability of $\{z_j = 1\}$ is chosen as d^{-1} so that on an average only $\mathcal{O}(1)$ covariates are included in the model. Let z^* denote the indicator vector associated with the truth β^* .

By viewing the latent variable indicator vector z as a parameter, we apply the block

mean-field approximation [13] by using the family

$$q(\beta, \sigma, z) = q_\sigma(\sigma) \prod_{j=1}^d q_{z_j, \beta_j}(z_j, \beta_j)$$

to approximate the joint α -fractional posterior distribution of $\theta = (\beta, \sigma, z)$ with $\hat{q}_\theta(\theta) = \hat{q}_\sigma(\sigma) \prod_{j=1}^d \hat{q}_{z_j, \beta_j}(z_j, \beta_j)$. Although we have a high-dimensional latent variable vector z , the latent variable is associated with the parameter β , and not with the observation Y^n . Consequently, this variational approximation still falls into our framework without latent variable, that is, $W^n = \theta = (z, \beta)$ and $\Delta_J \equiv 0$. It turns out that the spike and slab prior with Gaussian slab is particularly convenient for computation — it is “conjugate” in that the resulting variational approximation falls into the same spike and slab family [13]. An application of Theorem 3.3 leads to the following result.

Corollary 4.2. *Suppose $p_{\beta|z^*}$ is continuous and thick at $\beta_{z^*}^*$, and p_σ is continuous and thick at σ^* . If $s \log d/n \rightarrow 0$ as $n \rightarrow \infty$, then it holds with probability tending to one as $n \rightarrow \infty$ that*

$$\left\{ \int h^2 [p(\cdot|\theta) || p(\cdot|\theta^*)] \hat{q}_\theta(\theta) d\theta \right\}^{1/2} \lesssim \sqrt{\frac{s}{n \min\{\alpha, 1 - \alpha\}} \log(dn)}.$$

Corollary 4.2 implies a convergence rate $\sqrt{n^{-1} s \log(dn)}$ of the variational-Bayes estimator $\hat{\beta}_{\text{VB}, \alpha}$ under the restricted eigenvalue condition [5], which is the minimax rate up to log terms for high-dimensional sparse linear regression. To our knowledge, [31] is the only literature that studies the mean-field approximation to high-dimensional Bayesian linear regression with spike and slab priors. They show estimation consistency by directly analyzing an iterative algorithm for solving the variational optimization problem with $\alpha = 1$ and a specific prior. As before, Corollary 4.2 holds under very mild conditions on the prior and does not rely on having closed-form updates of any particular algorithm.

Here, we considered the block mean-field instead of the full mean-field approximation which further decomposes q_{z_j, β_j} into $q_{z_j} \otimes q_{\beta_j}$. In fact, the latter resembles a ridge regression estimator, and the KL term $\alpha^{-1} D(q_\theta || p_\theta)$ appearing in the upper bound in (3.2) cannot attain the minimax order $\sqrt{n^{-1} s \log d}$.

Example: (Gaussian approximation to regular parametric models) Consider a family of regular parametric models $\mathcal{P} = \{\mathbb{P}_\theta^{(n)} : \theta \in \Theta\}$ where n is the sample size, and the likelihood function $p_\theta^{(n)}$ is indexed by a parameter θ belonging to the parameter space $\Theta \subset \mathbb{R}^d$, which we assume to be compact. Let p_θ denote the prior density of over Θ , and

$Y^n = (Y_1, \dots, Y_n)$ be the observations from $\mathbb{P}_{\theta^*}^{(n)}$, with θ^* being the truth. We apply the Gaussian approximation by using the Gaussian family Γ_G (restricted to Θ)

$$q(\theta) \propto \mathcal{N}(\theta; \mu, \Sigma) I_{\Theta}(\theta), \quad \mu \in \mathbb{R}^d \text{ and } \Sigma \text{ is a } d \times d \text{ positive definite matrix.}$$

Details are postponed to the supplemental document for space constraints.

Example: (Mixture of Gaussian variational approximation to regular parametric models): Still consider the regular parametric model $\mathcal{P} = \{\mathbb{P}_{\theta}^{(n)} : \theta \in \Theta\}$ as in the previous example. Now we consider the more flexible variational family Γ_{MG} composed of

$$q(\theta) = \sum_{j=1}^J w_j \mathcal{N}(\theta; \mu_j, \Sigma_j), \quad \sum_{j=1}^J w_j = 1, \quad \mu_j \in \mathbb{R}^d, \quad \Sigma_j \in \mathbb{R}^{d \times d} \text{ is p.d.,}$$

as all mixtures of Gaussians with J components, where J is a pre-specified number. Let q_j denote the j th component $\mathcal{N}(\theta; \mu_j, \Sigma_j)$ of the variational density function q , and \mathbb{E}_p denotes the expectation under a density function p . Since any probability distribution can be approximated by a mixture of Gaussians within arbitrarily small error with sufficient large number J of components, this enlarged variational family may reduce the approximation error from using Γ_G and become capable of capturing multimodality and heavy tail behaviour of the posterior [47]. However, this additional flexibility in shape comes with the price of intractability of the entropy term $E_q[-\log q(\theta)]$. To facilitate computation, [47] conducted an additional application of Jensen's inequality

$$E_q[-\log q(\theta)] = \sum_{j=1}^J w_j \mathbb{E}_{q_j}[-\log q(\theta)] \geq - \sum_{j=1}^J w_j \log \mathbb{E}_{q_j}[q(\theta)],$$

yielding a lower bound to the ELBO as

$$L(q) = E_q \left[\log \frac{p(Y^n, \theta)}{q(\theta)} \right] \geq \underbrace{E_q[\log p(Y^n, \theta)] - \sum_{j=1}^J w_j \log \mathbb{E}_{q_j}[q(\theta)]}_{\bar{L}(q)}.$$

$\bar{L}(q)$ is more convenient to work with since its second term admits a simple analytic form. Using such a surrogate ELBO places us directly in the framework of § 3.3 and an application of Theorem 3.7 leads to the following bound.

Corollary 4.3. *For any measure $q_{\theta} = \sum_{j=1}^J w_j q_j \in \Gamma_{MG}$, it holds with probability at least*

$1 - \varepsilon$ that

$$\begin{aligned} \int \left\{ \frac{1}{n} D_\alpha^{(n)}(\theta, \theta^*) \right\} \widehat{q}_\theta(\theta) d\theta &\leq -\frac{\alpha}{n(1-\alpha)} \int_{\Theta} q_\theta(\theta) \log \frac{p(Y^n | \theta)}{p(Y^n | \theta^*)} d\theta \\ &+ \frac{1}{n(1-\alpha)} \sum_{j=1}^J w_j \left(\log \mathbb{E}_{q_j} [q_\theta(\theta)] - \mathbb{E}_{q_j} [\log p_\theta(\theta)] \right) + \frac{1}{n(1-\alpha)} \log(1/\varepsilon). \end{aligned} \quad (4.3)$$

In particular, under Assumption P, it holds with probability tending to one as $n \rightarrow \infty$ that

$$\left\{ \int h^2 [p(\cdot | \theta) || p(\cdot | \theta^*)] \widehat{q}_\theta(\theta) d\theta \right\}^{1/2} \lesssim \sqrt{\frac{d}{n \min\{\alpha, 1-\alpha\}} \log(dn)}.$$

As a concrete example of the development in Section 3.3, this corollary suggests that the additional Jensen gap due to the $E_q[-\log q(\theta)]$ term is reflected in the new variational inequality (4.3). More precisely, the KL divergence term $D[q_\theta || p_\theta]$ is replaced by its upper bound $\sum_{j=1}^J w_j (\log \mathbb{E}_{q_j} [q_\theta(\theta)] - \mathbb{E}_{q_j} [\log p_\theta(\theta)])$, which can be bounded by reducing it into a single Gaussian component case ($w_1 = 1$, and $w_j = 0$ for $j = 2, \dots, J$).

Example: (Mean field approximation to Gaussian mixture model) Suppose the true data generating model is the d -dimensional Gaussian mixture model with K components,

$$Y \sim \sum_{k=1}^K \pi_k \mathcal{N}(\mu_k, I_d),$$

where $\mu_k \in \mathbb{R}^d$ is the mean vector associated with the k th component and $\pi = (\pi_1, \dots, \pi_K) \in \mathcal{S}_K$ is the mixing probability. Here, for simplicity we assume the covariance matrix of each Gaussian component to be I_d . $\mu = (\mu_1, \dots, \mu_K)$ and π together forms the parameter $\theta = (\mu, \pi)$ of interest. By data augmentation, we can rewrite the model into the following hierarchical form by introducing the latent class variable S ,

$$S \sim \text{Categorical}(\pi_1, \pi_2, \dots, \pi_K), \quad Y | S = s \sim \mathcal{N}(\mu_s, I_d).$$

Let $Y^n = (Y_1, \dots, Y_n)$ be n i.i.d. copies of Y with parameter $\theta^* = (\mu^*, \pi^*)$, and $S^n = (S_1, \dots, S_n) \in \{1, \dots, K\}^n$ denote the corresponding latent variables. For simplicity, we assume that independent prior $p_\mu \otimes p_\pi$ are specified for (μ, π) .

We apply the mean field approximation by using the family of density functions of the

form

$$q(\pi, \mu, S^n) = q_\pi(\pi) q_\mu(\mu) q_{S^n}(s^n) = q_\pi(\pi) q_\mu(\mu) \prod_{i=1}^n q_{S_i}(s_i)$$

to approximate the joint α -fractional posterior distribution of (π, μ, S^n) , producing the α -mean-field approximation $\hat{q}_\theta \otimes \hat{q}_{S^n}$, where $(\hat{q}_\theta, \hat{q}_{S^n})$ are defined in (2.7). This variational approximation fits into the framework of Theorem 3.4. Therefore, an application of this theorem leads to the following result.

Assumption R: (regularity condition) There exists some constant $\delta_0 > 0$, such that each component of $\pi^* \in \mathcal{S}_K$ is at least δ_0 .

Corollary 4.4. *Suppose Assumption R holds, and the prior densities p_μ and p_π are thick and continuous at μ^* and π^* respectively. If $dK/n \rightarrow 0$ as $n \rightarrow \infty$, then it holds with probability tending to one as $n \rightarrow \infty$ that*

$$\left\{ \int h^2 [p(\cdot | \theta) || p(\cdot | \theta^*)] \hat{q}_\theta(\theta) d\theta \right\}^{1/2} \lesssim \sqrt{\frac{dK}{n \min\{\alpha, 1 - \alpha\}} \log(dn)}.$$

As a related result, [42] show that the with proper initialization, the coordinate descent algorithm for solving the variational optimization problem (2.7) with $\alpha = 1$ under conjugate priors converges to a local minimum that is $\mathcal{O}(n^{-1})$ away from the maximum likelihood estimate of (μ, π) by directly analyzing the algorithm using the contraction mapping theorem. In contrast, our proof does not require any structural assumptions on the priors, and can be easily extended to a broader class of mixture models beyond Gaussians.

Example: (Mean field approximation to latent Dirichlet allocation) As our final example, we consider Latent Dirichlet allocation (LDA, [11]), a conditionally conjugate probabilistic topic model [9] for uncovering the latent “topics” contained in a collection of documents. LDA treats documents as containing multiple topics, where a topic is a distribution over words in a vocabulary. Following the notation of [21], let K be a specific number of topics and V the size of the vocabulary. LDA defines the following generative process:

1. For each topic in $k = 1, \dots, K$,
 - (a) draw a distribution over words $\beta_k \sim \text{Dir}_V(\eta_\beta)$.
2. For each document in $d = 1, \dots, D$,

- (a) draw a vector of topic proportions $\gamma_d \sim \text{Dir}_K(\eta_\gamma)$.
- (b) For each word in $n = 1, \dots, N$,
 - i. draw a topic assignment $z_{dn} \sim \text{multi}(\gamma_d)$, then
 - ii. draw a word $w_{dn} \sim \text{multi}(\beta_{z_{dn}})$.

Here $\eta_\beta \in \mathbb{R}_+$ is a hyper-parameter of the symmetric Dirichlet prior on the topics β , and $\eta_\gamma \in \mathbb{R}_+^K$ are hyper-parameters of the Dirichlet prior on the topic proportions for each document. $z_{dn} \in \{1, \dots, K\}$ is the latent class variable over topics where $z_{dn} = k$ indicates the n th word in document d is assigned to the k th topic. Similarly, $w_{dn} \in \{1, \dots, V\}$ is the latent class variable over the words in the vocabulary where $w_{dn} = v$ indicates that the n th word in document d is the v th word in the vocabulary. To facilitate adaptation to sparsity using Dirichlet distributions when $V, K \gg 1$, we choose $\eta_\beta = 1/V^c$ and $\eta_\gamma = 1/K^c$ for some fixed number $c > 1$ [44].

