Variational Inference: "Does it work?"

Chris Stanton, Anna Menacher and Sahra Ghalebikesabi
Imperial College London and University of Oxford

Abstract

Whilst many diagnostic approaches exist for evaluating the accuracy of Markov Chain Monte Carlo (MCMC) posteriors, relatively few exist for Variational Inference (VI). In this paper, we summarize three recently proposed diagnostics for VI and illustrate their utility in a series of both parametric and non-parametric examples with real world data.

1 Introduction

Consider the classical Bayesian inference problem of estimating the posterior distribution of D dimensional parameter $\theta \in \Theta$, given data $\mathbf{x} \in \mathcal{X}$:

$$\pi(\theta) = p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})} \tag{1}$$

In the majority of practical use cases, the posterior is intractable, as one is unable to derive the marginal likelihood $p(\mathbf{x})$. An approximation of the posterior $\hat{\pi}$ is therefore required to conduct Bayesian inference. Variational Inference (VI) has gained popularity as an alternative to Markov chain Monte Carlo (MCMC) for generating this numerical approximation, as it is able to produce posterior estimates significantly faster, and thus scales better to more complex models. Many diagnostics exist for assessing the quality of posterior samples generated using MCMC: commonly used examples include sample autocorrelation, expected sample size and the Gelman-Rubin diagnostic (Gelman and Rubin, 1992). However, few diagnostics have been proposed for VI, and there is no consensus in the literature as to which of the proposed diagnostics are the most robust. In this report, we present and implement 4 different diagnostic approaches to variational inference that have recently been proposed on a real world dataset. We apply the diagnostics in both a parametric (Bayesian linear regression) and non-parametric (Gaussian Process) setting. The remainder of the paper is structured as follows: in section 2, we briefly introduce variational inference. In section 3, we outline the diagnostics that we will be evaluating in this paper. In section 4, we present practical implementations of these diagnostics on a real world dataset. In section 5, we compare these diagnostics in concluding remarks.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

2 Variational Inference

The approximation of the posterior $\hat{\pi}(\theta)$ is assumed to the member of the variational family of densities $\mathcal{Q}$ that minimises the Kullback-Leibler (KL) divergence between p and q:

$$\hat{\pi} = \underset{q \in \mathcal{Q}}{\operatorname{argmin}}\ D_{KL}(q(\theta)||\pi(\theta)) \tag{2}$$

The KL divergence between q and π cannot be evaluated exactly, as π is intractable. However, it is possible to evaluate the evidence lower bound (ELBO), given by:

$$\text{ELBO}(q) = \text{E}_q[\log(\pi(\theta))] - \text{E}_q[\log(q(\theta))] \tag{3}$$

The ELBO and the KL divergence between π and q are related by:

$$\text{ELBO}(q) = -D_{KL}(q(\theta)||\pi(\theta)) + \log(p(\mathbf{x})) \tag{4}$$

Therefore, minimising the KL divergence is equivalent to maximising the ELBO. As the KL divergence is always positive, the ELBO is a lower bound to the log-marginal likelihood (also known as the 'evidence'), which is used in Bayesian inference problems for model selection.

Selection of an appropriate $\mathcal{Q}$ for a particular inference problem is a trade-off between having a space that encompasses a large set of possible posteriors, such that $\mathcal{Q}$ contains densities that would be a reasonable representation of π, and having an optimisation problem that is sufficiently simple. Much of the research and

applications of variational inference assume the mean-field variational family for $\mathcal{Q}$:

$$\mathcal{Q} = \{q : q(\theta) = \prod_{i=1}^{D} q_i(\theta_i)\} \tag{5}$$

Imposing the condition of independence on the components of θ reduces the complexity of the optimisation problem, however if there are significant correlations between the components in the posterior, any density approximation from this variational family will not capture the dynamics of the posterior distribution.

