The Annals of Statistics
2020, Vol. 48, No. 2, 886–905
https://doi.org/10.1214/19-AOS1827

α-VARIATIONAL INFERENCE WITH STATISTICAL GUARANTEES

BY YUN YANG[1], DEBDEEP PATI[2,*] AND ANIRBAN BHATTACHARYA[2,**]

[1]*Department of Statistics, University of Illinois at Urbana–Champaign, yy84@illinois.edu*
[2]*Department of Statistics, Texas A&M University, *debdeep@stat.tamu.edu; **anirbanb@stat.tamu.edu*

We provide statistical guarantees for a family of variational approximations to Bayesian posterior distributions, called α-VB, which has close connections with variational approximations of *tempered posteriors* in the literature. 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, Gaussian mixture models and latent Dirichlet allocation.

1. Introduction and preliminaries. Variational inference [22, 34] 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, [14, 18]) 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 [26], graphical models [4, 34], mixture models [13, 20, 31] 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 [23, 24] and distributed optimization [1]. However, unlike MCMC that is guaranteed to produce (almost) exact samples from the target density for ergodic chains [29], 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 until around 2017; we discuss more recent work paralleling ours in a subsequent paragraph. Almost all

Received January 2018; revised January 2019.
MSC2010 subject classifications. Primary 62G07, 62G20; secondary 60K35.
Key words and phrases. Bayes risk, evidence lower bound, latent variable models, Rényi divergence, variational inference.

of these earlier 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 analyzes require certain structural assumptions on the priors such as conjugacy, and is not applicable to broader classes of priors.

This article studies first-order statistical optimality properties of a class of variational approximations, called α-VB, in a unified framework. The class of approximations introduces a fixed temperature parameter α inside the usual VB objective function which controls the relative tradeoff between model-fit and prior regularization. The usual VB approximation is retained as a special case corresponding to $\alpha = 1$. The general α-VB procedure inherits all the computational tractability and scalability from the $\alpha = 1$ case, and implementation-wise only requires minor modifications to existing variational algorithms such as the coordinate ascent variational inference (CAVI) algorithm [7, 10]; see the Supplementary Material [40] for specific examples. In the absence of latent variables, the α-VB approximation is identical to the variational approximation to tempered posteriors considered in [2, 3] and references therein. The α-VB objective function considered here provides a natural extension to their tempered variational approximation to models involving latent variables.

During the last year, there has been a surge of interest in the theoretical understanding of variational Bayes procedures. The reviewers directed our attention to the very interesting aforementioned preprints [2, 3] on the theoretical properties of the variational approximation to tempered posteriors. Through a number of examples, [2] demonstrate the applicability of their general theory dictating optimal first-order frequentist risk behavior of the tempered variational approximation. While this article was under review, we also came across the preprint [42] obtaining novel contraction results for the usual variational approximation (i.e., $\alpha = 1$) in models without latent variables. The main contribution of our article which separates itself from these recent works is its ability to handle latent variable models. On the other hand, [2] provide risk bounds under model misspecification and also combine their statistical bounds with algorithmic convergence in a particular case, none of which is considered here. Zhang and Gao [42] obtain an interesting connection between variational approximations and empirical Bayes procedures in the Gaussian sequence model, and provide a very interesting example where the variational approximation contracts faster than the true posterior.

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 [30] for a detailed review, and the present paper takes a step toward 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, Gaussian mixture models and latent Dirichlet allocation.

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. 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 [16] on the likelihood function detailed in Section 3.2.

It is a well-known fact [36, 38] 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 [32] 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 mismatch 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 $h(P \parallel Q) = (\int(\sqrt{dP/d\lambda} - \sqrt{dQ/d\lambda})^2 d\lambda)^{1/2}$ and $D(P \parallel Q) = \int \log(dP/dQ)\, dP$ denote the Hellinger distance and Kullback–Leibler divergence, respectively, between two probability measures P and Q that have Radon–Nikodym derivatives $dP/d\lambda$ and $dQ/d\lambda$ relative to a common dominating measure λ. Note that the value of the Hellinger distance does not depend on the choice of λ. We define an additional discrepancy measure $V(P \parallel Q) = \int \log^2(dP/dQ)\, dP$, 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

