Paper 7343 | Variational Bayes under Model Misspecification

We thank the reviewers for their positive and constructive comments. All reviewers agree that characterizing variational Bayes under model misspecification is an interesting addition to the theory of variational Bayes literature. We are glad that the reviewers also appreciate the clear and intuitive explanations of technical results in this work, which could serve pedagogical purposes to the community. Below we respond to the main comments.

R1 finds the presentation in Section 2.2 and Assumptions 4 & 5 in Section 2.3 repetitive.

Thank you for pointing it out. We will tighten up the explanation to make Theorem 2 clearer and more intuitive. We will also move Assumptions 4 & 5 into the main text to make Section 2.3 clearer.

R1 points out that the LAN assumption might not be satisfied in nonparametric models.

Thank you for pointing it out. There are a few nonparametric models that have been shown to satisfy the LAN assumption, including generalized linear mixed models (Hall et al., 2011), stochastic block models (Bickel et al., 2013), and mixture models (Westling & McCormick, 2015). In Section 2.4, we apply Theorems 1 and 2 to these specific models and characterize their VB posteriors under different forms of model misspecification.

That said, we agree that the Bernstein-von Mises phenomenon does not hold in many nonparametric models with infinite dimensional parameters (Freedman, 1999). In these models, there is no posterior contraction or in-fill asymptotics for either the exact posterior or the VB posterior. We will clarify this limitation in the paper.

R2 is concerned about the practical relevance of Bernstein-von Mises type results.

Thank you for pointing it out. We take the asymptotics perspective as a first step to understand the theoretical properties of VB posteriors and VB posterior predictive distributions. We were motivated by the empirical observation that variational Bayes predicts comparably with MCMC methods in large datasets. The results in this paper around the VB posterior predictive under model misspecification offers an explanation of this phenomenon. That said, we understand that Bernstein-von Mises type results can appear limited when the optimization complications of VB come in. We leave to future work the characterizations of how variational Bayes behaves in finite samples and how optimization complications affect the VB posterior.

R3 asks about how local optima in ELBO optimization fits into this story.

We agree with R3 that local optima is a real practical issue in VB. We will add a discussion about local optima in the paper. The results in this work assume that the ELBO optimization returns a global optima. These results provide the possibility for local optima to share these properties, though further research is needed to understand the properties of local optima. For particular models like stochastic block models, Zhang and Zhou (2017) shows that global optima of the ELBO can be reached under weak conditions of optimization initialization. We believe that combining this work with optimization guarantees could lead to a fruitful further characterization of variational Bayes.

R3 are interested in seeing more details about the simulations, in particular the code, HMC mixing issues, VB optimization issues, error bars, and higher maximum N values.

Thank you for the questions. We will include the figure-generation scripts in addition to current Stan code in the final version. Regarding practical HMC mixing and VB optimization issues, we follow the protocol as implemented in Stan. For HMC, we run four parallel chains and use 10,000 burn-in samples, and determine mixing using the R-hat convergence diagnostic (R-hat<1.01). For variational Bayes, we run optimization until convergence (i.e. a local optimum). We cannot confirm if the local optimal we reached is global. Further, we conduct multiple parallel runs under each simulation setup and report the mean and the standard deviation of "RMSE" or "Mean KL". The error bars in Figure 2 are the standard deviation across different runs of the same simulation. The simulation in the paper was conducted on a fairly small model, so the practical complications of HMC and VB did not appear detrimental. We agree that these practicalities are important, we will clarify these complications in the paper. We will also include higher maximum N values to approximate the infinite data limit closer.

R3 asks about how large a dataset should be to approximate the infinite data limit.

Thank you for the interesting question. When a model has more parameters or its convergence rate (δ_n in the LAN assumption) is lower, we should require a larger dataset to approximate the infinite data limit.

R3 asks about an example prior that fails the tail condition.

Extreme value distributions like Gumbel distribution can fail this tail condition.