Under the mean-field assumption, iterative updates to the marginal densities of each of the components of theta, or co-ordinates, can be derived as:

$$q_i(\theta_i) \propto \exp[E[\log(p(\theta, \mathbf{x})|\theta_i]] \tag{6}$$

A scheme that applies this update sequentially to each component and then repeats the process, is known as co-ordinate ascent variational inference (CAVI) (Blei et al., 2017). These updates are guaranteed to monotonically update the posterior, however the individual updates may not be analytically tractable or may require tedious calculations for each model. Stochastic gradient variational inference scales the co-ordinate ascent approach to modelling problems with large datasets by considering stochastic approximations of the natural gradient of the ELBO (Hoffman et al., 2013). Black box variational inference approaches have been proposed in the literature that are able to automatically implement variational inference without any tuning: Automatic Differentiation Variational Inference (ADVI) is arguably currently the most popular of these (Kucukelbir et al., 2017). The method transforms each component of the parameter such that its support is R. A Gaussian mean-field $\mathcal{Q}$ is then assumed on the transformed parameters. Updates are then made by conducting gradient descent on stochastic approximations of the ELBO and its gradient.

3 Diagnostics

In this section, we outline the VI diagnostics that we will be implementing on real world data. It is important to note from the outset that these diagnostics are to some extent not directly comparable, as they assess variational inference in different ways. For example, PSIS, through estimating the parameters of the generalised Pareto distribution of the largest importance weights, assesses the quality of a posterior estimate for a particular implementation of variational inference problem. However, VSBC considers whether variational inference is a reasonable approach for any possible set of responses simulated from the prior predictive distribution of a model. KLVI and CHIVI lie within a more rigorous framework, assessing whether the L_2 norm of the difference between estimators of relevant statistics of the posterior distribution derived from the VI approximation and the actual values of those statistics for the true posterior distribution are reasonable.

3.1 Pareto Smoothed Importance Sampling

Pareto Smoothed Importance Sampling (PSIS) (Yao et al., 2018) (Vehtari et al., 2015) both assesses whether a particular density approximation $\hat{\pi}$ to the true posterior π is reasonable, and provides an approach to estimating integrals under the true posterior with an optimal trade-off between bias and variance. If one aimed to estimate the integral $\mathrm{E}_\pi[h(\theta)]$, one could sample directly from $\theta^i \sim \hat{\pi}$ to obtain a Monte Carlo estimate:

$$\mathrm{E}_\pi[h(\theta)] \approx \frac{1}{S}\sum_{i=1}^{S} h(\theta^i) \tag{7}$$

however such an estimate would be both biased and inconsistent. One could use importance sampling to obtain a consistent estimator that directly samples from the posterior distribution. As the posterior weights are only known up to an additive constant, this necessitates a self-normalising importance sampling approach. Again, with $\theta^i \sim \hat{\pi}$:

$$\mathrm{E}_\pi[h(\theta)] \approx \frac{\sum_{i=1}^{S} w_i h(\theta^i)}{\sum_{i=1}^{S} w_i} \tag{8}$$

$$w_i = \frac{p(\theta^i, \mathbf{x})}{\hat{\pi}(\theta^i)} \tag{9}$$

This estimator achieves asymptotic consistency, however the variance of this estimator is dependent on the variance of the importance weights. Yao et al. (2018) argue that an estimator that remains asymptotically consistent can be obtained by 'smoothing' the weights. New weights r_i are obtained by taking the largest $\min(S/5, 3\sqrt{S})$ weights w_i, fitting the generalised Pareto distribution to the importance weights, and replacing the largest $\min(S/5, 3\sqrt{S})$ weights and replacing them with the first moment of the generalised Pareto distribution, based on the fitted parameters. The new importance weights r_i have an upper bound imposed of $\sup\{w_i, i \in \{1, \dots S\}\}$ The generalised Pareto distribution with shape parameter $k > 0$ and location-scale parameters (μ, σ), is given by:

$$p(y|k, \mu, \sigma) = \frac{1}{\sigma}(1 + k(\frac{y-\mu}{\sigma}))^{-1-\frac{1}{k}} \tag{10}$$