To apply our theory, we first identify all components in the model. For simplicity, we view N as the sample size, and D as the “dimension” of the parameters in the model. Under our vanilla notation, we are interested in learning parameters $\theta = (\pi, \mu)$, with $\pi = \{\gamma_d : d = 1, \dots, D\}$ and $\mu = \{\beta_k : k = 1, \dots, K\}$, from the posterior distribution $P(\pi, \mu, z | Y^N)$, where $S^N = \{S_n : n = 1, \dots, N\}$ with $S_n = \{z_{dn} : d = 1, \dots, D\}$ are latent variables, and $Y^N = \{Y_n : n = 1, \dots, N\}$ with $Y_n = \{w_{dn} : d = 1, \dots, D\}$ are the data, and the priors for (π, μ) are independent Dirichlet distributions $\text{Dir}_K(\eta_\gamma)$ and $\text{Dir}_V(\eta_\beta)$ whose densities are denoted by p_π and p_μ . The conditional distribution $p(Y^N | \mu, S^N)$ of the observation given the latent variable is

$$(w_{dn} | \mu, z_{dn}) \sim \text{multi}(\beta_{z_{dn}}), \quad d = 1, \dots, D \text{ and } n = 1, \dots, N.$$

Finally, the α -mean-field approximation considers using the family of probability density functions of forms

$$q(\mu, \pi, S^N) = q_\pi(\pi) q_\mu(\mu) \prod_{n=1}^N q_{S_n}(S_n) = \prod_{k=1}^K q_{\beta_k}(\beta_k) \prod_{d=1}^D \left(q_{\gamma_d}(\gamma_d) \prod_{n=1}^N q_{z_{dn}}(z_{dn}) \right)$$

to approximate the joint α -fractional posterior of (μ, π, S^N) . Since for LDA, each observation Y_n is composed of D independent observations, it is natural to present the variational inequality with the original loss function $D_\alpha[p(\cdot | \theta) || p(\cdot | \theta^*)] = \sum_{d=1}^D D_\alpha[p_d(\cdot | \theta) || p_d(\cdot | \theta^*)]$ re-scaling by a factor of D^{-1} , where $p_d(\cdot | \theta)$ denotes the likelihood function of the d th observation w_{dn} in Y_n . We make the following assumption.

Assumption S: (sparsity and regularity condition) Suppose for each k , β_k^* is $d_k \ll V$ sparse, and for each d , γ_d^* is $e_d \ll K$ sparse. Moreover, there exists some constant $\delta_0 > 0$, such that each nonzero component of β_k^* or γ_d^* is at least δ_0 .

Corollary 4.5. *Under Assumption S, it holds with probability at least $1 - C/(N \sum_{d=1}^D \varepsilon_{\gamma_d}^2 + N \sum_{k=1}^K \varepsilon_{\beta_k}^2)$ that*

$$\int \left\{ D^{-1} D_\alpha [p(\cdot | \theta) || p(\cdot | \theta^*)] \right\} \hat{q}_\theta(\theta) d\theta$$

$$\lesssim \frac{\alpha}{1-\alpha} \left\{ \frac{1}{D} \sum_{d=1}^D \varepsilon_{\gamma_d}^2 + \frac{1}{D} \sum_{k=1}^K \varepsilon_{\beta_k}^2 \right\} + \frac{1}{N(1-\alpha)} \left\{ \frac{1}{D} \sum_{d=1}^D e_d \log \frac{K}{\varepsilon_{\gamma_d}} + \frac{1}{D} \sum_{k=1}^K d_k \log \frac{V}{\varepsilon_{\beta_k}} \right\},$$

for any $\varepsilon_\gamma = (\varepsilon_{\gamma_1}, \dots, \varepsilon_{\gamma_d})$ and $\varepsilon_\beta = (\varepsilon_{\beta_1}, \dots, \varepsilon_{\beta_K})$. Therefore, if $(\sum_{d=1}^D e_d + \sum_{k=1}^K d_k)/(DN) \rightarrow 0$ as $N \rightarrow \infty$, then it holds with probability tending to one that as $N \rightarrow \infty$

$$\left\{ \int D^{-1} h^2 [p(\cdot | \theta) || p(\cdot | \theta^*)] \hat{q}_\theta(\theta) d\theta \right\}^{1/2}$$

$$\lesssim \sqrt{\frac{\sum_{d=1}^D e_d}{DN \min\{\alpha, 1-\alpha\}} \log(DKN) + \frac{\sum_{k=1}^K d_k}{DN \min\{\alpha, 1-\alpha\}} \log(KVN)}.$$

Corollary 4.5 implies estimation consistency as long as the “effective” dimensionality $\sum_{d=1}^D e_d + \sum_{k=1}^K d_k$ of the model is $o(DN)$ as the “effective sample size” $DN \rightarrow \infty$. In addition, the upper bound depends only logarithmically on the vocabulary size V due to the sparsity assumption.

5 Discussion

The primary motivation behind this work is to investigate whether point estimates obtained from mean-field or other variational approximations to a Bayesian posterior enjoy the same statistical accuracy as those obtained from the true posterior, and we answer the question in the affirmative for a wide range of statistical models. To that end, we have analyzed a class of variational objective functions indexed by a temperature parameter $\alpha \in (0, 1]$, with $\alpha = 1$ corresponding to the usual VB, and obtained risk bounds for the variational solution which can be used to show (near) minimax optimality of variational point estimates. Our theory was applied to a number of examples, including the mean-field approximation to Bayesian linear regression with and without variable selection, Gaussian mixture models, latent Dirichlet allocation, and (mixture of) Gaussian variational approximation in regular parameter models. This broader class of objective functions can be fitted in practice with

no additional difficulty compared to the usual VB. Hence, the proposed framework leads to a class of efficient variational algorithms with statistical guarantees.

The theory for the $\alpha < 1$ and the $\alpha = 1$ (usual VB) case lead to interesting contrasts. For $\alpha < 1$, a prior mass condition suffices to establish the risk bounds for the Hellinger (and more generally, Rényi divergences). However, the $\alpha = 1$ case requires additional conditions to be verified. When all conditions are met, there is no difference in terms of the rate of convergence for $\alpha < 1$ versus $\alpha = 1$. Hence, from a practical standpoint, the procedure with $\alpha < 1$ leads to theoretical guarantees with verification of fewer conditions. A comparison of second-order properties is left as a topic for future research, as is extension to models with dependent latent variates.

A Convention

As a convention, all equations defined in this supplementary document are numbered (S1), (S2), . . . , while equations cited from the main document retain their numbering in the main document. Similar for theorems, corollaries, lemmas etc.

In § S2, we provide an empirical study to compare the α -VB approach for $\alpha < 1$ to the usual VB in some of the models discussed in §4. In § S3, we illustrate applying our theory for continuous latent variable models. § S4 contains the Gaussian approximation example whose details were skipped in §4 of the main document. § S5 provides proofs of all theoretical results.

B Numerical Examples

In this section, we illustrate the α -VB procedure through several representative simulation examples. Since the objective functions $\Psi(q_\theta)$ and $\Psi(q_\theta, q_{S^n})$ differ from usual VB only through the presence of α , standard coordinate ascent variational inference (CAVI) algorithms[7, 10] can be implemented with simple modifications in the iterative updates. We implemented α -VB with different choices of α between 0.5 and 1 and the point estimates were fairly robust to the choice of α .

B.1 Bayesian high-dimensional linear regression

Consider sparsity inducing hierarchical prior $p_{\beta, z}$ over (β, z) as $\prod_{j=1}^d p_{\beta_j, z_j}$ with

$$p_{\beta_j, 1} = \mathcal{N}(\beta_j; 0, \nu_1 \sigma^2), \quad p_{\beta_j, 0} = \delta_0(\beta_j); \quad z_j \sim \text{Bernoulli}(1/d),$$

where δ_0 denotes the point mass measure at 0. Apply the variational approximation by using the family

$$q(\beta, \sigma, z) = q_\sigma(\sigma) \prod_{j=1}^d q_{z_j, \beta_j}(z_j, \beta_j)$$

where $q_{z_j, \beta_j}(z_j, \beta_j) = \prod_{j=1}^d N(\beta_j; \mu_j, \sigma_j^2)^{z_j} \delta_0(\beta_j)^{1-z_j} [q_{z_j}(1)]^{z_j} [q_{z_j}(0)]^{1-z_j}$. Let $\phi_j = q_{z_j}(1)$ for $j = 1, \dots, p$.

An implementation of the α -VB algorithm for Bayesian high-dimensional linear regression (α -VB-HDR) is described in Algorithm 1 and follows the batch-wise variational Bayes algorithm in Algorithm 2 of [22]. We sample $n = 100$ observations from the linear regression

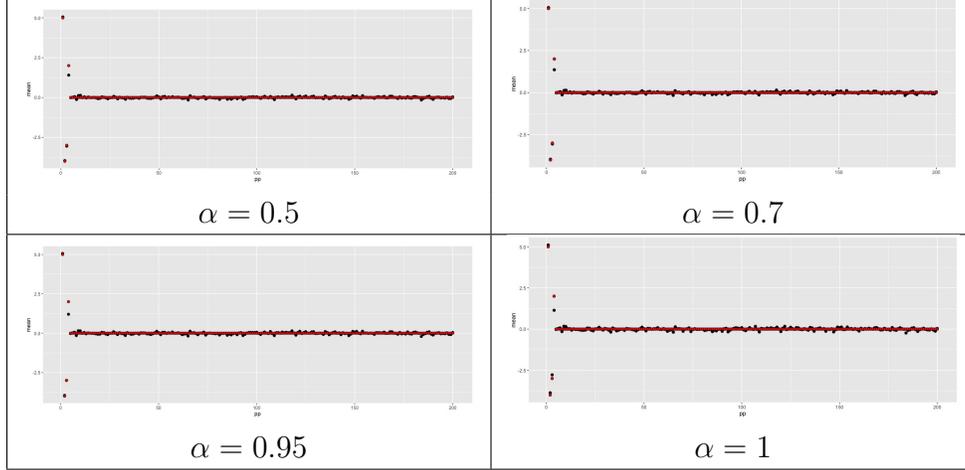


Figure 1: Plot of the 500 coefficients; true β in red and estimated means of β using α -VB-HDR in black.

model with $d = 500$, with the entries of the covariate matrix \mathbf{X} sampled independently from $N(0, 1.5^2)$, and error standard deviation $\sigma = 1$. The first 4 coefficients are non-zero and are set equal to $(5, -4, -3, 2)$. Figure 1 illustrates the performance of α -VB-HDR for different values of α . In all the cases, the convergence of ELBO occurs within less than 20 iterates.