$$D_\alpha(P \parallel Q) = \frac{1}{\alpha - 1} \log \int \left(\frac{dP}{dQ}\right)^\alpha dQ \tag{1.1}$$

denote the Rényi divergence of order α. Jensen's inequality implies that $D_\alpha(P \parallel Q) \geq 0$ 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 $D_{1/2}(P \parallel Q) = -2\log\{1 - (1/2)h^2(P \parallel Q)\} \geq h^2(P \parallel Q)$ using the inequality $\log(1 + t) < t$ for $t > -1$. More details and properties of the α-divergence can be found in [33]. We will also interchangeably use notation $h(p \parallel q)$, $D(p \parallel q)$, $V(p \parallel q)$ and $D_\alpha(p \parallel q)$ to denote these discrepancy measures when density functions $p = dP/d\lambda$ and $q = dQ/d\lambda$ are clear from the context.

1.2. *Review of variational inference.* Suppose we have observations $Y^n = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$ 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 a common dominating 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) $p_\theta^{(n)}$. For example, when Y^n is discrete (continuous) and the common dominating measure is the counting (Lebesgue) measure, $p_\theta^{(n)}$ corresponds to the probability mass (density) function of Y^n given θ. Assume additionally that the likelihood $p(Y^n|\theta)$ can be represented as

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

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 probabilistic 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 Supplementary Material [40].

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

$$p\big(W^n|Y^n\big) = p\big(\theta, S^n|Y^n\big) \propto p\big(Y^n|\theta, S^n\big)p\big(S^n|\theta\big)p_\theta(\theta). \tag{1.2}$$

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 prespecified family of density functions over $\mathscr{W}^n$ 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

$$\begin{aligned}
\widehat{q}_{W^n} &:= \operatorname*{argmin}_{q_{W^n}\in\Gamma} D\big[q_{W^n}(\cdot) \parallel p\big(\cdot|Y^n\big)\big] \\
&= \operatorname*{argmin}_{q_{W^n}\in\Gamma}\bigg\{-\int_{\mathscr{W}^n} q_{W^n}\big(w^n\big)\log\frac{p(w^n|Y^n)}{q_{W^n}(w^n)}\,dw^n\bigg\} \\
&= \operatorname*{argmin}_{q_{W^n}\in\Gamma}\Bigg\{\underbrace{-\int_{\mathscr{W}^n} q_{W^n}\big(w^n\big)\log\frac{p(Y^n|w^n)p_{W^n}(w^n)}{q_{W^n}(w^n)}\,dw^n}_{L(q_{W^n})}\Bigg\},
\end{aligned} \tag{1.3}$$

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 $\log p(Y^n)$,

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

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 [22]—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 $\widehat{q}_{W^n}$ exactly recovers $p(W^n|Y^n)$. In general, the variational family Γ is chosen to balance between computational tractability and 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 [32]), 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 [43], 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 $W^n = (W_1, W_2, \ldots, W_m)$ 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

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

where each variational factor q_{W_j} is a density function over $\mathscr{W}_j$ for $j = 1, \ldots, m$. 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 (e.g., the mean-field family if θ is multidimensional), which will be denoted by Γ_θ in the sequel.

The rest of the paper is organized as follows. In Section 2, we introduce the α-VB objective function and relate it to usual VB. Section 3 presents our general theoretical results concerning finite sample risk bounds for the α-VB solution. In Section 4, we apply the theory to concrete examples. We conclude with a discussion in Section 5. All proofs and some additional discussions are provided in the Supplementary Material [40], which 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. We decompose $\theta = (\mu, \pi)$, with the assumption that the joint distribution of (Y^n, S^n) conditional on θ is

expressed as $p(Y^n, S^n = s^n|\theta) = p(Y^n|S^n = s^n, \mu) \times p(S^n = s^n|\pi)$. 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 π_{s^n} characterizes the distribution $P(S^n|\pi)$ of the latent variables. Denote $p(S^n = s^n|\pi)$ by π_{s^n}. When S^n is discrete taking values in a countable set $\mathcal{C}^n$, $\pi = \{\pi_{s^n} : s^n \in \mathcal{C}^n\}$. 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

$$\begin{aligned} &D\big[q_{W^n}(\cdot) \parallel p(\cdot|Y^n)\big] \\ &\quad = \log p(Y^n) + D(q_\theta \parallel p_\theta) \\ &\qquad - \int_\Theta \underbrace{\Big[\sum_{s^n} q_{S^n}(s^n) \log \frac{p(Y^n|\mu, s^n)\pi_{s^n}}{q_{S^n}(s^n)}\Big]}_{\widehat{\ell}_n(\theta)} q_\theta(d\theta). \end{aligned} \tag{2.2}$$

