STAT547 Final Project: A Variational Bernstein von–Mises type result

Johnny Xi

December 10, 2020

1 Introduction

Bayesian inference is a elegant framework for statistical analysis when the analyst wishes to embed probabilistic prior beliefs into a model. These prior beliefs are then updated according to the data to produce a posterior distribution of the parameters. For much of its history, Bayesian analysis was merely a clever idea that was hindered by the analytic tractability of complex high dimensional integrals in normalizing constant of the fitted posterior distribution. However more recent computational advancements have allowed for the increased applicability of the general Bayesian framework for highly expressive statistical modeling. Today, Bayesian methods are deployed in various scientific and machine learning disciplines as the state of the art, from clinical trials to text and image analysis.

One computational strategy that has gained popularity recently is known as variational inference (VI). VI is a member of a broader class of methods known as approximate inference. Instead of attempting to solve the normalizing integral, VI proposes a family of simpler distributions and finds a best approximation to the posterior, given the data and prior. In most cases, the variational family does not include the true posterior, and so it is truly an approximation, with no aim to recover exact posteriors. VI uses tools from optimization theory and has been proven to be faster than its counterparts and is often the only feasible strategy for extremely complex cases. However, the statistical propeties of VI outputs are not as well understood as it is often unable to directly leverage the elegance of Bayesian analysis due to the necessarily imperfect approximation.

The theoretical development of VI in the scope of classic statistical theory is hence a much needed task in the modern landscape. The popularity of VI suggests that it performs well in many relevant real-world scenarios, and that in turn suggests that theoretical properties are not necessarily absent, but rather simply underdeveloped. In a first step towards understanding the theoretical properties of VI, [WB19b] show a version of the classical Bernstein–von Mises Theorem for variational approximations to a parametric Bayesian posterior. Such a result can not only justify the current use of variational inference, but also opens the door to further theoretical developments, much like the impact of the original Bernstein–von Mises result.

The rest of the report will consist of three main sections. We will begin with a brief discussion of Bayesian asymptotics and hence the original Bernstein–von Mises Theorem, then a short introduction to VI and the mean-field family. Then, we will describe the chain of implications, written as four lemmata, that eventually leads to the variational Bernstein von–Mises result. Finally, we discuss the future research directions that this result inspires, and some specific open questions. We also provide two short exercises in the appendix, relevant to the understanding of the report. The actual proof of the results are highly technical, and so due to the expository nature of the work and limited space, will be described largely by proof sketches.

2 Background

2.1 Bayesian Inference

We begin with a brief tour of Bayesian inference which provides the setting for the rest of the report. I borrow heavily from the first chapter of [Sch95] for this sub-section. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the background probability space, and let $(F, \mathcal{F})$ (the parameter space) and $(E, \mathcal{E})$ (the sample space) be standard measurable spaces. Then, the parameter Θ and data X are random variables taking values in $(F, \mathcal{F})$ and $(E, \mathcal{E})$ respectively, and regular versions of their conditional distributions given each other exist. The statistical model at hand is then defined by the family of conditional distributions of X given $\Theta(\omega) = \theta$, seen as probability measures over $(E, \mathcal{E})$, and denoted $P_\Theta = \{P_{X|\theta}; \theta \in F\}$. For this report, we assume that $(F, \mathcal{F}) = (\mathbb{R}^d, \mathbf{B}(\mathbb{R}^d))$ and all densities defined here are with respect to the Lebesgue measure for simplicity. Now suppose a prior distribution μ_0 with density p is specified for Θ, and X has conditional and marginal densities with respect to a dominating measure ν. Then, Bayes' Theorem ([Sch95], Theorem 1.31) gives the conditional density of

Θ given X with respect to μ_0:

$$\frac{d\mu_{\Theta|X}}{d\mu_0} = \frac{p_{X|\Theta}(x|\theta)p(\theta)}{\int_{\mathcal{F}} p_{X|\theta}(x|\theta)p(\theta)(d\theta)} \tag{1}$$

where $p_{X|\Theta}(x|\theta)$ is the conditional density at $\theta = \Theta(\omega)$ and the denominator is the marginal density of X. The distribution $\mu_{\Theta|X}$ is the main object of interest in Bayesian inference, and is known as the posterior distribution of the statistical parameters. The more practical setting concerns a random sample of size n from the distribution $P_{X|\Theta}$, in which the posterior density has the form

$$\frac{\prod^n p_{X|\Theta}(x_i|\theta)p(\theta)}{\int_{\mathcal{F}} \prod^n p_{X|\theta}(x_i|\theta)p(\theta)(d\theta)}$$

We typically refer to the conditional density $\ell(\theta; x) = \prod^n p_{X|\Theta}(x_i|\theta)$ as the likelihood. Denote the corresponding posterior distribution for n samples as $P^n_{\Theta|X_n}$.

2.2 Bayesian Asymptotics

The previous setting philosophically differs from frequentist statistics, where the parameters are assumed to have an atomic *true* value θ_0. However, Bayesian procedures are still understood in the domain of frequentist estimation, and a rich literature exists on their "frequentist" asymptotic properties. In the main reference for this sub-section [Vaa00], van der Vaart simply states that "Bayes estimators are studied from a frequentist perspective". A remarkably general result is sometimes referred to as Doob's Theorem [Doo49], which states that posterior contracts to a true parameter value μ_0-almost surely. Roughly speaking, if the true value is not only contained in null sets under the prior, then the posterior is consistent in the frequentist sense. Although this is a theoretically strong result, its practical application is limited as the null set under the prior can be large, particularly in the case of nonparametric models. A weaker, but more applicable result is known as the Bernstein von–Mises theorem (BvM). We will first discuss the assumptions of the theorem.