Algorithm 1: α -VB-HDR

- 1 Set $\tilde{\sigma} = \sigma/\sqrt{\alpha}$.
- 2 Initialize $(\mu_1, \dots, \mu_d), (\sigma_1^2, \dots, \sigma_d^2), (\phi_1, \dots, \phi_d) = (1, \dots, 1)'$ and $\Phi = \text{Diag}(\phi_1, \dots, \phi_d)$.
- 3 **while** *ELBO does not converge* **do**
- 4 $(\mu_1, \dots, \mu_d)' = (\mathbf{X}'\mathbf{X} + \Phi/\nu_1)^{-1}\mathbf{X}'\mathbf{Y}$
- 5 **for** $j = 1, \dots, d$ **do**
- 6
$$\frac{1}{2\sigma_j^2} = \frac{\text{Diag}(\mathbf{X}'\mathbf{X})_j}{2\tilde{\sigma}^2} + \frac{\phi_j}{2\nu_1\tilde{\sigma}^2}$$

$$\phi_j = \text{Logit}^{-1}\left\{\text{Logit}(1/d) + \frac{1}{2}\log\left(\frac{\sigma_j^2}{\nu_1\tilde{\sigma}^2}\right) + \frac{\mu_j^2}{2\sigma_j^2}\right\}$$
- 7 **end**
- 8 **end**

B.2 Gaussian mixture models

We sample $n = 1000$ bi-variate observations from

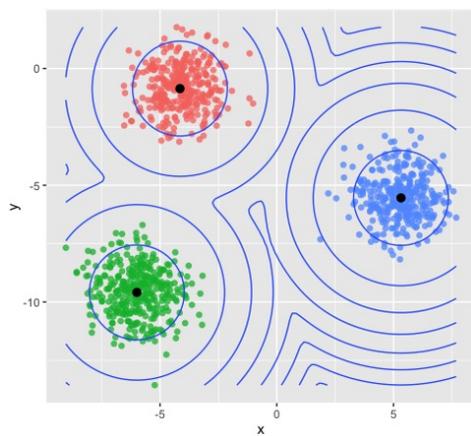
$$Y \sim \sum_{k=1}^K \pi_k \mathcal{N}(\mu_k, I_2),$$

for $\pi_k = 1/3$, $k = 1, \dots, K = 3$. μ_k are drawn from $\mathcal{N}_2(0, 50I_2)$ for $k = 1, \dots, 3$. Let $\pi(\mu_k) = \mathcal{N}_2(\mu_0, \sigma_0^2 I_2)$. We use $\mu_0 = (0, 0)'$ and $\sigma_0^2 = 50$. For simplicity, we assume π_k to be known in the study. We apply the mean field approximation by using the family of density functions of the form

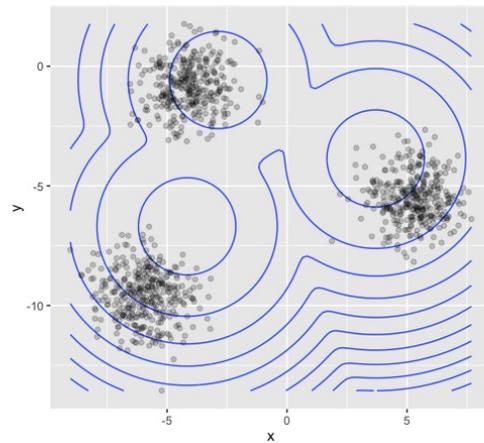
$$q(\mu, S^n) = q_\mu(\mu) q_{S^n}(s^n) = q_\mu(\mu) \prod_{i=1}^n q_{S_i}(s_i)$$

Following [10], we develop α -VB algorithm for Gaussian mixture models (α -VB-GMM), described in Algorithm 2. The derivation follows very closely to the case when $\alpha = 1$ and hence the details are omitted. Numerical results are shown in Figure 2. In all the cases, the convergence of ELBO occurs within less than 10 iterates. It is evident that for α close to 1, α -VB-GMM can recover the true density almost perfectly.

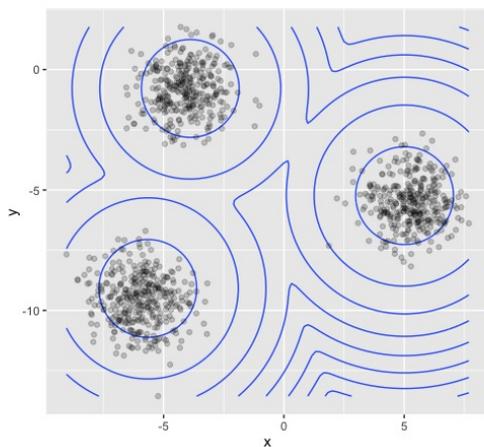
Algorithm 2: α -VB-GMM	
1	Initialize $\tilde{\mu}_k, \tilde{\sigma}_k, k = 1, \dots, K$ and $s_i, i = 1, \dots, n$.
2	while <i>ELBO does not converge</i> do
3	for $i = 1, \dots, n$ do
4	$q_{S_i}(s_i) \propto \exp\{\alpha \log \pi_{s_i} + \alpha y_{s_i} E(\mu_{s_i}) - \alpha E(\mu_{s_i}^2/2)\}$
5	end
6	for $k = 1, \dots, K$ do
7	Update
	$\tilde{\mu}_k = \frac{\mu_0/\sigma_0^2 + \sum_{i=1}^k q_{S_i}(s_i = k)y_{s_i}}{1/\sigma_0^2 + (1/\alpha) \sum_{i=1}^k q_{S_i}(s_i = k)}, \quad \tilde{\sigma}_k^2 = \frac{1}{1/\sigma_0^2 + (1/\alpha) \sum_{i=1}^k q_{S_i}(s_i = k)}$
	Set $q_{\mu_k} = \mathcal{N}_2(\mu_k; \tilde{\mu}_k, \tilde{\sigma}_k^2 I_2)$
8	end
9	end



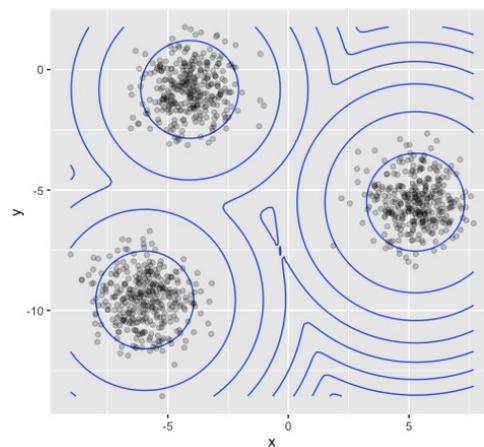
(a) True mixture density



(b) $\alpha = 0.7$



(c) $\alpha = 0.95$



(d) $\alpha = 1$

Figure 2: Contour plots for the true and predicted density using α -VB-GMM. The colors in Figure 2a represent different cluster components.

B.3 Latent Dirichlet Allocation

We implemented a version of LDA which is exactly same as Section 5.2 of [11]. The approach is the same as the one described here with one minor difference. The parameter η_γ is set to $1/K$, but η_β is estimated using an empirical Bayes approach described in Section 5.3 of [11] instead of fixing it to be $1/V$. To implement α -VB, we note that the only change required will be to Equation (6) of Section 5.2 where we replace $\phi_{ni} \propto \beta_{iw_n} \exp\{E_q[\log \theta_i|\gamma]\}$ to $\phi_{ni} \propto \beta_{iw_n} \exp\{\alpha E_q[\log \theta_i|\gamma]\}$. We provide an illustrative example of the use of an LDA model on a real data comprising of the first 5 out of 16,000 documents from a subset of the TREC AP corpus [19]. The maximum number of topics is set to 10. The top words from some of the resulting multinomial distributions $p(w|z)$ are illustrated in Table 1. The distributions seem to capture some of the underlying topics in the corpus with decreasing word similarity as α decreases.

Table 1: Top 5 words for each of the 10 extracted topics for $\alpha = 0.5, 0.7, 0.95, 1$ with the top 5 rows corresponding to $\alpha = 0.5$, next 5 rows corresponding to $\alpha = 0.7$ and so on.

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
history	police	year	peres	liberace	school	classroom	i	peres	first
ago	shot	people	israel	back	teacher	teacher	mrs	official	year
york	gun	get	bechtel	mrs	guns	boy	jewelry	rappaport	minister
president	students	first	offer	museum	boys	shot	museum	pipeline	new
today's	door	just	memo	man	saturday	baptist	bloomberg	offer	invasion
history	police	year	peres	liberace	school	shot	i	peres	year
president	students	people	offer	mrs	teacher	classroom	police	official	first
ago	school	get	bechtel	bloomberg	guns	teacher	mrs	rappaport	new
york	gun	volunteers	memo	back	shot	baptist	museum	pipeline	invasion
today's	yearold	mail	israel	door	boys	marino	jewelry	offer	minister
history	school	first	memo	liberace	first	shot	mrs	peres	people
president	police	year	effect	door	year	baptist	i	offer	get
ago	teacher	just	wage	back	day	marino	police	official	year
first	students	died	quoted	mrs	died	teacher	museum	rappaport	thompson
year	boys	day	bechtel	bloomberg	people	kids	bloomberg	bechtel	program
ago	police	get	memo	liberace	teacher	shot	i	peres	people
president	school	volunteers	bechtel	mrs	school	police	police	official	year
history	students	year	peres	bloomberg	shot	baptist	mrs	offer	thompson
first	teacher	mail	offer	back	guns	teacher	museum	rappaport	program
year	boys	people	israel	door	students	classroom	jewelry	pipeline	get

C Extension to continuous latent variables

As discussed in the main draft, we extend results on mean field approximations to models from discrete latent variables to continuous latent variables. For simplicity, we only focus on the $\alpha < 1$ case. All the proofs proceed in a similar way—the only difference is replacing all sums over latent variables with integrations. Specifically, in the definition of Ψ_α in (2.6),

the only change takes place in the quantity Δ_J , where the approximation to the likelihood is now made with continuous latent variables. We present a version where i.i.d. copies $T^n = (T_1, \dots, T_n)$ of the latent variable $T \in \mathcal{T}$ are continuous and there is no restrictions on the variational factor q_{T_i} for each latent variable T_i . In this setting, the α -VB objective function is simplified to

$$\begin{aligned} \Psi_\alpha(q_\theta, q_{T^n}) = & \\ & - \int_{\Theta} q_\theta(\theta) \sum_{i=1}^n \int_{\mathcal{T}} q_{T_i}(t_i) \log \frac{p(Y_i | \mu, t_i) p_{T_i}(t_i | \pi)}{p(Y_i | \theta^*) q_{T_i}(t_i)} dt_i d\theta + \alpha^{-1} D(q_\theta || p_\theta), \end{aligned} \quad (\text{C.1})$$

where we assume that the distribution family $p_T(\cdot | \pi)$ for the latent variable is indexed by its own parameter π , and recall that μ is the parameter in the likelihood function $p(Y | \mu, T)$ of response Y given the latent variable T , and $\theta = (\pi, \mu)$ are the parameters.

Similar to the discrete case, for continuous latent variables, we define the following two KL neighborhoods of π^* and μ^*

$$\begin{aligned} \mathcal{B}_n^{\text{con}}(\pi^*, \varepsilon_\pi) &= \left\{ D[p_T(\cdot | \pi^*) || p_T(\cdot | \pi)] \leq \varepsilon_\pi^2, \quad V[p_T(\cdot | \pi^*) || p_T(\cdot | \pi)] \leq \varepsilon_\pi^2 \right\}, \\ \mathcal{B}_n^{\text{con}}(\mu^*, \varepsilon_\mu) &= \left\{ \sup_t D[p(\cdot | \mu^*, t) || p(\cdot | \mu, t)] \leq \varepsilon_\mu^2, \right. \\ & \quad \left. \sup_t V[p(\cdot | \mu^*, t) || p(\cdot | \mu, t)] \leq \varepsilon_\mu^2 \right\}. \end{aligned}$$

We now state a theorem with the same combined conclusions of Corollary 3.2 and Theorem 3.4 for the continuous case. The proof is similar and hence omitted.

Theorem C.1. *For any measure q_θ over θ satisfying $q_\theta \ll p_\theta$, it holds with probability at least $(1 - \zeta)$ that*

$$\begin{aligned} & \int \left\{ D_\alpha[p(\cdot | \theta) || p(\cdot | \theta^*)] \right\} \hat{q}_{\theta, \alpha}(\theta) d\theta \\ & \leq -\frac{\alpha}{n(1-\alpha)} \int_{\Theta} q_\theta(\theta) \sum_{i=1}^n \left\{ \int_{\mathcal{T}} \tilde{q}_{T_i}(t_i) \log \frac{p(Y_i | \mu, t_i) p_{T_i}(t_i | \pi)}{p(Y_i | \mu^*, t_i) p_{T_i}(t_i | \pi^*)} dt_i \right\} d\theta \\ & \quad + \frac{1}{n(1-\alpha)} D(q_\theta || p_\theta) + \frac{1}{n(1-\alpha)} \log(1/\zeta), \end{aligned} \quad (\text{C.2})$$

where \tilde{q}_{T_i} is a probability distribution over \mathcal{T} satisfying

$$\tilde{q}_{T_i}(t_i) = \frac{p_{T_i}(t_i | \pi^*) p(Y_i | \mu^*, t_i)}{p(Y_i | \theta^*)}, \quad t_i \in \mathcal{T}. \quad (\text{C.3})$$

Moreover, for any fixed $(\varepsilon_\pi, \varepsilon_\mu) \in (0, 1)^2$, with \mathbb{P}_{θ^*} probability at least $1 - 5/\{(D-1)^2 n (\varepsilon_\pi^2 + \varepsilon_\mu^2)\}$, it holds that

$$\int \left\{ D_\alpha [p(\cdot | \theta) || p(\cdot | \theta^*)] \right\} \widehat{q}_{\theta, \alpha}(\theta) d\theta \leq \frac{D\alpha}{1-\alpha} (\varepsilon_\pi^2 + \varepsilon_\mu^2) + \left\{ -\frac{1}{n(1-\alpha)} \log P_\pi [\mathcal{B}_n^{\text{con}}(\pi^*, \varepsilon_\pi)] \right\} + \left\{ -\frac{1}{n(1-\alpha)} \log P_\mu [\mathcal{B}_n^{\text{con}}(\mu^*, \varepsilon_\mu)] \right\}. \quad (\text{C.4})$$

In presence of continuous latent variables, if the mean-field variational family is further constrained by restricting each factor q_{T_i} corresponding to the latent variable T_i to belong to a parametric family Γ_{T_i} , such as the exponential family, then the Bayes risk bound of Theorem C.1 still applies as long as the family Γ_{T_i} for q_{T_i} contains densities of form (C.3)—which is the case if the conditional distribution $p(T_i | \pi)$ also belongs to Γ_{T_i} and the model $p(Y_i | \mu, \tau_i)$ is conjugate with respect to family Γ_{T_i} .