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

$$\begin{aligned} \log p(Y^n|\theta) &= \log\Big[\sum_{s^n} q_{S^n}(s^n) \frac{p(Y^n|\mu, s^n)\pi_{s^n}}{q_{S^n}(s^n)}\Big] \\ &\geq \sum_{s^n} q_{S^n}(s^n) \log \frac{p(Y^n|\mu, s^n)\pi_{s^n}}{q_{S^n}(s^n)}. \end{aligned}$$

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,

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

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

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

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),

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

which readily follows since

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

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

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

and observe that Ψ_n differs from the KL divergence $D[q_{W^n}(\cdot) \parallel p(\cdot|Y^n)]$ in (2.3) only by $\ell_n(\theta^*)$ which does not involve the variational densities. Hence, minimizing $D[q_{W^n}(\cdot) \parallel p(\cdot|Y^n)]$ 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 tradeoff between model-fit and fidelity to the prior underlying a variational approximation, which is akin to the classical bias-variance tradeoff 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 overfitting 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,

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

and define the α-VB solution as

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

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 $\widehat{q}_\theta$ 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 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 [5] of (θ, S^n),

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

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 [21] 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*

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

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 proof of Proposition 2.1 is straightforward, and hence omitted. 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 decoupled. 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 $\mathbb{P}^{(n)}_{\theta^*}$) to the variational risk

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

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 $\widehat{\theta}_{\mathrm{VB},\alpha} = \int \theta\widehat{q}_{\theta,\alpha}(d\theta)$ using Jensen's inequality:

$$r\big(\widehat{\theta}_{\mathrm{VB},\alpha}, \theta^*\big) \le \int r(\theta, \theta^*)\widehat{q}_{\theta,\alpha}(d\theta).$$

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

3.1. *Risk bounds for the $\alpha < 1$ case.* We use the shorthand

$$\frac{1}{n}D^{(n)}_\alpha(\theta, \theta^*) := \frac{1}{n}D_\alpha\big[p^{(n)}_\theta \parallel p^{(n)}_{\theta^*}\big]$$

to denote the averaged α-divergence between $\mathbb{P}^{(n)}_\theta$ and $\mathbb{P}^{(n)}_{\theta^*}$. We adopt the theoretical framework of [5] 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 $D_\alpha[p_\theta \parallel p_{\theta^*}]$, which is stronger than the squared Hellinger distance $h^2[p_\theta \parallel p_{\theta^*}]$ 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 $\mathbb{P}^n_{\theta^*}$ 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,*

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

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 as minimizing an upper bound on 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 [35] indicates that this mean-field approximation for the latent variables may not generally work for nonindependent observations with nonindependent latent variables. For this reason, we assume that the observation latent variable pair (S_i, Y_i) are mutually independent across $i = 1, 2, \ldots, n$. 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 $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$ 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, for example, $\pi_{S^n} = \prod_{i=1}^n \pi_{s_i}$, $p(Y^n, S^n|\mu) = \prod_{i=1}^n \pi_{s_i} p(Y_i|\mu, S_i)$, and $p(Y^n|\theta) = \prod_{i=1}^n p(Y_i|\theta)$. Moreover, under these assumptions, $n^{-1}D_\alpha^{(n)}(\theta,\theta^*) = D_\alpha[p(\cdot|\theta) \parallel p(\cdot|\theta^*)]$.

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 $\widetilde{q}_{S^n}$ which controls the Jensen gap Δ_J, and then make a choice of q_θ which controls $\Psi_{n,\alpha}(q_\theta, \widetilde{q}_{S^n})$. 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, that is,

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

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 $\widetilde{q}_{S_i} \propto \pi^*_{s_i} p(Y_i|\mu^*, S_i)$. This leads us to the following corollary.