Definition 2.1 (Differentiable in Quadratic Mean (DQM))**.** *Let p_{θ_0} denote the density with respect to ν of P_{θ_0}. The model is said to be differentiable in quadratic mean at θ_0 if there exists a measurable function l'_{θ_0} such that*

$$\int \left(\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \frac{1}{2}h^T l'_{\theta_0}\sqrt{p_{\theta_0}}\right)^2 d\nu = o(|h|^2) \tag{2}$$

as $h \to 0$.

Proposition 2.2 (Theorem 7.2, [Vaa00])**.** *Suppose the model P_Θ is DQM at θ_0 and that Θ is an open subset of $\mathbb{R}^k$. Then, $E_{\theta_0}[l'_{\theta_0}] = 0$ and the Fisher information $E_{\theta_0}[l'_{\theta_0}(l'_{\theta_0})^T]$ exists. Furthermore, for any sequence $h_n \to h$, as $n \to \infty$,*

$$\sum_{i=1}^n \log\left(\frac{p_{\theta_0+h_n/\sqrt{n}}}{p_{\theta_0}}(x_i)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n h^T l'_{\theta_0}(x_i) - \frac{1}{2}h^T I_{\theta_0} h + o_{P_{\theta_0}}(1) \tag{3}$$

- (A1) (Prior Mass): The prior distribution μ_0 has a continuous positive density in a neighbourhood of θ_0.
- (A2) (Consistent Testability) Let $P^n_{\theta 0}$ be the conditional distribution of the data $X_1, \dots X_n$ given θ_0. For every $\epsilon > 0$, there exists a sequence of measurable functions $\phi_n : E \to [0, 1]$ (known as tests) such that

$$P^n_{\theta_0}\phi_n \to 0 \qquad \sup_{|\theta-\theta_0|\geq\epsilon} P^n_\theta(1 - \phi_n) \to 0 \tag{4}$$

- (A3) (Local Asymptotic Normality): The statement in proposition 2.2 is known as local asymptotic normality (LAN). It roughly states that the log likelihood ratio converges under the true model to a quadratic expression, suggesting its normality. For the following theorem, also suppose that the Fisher information is nonsingular.

Heuristically, (A1) is necessary for the result as the posterior has no hope of contracting near a region of zero density under the prior. (A2) essentially states that there is a sequence of tests to separate a null hypothesis $\theta = \theta_0$ and its global alternative. In practice, this is a mild condition and is satisfied for example if the parameter space is compact and the model is identifiable and continuous in the parameters. (A3) is necessary for establishing the normality result that is implied by the BvM, and is essentially stating that the likelihood ratio is locally dominated by the quadratic terms of its Taylor expansion. We now state the result.

Theorem 2.3 (Bernstein von–Mises Theorem, Theorem 10.1 [Vaa00]). *Suppose (A1) – (A3) are satisfied. Let*

$$\Delta_{n,\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{\theta_0}^{-1} l'_{\theta_0}(x_i) \tag{5}$$

be a sequence of random vectors. Then, the sequence of posterior distributions satisfy

$$\| P^n_{\sqrt{n}(\theta-\theta_0)} - N(\Delta_{n,\theta_0}, I_{\theta_0}^{-1}) \|_{TV} \to 0$$

where the total variation norm of a distribution $\mu \in \mathcal{F}$ is equal to $\sup_{B \in \mathcal{F}} \mu(B)$.

The proof can be found in [Vaa00]. Theorem 2.3 and its extensions are considered to be the main result in the intersection of asymptotic statistics and Bayesian inference. The restriction that the prior density need only be positive and continuous at θ_0 is much weaker than that in Doob's theorem. The statement however is not necessarily straight forward. Since the posterior is a conditional distribution, it is really a random measure. The target Normal is also a random measure – it is centered at Δ_{n,θ_0}, which is a function of the random data x_i. However, the result is still useful in the sense that a point estimator derived from minimizing a suitable expected loss function under the posterior converges to the minimizer of the same loss function under what is almost a centered normal. Most specifically, one can expect that the mean of the scaled and centered posterior converges to 0. For more details, see Theorem 10.8 in [Vaa00].

2.3 Variational Inference (VI)

Modern Bayesian modelling often assumes the existence of *local* latent variabes z_i, for each data observation. These can be considered additional dimensions of Θ in the context above, so that if the model has d dimensions, then $\dim(\Theta) = n + d$. To avoid carrying around extra notation, denote the model parameters as θ, and we refer to them as global latent variables. Note that only the global latent variables carry prior distributions. The posterior density $p(\theta, z|x)$ is analytically intractable for many useful models. A widely popular scheme to perform Bayesian inference involves drawing approximate samples from the posterior, and performing Monte carlo analysis. These strategies typically guarantee that samples resemble the posterior asympotically, but this rate of convergence can be slow for models with many parameters. VI, (or variational Bayes (VB)) is an alternative strategy for explicit approximate inference.

The idea is that instead of attempting to find the true posterior, a convenient variational family of distributions Q is defined, and the best approximator $q^* \in Q$ is taken as a surrogate to the posterior p. The focus here is on a special case of VB, where Q is restricted to have a factorizable density and optimality is defined in terms of the Kullback–Leibler (KL) divergence. This is often referred to as mean-field variational Bayes (MVFB) and Q is called the mean-field family. Specifically, let $\mathrm{KL}(\cdot||\cdot)$ be the KL divergence between two distributions, expressed in terms of their densities. Then,

$$Q^{n+d} = \left[q; q(\theta, z) = \prod_{i=1}^{d} q_{\theta_i}(\theta_i) \prod_{i=i}^{n} q_{z_j}(z_j) \right] \tag{6}$$

$$q^*(\theta, z) = \arg\min_{Q^{n+d}} \mathrm{KL}(q(\theta, z) \| p(\theta, z|x)) \tag{7}$$

Note that the KL divergence is not symmetric, and it is a choice of convenience to order the distributions like so. If the terms were flipped, the KL divergence would require to compute an expectation over the intractable posterior. Under this mean field family, notice that the variational distribution may be factorized as

$$q(\theta, z) = q_\theta(\theta) q_z(z) \tag{8}$$

where each component does not depend on the other parameter. Often, the local latents are considered to be nuisance parameters, and we call the optimal q_θ^* the VB posterior. The target is then the marginal posterior $p(\theta|x) = \int dz p(\theta, z|x)$. Denote Q^d and Q^n to be the variational families in which q_θ and q_z are defined, respectively. There is an immediate limitation in the expressiveness of the mean-field family. The factorizable densities implies that the marginal distributions are independent. This is not the case for many distributions of interest, and hence the mean-field approximation can be very poor if dependencies across parameters are strong.