D Gaussian approximation to regular parametric models

We discuss the details of this example from §4 which were skipped in the main document. For sake of completeness, we remind the readers of the setting.

Consider a family of regular parametric models $\mathcal{P} = \{\mathbb{P}_\theta^{(n)} : \theta \in \Theta\}$ where n is the sample size, and the likelihood function $p_\theta^{(n)}$ is indexed by a parameter θ belonging to the parameter space $\Theta \subset \mathbb{R}^d$, which we assume to be compact. Let p_θ denote the prior density of over Θ , and $Y^n = (Y_1, \dots, Y_n)$ be the observations from $\mathbb{P}_{\theta^*}^{(n)}$, with θ^* being the truth. We apply the Gaussian approximation by using the Gaussian family Γ_G (restricted to Θ)

$$q(\theta) \propto \mathcal{N}(\theta; \mu, \Sigma) I_\Theta(\theta), \quad \mu \in \mathbb{R}^d \text{ and } \Sigma \text{ is a } d \times d \text{ positive definite matrix.}$$

The Gaussian variational approximation \widehat{q}_θ as

$$\widehat{q}_\theta := \operatorname{argmin}_{q_\theta \in \Gamma_G} \left\{ -\alpha \int_\Theta \int q_\theta(\theta) \log p_\theta^{(n)}(Y^n) d\theta + D(q_\theta || p_\theta) \right\}.$$

we make the following assumption.

Assumption P: (prior thickness and regularity condition) The prior density p_θ satisfies $\inf_{\theta \in \Theta} p_\theta(\theta) > 0$, and there exists some constant C such that $D[p(\cdot | \theta_1) || p(\cdot | \theta_2)] \leq C$.

$C \|\theta_1 - \theta_2\|^2$ and $V[p(\cdot | \theta_1) || p(\cdot | \theta_2)] \leq C \|\theta_1 - \theta_2\|^2$ holds for all $(\theta_1, \theta_2) \in \Theta^2$.

Corollary D.1. *Under Assumption P, it holds with probability tending to one as $n \rightarrow \infty$ that*

$$\left\{ \int h^2[p(\cdot | \theta) || p(\cdot | \theta^*)] \hat{q}_\theta(\theta) d\theta \right\}^{1/2} \lesssim \sqrt{\frac{d}{n \min\{\alpha, 1 - \alpha\}} \log(dn)}.$$

Under the model identifiability condition $h^2[p(\cdot | \theta) || p(\cdot | \theta^*)] \gtrsim \|\theta - \theta^*\|^2$, Corollary D.1 implies a convergence rate $\sqrt{n^{-1} d \log(dn)}$ for the variational-Bayes estimator $\hat{\theta}_B$ of θ . By examining the proof of the corollary, we find that the normality form in the variational approximation does not play a critical role in the proof—similar Bayesian risk upper bounds hold under some additional conditions for a broader class of variational distributions as well, such as any location-scale distribution family with sub-exponential tails. It is a well-known fact [41, 43] that the covariance matrices from the variational approximations are typically “too small” compared with those for the sampling distribution of the maximum likelihood estimator, which combined with the Bernstein von-mises theorem implies that the variational approximation \hat{q}_θ may not converge to the true posterior distribution. This fact combined with the result in Corollary D.1 indicates: 1. minimizing the KL divergence over the variational family forces the variational distribution \hat{q}_θ to concentrate around the truth θ^* at the optimal rate (due to the heavy penalty on the tails in the KL divergence); 2. however, the local shape of \hat{q}_θ around θ^* can be far away from that of the true posterior due to dis-match between the distributions in the variational family and the true posterior.

E Proofs

In this section, we present proofs of all technical results in the main document.

E.1 Proof of Theorem 3.1

We first state a key variational lemma that plays a critical role in the proof.

Lemma E.1. *Let μ_θ be a probability measure over θ and μ_{S^n} be a probability measure over S^n , and $h(\theta, S^n)$ a measurable function such that for any fixed S^n , $e^{h(\cdot, S^n)} \in L_1(\mu_\theta)$. Then,*

$$\begin{aligned} \log \int \sum_{s^n} e^{h(\theta, s^n)} \mu_{S^n}(s^n) \mu_\theta(d\theta) \\ = \sup_{\rho_n(\theta, S^n)} \left[\int \sum_{s^n} h(\theta, s^n) \rho_n(d\theta, s^n) - D(\rho_n(\theta, S^n) || \mu_\theta \otimes \mu_{S^n}) \right], \end{aligned}$$

where the supremum is over all probability measures $\rho_n(\theta, S^n) \ll \mu_\theta \otimes \mu_{S^n}$. Further, the supremum on the right hand side is attained when

$$\frac{\rho_n(d\theta, s^n)}{\mu(d\theta) \mu_{S^n}(s^n)} = \frac{e^{h(\theta, S^n)}}{\int \sum_{s^n} e^{h(\theta, s^n)} \mu_{S^n}(s^n) \mu(d\theta)}.$$

Proof. Use the well-known variational dual representation of the KL divergence (see, e.g., Corollary 4.15 of [12]) which states that for any probability measure μ and any measurable function h with $e^h \in L_1(\mu)$, one has

$$\log \int e^{h(\eta)} \mu(d\eta) = \sup_{\rho \ll \mu} \left[\int h(\eta) \rho(d\eta) - D(\rho || \mu) \right],$$

where the supremum is over all probability distributions $\rho \ll \mu$, and equality is attained when $d\rho/d\mu \propto e^h$. This fact simply follows upon an application of Jensen's inequality. In the current context, set $\eta = (\theta, s^n)$, $\mu = \mu_\theta \otimes \mu_{S^n}$ and $\rho(d\eta) = \rho_n(d\theta, s^n)$ to obtain the conclusion of Lemma E.1. \square

Return to the proof of the theorem. By applying Jensen's inequality to function $x \mapsto x^\alpha$ ($\alpha < 1$), we obtain that, for any (possibly data dependent) measure q_{S^n} ,

$$\begin{aligned} \mathbb{E}_{\theta^*} \left[\sum_{s^n} q_{S^n}(s^n) \exp \left\{ \alpha \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} \right\} \right] \\ &= \int_{\mathbb{R}^n} \sum_{s^n} q_{S^n}(s^n) \left\{ \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} \right\}^\alpha p(Y^n | \theta^*) dY^n \\ &\leq \int_{\mathbb{R}^n} \left\{ \sum_{s^n} q_{S^n}(s^n) \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} \right\}^\alpha p(Y^n | \theta^*) dY^n \\ &\leq \int \left\{ \frac{p(Y^n | \theta)}{p(Y^n | \theta^*)} \right\}^\alpha p(Y^n | \theta^*) dY^n \\ &= e^{-(1-\alpha) D_\alpha^{(n)}(\theta, \theta^*)}, \end{aligned}$$

with $D_\alpha^{(n)}(\theta, \theta^*)$ defined in the first display of §3.1. Thus, for any $\zeta \in (0, 1)$, we have

$$\begin{aligned} \mathbb{E}_{\theta^*} \left[\sum_{s^n} q_{S^n}(s^n) \exp \left\{ \alpha \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} \right. \right. \\ \left. \left. + (1 - \alpha) D_\alpha^{(n)}(\theta, \theta^*) - \log(1/\zeta) \right\} \right] \leq \zeta. \end{aligned}$$

Integrating both side of this inequality with respect to p_θ and interchanging the integrals

using Fubini's theorem, we obtain

$$\mathbb{E}_{\theta^*} \left[\int_{\Theta} \sum_{s^n} p_{\theta}(\theta) q_{S^n}(s^n) \exp \left\{ \alpha \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} \right. \right. \\ \left. \left. + (1 - \alpha) D_{\alpha}^{(n)}(\theta, \theta^*) - \log(1/\zeta) \right\} d\theta \right] \leq \zeta.$$

Now, apply Lemma E.1 with $\mu_{\theta} = p_{\theta}$, $\mu_{S^n} = q_{S^n}$ and

$$h(\theta, s^n) = \alpha \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} + (1 - \alpha) D_{\alpha}^{(n)}(\theta, \theta^*) - \log(1/\zeta),$$

to obtain that

$$\mathbb{E}_{\theta^*} \exp \sup_{\rho(\theta, S^n)} \left[\int_{\Theta} \sum_{s^n} \left\{ \alpha \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} + (1 - \alpha) D_{\alpha}^{(n)}(\theta, \theta^*) \right. \right. \\ \left. \left. - \log(1/\zeta) \right\} \rho(d\theta, s^n) - D(\rho || p_{\theta} \otimes q_{S^n}) \right] \leq \zeta.$$

If we choose $\rho = q_{\theta} \otimes q_{S^n}$ in the preceding display for any (possibly data dependent) $q_{\theta} \ll p_{\theta}$, then

$$\mathbb{E}_{\theta^*} \exp \left[\int_{\Theta} \sum_{s^n} \left\{ \alpha \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} + (1 - \alpha) D_{\alpha}^{(n)}(\theta, \theta^*) \right. \right. \\ \left. \left. - \log(1/\zeta) \right\} q_{\theta}(d\theta) q_{S^n}(s^n) - D(q_{\theta} || p_{\theta}) \right] \leq \zeta.$$

By applying Markov's inequality, we further obtain that with \mathbb{P}_{θ^*} probability at least $(1 - \zeta)$,

$$(1 - \alpha) \int D_{\alpha}^{(n)}(\theta, \theta^*) q_{\theta}(d\theta) \\ \leq -\alpha \int_{\Theta} \sum_{s^n} \left\{ \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} \right\} q_{\theta}(d\theta) q_{S^n}(s^n) + D(q_{\theta} || p_{\theta}) + \log(1/\zeta) \\ = \alpha \Psi_{\alpha}(q_{\theta}, q_{S^n}) + \log(1/\zeta),$$

since, from (2.2) – (2.3) and (2.6),

$$\Psi_{\alpha}(q_{\theta}, q_{S^n}) = - \sum_{s^n} \left[q_{S^n}(s^n) \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} \right] q_{\theta}(d\theta) + \alpha^{-1} D(q_{\theta} || p_{\theta}).$$

Since the inequality in the penultimate display holds for any (possibly data dependent)

$q_\theta \ll p_\theta$ and q_{S^n} , we obtain, in particular,

$$(1 - \alpha) \int D_\alpha^{(n)}(\theta, \theta^*) \widehat{q}_{\theta, \alpha}(d\theta) \leq \alpha \Psi_\alpha(\widehat{q}_{\theta, \alpha}, \widehat{q}_{S^n, \alpha}) + \log(1/\zeta).$$

The conclusion of the Theorem follows since $\Psi_\alpha(\widehat{q}_{\theta, \alpha}, \widehat{q}_{S^n, \alpha}) \leq \Psi_\alpha(q_\theta, q_{S^n})$ for any $q_\theta \ll p_\theta$ and q_{S^n} .

E.2 Proof of Theorem 3.4

We choose q_θ as the probability density function q_θ^* of

$$Q_\theta^* = \frac{P_\pi[\cdot \cap \mathcal{B}_n(\pi^*, \varepsilon_\pi)] \otimes P_\mu[\cdot \cap \mathcal{B}_n(\mu^*, \varepsilon_\mu)]}{P_\pi[\mathcal{B}_n(\pi^*, \varepsilon_\pi)] \cdot P_\mu[\mathcal{B}_n(\mu^*, \varepsilon_\mu)]},$$

the product measure of restrictions of the priors (P_π, P_μ) for (π, μ) to two KL neighborhoods $\mathcal{B}_n(\pi^*, \varepsilon_\pi)$ and $\mathcal{B}_n(\mu^*, \varepsilon_\mu)$ around (π^*, μ^*) .