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

$$\begin{aligned}
&\int \{D_\alpha[p(\cdot|\theta) \parallel p(\cdot|\theta^*)]\}\widehat{q}_{\theta,\alpha}(\theta)\,d\theta \\
&\quad\le \frac{\alpha}{n(1-\alpha)}\Psi_{n,\alpha}(q_\theta, \widetilde{q}_{S^n}) + \frac{1}{n(1-\alpha)}\log(1/\zeta), \\
&\quad= \frac{\alpha}{n(1-\alpha)}\Bigg[-\int_\Theta \sum_{i=1}^n \sum_{s_i} \widetilde{q}_{S_i}(s_i)\log\frac{p(Y_i|\mu,s_i)\pi_{s_i}}{p(Y_i|\mu^*,s_i)\pi^*_{s_i}} q_\theta(d\theta) \\
&\qquad + \frac{D(q_\theta \parallel p_\theta)}{\alpha} + \frac{\log(1/\zeta)}{\alpha}\Bigg],
\end{aligned} \tag{3.2}$$

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

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

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

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

After choosing $\widetilde{q}_{S^n}$ as (3.3) in Corollary 3.2, we can make the first term in the right-hand side 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 Ridgway [2] where they exploit risk bounds for fractional posteriors developed in [5] 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 $\mathbb{P}^n_{\theta^*}$ probability at least $(1-\zeta)$ that for any probability measure $q_\theta \in \Gamma_\theta$ with $q_\theta \ll p_\theta$,*

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

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

$$\begin{aligned}\mathcal{B}_n(\pi^*, \varepsilon_\pi) &= \{D(\pi^* \parallel \pi) \le \varepsilon_\pi^2, V(\pi^* \parallel \pi) \le \varepsilon_\pi^2\},\\ \mathcal{B}_n(\mu^*, \varepsilon_\mu) &= \Big\{\sup_s D\big[p(\cdot|\mu^*, s) \parallel p(\cdot|\mu, s)\big] \le \varepsilon_\mu^2,\\ &\qquad \sup_s V\big[p(\cdot|\mu^*, s) \parallel p(\cdot|\mu, s)\big] \le \varepsilon_\mu^2\Big\},\end{aligned} \tag{3.5}$$

where we used the shorthand $D(\pi^* \parallel \pi) = \sum_s \pi_s^* \log(\pi_s^*/\pi_s)$ 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, that is, $p_\theta = p_\mu \otimes p_\pi$, 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*

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

Although the results in this section assume discrete latent variables, similar results can be seamlessly obtained for continuous latent variables; see the Supplementary Material [40] for more details. We will apply this theorem for analyzing 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, that is, 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, for example, [5], 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 $\varepsilon \ge \varepsilon_n$, there exists a sieve set $\mathcal{F}_{n,\varepsilon} \subset \Theta$ and a test function $\phi_{n,\varepsilon} : \mathcal{Y}^n \to [0, 1]$ such that

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

$$\mathbb{E}_{\theta^*}[\phi_{n,\varepsilon}] \le e^{-cn\varepsilon_n^2}, \tag{3.8}$$

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

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_{θ^*}, such a test $\phi_{n,\varepsilon}$ always exists [17]. A similar set of assumptions are used for showing the concentration of the usual posterior (e.g., see [16] 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 $\mathbb{P}^{(n)}_{\theta^*}$ probability at least $1 - 2e^{-cn\varepsilon_n^2/2}$, 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*

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

The first term on the left-hand side of inequality (3.10) relates the variational complementary probability $\widehat{Q}_\theta(\mathcal{F}^c_{n,\varepsilon})$ to the prior complementary probability $P_\theta(\mathcal{F}^c_{n,\varepsilon})$. 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 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 [28], 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 $e^{-Cn\varepsilon_n^2}$. Under those assumptions, the result in [28] 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.

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 neighborhoods $\mathcal{B}_n(\pi^*, \varepsilon)$ and $\mathcal{B}_n(\mu^*, \varepsilon)$ defined in (3.5).