The key topic of the report is whether a BvM-type phenomenon exists for the global variables θ on the variational approximation q^* to the posterior. The literature of BvM adaptations is rich, including under model misspecification ([KV12]), so it is not unreasonable to suggest that some element of BvM also survives the VB approximation. Indeed, [WB19b] answer this in the positive. An exposition of this work, and in particular Theorem 5 represents the body of this report.

2.3.1 Ideal VB

To proceed with the exposition, we must define and translate some ideas to the VI framework. We begin with the frequentist view – define the variational log likelihood:

$$M_n(\theta; x) = \max_{Q^n} \mathbb{E}_{q_z}[\log p(x, z|\theta) - \log(q_z(z))] \tag{9}$$

in contrast to equation 7, this is a function of θ (viewed as a constant), and the optimization on q is only over the local variables z. Note that in relation with the usual hierarchical log-likelihood, $q(z)$ plays the role of approximating the conditional density $p(z|\theta)$. In the special case that there are not local latent variables z_i, equation 9 corresponds exactly to the usual log likelihood $\log p(x|\theta)$. The optimal variational distribution $q_z^\dagger \in Q^n$ for each θ is said to solve the local VI problem. Now, the variational frequentist estimator is defined as

$$\hat{\theta_n} = \arg\max_\theta M_n(\theta; x) \tag{10}$$

Now, suppose a Bayesian model is fit to a random sample x, with prior density $p(\theta)$ and likelihood $p(x, z|\theta) = p(x|z, \theta)p(z|\theta)$. Define the VB ideal density to be

$$\pi^*(\theta|x) = \frac{p(\theta)\exp(M_n(\theta; x))}{\int p(\theta)\exp(M_n(\theta; x))d\theta} \tag{11}$$

Note that this is a highly complicated expression that is defined only for theoretical purposes, owing to the inner optimization for each θ. As mentioned, with no latents z_i, the variational log-likelihood reduces to the usual log likelihood and hence equation 11 is the exact posterior. Otherwise, it involves the variational approximation of the conditional density $p(z|x, \theta)$. This suggests that the VB ideal is in some sense between the exact posterior, from which it differs by the error in q_z, and the VB posterior, from which is it differs by the error in q_θ. This is explored in slightly more depth in exercise A.2.

2.3.2 Evidence Lower Bound

In most practical settings, KL optimization requires iterative procedures and hence the posterior density must be evaluated at each iteration. This is clearly not feasible in the cases where VI would be useful, so instead an alternative objective is defined.

Definition 2.4 (Evidence Lower Bound (ELBO), [BKM17], section 2.2)**.** *In the notation above, the ELBO between the posterior and the variational distribution q is defined as*

$$ELBO(q) = \mathbb{E}_q[\log(p(\theta, z, x))] - \mathbb{E}_q[\log(q(\theta, z))]$$

Note the first term in the ELBO is the joint density, which side-steps the usual difficulty of computing the marginal distribution $p(x)$ (the denominator of equation 1). The ELBO satisfies the following equation

$$ELBO(q) = -\mathrm{KL}(q(\theta, z) \| p(\theta, z|x)) + \log p(x) \tag{12}$$

Since the marginal distribution does not depend on q, maximizing the ELBO over the variational family solves an equivalent problem to minimizing the KL divergence. This is the alternative optimization used most commonly in practice.

3 The Variational Bernstein von–Mises Theorem

We now present the Variational Bernstein von–Mises theorem. In this form, the BvM states that the VB posterior converges in total variation to the variational approximation to a normal distribution, i.e., the KL minimizer within the variational family to some normal. The proof of the theorem follows directly from four preceding lemmata. The notation for the rest of the report will be in the style of [WB19b], with translatations from [Vaa00] when necessary. As before, assume that there exists a true atomic value of the parameter θ_0, and further suppose that the model also depends locally on a latent parameter z. That is, the model has density

$$\int p(x, z|\theta = \theta_0) dz \tag{13}$$

while each draw from the model implicitly carries a realization of the latent state.

3.1 VB Ideal

Define the frequentist variational model to have likelihood

$$\ell_n(\theta; x) \propto \exp(M_n(\theta; x)) \tag{14}$$

then, the VB ideal $\pi^*(\theta|x)$ is the usual posterior density under this model. Denote the associated VB posterior distribution to be Π^*. Importantly, the frequentist variational model is typically misspecified for the true model in equation 13, since the maximum and expectation operators in equation (8) does not recover the model log-likelihood unless the local optimal q_z recovers $p(z|x, \theta)$ exactly. As mentioned, a BvM result exists for model misspecification [KV12], which will be applied in the result here. Recall that in Theorem 2.3, the posterior distributions were centered at θ_0 and scaled by $\sqrt{n}$. We also define a similar transformation here:

$$\tilde{\theta} = \delta_n^{-1}(\theta - \theta_0) \tag{15}$$

for some diagonal matrix δ_n. Then, we have the transformed VB ideal $\tilde{\Pi}$ with density:

$$\tilde{\pi}(\tilde{\theta}|x) = \pi^*(\theta_0 + \delta_n\tilde{\theta}|x)|det(\delta_n)| \tag{16}$$

We will now briefly discuss the conditions of the variational BvM in the context of (A1) – (A3) in the background section. For the rest of the report, assume the following conditions:

- (VA1) (Prior Mass) The prior distribution has a density $p(\theta)$ that is continuous and positive on a neighbourhood of θ_0. There exists a constant $M_p > 0$ such that $|(\log p(c))''| \leq M_p e^{|c|^2}$ for all $c \in \mathbb{R}$.
- (VA2) (Consistent Testability) For every $\epsilon > 0$, there exists a sequence of tests ϕ_n such that

$$\int \phi_n(x) p(x, z|\theta_0) dx dz$$

$$\sup_{\theta: \|\theta - \theta_0\| \geq \epsilon} \int (1 - \phi_n(x)) \frac{\ell_n(\theta; x)}{\ell_n(\theta_0; x)} p(x, z|\theta_0) dx dz$$

- (VA3) (LAN) For every compact set $K \in \mathbb{R}^d$, there exists random vectors Δ_{n,θ_0} bounded in probability and nonsingular matrices V_{θ_0} such that for $\delta_n \to 0$ a diagonal matrix, we have

$$\sup_{h \in K} |M_n(\theta + \delta_n h; x) - M_n(\theta; x) - h^T V_{\theta_0} \Delta'_{n,\theta_0} + \frac{1}{2} h^T V_{\theta_0} h| \to 0 \tag{17}$$

  in probability under the 'true' data distribution P_{θ_0}. Note the notation Δ'_{n,θ_0} is slightly changed from [WB19b] to contrast with the standard (A3). Δ'_{n,θ_0} plays the same role as Δ_{n,θ_0} in the traditional BvM, but here it is only a bounded random vector while in Theorem 2.3 it is exactly specified.

Note that (VA1) and (A1) are almost equivalent conditions - there is only a mild technical addition on the hessian of the prior density that is true for most non-heavy tailed distributions. (VA3) is equivalent to (A3) by replacing the model with the frequentist variational model, and its truth is checked case-by-case by analyzing M_n. As mentioned, in the case where there are no local latent variables, (A3) corresponds to (VA3) exactly. There are conditions that may be checked for when the LAN expansions in (A3) and (VA3) have an exact correspondence, see ([WB19b], section 3.4) for a detailed discussion. Connecting (VA2) and (A2) requires a bit more care. (VA2) roughly states that there exists a sequence of tests for identifying the events $\|\theta - \theta_0\| \geq \epsilon$ and $\theta = \theta_0$ given samples from the true model, assessed under the likelihood ratio of the misspecified variational model. This is shown to be equivalent to the testability condition in [KV12], which implies the misspecified BvM.

Lemma 3.1 (Lemma 1, [WB19b]). *Under assumptions (VA1) – (VA3), we have*

$$\|\tilde{\Pi} - N(\Delta'_{n,\theta_0}, V_{\theta_0}^{-1})\|_{TV}$$

The proof of Lemma 3.1 boils down to checking that assumption (VA2) implies the consistent testability condition for misspecified BvM ([KV12], (2.3)), then applying this result to the rescaled VB ideal posterior. Notice that this is almost the statement in Theorem 2.3. The key difference is that we use Δ'_{n,θ_0} and V_{θ_0} in place of the more specifically defined Δ_{n,θ_0} and I_{θ_0}.

3.2 KL Minimizer of the VB Ideal

Recall that the VB ideal is highly complex and of little practical interest. To relate it to practical VI, a second variational approximation must be made to the VB ideal itself. Since the VB ideal is parametrized only by θ, the mean-field family is just Q^d. Heuristically, the scaled VB ideal converges to a distribution with Normal spread, so from traditional asymptotics it is expected that the unscaled version converges instead to a point mass. We show that the variational approximation converges to the point mass at θ_0, i.e., it is consistent.

Lemma 3.2 (Lemma 2, [WB19b]). *Let δ_{θ_0} be the dirac probability measure at θ_0. Then, the following event occurs almost surely under the probability measure P_{θ_0}:*

$$\arg\min_{Q^d} KL(q_\theta(\theta) \| \pi^*(\theta|x)) \to \delta_{\theta_0}$$

in distribution.

A heuristic sketch of the proof is that mean-field families include point masses in general (the dirac delta on each dimension), so in this general setting, the lemma holds and hence we have consistency. Recall the KL divergence between two distributions with point masses is either 0, at which point they coincide, otherwise it is $-\infty$. The actual justification is much more technical, and can be found in the appendix of [WB19b].

Now, we provide a result on the KL minimizer of the scaled VB ideal. We first state a technical condition for the lemma. Recall that we are working with the mean-field family. Suppose that under the transformation $\tilde{\theta} = \delta_n^{-1}(\theta - \mu)$ for some μ, the densities have the form

$$q_\theta(\theta) = \prod_{i=1}^{d} \delta_{n,ii}^{-1} q_{h,i}(\tilde{\theta})$$

where the following assumptions are true:

- $q_{h,i}$ have continuous densities.
- $-\int d\theta q_{h,i}(\theta) \log q_{h,i}(\theta)$, i.e., the differential entropy is finite and positive.
- $\int d\theta (q_{h,i}(\theta))'$, the first derivative is integrable.

Lemma 3.3 (Lemma 3, [WB19b]). *Under the above technical conditions, we have that*

$$\| \arg\min_{Q^d} KL(q_\theta(\theta) \| \tilde{\pi}(\theta|x)) - \arg\min_{Q^d} KL(q_\theta(\theta) \| N(\theta; \Delta'_{n,\theta_0}, V_{\theta_0}^{-1})) \|_{TV} \to 0$$

in probability, where we write $N(\theta; \cdot, \cdot)$ to represent the normal density.

Lemma 3.1 establishes this result for the optimization targets, so Lemma 3.3 can be understood as a continuity argument for the KL minimizer with respect to Q^d, see Figure 3.2 for a visual representation. This is close to the final result, which replaces the first term in the norm with the VB posterior.