Next, we will characterize the first two moments of the first term on the r.h.s. in inequality (3.2) under this choice of q_θ^* . By applying Fubini's theorem, we have

$$\begin{aligned} \mathbb{E}_{\theta^*} \left[\int_{\Theta} q_\theta^*(\theta) \sum_{i=1}^n \sum_{s_i} \widetilde{q}_{S_i}(s_i) \log \frac{p(Y_i | \mu, s_i) \pi_{s_i}}{p(Y_i | \mu^*, s_i) \pi_{s_i}^*} d\theta \right] \\ = \int_{\Theta} \mathbb{E}_{\theta^*} \left[\sum_{i=1}^n \sum_{s_i} \widetilde{q}_{S_i}(s_i) \log \frac{p(Y_i | \mu, s_i) \pi_{s_i}}{p(Y_i | \mu^*, s_i) \pi_{s_i}^*} \right] q_\theta^*(\theta) d\theta. \end{aligned}$$

By plugging-in the expression of $\widetilde{q}_{S_i}(s_i)$ and applying Fubini's theorem, we obtain

$$\begin{aligned} \mathbb{E}_{\theta^*} \left[\sum_{i=1}^n \sum_{s_i} \widetilde{q}_{S_i}(s_i) \log \frac{p(Y_i | \mu, s_i) \pi_{s_i}}{p(Y_i | \mu^*, s_i) \pi_{s_i}^*} \right] \\ = n \mathbb{E}_{\theta^*} \left[\widetilde{q}_S(s) \log \frac{p(Y | \mu, s) \pi_s}{p(Y | \mu^*, s) \pi_s^*} \right] \\ = -n D(\pi^* || \pi) - n \sum_s \pi_s^* D[p(\cdot | \mu^*, s) || p(\cdot | \mu, s)], \end{aligned}$$

where recall shorthand $D(\pi^* || \pi) = \sum_s \pi_s^* \log(\pi_s^*/\pi_s)$ as the KL divergence between categorical distributions with parameters π^* and π . Combining the two preceding displays and

invoking the definitions of $B_n(\pi^*, \varepsilon_\pi)$ and $B_n(\mu^*, \varepsilon_\mu)$, we obtain

$$\mathbb{E}_{\theta^*} \left[- \int_{\Theta} q_{\theta}^*(\theta) \sum_{i=1}^n \sum_{s_i} \tilde{q}_{S_i}(s_i) \log \frac{p(Y_i | \mu, s_i) \pi_{s_i}}{p(Y_i | \mu^*, s_i) \pi_{s_i}^*} d\theta \right] \leq n \varepsilon_\pi^2 + n \varepsilon_\mu^2.$$

Similarly, by applying Fubini's theorem, we have

$$\begin{aligned} \text{Var}_{\theta^*} \left[\int_{\Theta} q_{\theta}^*(\theta) \sum_{i=1}^n \sum_{s_i} \tilde{q}_{S_i}(s_i) \log \frac{p(Y_i | \mu, s_i) \pi_{s_i}}{p(Y_i | \mu^*, s_i) \pi_{s_i}^*} d\theta \right] \\ &= n \text{Var}_{\theta^*} \left[\int_{\Theta} q_{\theta}^*(\theta) \sum_s \tilde{q}_S(s) \log \frac{p(Y | \mu, s) \pi_s}{p(Y | \mu^*, s) \pi_s^*} d\theta \right] \\ &\leq n \mathbb{E}_{\theta^*} \left[\int_{\Theta} q_{\theta}^*(\theta) \sum_s \tilde{q}_S(s) \log \frac{p(Y | \mu, s) \pi_s}{p(Y | \mu^*, s) \pi_s^*} d\theta \right]^2 \\ &\stackrel{(i)}{\leq} n \int_{\Theta} \mathbb{E}_{\theta^*} \left[\sum_s \tilde{q}_S(s) \log \frac{p(Y | \mu, s) \pi_s}{p(Y | \mu^*, s) \pi_s^*} \right]^2 q_{\theta}^*(\theta) d\theta \\ &\stackrel{(ii)}{\leq} n \int_{\Theta} \mathbb{E}_{\theta^*} \left[\sum_s \tilde{q}_S(s) \log^2 \frac{p(Y | \mu, s) \pi_s}{p(Y | \mu^*, s) \pi_s^*} \right] q_{\theta}^*(\theta) d\theta, \end{aligned}$$

where steps (i) and (ii) follows by Jensen's inequality and Fubini's theorem. By plugging-in the expression of $\tilde{q}_{S_i}(s_i)$ and applying Fubini's theorem, we obtain

$$\begin{aligned} \mathbb{E}_{\theta^*} \left[\sum_s \tilde{q}_S(s) \log^2 \frac{p(Y | \mu, s) \pi_s}{p(Y | \mu^*, s) \pi_s^*} \right] q_{\theta}^*(\theta) d\theta \\ \leq 2n V(\pi^* || \pi) + 2n \sum_s \pi_s^* V[p(\cdot | \mu^*, s) || p(\cdot | \mu, s)], \end{aligned}$$

where recall the shorthand $V(\pi^* || \pi) = \sum_s \pi_s^* \log^2(\pi_s^*/\pi_s)$ to denote the V -divergence between categorical distributions with parameters π^* and π , and we applied the inequality $(x + y)^2 \leq 2x^2 + 2y^2$. By combining the two preceding displays and invoking the definitions of $B_n(\pi^*, \varepsilon_\pi)$ and $B_n(\mu^*, \varepsilon_\mu)$, we obtain

$$\text{Var}_{\theta^*} \left[\int_{\Theta} q_{\theta}^*(\theta) \sum_{i=1}^n \sum_{s_i} \tilde{q}_{S_i}(s_i) \log \frac{p(Y_i | \mu, s_i) \pi_{s_i}}{p(Y_i | \mu^*, s_i) \pi_{s_i}^*} d\theta \right] \leq 2n \varepsilon_\pi^2 + 2n \varepsilon_\mu^2.$$

Putting piece together, we obtain by applying Chebyshev's inequality that

$$\begin{aligned}
& \mathbb{P}_{\theta^*} \left\{ \int_{\Theta} q_{\theta}^*(\theta) \sum_{i=1}^n \sum_{s_i} \tilde{q}_{S_i}(s_i) \log \frac{p(Y_i | \mu, s_i) \pi_{s_i}}{p(Y_i | \mu^*, s_i) \pi_{s_i}^*} d\theta \leq -D n(\varepsilon_{\pi}^2 + \varepsilon_{\mu}^2) \right\} \\
& \stackrel{(i)}{\leq} \mathbb{P}_{\theta^*} \left\{ \int_{\Theta} q_{\theta}^*(\theta) \sum_{i=1}^n \sum_{s_i} \tilde{q}_{S_i}(s_i) \log \frac{p(Y_i | \mu, s_i) \pi_{s_i}}{p(Y_i | \mu^*, s_i) \pi_{s_i}^*} d\theta \right. \\
& \quad \left. - \mathbb{E}_{\theta^*} \left[\int_{\Theta} q_{\theta}^*(\theta) \sum_{i=1}^n \sum_{s_i} \tilde{q}_{S_i}(s_i) \log \frac{p(Y_i | \mu, s_i) \pi_{s_i}}{p(Y_i | \mu^*, s_i) \pi_{s_i}^*} d\theta \right] \leq -(D-1) n(\varepsilon_{\pi}^2 + \varepsilon_{\mu}^2) \right\} \\
& \leq \frac{\text{Var}_{\theta^*} \left[\int_{\Theta} q_{\theta}^*(\theta) \sum_{i=1}^n \sum_{s_i} \tilde{q}_{S_i}(s_i) \log \frac{p(Y_i | \mu, s_i) \pi_{s_i}}{p(Y_i | \mu^*, s_i) \pi_{s_i}^*} d\theta \right]}{(D-1)^2 n^2 (\varepsilon_{\pi}^2 + \varepsilon_{\mu}^2)^2} \stackrel{(ii)}{\leq} \frac{4}{(D-1)^2 n (\varepsilon_{\pi}^2 + \varepsilon_{\mu}^2)},
\end{aligned}$$

where in steps (i) and (ii), we have respectively used the derived first and second moment bounds.

Finally, we have

$$D(q_{\theta}^* || p_{\theta}) = - \left[\log P_{\pi}[\mathcal{B}_n(\pi^*, \varepsilon_{\pi})] + \log P_{\mu}[B_n(\mu^*, \varepsilon_{\mu})] \right],$$

since for any probability measure μ , a measurable set A with $\mu(A) > 0$, and $\tilde{\mu}(\cdot) = \mu(\cdot \cap A)/\mu(A)$ the restriction of μ to A , $D(\tilde{\mu} || \mu) = -\log \mu(A)$.

The claimed bound in the theorem is now a direct consequence of the preceding two displays and Corollary 3.2 with the choice $q_{\theta} = q_{\theta}^*$.

E.3 Proof of Theorem 3.5

Recall that $\ell_n(\theta) = \log p(Y^n | \theta)$ is the marginal log-likelihood function (after marginalizing out latent variables), and $\ell_n(\theta, \theta^*) = \ell_n(\theta) - \ell_n(\theta^*)$ the log-likelihood ratio function. Clearly, $\mathbb{E}_{\theta^*} \exp\{\ell_n(\theta, \theta^*)\} = 1$. The type II error bound (3.9) in Assumption T implies for fixed $\varepsilon > \varepsilon_n$, any $\theta \in \mathcal{F}_{n,\varepsilon}$, and any (possibly data dependent) probability measure q_{S^n} ,

$$\begin{aligned}
& \mathbb{E}_{\theta^*} \left[\sum_{s^n} q_{S^n}(s^n) \exp \left\{ \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} \right\} (1 - \phi_{n,\varepsilon}) \right] \\
& = \mathbb{E}_{\theta^*} \left[\exp \{ \ell_n(\theta, \theta^*) \} (1 - \phi_{n,\varepsilon}) \right] \leq \exp \{ -c n r(\theta, \theta^*) \mathbb{I}[r(\theta, \theta^*) \geq \varepsilon^2] \}.
\end{aligned}$$

Thus, for any $\eta \in (0, 1)$, we have

$$\mathbb{E}_{\theta^*} \left[\sum_{s^n} q_{S^n}(s^n) \exp \left\{ \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{s^n}(s^n)} + c n r(\theta, \theta^*) \mathbb{I}[r(\theta, \theta^*) \geq \varepsilon^2] - \log(1/\eta) \right\} (1 - \phi_{n,\varepsilon}) \right] \leq \eta.$$

Let $P_{\theta, \mathcal{F}_{n,\varepsilon}}(\cdot) = P_\theta(\cdot \cap \mathcal{F}_{n,\varepsilon})/P_\theta(\mathcal{F}_{n,\varepsilon})$ denote the restriction of the prior P_θ on $\mathcal{F}_{n,\varepsilon}$. Integrating both side of this inequality with respect to $P_{\theta, \mathcal{F}_{n,\varepsilon}}$ on $\mathcal{F}_{n,\varepsilon}$ and interchanging the integrals using Fubini's theorem, we obtain

$$\mathbb{E}_{\theta^*} \left[(1 - \phi_{n,\varepsilon}) \int_{\mathcal{F}_{n,\varepsilon}} \sum_{s^n} q_{S^n}(s^n) \exp \left\{ \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{s^n}(s^n)} + c n r(\theta, \theta^*) \mathbb{I}[r(\theta, \theta^*) \geq \varepsilon^2] - \log(1/\eta) \right\} P_{\theta, \mathcal{F}_{n,\varepsilon}}(d\theta) \right] \leq \eta.$$

Now, Lemma E.1 implies for any $\rho \ll P_{\theta, \mathcal{F}_{n,\varepsilon}}$,

$$\mathbb{E}_{\theta^*} \left[(1 - \phi_{n,\varepsilon}) \exp \left\{ \int_{\mathcal{F}_{n,\varepsilon}} \sum_{s^n} q_{S^n}(s^n) \left(\log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{s^n}(s^n)} + c n r(\theta, \theta^*) \mathbb{I}[r(\theta, \theta^*) \geq \varepsilon^2] - \log(1/\eta) \right) \rho(d\theta) - D(\rho || P_{\theta, \mathcal{F}_{n,\varepsilon}}) \right\} \right] \leq \eta.$$

Take ρ to be the restriction $\widehat{Q}_{\mathcal{F}_{n,\varepsilon}}$ of \widehat{Q} over $\mathcal{F}_{n,\varepsilon}$, we obtain

$$\mathbb{E}_{\theta^*} \left[(1 - \phi_{n,\varepsilon}) \exp \left\{ \frac{1}{\widehat{Q}(\mathcal{F}_{n,\varepsilon})} \int_{\mathcal{F}_{n,\varepsilon}} \sum_{s^n} q_{S^n}(s^n) \left(\log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{s^n}(s^n)} + c n r(\theta, \theta^*) \mathbb{I}[r(\theta, \theta^*) \geq \varepsilon^2] - \log(1/\eta) \right) \widehat{Q}(d\theta) - D(\widehat{Q}_{\mathcal{F}_{n,\varepsilon}} || P_{\theta, \mathcal{F}_{n,\varepsilon}}) \right\} \right] \leq \eta.$$