ASSUMPTION P (Prior concentration). There exists some constant $C > 0$ such that

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

Under Assumptions T and P, Theorem 3.5 leads to a high probability bound on the 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* $c' > 0$),

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

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

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

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

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

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 [28]) that there is a global test function $\phi_n : \mathcal{Y}^n \to [0, 1]$, such that the type I error bound (3.8) holds, and the following type II error bound holds for all $\theta \in \Theta$ satisfying $r(\theta, \theta^*) \geq \varepsilon_n^2$,

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

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 $\bar{D}[p \,\|\, q] \geq D[p \,\|\, q]$, for example, χ^2 divergence and Rényi divergence [25], and the corresponding variational approximation

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

As another example, in some applications of variational inference [43], 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 divergence

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

The following theorem provides a variational Bayes risk upper bound to $\bar{q}_\theta$. 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

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

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 $\mathbb{P}^n_{\theta^*}$ 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,*

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

This theorem provides a simple replacement rule for α-VB—if the α-VB objective function $\Psi_{n,\alpha}$ is replaced with a 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 apply this replacement rule to obtain a minimax variational risk bound in the mixture of Gaussian variational approximation example provided in Section S6 of the Supplementary Material [40].

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, 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 $\widehat{q}_{\theta,\alpha}$ is dropped. Extensions to the $\alpha = 1$ case are discussed in the Supplementary Material [40]. We point out here that the Hellinger risk bounds in Corollaries 4.1–4.4 may actually depend on the unknown true parameter θ^*. The dependence is suppressed using $\precsim$ notation for clarity of exposition.

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

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

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 nonindependent 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 $\widehat{q}_\theta = \widehat{q}_\beta \otimes \widehat{q}_\sigma$. This falls into our framework when there is no latent variable and $W^n = \theta$. Computationalwise, a normal prior for θ and an inverse gamma prior for σ^2 are attractive since they are "conjugate" priors—the resulting variational densities $\widehat{q}_\beta$ and $\widehat{q}_\sigma$ 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* $d/n \to 0$ *as* $n \to \infty$, *then with probability tending to one as* $n \to \infty$,

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

The convergence rate of $\mathcal{O}(\sqrt{n^{-1}d\log(dn)})$ under the Hellinger distance implies that the α-VB estimator $\widehat{\beta}_{\mathrm{VB},\alpha} = \int \beta\widehat{q}_\beta(\beta)\,d\beta$ converges toward β^* relative to the ℓ_2 norm at rate $\sqrt{n^{-1}d\log(dn)}$ (under the condition that $n^{-1}X^TX$ 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 [41] 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, that is, the number of nonzero coefficients, of the true regression parameter β^*.

We consider the popularly used spike and slab priors [15] on β. Following [15], we introduce a latent indicator variable $z_j = I(\beta_j \neq 0)$ for each β_j to indicate whether the jth covariate X_j is included in the model, and call $z = (z_1, \ldots, z_d) \in \{0,1\}^d$ 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):

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

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 [12] 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 $\widehat{q}_\theta(\theta) = \widehat{q}_\sigma(\sigma)\prod_{j=1}^{d}\widehat{q}_{z_j,\beta_j}(z_j, \beta_j)$. 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 [12]. 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 $s\log d/n \to 0$ as $n \to \infty$, then it holds with probability tending to one as $n \to \infty$ that*

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

Corollary 4.2 implies a convergence rate $\sqrt{n^{-1}s\log(dn)}$ of the variational-Bayes estimator $\widehat{\beta}_{\mathrm{VB},\alpha}$ under the restricted eigenvalue condition [6], which is the minimax rate up to log terms for high-dimensional sparse linear regression. To our knowledge, [27] 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 $\alpha^{-1}D(q_\theta \parallel p_\theta)$ appearing in the upper bound in (3.2) cannot attain the minimax order $\sqrt{n^{-1}s\log d}$.

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 kth component and $\pi = (\pi_1, \ldots, \pi_K) \in \mathcal{S}_K$ is the mixing probability. Here, for simplicity we assume the covariance matrix of each Gaussian component to be I_d. $\mu = (\mu_1, \ldots, \mu_K)$ 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:

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

Let $Y^n = (Y_1, \ldots, Y_n)$ be n i.i.d. copies of Y with parameter $\theta^* = (\mu^*, \pi^*)$, and $S^n = (S_1, \ldots, S_n) \in \{1, \ldots, K\}^n$ 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

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

to approximate the joint α-fractional posterior distribution of (π, μ, S^n), producing the α-mean-field approximation $\widehat{q}_\theta \otimes \widehat{q}_{S^n}$, where $(\widehat{q}_\theta, \widehat{q}_{S^n})$ 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.