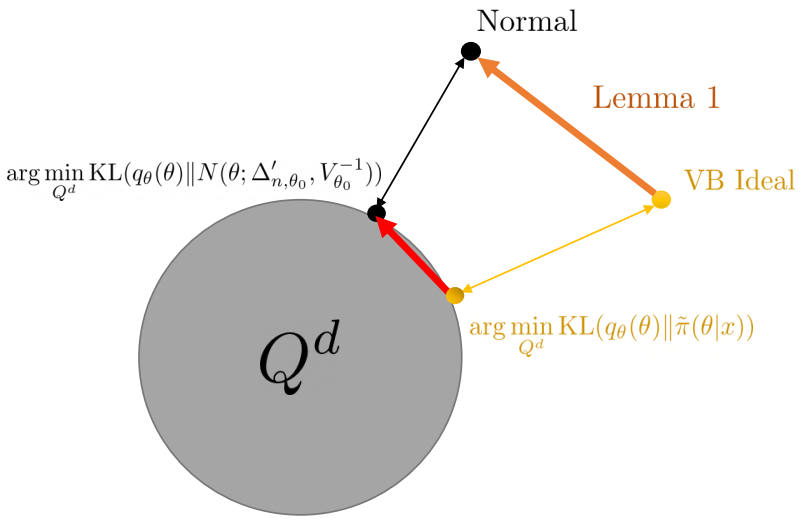

Figure 1: Schematic of Lemma 3.3. The red arrow indicates the convergence implied.

3.3 Connecting the VB Posterior

As mentioned before, we want to derive a BvM result on the global portion of the VB posterior, i.e., q_θ. Recall the factorization from equation (8). We can expand the ELBO according to this:

$$\begin{aligned} ELBO(q_\theta(\theta) q_z(z)) &= \mathbb{E}_q[\log(p(\theta, z, x))] - \mathbb{E}_q[\log(q_\theta(\theta) q_z(z))] \\ &= \mathbb{E}_q[\log(p(x, z|\theta) p(\theta)) - \log(q_\theta(\theta) q_z(z))] \\ &= \int \int d\theta dz q_\theta(\theta) q_z(z) \log\left(\frac{p(\theta) p(x, z|\theta)}{q_\theta(\theta) q_z(z)}\right) \\ = \int d\theta q_\theta(\theta) \log\left(\frac{p(\theta)}{q_\theta(\theta)}\right) &+ \int d\theta q_\theta(\theta) \int dz q_z(z) \log\left(\frac{p(x, z|\theta)}{q_z(z)}\right) \end{aligned}$$

In this form, we can use the strategy of profiling to obtain the optimal distribution q_θ, viewing the local factor as a nuisance parameter. That is, if q_z is viewed as fixed, we can derive

$$q'_\theta = \arg\max_{Q^d} ELBO(q_\theta(\theta) q_z(z))$$

Now, the factor q_θ^* of the global optimum, i.e., the VB posterior, is obtained simply by taking the maximum value of the ELBO with respect to q_z first. That is,

$$q_\theta^*(\theta) = \arg\max_{Q^d} \left(\sup_{Q^n} ELBO(q_\theta(\theta) q_z(z)) \right) \tag{18}$$

Note the inner supremum still takes on a different value for each q_θ. If we denote the profiled ELBO $\sup_{Q^n} ELBO(q_\theta(\theta)q_z(z)) = ELBO_p(q_\theta(\theta))$, then we can rewrite Equation 18 as $q_\theta^*(\theta) = \arg\max_{Q^d} ELBO_p(q_\theta(\theta))$, where

$$ELBO_p(q_\theta) = \sup_{Q^n} \int d\theta q_\theta(\theta) \left(-\log q_\theta(\theta) \log \left[p(\theta) \exp \left[\int dz q_z(z) \log \left(\frac{p(x, z|\theta)}{q_z(z)} \right) \right] \right] \right) \tag{19}$$

We now have the notation to state the final lemma.

Lemma 3.4 (Lemma 4, [WB19b]). *Under some additional technical conditions on the variational densities in Q^d (see exercise 1), we have that for any $q_\theta(\theta) \in Q^d$,*

$$ELBO_p(q_\theta(\theta)) = -KL(q_\theta(\theta) \| \pi^*(\theta|x)) + C_n + o_p(1)$$

where $o_p(1)$ indicates that a remainder term converges to 0 in probability under the probability measures in Q^d and C_n is bounded.

This lemma suggests that the maximizer of the profiled ELBO, i.e., the VB posterior, is also the minimizer of the KL to the VB ideal. In view of Lemma 3.3, this is a profound result. The proof sketch of this Lemma is appealing and insightful. Notice that the KL divergence is to the VB ideal, which implicitly involves a maximization over q, much like the profiled ELBO. We can write the negative KL-divergence as

$$-KL(q_\theta(\theta) \| \pi^*(\theta|x)) = \int d\theta q(\theta) \log \left(\frac{p(\theta) \exp(M_n(\theta; x))}{q_\theta(\theta)} \right) + C_n$$

where the constant C_n is the log of the denominator in the VB ideal. Then,

$$-KL(q_\theta(\theta) \| \pi^*(\theta|x)) = \tag{20}$$

$$\int d\theta q_\theta(\theta) \left(-\log q_\theta(\theta) + \log \left[p(\theta) \exp \left[\sup_{Q^n} \int dz q_z(z) \log \left(\frac{p(x, z|\theta)}{q_z(z)} \right) \right] \right] \right) + C_n$$

Notice that the term which depends on q_θ differs from the profiled ELBO only by the location of the supremum. Given a q_θ, the profiled ELBO implicitly solves the optimization over q_z once. Since the derivation involves solving the joint optimization over both q_z and q_θ, it is reasonable to expect that these are simultaneously determined. In equation 20, every value of θ in q_θ can result in a different optimal q_z determining the supremum. To avoid this issue, if q_θ is restricted to be a point mass, then varying θ does not change the expression and the profiled ELBO and equation 20 are equivalent. Recall that Lemma 3.1 suggests the VB ideal converges to a point mass. Since the KL divergence between two distributions is infinite if the do not share the same support, the minimizer of the KL divergence to the VB ideal should also converge to a point mass. A portion of the proof is included as Exercise A.1. For a full technical justification, see the appendix in [WB19b].