By applying Markov's inequality, we further obtain that with \mathbb{P}_{θ^*} probability at least $(1 - \sqrt{\eta})$,

$$(1 - \phi_{n,\varepsilon}) \exp \left\{ \frac{1}{\widehat{Q}(\mathcal{F}_{n,\varepsilon})} \int_{\mathcal{F}_{n,\varepsilon}} \sum_{s^n} q_{S^n}(s^n) \left(\log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{s^n}(s^n)} + c n r(\theta, \theta^*) \mathbb{I}[r(\theta, \theta^*) \geq \varepsilon^2] - \log(1/\eta) \right) \widehat{Q}(d\theta) - D(\widehat{Q}_{\mathcal{F}_{n,\varepsilon}} || P_{\theta, \mathcal{F}_{n,\varepsilon}}) \right\} \leq \eta^{-1/2}.$$

Denote the big exponential term in the above display by A_n . Then the above display is equivalent to

$$(1 - \phi_{n,\varepsilon}) A_n \leq \eta^{-1/2}.$$

The type I error bound (3.8) in Assumption T implies, by Markov's inequality, that $\phi_{n,\varepsilon} \leq e^{-cn\varepsilon_n^2/2}$ holds with \mathbb{P}_{θ^*} probability at least $(1 - e^{-cn\varepsilon_n^2/2})$, implying

$$\phi_{n,\varepsilon} A_n \leq e^{-cn\varepsilon_n^2/2} A_n.$$

Combining the two preceding displays, we obtain that with \mathbb{P}_{θ^*} probability at least $(1 - 2e^{-cn\varepsilon_n^2/2})$ (taking $\eta = e^{-cn\varepsilon_n^2/2}$),

$$A_n = (1 - \phi_{n,\varepsilon})A_n + \phi_{n,\varepsilon} A_n \leq e^{cn\varepsilon_n^2/2} + e^{-cn\varepsilon_n^2/2} A_n,$$

leading to the following bound for A_n as

$$A_n \leq \frac{1}{1 - e^{-cn\varepsilon_n^2/2}} e^{cn\varepsilon_n^2/2} \leq 2e^{cn\varepsilon_n^2/2}.$$

Consequently, using the definition of A_n , we get

$$\begin{aligned} \frac{1}{\widehat{Q}(\mathcal{F}_{n,\varepsilon})} \int_{\mathcal{F}_{n,\varepsilon}} \sum_{s^n} q_{S^n}(s^n) \left(\log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{s^n}(s^n)} + cn r(\theta, \theta^*) \mathbb{I}[r(\theta, \theta^*) \geq \varepsilon^2] \right) \widehat{Q}(d\theta) \\ - D(\widehat{Q}_{\mathcal{F}_{n,\varepsilon}} \| P_{\theta, \mathcal{F}_{n,\varepsilon}}) \leq cn\varepsilon_n^2/2 + \log 2. \end{aligned}$$

Rearranging terms, we obtain

$$\begin{aligned} cn \int_{\theta \in \mathcal{F}_{n,\varepsilon}, r(\theta, \theta^*) \geq \varepsilon^2} r(\theta, \theta^*) \widehat{Q}(d\theta) - \widehat{Q}(\mathcal{F}_{n,\varepsilon}) D(\widehat{Q}_{\mathcal{F}_{n,\varepsilon}} \| P_{\theta, \mathcal{F}_{n,\varepsilon}}) \\ \leq \int_{\mathcal{F}_{n,\varepsilon}} - \sum_{s^n} q_{S^n}(s^n) \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{s^n}(s^n)} \widehat{Q}(d\theta) + (cn\varepsilon_n^2/2 + \log 2) \widehat{Q}(\mathcal{F}_{n,\varepsilon}). \end{aligned} \tag{E.1}$$

Similarly, for each $\theta \in \mathcal{F}_{n,\varepsilon}^c$, from the identity $\mathbb{E}_{\theta^*} [\exp \{ \ell_n(\theta, \theta^*) \}] = 1$ and Lemma E.1, we can obtain that for any measure $\rho \ll P_{\theta, \mathcal{F}_{n,\varepsilon}^c}$,

$$\begin{aligned} \mathbb{E}_{\theta^*} \left[\exp \left\{ \int_{\mathcal{F}_{n,\varepsilon}^c} \sum_{s^n} q_{S^n}(s^n) \left(\log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{s^n}(s^n)} \right. \right. \right. \\ \left. \left. \left. - \log(1/\eta) \right) \rho(d\theta) - D(\rho \| P_{\theta, \mathcal{F}_{n,\varepsilon}^c}) \right\} \right] \leq \eta. \end{aligned}$$

Take ρ to be the restriction $\widehat{Q}_{\mathcal{F}_{n,\varepsilon}^c}$ of \widehat{Q} over $\mathcal{F}_{n,\varepsilon}^c$ and $\eta = e^{-cn\varepsilon_n^2}$, we can get that with \mathbb{P}_{θ^*}

probability at least $(1 - 2e^{-cn\varepsilon_n^2/2})$

$$\frac{1}{\widehat{Q}(\mathcal{F}_{n,\varepsilon}^c)} \left\{ \int_{\mathcal{F}_{n,\varepsilon}^c} \sum_{s^n} q_{S^n}(s^n) \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} \widehat{Q}(d\theta) - D(\widehat{Q}_{\mathcal{F}_{n,\varepsilon}^c} \| P_{\theta, \mathcal{F}_{n,\varepsilon}^c}) \right\} \leq cn\varepsilon_n^2/2,$$

which implies

$$0 \leq \int_{\mathcal{F}_{n,\varepsilon}^c} - \sum_{s^n} q_{S^n}(s^n) \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} \widehat{Q}(d\theta) + \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c) D(\widehat{Q}_{\mathcal{F}_{n,\varepsilon}^c} \| P_{\theta, \mathcal{F}_{n,\varepsilon}^c}) \quad (\text{E.2})$$

$$+ (cn\varepsilon_n^2/2 + \log 2) \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c). \quad (\text{E.3})$$

Finally, by combining equations (E.1) and (E.2), and using the identity

$$\begin{aligned} D(\widehat{Q} \| P_\theta) &= \int \widehat{q}(\theta) \log \frac{\widehat{q}(\theta)}{\pi(\theta)} d\theta \\ &= \widehat{Q}(\mathcal{F}_{n,\varepsilon}) \int_{\mathcal{F}_{n,\varepsilon}} \widehat{q}_{\mathcal{F}_{n,\varepsilon}}(\theta) \log \frac{\widehat{q}_{\mathcal{F}_{n,\varepsilon}}(\theta)}{\pi_{\mathcal{F}_{n,\varepsilon}}(\theta)} d\theta + \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c) \int_{\mathcal{F}_{n,\varepsilon}^c} \widehat{q}_{\mathcal{F}_{n,\varepsilon}^c}(\theta) \log \frac{\widehat{q}_{\mathcal{F}_{n,\varepsilon}^c}(\theta)}{\pi_{\mathcal{F}_{n,\varepsilon}^c}(\theta)} d\theta \\ &\quad + \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c) \log \frac{\widehat{Q}(\mathcal{F}_{n,\varepsilon}^c)}{P_\theta(\mathcal{F}_{n,\varepsilon}^c)} + (1 - \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c)) \log \frac{1 - \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c)}{1 - P_\theta(\mathcal{F}_{n,\varepsilon}^c)}, \end{aligned}$$

we have that with \mathbb{P}_{θ^*} probability at least $(1 - 2e^{-cn\varepsilon_n^2/2})$,

$$\begin{aligned} &cn \int_{\theta \in \mathcal{F}_{n,\varepsilon}, r(\theta, \theta^*) \geq \varepsilon^2} r(\theta, \theta^*) \widehat{Q}(d\theta) \\ &\quad + \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c) \log \frac{\widehat{Q}(\mathcal{F}_{n,\varepsilon}^c)}{P_\theta(\mathcal{F}_{n,\varepsilon}^c)} + (1 - \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c)) \log \frac{1 - \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c)}{1 - P_\theta(\mathcal{F}_{n,\varepsilon}^c)} \quad (\text{E.4}) \\ &\leq \int - \sum_{s^n} q_{S^n}(s^n) \log \frac{p(Y^n | \mu, s^n) \pi_{s^n}}{p(Y^n | \theta^*) q_{S^n}(s^n)} \widehat{Q}(d\theta) + D(\widehat{Q} \| P_\theta) + cn\varepsilon_n^2/2 + \log 2 \\ &= \Psi(\widehat{q}_\theta, q_{S^n}) + cn\varepsilon_n^2/2 + \log 2. \end{aligned}$$

As a consequence, the first claimed bound follows by taking $q_{S^n} = \widehat{q}_{S^n}$ and the definition of \widehat{q}_θ and \widehat{q}_{S^n} that minimizes $\Psi(q_\theta, q_{S^n})$ over the variational family.

E.4 Proof of Theorem 3.6

Similar to the proof of Theorem 3.4, under Assumption P, there exists a event \mathcal{A}_n satisfying $\mathbb{P}_{\theta^*}(\mathcal{A}_n) \geq 1 - \frac{2}{(D-1)^2 n \varepsilon_n^2}$ and measures $(Q_{\theta^*}^*, q_{S^n}^*)$, such that under this event,

$$\Psi(Q_{\theta^*}^*, q_{S^n}^*) \leq 2D n \varepsilon_n^2.$$

For any fixed $\varepsilon \geq \varepsilon_n$, denote the event under which the result of Theorem 3.5 holds as \mathcal{B}_ε . Consequently, $\mathbb{P}_{\theta^*}(\mathcal{B}_\varepsilon) \geq 1 - 2e^{-cn\varepsilon_n^2}$, and under event $\mathcal{A}_n \cap \mathcal{B}_\varepsilon$, we have

$$\begin{aligned} & \left\{ \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c) \log \frac{\widehat{Q}(\mathcal{F}_{n,\varepsilon}^c)}{P_\theta(\mathcal{F}_{n,\varepsilon}^c)} + (1 - \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c)) \log \frac{1 - \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c)}{1 - P_\theta(\mathcal{F}_{n,\varepsilon}^c)} \right\} \\ & + cn \int_{\theta \in \mathcal{F}_{n,\varepsilon}, r(\theta, \theta^*) \geq \varepsilon^2} r(\theta, \theta^*) \widehat{Q}(d\theta) \leq C n \varepsilon_n^2, \end{aligned}$$

where $C > 0$ is some constant independent of n and ε . Since both terms on the l.h.s. of the above is nonnegative, we obtain that

$$\begin{aligned} \widehat{Q}(\theta \in \mathcal{F}_{n,\varepsilon}, r(\theta, \theta^*) \geq \varepsilon^2) & \leq \varepsilon^{-2} \int_{\theta \in \mathcal{F}_{n,\varepsilon}, r(\theta, \theta^*) \geq \varepsilon^2} r(\theta, \theta^*) \widehat{Q}(d\theta) \leq C' \frac{\varepsilon_n^2}{\varepsilon^2}, \\ \text{and } \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c) & \leq C'' \frac{\varepsilon_n^2}{\varepsilon^2}, \quad \text{for some constants } C', C'' > 0. \end{aligned}$$

Here, the second inequality holds by using $P_\theta(\mathcal{F}_{n,\varepsilon}^c) \leq e^{-cn\varepsilon^2}$ (Assumption T), and the inequality $x \log x + (1-x) \log(1-x) \geq -\log 2$ ($x \in (0, 1)$).