COROLLARY 4.3. *Suppose the prior densities p_μ and p_π are thick and continuous at μ^* and π^* respectively. If $dK/n \to 0$ as $n \to \infty$, then it holds with probability tending to one as $n \to \infty$ that*

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

As a related result, [37] 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. Our current analysis opens the door for analyzing the optimization algorithms using 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 [19], 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 $k = 1, \ldots, K$,
   (a) draw a distribution over words $\beta_k \sim \mathrm{Dir}_V(\eta_\beta)$.
2. For each document in $d = 1, \ldots, D$,
   (a) draw a vector of topic proportions $\gamma_d \sim \mathrm{Dir}_K(\eta_\gamma)$.
   (b) For each word in $n = 1, \ldots, N$,
      i. draw a topic assignment $z_{dn} \sim \mathrm{multi}(\gamma_d)$, then
      ii. draw a word $w_{dn} \sim \mathrm{multi}(\beta_{z_{dn}})$.

Here, $\eta_\beta \in \mathbb{R}_+$ is a hyperparameter of the symmetric Dirichlet prior on the topics β, and $\eta_\gamma \in \mathbb{R}_+^K$ are hyperparameters of the Dirichlet prior on the topic proportions for each document. $z_{dn} \in \{1, \ldots, K\}$ is the latent class variable over topics where $z_{dn} = k$ indicates the nth word in document d is assigned to the kth topic. Similarly, $w_{dn} \in \{1, \ldots, V\}$ is the latent class variable over the words in the vocabulary where $w_{dn} = v$ indicates that the nth word in document d is the vth 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 $c > 1$ [39].

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 $\pi = \{\gamma_d : d = 1, \ldots, D\}$ and $\mu = \{\beta_k : k = 1, \ldots, K\}$, from the posterior distribution $P(\pi, \mu, z|Y^n)$, where $S^N = \{S_n : n = 1, \ldots, N\}$ with $S_n = \{z_{dn} : d = 1, \ldots, D\}$ are latent variables, and $Y^N = \{Y_n : n = 1, \ldots, N\}$ with $Y_n = \{w_{dn} : d = 1, \ldots, D\}$ are the data, and the priors for (π, μ) are independent Dirichlet distributions $\mathrm{Dir}_K(\eta_\gamma)$ and $\mathrm{Dir}_V(\eta_\beta)$ 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

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

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

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

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 $D_\alpha[p(\cdot|\theta) \parallel p(\cdot|\theta^*)] = \sum_{d=1}^{D} D_\alpha[p_d(\cdot|\theta) \parallel p_d(\cdot|\theta^*)]$

rescaling by a factor of D^{-1}, where $p_d(\cdot|\theta)$ denotes the likelihood function of the dth 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 4.4. *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*

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

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

$$\begin{aligned}&\left\{\int D^{-1} h^2[p(\cdot|\theta) \parallel p(\cdot|\theta^*)]\widehat{q}_\theta(\theta)\,d\theta\right\}^{1/2} \\ &\lesssim \sqrt{\frac{\sum_{d=1}^D e_d}{DN\min\{\alpha, 1-\alpha\}}\log(DKN) + \frac{\sum_{k=1}^K d_k}{DN\min\{\alpha, 1-\alpha\}}\log(KVN)}.\end{aligned}$$

Corollary 4.4 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" $DN \to \infty$. 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 parametric 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 variables.

Acknowledgments. The first author was supported by NSF Grant DMS-1810831.

The second author was supported by NSF Grant DMS-1613156 and NSF CAREER Grant DMS-1653404.

The third author was supported by NSF Grant DMS-1613156.

SUPPLEMENTARY MATERIAL

Supplement to "α-variational inference with statistical guarantees" (DOI: 10.1214/19-AOS1827SUPP; .pdf). Supplementary information.

REFERENCES

[1] AHMED, A., ALY, M., GONZALEZ, J., NARAYANAMURTHY, S. and SMOLA, A. (2012). Scalable inference in latent variable models. In *International Conference on Web Search and Data Mining* (*WSDM*) **51** 1257–1264.