3.4 Main Result

Prior to stating the main result, we will discuss the chain of implications of the four lemmata above. In Lemma 3.1, we established the classical BvM result for the rescaled VB ideal, using an extension to misspecified models to account for the variational approximation. This is a precursor to the next two lemmata. We argued heuristically that this meant unscaled VB ideal converges to a point mass. In Lemma 3.2, we showed that the KL minimizer in the variational family to the VB ideal converges to the point mass at θ_0, i.e., it is consistent. Lemma 3.3 shows that the KL minimizer, or variational approximation to the scaled VB ideal and the variational approximation to a normal are equal in the limit in total variation. In other words, the result from Lemma 3.1 survives the variational approximation on both terms. Finally,

Lemma 3.4 states that the maximizer of the ELBO converges to the variational approximation to the VB ideal. Combining with Lemma 3.3 essentially gives the variational BvM result. We now state the main result.

Theorem 3.5 (Variational Bernstein von–Mises Theorem, Theorem 5 [WB19b]). *Let q^*_θ denote the density of the variational approximation Q^*_θ to the posterior distribution of the global latent variables. Suppose the technical conditions of Lemma 3.3 and Lemma 3.4 hold.*

1. *(Consistency) The following event occurs almost surely under the true probability measure P_{θ_0}*

$$Q^*_\theta(\theta) \to \delta_{\theta_0}$$

in distribution.

2. *(Asymptotic Normality) The VB posterior converges to the variational approximation of a normal distribution in total variation. Denote the scaled variational approximation by $\tilde{Q}_\theta$ with density $\tilde{q}_\theta(\tilde{\theta}) = q^*_\theta(\theta_0 + \delta_n\tilde{\theta})|det(\delta_n)|$. Then,*

$$\|\tilde{Q}_\theta - \arg\min_{Q^d} KL(q_\theta(\theta)\|N(\theta; \Delta'_{n,\theta_0}, V^{-1}_{\theta_0}))\|_{TV} \to 0$$

Recall that normal marginals are independent if and only if its covariance terms are 0. This special fact means that the mean-field variational family includes normal distributions, so long as the covariance matrix is diagonal. Indeed, we can show that

$$\arg\min_{Q^d} \mathrm{KL}(q_\theta(\theta)\|N(\theta; \Delta'_{n,\theta_0}, V^{-1}_{\theta_0})) = N(\theta; \Delta'_{n,\theta_0}, V'^{-1}_{\theta_0}) \tag{21}$$

where V'_{θ_0} is equal to $V_{\theta 0}$ with all off-diagonal terms set to 0. This means that the mean-field scaled VB posterior converges to an uncorrelated normal distribution. The proof of equation 21 can be found in ([Bis06], section 10.1).

Although we restricted the exposition to the mean-field family, the authors emphasize that the results hold as long as lemmata 3.1-3.4 are satisfied. In particular, lemmata 3.1 and 3.4 are general in that the proof does not assume the mean-field. Lemma 3.2 holds as long as the variational family at least includes the mean-field. In [WB19b], the proof for lemma 3.3 assumes the mean-field family, but it can also be shown for the full-rank Gaussian family. In other words, checking Theorem 3.5 for a variational family that is extended beyond the mean-field is equivalent to checking lemma 3.3.

4 Open questions and research directions

This report was an exposition on the Variational Bernstein von–Mises theorem, as initially proven in [WB19b]. Almost concurrently, the same authors extended the result to misspecified models, i.e., the model P_Θ does not include P_{θ_0} [WB19a]. The results here are appealing for a number of reasons. Bayesian statisticians have long used BvM-type results to justify the validity of their procedures asymptotically. Modern Bayesian inference is typically fit using Markov Chain Monte carlo (MCMC), which has nice theoretical properties in the limit, as samples from MCMC algorithms can be viewed as being from the exact posterior. The use of VI, which may never recover the exact posterior, is hence not as appealing for deriving theoretical guarantees. Nonetheless, for highly expressive models with large dimension and sample size, VI is much faster than MCMC and often the only feasible choice. Obtaining a theory for VI is a necessary precursor to obtain theoretical guarantees on these models, which have seen major applications in the real world.

The applicability of the results here are less obvious in the real world use-cases of VI. Specifically, it is unlikely that checking 3.3 will be straight forward for most models that applied users are interested in. This is not a fault of the variational BvM, in fact the original BvM often does not hold for many popular models, including many non-parametric Bayesian models. Much research is still being done to extend the standard BvM to these cases, and in many cases entire articles are dedicated to establishing these asymptotics even in model specific cases (see for example, [CN13] and [CR15], amongst many others). In this sense, the upcoming research directions are extremely vast for the variational BvM, and the open questions are plentiful.

A more focused research direction is one that is complementary to the recent theoretical developments of VI. [GBJ18] examines the infinitesimal sensitivity of mean-field VB to changes in the model and prior, and it would be natural to consider the consequences for asymptotics as well. There has also been an interest in designing alternative objectives for VI beyond the standard KL divergence and ELBO dual. For example, the α-divergence [LT16] and f-divergence [WLH20] are progressively more general families that include the KL divergence. All the results in [WB19b] technically rests on the KL objective, but the main ideas, even within each lemma, extend to any variational objective.

The VI optimization is often approximate in practice, using techniques such as automatic differentiation and stochastic optimization. It would be an interesting research item to examine lemma 3.3 under approximate solutions. The proof of this lemma already rests on the idea of Γ-convergence, which is a classical technique in optimization. It is not unreasonable to think that some results could be obtained in connection with modern developments in optimization.