Applying above results to $\varepsilon = k \varepsilon_n$ with $k = 1, 2, \dots, e^{cn\varepsilon_n^2/4}$, and using a union bound, we obtain that the following holds with probability at least $1 - \frac{2}{(D-1)^2 n \varepsilon_n^2} - 2e^{-cn\varepsilon_n^2/4} \geq 1 - \frac{3}{(D-1)^2 n \varepsilon_n^2}$,

$$\widehat{Q}(\theta \in \mathcal{F}_{n,\varepsilon}, r(\theta, \theta^*) \geq \varepsilon^2) \leq C' \frac{\varepsilon_n^2}{\varepsilon^2}, \quad \text{and } \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c) \leq C'' \frac{\varepsilon_n^2}{\varepsilon^2},$$

for all $\varepsilon = k \varepsilon_n$ with $k = 1, 2, \dots, e^{cn\varepsilon_n^2/4}$. Note that the preceding display implies

$$\widehat{Q}(r(\theta, \theta^*) \geq \varepsilon^2) \leq \widehat{Q}(\theta \in \mathcal{F}_{n,\varepsilon}, r(\theta, \theta^*) \geq \varepsilon^2) + \widehat{Q}(\mathcal{F}_{n,\varepsilon}^c) \leq (C' + C'') \frac{\varepsilon_n^2}{\varepsilon^2}.$$

For general $\varepsilon \in [\varepsilon_n, e^{cn\varepsilon_n^2/4} \varepsilon_n)$, we can always find an integer k^* such that $k^* \varepsilon_n \leq \varepsilon <$

$(k^* + 1)\varepsilon_n$. Using the monotonicity of $\widehat{Q}(r(\theta, \theta^*) \geq \varepsilon^2)$ in ε , we can obtain

$$\widehat{Q}(r(\theta, \theta^*) \geq \varepsilon^2) \leq \widehat{Q}(r(\theta, \theta^*) \geq (k^*\varepsilon_n)^2) \leq (C' + C'') \frac{1}{(k^*)^2} \leq C_1 \frac{\varepsilon_n^2}{\varepsilon^2}.$$

The second claimed bound follows by

$$\begin{aligned} \int_{\theta: r(\theta, \theta^*) \leq R^2} r(\theta, \theta^*) \widehat{Q}(d\theta) &= \int_0^{R^2} \widehat{Q}(r(\theta, \theta^*) \geq t) dt \\ &\leq \varepsilon_n^2 + 2 \int_{\varepsilon_n}^R \varepsilon \widehat{Q}(r(\theta, \theta^*) \geq \varepsilon^2) d\varepsilon \\ &\leq \varepsilon_n^2 \left(1 + 2C_1 \int_{\varepsilon_n}^R \frac{1}{t} dt \right) \leq C_2 \varepsilon_n^2 (1 + \log(R/\varepsilon_n)). \end{aligned}$$

E.5 Proof of Theorem 3.7

According to the proof of Theorem 3.1, we have that for any $\varepsilon \in (0, 1)$, and any (q_θ, q_{S^n}) in the variational family,

$$\int \frac{1}{n} D_\alpha^{(n)}(\theta, \theta^*) q_\theta(\theta) d\theta \leq \frac{\alpha}{n(1-\alpha)} \Psi_\alpha(q_\theta, q_{S^n}) + \frac{1}{n(1-\alpha)} \log(1/\zeta).$$

Since $\bar{\Psi}_\alpha$ is an upper bound of Ψ_α , the above implies

$$\int \frac{1}{n} D_\alpha^{(n)}(\theta, \theta^*) q_\theta(\theta) d\theta \leq \frac{\alpha}{n(1-\alpha)} \bar{\Psi}_\alpha(q_\theta, q_{S^n}) + \frac{1}{n(1-\alpha)} \log(1/\zeta).$$

Now choosing (q_θ, q_{S^n}) as $(\bar{q}_\theta, \bar{q}_{S^n})$ in the above, and the claimed bound follows since

$$(\bar{q}_\theta, \bar{q}_{S^n}) = \operatorname{argmin}_{q_\theta, q_{S^n}} \bar{\Psi}_\alpha(q_\theta, q_{S^n}).$$

E.6 Proof of Corollary 4.1

For the linear model, we have $\mathcal{B}_n(\theta^*, \varepsilon) \supset \{\theta = (\beta, \sigma) : (2n\sigma^2)^{-1} \|X(\beta - \beta^*)\|^2 + ((\sigma^*)^2/\sigma^2 - 1 - \log[(\sigma^*)^2/\sigma^2])/2 \leq 2\varepsilon^2\}$. Therefore, we may take the neighborhood $\mathcal{N}_n(\theta^*, \varepsilon)$ as the product set $\{\beta : n^{-1}(\sigma^*)^{-2} \|X(\beta - \beta^*)\|^2 \leq c_1 \varepsilon^2\} \times \{\sigma : |\sigma - \sigma^*| \leq c_2 \varepsilon\}$, for some sufficiently small constants (c_1, c_2) such that $\mathcal{N}_n(\theta^*, \varepsilon) \subset \mathcal{B}_n(\theta^*, \varepsilon)$. In addition, due to the product form of $\mathcal{N}_n(\theta^*, \varepsilon)$, probability density function q_θ^* defined as $q_\theta^* \propto I_{\mathcal{N}_n(\theta^*, \varepsilon)}$ belongs to the mean field approximation family Γ . Consequently, we may apply Theorem 3.3 to obtain (noting

that the volume of the neighborhood $\{\beta : n^{-1}(\sigma^*)^{-2} \|X(\beta - \beta^*)\|^2 \leq c_1 \varepsilon^2\}$ is $\mathcal{O}[(d/\varepsilon)^{-d}]$

$$\int \left\{ \frac{1}{n} D_{\theta^*, \alpha}^{(n)}(\theta, \theta^*) \right\} \widehat{q}_\theta(\theta) d\theta \lesssim \frac{\alpha}{1-\alpha} \varepsilon^2 + \frac{d}{n(1-\alpha)} \log \frac{d}{\varepsilon}.$$

Setting $\varepsilon = \sqrt{d/n}$ in the preceding inequality and using the fact that $\max\{1, (1-\alpha)^{-1} \alpha\} h^2(p \| q) \leq D_\alpha(p \| q)$ for any density p and q yields the claimed bound.

E.7 Proof of Corollary 4.2

Similar to the proof of Corollary 4.1, we choose $\mathcal{N}_n(\theta^*, \varepsilon)$ as the product set $\{\beta : \beta_{(z^*)^c} = 0, n^{-1}(\sigma^*)^{-2} \|X(\beta - \beta^*)\|^2 \leq c_1 \varepsilon^2\} \times \{\sigma : |\sigma - \sigma^*| \leq c_2 \varepsilon\}$, and define the joint measure

$$q_\theta^* \otimes q_{z^*}^* \propto I_{\mathcal{N}_n(\theta^*, \varepsilon)} \otimes \delta_{z^*},$$

which belongs to the mean field approximation family Γ . Now, by applying Theorem 3.3 with parameter $\theta = (\beta, \sigma, z)$, we obtain (by replacing d with s in the proof of Corollary 4.1 for the β part) that

$$\int \left\{ \frac{1}{n} D_{\theta^*, \alpha}^{(n)}(\theta, \theta^*) \right\} \widehat{q}_\theta(\theta) d\theta \lesssim \frac{\alpha}{1-\alpha} \varepsilon^2 + \frac{s}{n(1-\alpha)} \log \frac{s}{\varepsilon} + \frac{1}{n(1-\alpha)} s \log d,$$

where the last term is due to $-\log p_z(z^*) \asymp s \log d$. Setting $\varepsilon = \sqrt{s/n}$ leads to the claimed bound.

E.8 Proof of Corollary 4.3

The first claimed bound is a direct consequence by applying Theorem 3.7 (with no latent variables) to the new ELBO $\bar{L}(q)$. The second bound can be obtained by applying the first claimed inequality (4.3) (taking $w_1 = 1, w_2 = \dots = w_J = 0$ to reduce the bound to that of the single Gaussian variational approximation) and the arguments in Corollary D.1 (for a single Gaussian variational approximation).

E.9 Proof of Corollary 4.4

It is easy to verify that under Assumption R, there exists some constant C_1 depending only on δ_0 such that $\mathcal{B}_n(\pi^*, \sqrt{K} \varepsilon) \supset \{\pi : \max_k |\pi_k - \pi_k^*| \leq C_1 \varepsilon\}$ (by using the inequality $D(p \| q) \geq 2 h^2(p \| q)$). In addition, for Gaussian mixture model, it is easy to verify that the KL neighborhood $\mathcal{B}_n(\mu^*, \varepsilon)$ defined before Theorem 3.4 contains the set $\{\mu : \max_k \|\mu_k -$

$\mu_k^* \|\leq 2\varepsilon\}$. As a consequence, a direct application of Theorem 3.4 with $\varepsilon_\pi = \sqrt{K}\varepsilon$ and $\varepsilon_\mu = \varepsilon$ yields (using the prior thickness assumption and the fact that the volumes of $\{\pi : \max_k |\pi_k - \pi_k^*| \leq C_1 \varepsilon\}$ and $\{\mu : \max \|\mu_k - \mu_k^*\| \leq C_2 \varepsilon\}$ are at least $\mathcal{O}(\varepsilon^{-K})$ and $\mathcal{O}((\sqrt{d}/\varepsilon)^{dK})$ respectively) that with probability tending to one as $n \rightarrow \infty$,

$$\int \left\{ D_\alpha [p(\cdot | \theta) || p(\cdot | \theta^*)] \right\} \hat{q}_\theta(\theta) d\theta \lesssim \frac{\alpha}{1-\alpha} K \varepsilon^2 + \frac{dK}{n(1-\alpha)} \log \frac{d}{\varepsilon}.$$

Choosing $\varepsilon = \sqrt{d/n}$ in the above display yields the claimed bound.

E.10 Proof of Corollary 4.5

Under the notation of Theorem 3.1 and Corollary 3.2, for each $n = 1, \dots, N$, the latent variable $S_n = \{z_{dn} : d = 1, \dots, D\}$, we will use an extended version of Corollary 3.2 from 1 latent variable per observation to D independent latent variable per observation. In fact, similar arguments as the proofs of Theorem 3.1 and Corollary 3.2 yield that for the ensemble of KL neighborhoods $\{\mathcal{B}_N(\gamma_d^*; \varepsilon_{\gamma_d}) : d = 1, \dots, D\}$ of $\{\gamma_d^* : d = 1, \dots, D\}$ where $\mathcal{B}_N(\gamma_d^*; \varepsilon_{\gamma_d}) := \{D(\gamma_d^* || \gamma_d) \leq \varepsilon_{\gamma_d}^2, V(\gamma_d^* || \gamma_d) \leq \varepsilon_{\gamma_d}^2\}$, for $d = 1, \dots, D$, it holds with probability tending to one as $N \rightarrow \infty$ that

$$\begin{aligned} & \int \left\{ D_\alpha [p(\cdot | \theta), p(\cdot | \theta^*)] \right\} \hat{q}_\theta(\theta) d\theta \\ & \leq \frac{D\alpha}{1-\alpha} \left(\sum_{d=1}^D \varepsilon_{\gamma_d}^2 + \varepsilon_\mu^2 \right) + \left\{ -\frac{1}{N(1-\alpha)} \sum_{d=1}^D \log P_{\gamma_d} [\mathcal{B}_N(\gamma_d^*, \varepsilon_{\gamma_d})] \right\} \\ & \quad + \left\{ -\frac{1}{N(1-\alpha)} \log P_\mu [B_N(\mu^*, \varepsilon_\mu)] \right\}, \end{aligned}$$

where $B_N(\mu^*, \varepsilon_\mu) = \{\max_{S_n} D[p(\cdot | \mu^*, S_n) || p(\cdot | \mu, S_n)] \leq \varepsilon_\mu^2, \max_{S_n} V[p(\cdot | \mu^*, S_n) || p(\cdot | \mu, S_n)] \leq \varepsilon_\mu^2\}$. Recall that each observation Y_n composed of i.i.d. observations $\{w_{dn} : d = 1, \dots, D\}$, where the conditional distribution of w_{dn} given latent variable $\{z_{dn} = k\}$ only depends on β_k for $d = 1, \dots, D$ and $k = 1, \dots, K$. Therefore, when applied to LDA, the preceding display can be further simplified into

$$\begin{aligned} & \int \left\{ D_\alpha [p(\cdot | \theta), p(\cdot | \theta^*)] \right\} \hat{q}_\theta(\theta) d\theta \\ & \leq \frac{D\alpha}{1-\alpha} \left(\sum_{d=1}^D \varepsilon_{\gamma_d}^2 + \sum_{k=1}^K \varepsilon_{\beta_k}^2 \right) + \left\{ -\frac{1}{N(1-\alpha)} \sum_{d=1}^D \log P_{\gamma_d} [\mathcal{B}_N(\gamma_d^*, \varepsilon_{\gamma_d})] \right\} \quad (\text{E.5}) \\ & \quad + \left\{ -\frac{1}{N(1-\alpha)} \sum_{k=1}^K \log P_\mu [B_N(\beta_k^*, \varepsilon_{\beta_k})] \right\}, \end{aligned}$$

where $B_N(\beta_k^*, \varepsilon_{\beta_k}) = \{ \max_k D[p(\cdot | \beta_k^*, k) || p(\cdot | \beta_k, k)] \leq \varepsilon_{\beta_k}^2, \max_{S_n} V [p(\cdot | \beta_k^*, k) || p(\cdot | \beta_k, k)] \leq \varepsilon_{\beta_k}^2 \}$.