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

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

[4] ATTIAS, H. (2000). A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems* 209–215.

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

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

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

[8] BISSIRI, P. G., HOLMES, C. C. and WALKER, S. G. (2016). A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 1103–1130. MR3557191 https://doi.org/10.1111/rssb.12158

[9] BLEI, D. M. (2012). Probabilistic topic models. *Commun. ACM* **55** 77–84.

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

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

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

[13] CORDUNEANU, A. and BISHOP, C. M. (2001). Variational Bayesian model selection for mixture distributions. In *Artificial Intelligence and Statistics* **2001** 27–34. Morgan Kaufmann, Waltham, MA.

[14] GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

[15] GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

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

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

[18] HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109. MR3363437 https://doi.org/10.1093/biomet/57.1.97

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

[20] HUMPHREYS, K. and TITTERINGTON, D. (2000). Approximate Bayesian inference for simple mixtures. In *Proc. Computational Statistics* 2000 331–336.
[21] JIANG, W. and TANNER, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.* **36** 2207–2231. MR2458185 https://doi.org/10.1214/07-AOS547
[22] JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.
[23] KINGMA, D. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at arXiv:1412.6980.
[24] KUSHNER, H. J. and YIN, G. G. (1997). *Stochastic Approximation Algorithms and Applications. Applications of Mathematics* (*New York*) **35**. Springer, New York. MR1453116 https://doi.org/10.1007/978-1-4899-2696-8
[25] LI, Y. and TURNER, R. E. (2016). Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*.
[26] MACKAY, D. J. (1997). Ensemble learning for hidden Markov models.
[27] ORMEROD, J. T., YOU, C. and MULLER, S. (2017). A variational Bayes approach to variable selection. *Electron. J. Statist.* **11** 3549–3594.
[28] PATI, D., BHATTACHARYA, A. and YANG, Y. (2018). On statistical optimality of variational Bayes. In *International Conference on Artificial Intelligence and Statistics* 1579–1588.
[29] ROBERT, C. P. (2004). *Monte Carlo Methods*. Wiley Online Library.
[30] ROUSSEAU, J. (2016). On the frequentist properties of Bayesian nonparametric methods. *Annual Review of Statistics and Its Application* **3** 211–231.
[31] UEDA, N. and GHAHRAMANI, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Netw.* **15** 1223–1241.
[32] VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256
[33] VAN ERVEN, T. and HARREMOËS, P. (2014). Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. Inform. Theory* **60** 3797–3820. MR3225930 https://doi.org/10.1109/TIT.2014.2320500
[34] WAINWRIGHT, M. J., JORDAN, M. I. et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* **1** 1–305.
[35] WANG, B. and TITTERINGTON, D. (2004). Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Processing Letters* **20** 151–170.
[36] WANG, B. and TITTERINGTON, D. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *AISTATS*.
[37] WANG, B. and TITTERINGTON, D. M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Anal.* **1** 625–649. MR2221291 https://doi.org/10.1214/06-BA121
[38] WESTLING, T. and MCCORMICK, T. H. (2015). Establishing consistency and improving uncertainty estimates of variational inference through M-estimation. Preprint. Available at arXiv:1510.08151.
[39] YANG, Y. and DUNSON, D. B. (2014). Minimax optimal Bayesian aggregation. Preprint. Available at arXiv:1403.1345.
[40] YANG, Y., PATI, D. and BHATTACHARYA, A. (2020). Supplement to "α-variational inference with statistical guarantees." https://doi.org/10.1214/19-AOS1827SUPP.
[41] YOU, C., ORMEROD, J. T. and MÜLLER, S. (2014). On variational Bayes estimation and variational information criteria for linear regression models. *Aust. N. Z. J. Stat.* **56** 73–87. MR3200293 https://doi.org/10.1111/anzs.12063
[42] ZHANG, F. and GAO, C. (2017). Convergence Rates of Variational Posterior Distributions. Preprint. Available at arXiv:1712.02519.
[43] ZOBAY, O. (2014). Variational Bayesian inference with Gaussian-mixture approximations. *Electron. J. Stat.* **8** 355–389. MR3195120 https://doi.org/10.1214/14-EJS887