We conclude that there are a number of future research directions not only for the asymptotic study of variational Bayes, but for the theoretical qualities of variational inference in general. In particular, there is value in showing classical qualities such as asymptotic normality to improve the applicability of modern techniques. [WB19b] and [WB19a] provide a strong start, but there is an entire literature of just BvM-type results to adapt to VI. At the same time, as VI continues to adapt to new problems, theoretical study needs to keep pace with these developments as well. We expect these theoretical advances to be plentiful in the near future.

A Exercises

A.1 Exercise 1

Exercise 1 is a portion of the proof of Lemma 4 in [WB19b]. Recall the parameter space is R^d, and latent space is R^n. Let γ_n be a $d \times d$ diagonal matrix with all entries $\gamma_{n,ii} \to 0$. Suppose the variational density for θ under the shifting and rescaling $\theta = \gamma_n^{-1}(\theta - \mu)$ can be expressed as

$$q(\theta) = |det\gamma_n|^{-1} q_h(\gamma_n^{-1}(\theta - \mu))$$

for some $\mu \in \mathbb{R}^d$, and such that $\int q_h(h)dh = 0$. Suppose that q_h has zero mean, i.e., $\int h q_h(h)dh = 0$. Now further suppose that $\int h^2 q_h(h)dh < \infty$, $\sup_{z,x} \| \log(p(z,x|\theta))'' \| \leq A q_h(\theta)^{-B}$ for positive constants A, B and some induced p-norm $p \in [1, \infty]$. To reduce clutter, denote $q_\theta(\theta)$ and $q_z(z)$ simply as $q(\theta)$ and $q(z)$, they are identified by their arguments. Prove that the first term in equation 20 satisfies

$$\int d\theta q(\theta) \left(-\log q(\theta) + \log \left[p(\theta) \exp \left[\sup_{Q^n} \int dz q(z) \log \left(\frac{p(x,z|\theta)}{q_z(z)} \right) \right] \right] \right)$$
$$\leq \int d\theta q(\theta) \log(p(\theta)) - \int d\theta q(\theta) \log q(\theta) + \sup_{Q^n} \int dz q(z) \log \frac{p(x,z|\mu)}{q(z)} + o(1)$$

where $o(1)$ refers to a term going to 0 as $n \to \infty$. Note this essentially shows that the supremum in the negative KL term can be bounded by the supremum in the profiled ELBO term in the limit. The rest of the proof of lemma 3.4 rests on showing a similar inequality for the profiled ELBO.

Solution:

$$\int d\theta q(\theta) \left(-\log q(\theta) + \log \left[p(\theta) \exp \left[\sup_{Q^n} \int dz q(z) \log \left(\frac{p(x,z|\theta)}{q_z(z)} \right) \right] \right] \right) =$$
$$\int d\theta q(\theta) \log p(\theta) - \int d\theta q(\theta) \log q(\theta) + \int d\theta q(\theta) \left[\sup_{Q^n} \int dz q(z) \log \left(\frac{p(x,z|\theta)}{q_z(z)} \right) \right]$$

Now, we simply require to show the inequality

$$\int d\theta q(\theta) \left[\sup_{Q^n} \int dz q(z) \log \left(\frac{p(x,z|\theta)}{q_z(z)} \right) \right] \leq \sup_{Q^n} \int dz q(z) \log \frac{p(x,z|\mu)}{q(z)} + o(1)$$

We manipulate the left side by considering the Taylor expansion in θ about μ, with a mean value $\theta^\dagger$ remainder

$$\int d\theta q(\theta) \left[\sup_{Q^n} \int dz q(z) \log \left(\frac{p(x,z|\mu)}{q(z)} \right) \right] + (\theta - \mu) \left[\sup_{Q^n} \int dz q(z) \log \left(\frac{p(x,z|\mu)}{q(z)} \right) \right]'$$
$$+ \frac{1}{2} (\theta - \mu)^T \left(\sup_{Q^n} \int dz q(z) \log \left(\frac{p(x,z|\theta^\dagger)}{q(z)} \right)'' (\theta - \mu) \right]$$

The first term is free of θ and may come out of the integral, leaving $\int d\theta q(\theta) = 1$. The second term is equal to 0 by subbing in $\tilde{\theta} = \gamma_n^{-1}(\theta - \mu)$:

$$\int d\theta q(\theta) \left((\theta - \mu) \left[\sup_{Q^n} \int dz q(z) \log \left(\frac{p(x,z|\mu)}{q(z)} \right) \right]' \right)$$
$$= \int d\theta |det\gamma_n|^{-1} q_h(\tilde{\theta}) \left(\gamma(\tilde{\theta}) \left[\sup_{Q^n} \int dz q(z) \log \left(\frac{p(x,z|\mu)}{q(z)} \right) \right]' \right) = 0$$

since the inner portion under the sup is free of θ, and the mean of q_h is assumed to be 0. This leaves the final term. Let $\|\cdot\|$ represent some vector p-norm $p \in [1, \infty]$ and the associated induced matrix norm. Then,

$$\begin{aligned}
&\int d\theta q(\theta) \left[\frac{1}{2}(\theta - \mu)^T \left(\sup_{Q^n} \int dz q(z) \log \left(\frac{p(x, z|\theta^\dagger)}{q(z)} \right) \right)'' (\theta - \mu) \right] \\
&\leq \int d\theta \frac{1}{2} q(\theta) \left[\|\theta - \mu\|^2 \sup_{x,z} \left\| \sup_{Q^n} \int dz q(z) \log \left(\frac{p(x, z|\theta^\dagger)}{q(z)} \right)'' \right\| \right] \\
&\leq A \int d\theta q(\theta) \|\theta - \mu\|^2 q_h(\theta)^{-B}
\end{aligned}$$

by a basic inequality for quadratic forms and the assumption. Proceeding by the same substitution $\tilde{\theta}$,

$$\begin{aligned}
&A \int d\theta q(\theta) \|\theta - \mu\|^2 q_h(\theta)^{-B} \\
&= C \int q_h(\tilde{\theta}) \|\gamma_n \tilde{\theta}\|^2 q_h(\mu + \gamma_n \tilde{\theta})^{-B} \\
&\leq C' \max_i (\gamma_{n,ii}^2)
\end{aligned}$$