Return to the proof of the theorem. Let S_k^β denote the index set corresponding to the non-zero components of β_k for $k = 1, \dots, K$, and S_d^γ the index set corresponding to the non-zero components of γ_d for $d = 1, \dots, D$. Under Assumption S, it is easy to verify that for some sufficiently small constants $c_1, c_2 > 0$, it holds for all $d = 1, \dots, D$ that $\mathcal{B}_N(\gamma_d^*, \varepsilon_{\gamma_d}) \supset \{ \|(\gamma_d)_{(S_d^\gamma)^c}\|_1 \leq c_1 \varepsilon_{\gamma_d}, \|(\gamma_d)_{S_d^\gamma} - (\gamma_d^*)_{S_d^\gamma}\|_\infty \leq c_1 \varepsilon_{\gamma_d} \}$, and for all $k = 1, \dots, K$ that $B_N(\beta_k^*, \varepsilon_{\beta_k}) \supset \{ \|(\beta_k)_{(S_k^\beta)^c}\|_1 \leq c_2 \varepsilon_{\beta_k}, \|(\beta_k)_{S_k^\beta} - (\beta_k^*)_{S_k^\beta}\|_\infty \leq c_2 \varepsilon_{\beta_k} \}$. Applying Theorem 2.1 in [44], we obtain the following prior concentration bounds for high-dimensional Dirichlet priors

$$\begin{aligned} P_{\gamma_d} \{ & \|(\gamma_d)_{(S_d^\gamma)^c}\|_1 \leq c_1 \varepsilon_{\gamma_d}, \|(\gamma_d)_{S_d^\gamma} - (\gamma_d^*)_{S_d^\gamma}\|_\infty \leq c_1 \varepsilon_{\gamma_d} \} \\ & \gtrsim \exp \left\{ -C e_d \log \frac{K}{\varepsilon_{\gamma_d}} \right\}, \quad d = 1, \dots, D; \\ P_{\beta_k} \{ & \|(\beta_k)_{(S_k^\beta)^c}\|_1 \leq c_2 \varepsilon_{\beta_k}, \|(\beta_k)_{S_k^\beta} - (\beta_k^*)_{S_k^\beta}\|_\infty \leq c_2 \varepsilon_{\beta_k} \} \\ & \gtrsim \exp \left\{ -C d_k \log \frac{V}{\varepsilon_{\beta_k}} \right\}, \quad k = 1, \dots, K. \end{aligned}$$

Putting pieces together, we obtain

$$\begin{aligned} & \int \left\{ D_\alpha [p(\cdot | \theta), p(\cdot | \theta^*)] \right\} \hat{q}_\theta(\theta) d\theta \\ & \lesssim \frac{\alpha}{1-\alpha} \left(\sum_{d=1}^D \varepsilon_{\gamma_d}^2 + \sum_{k=1}^K \varepsilon_{\beta_k}^2 \right) + \frac{1}{N(1-\alpha)} \sum_{d=1}^D e_d \log \frac{K}{\varepsilon_{\gamma_d}} + \frac{1}{N(1-\alpha)} \sum_{k=1}^K d_k \log \frac{V}{\varepsilon_{\beta_k}}, \end{aligned}$$

which is the desired result.

F Extension of examples to $\alpha = 1$

In this section, we briefly discuss the verification of Assumption T (choice of loss function and constructions of test function $\phi_{n,\varepsilon}$ and sieve $\mathcal{F}_{n,\varepsilon}$) in the examples of the paper, which implying the variational risk bound through applying Theorem 3.6.

Mean field approximation to low-dimensional Bayesian linear regression: To simplify the presentation, we assume the priors on β and σ satisfy $P_\beta(\|\beta\| \geq R) \leq CR^{-cd}$ and $P_\sigma(\sigma \in [a, b]) = 1$, where $[a, b]$ contains the truth σ^* . In addition, the design matrix X satisfies that the minimal eigenvalue of $n^{-1} X^T X$ is bounded away from zero. Recall that

in this example, $\theta = (\beta, \sigma)$. Under these two assumptions, it can be proved that Assumption T holds with $\phi_{n,\varepsilon}$ being the likelihood ratio test $\phi_{n,\varepsilon} = \mathbb{I}(\ell(\beta, \beta^*) \geq C' n \varepsilon^2)$, sieve $\mathcal{F}_{n,\varepsilon} = \{\|\beta\| \leq \exp(C'' d^{-1} n \varepsilon^2)\} \times [a, b]$, and loss function $r(\beta, \beta^*) = \|\beta - \beta^*\|^2$, for all $\varepsilon^2 \geq \varepsilon_n^2 = d \log n/n$, and sufficiently large constant $C', C'' > 0$.

Mean field approximation to high-dimensional Bayesian linear regression with sparse priors: Similar to the previous example, we make the assumption that given z , the conditional prior of β satisfies $P_{\beta|z}(\|\beta\| \geq R | z) \leq CR^{-c|z|}$, where $|z|$ is the size of binary vector z , and the prior on σ satisfies $P_\sigma(\sigma \in [a, b]) = 1$. In addition, we make the sparse eigenvalue assumption that there exists some sufficiently large $C > 0$, such that for any Cs sparse vector u , $n^{-1}\|Xu\|^2/\|u\|^2 \geq \mu > 0$. Recall that in this example, $\theta = (z, \beta, \sigma)$. Under these assumptions, it can be verified that Assumption T holds with $\phi_{n,\varepsilon}$ being the likelihood ratio test $\phi_{n,\varepsilon} = \mathbb{I}(\ell(\beta, \beta^*) \geq C' n \varepsilon^2)$, sieve $\mathcal{F}_{n,\varepsilon} = \bigcup_{z: |z| \leq C''s} \left[\{z\} \times \{\beta_{z^c} = 0, \|\beta - \beta^*\| \leq \exp(C''' s^{-1} n \varepsilon^2)\} \times [a, b] \right]$, and loss function $r(\beta, \beta^*) = \|\beta - \beta^*\|^2$, for all $\varepsilon^2 \geq \varepsilon_n^2 = s \log(nd)/n$, and sufficiently large constant $C', C'', C''' > 0$.

Remaining examples: In all the remaining examples, the parameter space Θ of θ is compact. In this case, we can simply take $\mathcal{F}_{n,\varepsilon} = \Theta$ for all ε (so $P_\theta(\mathcal{F}_{n,\varepsilon}^c) = 0$), and apply a general recipe [18] to construct such tests: (i) construct an $\varepsilon/2$ -net $\mathcal{N} = \{\theta_1, \dots, \theta_N\}$ such that for any θ with $r(\theta, \theta^*) > \varepsilon^2$, there exists $\theta_j \in \mathcal{N}$ with $r(\theta, \theta_j) < \varepsilon^2/2$, (ii) construct a test $\phi_{n,j}$ for $H_0 : \theta = \theta^*$ versus $H_1 : \theta = \theta_j$ with type-I and II error rates as in Assumption T, and (iii) set $\phi_n = \max_{1 \leq j \leq N} \phi_{n,j}$. The type-II error of ϕ_n retains the same upper bound, while the type-I error can be bounded by $N e^{-2n\varepsilon^2}$. Since N can be further bounded by $N(\Theta, \varepsilon^2/2, r)$, the covering number of Θ by r -balls of radius $\varepsilon^2/2$, it suffices to show that $N(\Theta, \varepsilon^2/2, r) \lesssim e^{n\varepsilon^2}$. When Θ is a compact subset of \mathbb{R}^d and $r(\theta, \theta^*) \gtrsim \|\theta - \theta^*\|^2$ (the squared Euclidean metric), then $N(\Theta, \varepsilon^2, r) \lesssim \varepsilon^{-d} \lesssim e^{n\varepsilon^2}$ as long as $\varepsilon \gtrsim \sqrt{\log n/n}$. More generally, if Θ is a space of densities and r the squared Hellinger/ L_1 metric, then construction of the point-by-point tests in (i) from the likelihood ratio test statistics follows from the classical Birgé-Lecam testing theory [6, 28]; see also [18].

To summarize, in these examples with compact parameter space, Assumption T holds with $\mathcal{F}_{n,\varepsilon} = \Theta$, $r(\theta, \theta^*) = h^2(p(\cdot | \theta), p(\cdot | \theta^*))$, the squared Hellinger distance between $p(\cdot | \theta)$ and $p(\cdot | \theta^*)$, and $\phi_{n,\varepsilon}$ the likelihood ratio test function, for all $\varepsilon \gtrsim \sqrt{\log n/n}$. Moreover, the rate ε_n is determined by their respective prior concentration Assumption P.

References

- [1] Amr Ahmed, Mohamed Aly, Joseph Gonzalez, Shравan Narayanamurthy, and Alexander Smola. Scalable inference in latent variable models. In *International conference on Web search and data mining (WSDM)*, volume 51, pages 1257–1264, 2012.
- [2] Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *arXiv preprint arXiv:1706.09293*, 2017.
- [3] Hagai Attias. A variational bayesian framework for graphical models. In *Advances in neural information processing systems*, pages 209–215, 2000.
- [4] Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. *arXiv preprint arXiv:1611.01125*, 2016.
- [5] Peter J Bickel, Ya’acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.
- [6] Lucien Birgé. Approximation dans les espaces métriques et théorie de l’estimation. *Probability Theory and Related Fields*, 65(2):181–237, 1983.
- [7] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.
- [8] P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.
- [9] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
- [10] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted), 2017.
- [11] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [12] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.
- [13] Peter Carbonetto and Matthew Stephens. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108, 2012.

- [14] Adrian Corduneanu and Christopher M Bishop. Variational bayesian model selection for mixture distributions. In *Artificial intelligence and Statistics*, volume 2001, pages 27–34. Waltham, MA: Morgan Kaufmann, 2001.
- [15] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.
- [16] Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.
- [17] S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223, 2007.
- [18] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.
- [19] D Harmon. Overview of the first text retrieval conference (trec-1). *NIST Special Publication*, pages 500–207.
- [20] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- [21] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.
- [22] Xichen Huang, Jin Wang, and Feng Liang. A variational algorithm for bayesian variable selection. *arXiv preprint arXiv:1602.07640*, 2016.
- [23] K Humphreys and DM Titterington. Approximate bayesian inference for simple mixtures. In *Proc. Computational Statistics 2000*, pages 331–336, 2000.
- [24] W. Jiang and M. A. Tanner. Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.*, pages 2207–2231, 2008.
- [25] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.
- [26] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- [27] Harold J Kushner and G George Yin. Stochastic approximation algorithms and applications. 1997.
- [28] Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.
- [29] Yingzhen Li and Richard E. Turner. Rényi divergence variational inference. In *Advances in neural information processing systems*, 2016.
- [30] David JC MacKay. Ensemble learning for hidden markov models.
- [31] JOHN T Ormerod, Chong You, and SAMUEL Muller. A variational bayes approach to variable selection.
- [32] Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational bayes. *AISTATS, To Appear*, 2017.
- [33] Christian P Robert. *Monte carlo methods*. Wiley Online Library, 2004.
- [34] Judith Rousseau. On the frequentist properties of bayesian nonparametric methods. *Annual Review of Statistics and Its Application*, 3:211–231, 2016.
- [35] Naonori Ueda and Zoubin Ghahramani. Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, 15(10):1223–1241, 2002.
- [36] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- [37] T. Van Erven and P. Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.
- [38] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- [39] S. Walker and N. L. Hjort. On Bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821, 2001.
- [40] Bo Wang and DM Titterington. Lack of consistency of mean field and variational bayes approximations for state space models. *Neural Processing Letters*, 20(3):151–170, 2004.
- [41] Bo Wang and DM Titterington. Inadequacy of interval estimates corresponding to variational bayesian approximations. In *AISTATS*, 2005.

- [42] Bo Wang, DM Titterington, et al. Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650, 2006.
- [43] Ted Westling and Tyler H McCormick. Establishing consistency and improving uncertainty estimates of variational inference through m-estimation. *arXiv preprint arXiv:1510.08151*, 2015.
- [44] Yun Yang and David B Dunson. Minimax optimal bayesian aggregation. *arXiv preprint arXiv:1403.1345*, 2014.
- [45] Chong You, John T Ormerod, and Samuel Müller. On variational bayes estimation and variational information criteria for linear regression models. *Australian & New Zealand Journal of Statistics*, 56(1):73–87, 2014.
- [46] Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *arXiv preprint arXiv:1712.02519*, 2017.
- [47] O Zobay et al. Variational bayesian inference with gaussian-mixture approximations. *Electronic Journal of Statistics*, 8(1):355–389, 2014.