Since we assumed that $\int h^2 q_h(h) dh < \infty$. Now, since $\gamma_{n,ii} \to 0$ for all i, we have that the third term also goes to 0, i.e., it is equal to $o(1)$. To recap, we showed that the first term of the expansion is free of θ and hence equal to just

$$\sup_{Q^n} \int dz q(z) \log \left(\frac{p(x, z|\mu)}{q_(z)} \right)$$

while the second term was equal to 0, and the third term is dominated by a term that is $o(1)$. This shows the desired inequality.

A.2 Exercise 2

This exercise was conceived without external references. Recall from elementary probability that the joint density of a Bayesian model with data x, global latents θ and local latents z can be written as:

$$p(x, z, \theta) = p(x|z, \theta) p(z|\theta) p(\theta)$$

In this case, we can write the density for the posterior distribution $P_{\Theta, Z|X}$ succinctly as

$$p(\theta, z|x) \propto p(x|z, \theta) p(z|\theta) p(\theta)$$

by integrating out the local latents z, we have

$$p(\theta|x) \propto p(x|z, \theta) p(\theta)$$

Show that the KL divergence between this posterior and the VB ideal $KL(p(\theta|x) \| \pi^*(\theta|x))$ is dependent on the expected (under $p(\theta|x)$) quality of the expected local variational approximation ($\log q_z^\dagger$) to $\log p(z|x, \theta)$ (i.e., $\mathbb{E}_q^\dagger[\log q_z^\dagger(z)] - \log p(z|x, \theta)$).

Note: This exercise merely wants to show that the quality of the local approximation impacts the KL divergence, not that it is the only term.

Solution Note that $p(x|z,\theta) = \frac{p(x,z|\theta)}{p(z|x,\theta)}$. Also note that the denominators in the posterior and VB ideal are constants in θ. Denote by expectation over θ with respect to the posterior $P_{\Theta|x}$ as $E_p[\cdot]$. Then, the KL divergence can be written as:

$$\begin{aligned} KL(p(\theta|x)\|\pi^*(\theta|x)) &= \mathbb{E}_p[\log(p(x|z,\theta)p(\theta)) - \log(p(\theta)\exp(M_n(\theta;x)))] + C \\ &= \mathbb{E}_p[\log p(x|z,\theta) - M_n(\theta;x)] + C \\ &= \mathbb{E}_p[\log p(x,z|\theta) - \log p(z|x,\theta) - \max_{Q^n} \mathbb{E}_{q(z)}[\log p(x,z|\theta) - \log q_z(z)]] + C \end{aligned}$$

Now, recall that we denote the arg max density in the inner expectation $q_z^\dagger$. Then,

$$\begin{aligned} KL(p(\theta|x)\|\pi^*(\theta|x)) &= \mathbb{E}_p[\log p(x,z|\theta) - \log p(z|x,\theta) - \mathbb{E}_{q(z)}[\log p(x,z|\theta) - \log q_z^\dagger(z)]] + C \\ &= \mathbb{E}_p[(\log p(x,z|\theta) - \mathbb{E}_q^\dagger[\log p(x,z|\theta)])] + \mathbb{E}_p[\mathbb{E}_q^\dagger[\log q_z^\dagger(z)] - \log p(z|x,\theta)] + C \end{aligned}$$

where the second term measures the quality of the approximation.

References

[Bis06] C. M. Bishop. In: *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006, pp. 466–467.

[BKM17] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112 (2017), pp. 859–877.

[CN13] I. Castillo and R. Nickl. "Nonparametric Bernstein–von Mises theorems in Gaussian white noise". In: *Annals of Statistics* 41 (2013), pp. 1999–2028.

[CR15] I. Castillo and J. Rousseau. "A Bernstein–von Mises theorem for smooth functionals in semiparametric models". In: *Annals of Statistics* 43 (2015), pp. 2353–2383.

[Doo49] J. L. Doob. "Application of The Theory of Martingales". In: *Actes du Colloque International Le Calcul des Probabilités et ses applications*. 1949.

[GBJ18] R. Giordano, T. Broderick, and M. I. Jordan. "Covariances, Robustness and Variational Bayes". In: *J. Mach. Learn. Res.* 19.1 (2018), pp. 1981–2029.

[KV12] B. Kleijn and A. van der Vaart. "The Bernstein-Von-Mises theorem under misspecification". In: *Electron. J. Statist.* 6 (2012), pp. 354–381.

[LT16] Y. Li and R. E. Turner. "Rényi Divergence Variational Inference". In: *Advances in Neural Information Processing Systems 29*. 2016, pp. 1073–1081.

[Sch95] M. J. Schervish. In: *Theory of Statistics*. Springer Series in Statistics. Springer-Verlag New York, 1995, pp. 1–24.

[Vaa00] A. van der Vaart. "Asymptotic Statistics". In: *Asymptotic Statistics*. Cambridge Series in Statistical and Probablistic Mathematics. Cambridge University Press, 2000, pp. 138–152.

[WLH20] N. Wan, D. Li, and N. Hovakimyan. *f-Divergence Variational Inference*. 2020. arXiv, to appear in NeurIPS 2020: 2009.13093 (cs.LG).

[WB19a] Y. Wang and D. Blei. "Variational Bayes under Model Misspecification". In: *Advances in Neural Information Processing Systems 32*. 2019, pp. 13357–13367.

[WB19b] Y. Wang and D. M. Blei. "Frequentist Consistency of Variational Bayes". In: *Journal of the American Statistical Association* 114.527 (2019), pp. 1147–1161.