The Pareto smoothed importance sampling estimator is thus given by:

$$\mathrm{E}_\pi[h(\theta)] \approx \frac{\sum_{i=1}^{S} r_i h(\theta^i)}{\sum_{i=1}^{S} r_i} \tag{11}$$

Extreme value theory formalises a link between the value of k for the extreme value distribution of a random variable, and the threshold between finite and infinite moments for that random variable. Particularly:

$$k = \inf\{c \in R : \mathrm{E}_{\hat{\pi}}[(\frac{p(\theta, \mathbf{x})}{\hat{\pi}(\theta)})^{\frac{1}{c}}] < \infty\} \tag{12}$$

Taking the logarithm of the integral $\mathrm{E}_q[(\frac{p(\theta,\mathbf{x})}{q(\theta)})^{\frac{1}{c}}]$ and applying a linear transformation gives the Rényi divergence (Rényi, 1961) $D_\alpha(\pi||\hat{\pi}(\theta))$:

$$\frac{k}{1-k}\log(\mathrm{E}_{\hat{\pi}}[(\frac{p(\theta,\mathbf{x})}{\hat{\pi}(\theta)})^{\frac{1}{c}}]) - \frac{1}{1-k}\log(p(\mathbf{x})) = D_{\frac{1}{k}}(\pi||\hat{\pi}) \tag{13}$$

Thus, the value of $\hat{k}$ is indicative of the values of α at which the Rényi divergence is finite. Smaller values of k indicate finiteness of the Rényi divergence for higher values of α, and therefore suggesting that q is a reasonable approximation to the posterior distribution. In particular, $k = 1$ implies that KL divergence is itself infinite, indicating that the VI approximation $\hat{\pi}(\theta)$ is totally unreliable. Yao et al. (2018) propose that as a heuristic, $\hat{k} < 0.7$ indicates a sufficiently reliable VI approximation, whereas $k > 0.7$ indicates an unreliable approximation. Note that $\hat{k} < 0.5$ implies that second moments of the importance weights exists, and thus the central limit theorem holds for both the IS and the PSIS approaches, resulting in fast convergence rates.

3.2 Variation Simulation Based Calibration

Variation Simulation Based Calibration (VSBC) (Yao et al., 2018) assesses whether for a defined Bayesian model, a particular approach to VI approximation performs well in approximating the posterior $\pi(\theta)$ for any possible simulated data $\mathbf{x}$ generated from the prior predictive $p(\mathbf{x})$. Given a defined Bayesian model, with D dimensional parameter $\theta = (\theta_1, \ldots, \theta_D)^T$ the following steps are conducted:

1) Generate parameter θ^i from the prior $p(\theta)$
2) Generate dataset $\mathbf{x}^i$ from the likelihood $p(\mathbf{x}|\theta^i)$ - the resulting values $(\theta^i, \mathbf{x}^i)$ are a sample from the joint distribution $p(\mathbf{x}^i, \theta)$, and therefore θ^i is a sample from $p(\theta|\mathbf{x}^i)$
3) Use a VI approach to approximate $p(\theta|\mathbf{x}^i)$ with $\hat{\pi}_i(\theta)$
4) Generate sufficiently large S samples of $\theta^{ij} \sim \hat{\pi}_i(\theta)$
5) For each parameter component, record $U_d^i = \hat{F}_d^i(\theta_d^i)$, where $\hat{F}_d^i(c) = \frac{1}{S}\sum_{j=1}^{S} I[\theta_d^{ij} \leq c]$

It is easy to show that if the VI approximation $\hat{\pi}_i(\theta)$ to posterior $p(\theta|\mathbf{x}^i)$, the random variables $U_d^i, d \in \{1, \ldots, D\}$ converge to the uniform distribution as $S \to \infty$ (Cook et al., 2006). The above set of steps are run numerous times, to generate many realisations of U_d^i. The resulting samples in each dimension can then be tested for uniformity, both quantitatively with the Kolmogorov-Smirnov test, and qualitatively with histograms.

3.3 Posterior Error Bounds on Variational Objectives

The error bounds on posterior quantities as proposed by Huggins et al. (2019) introduce a computationally efficient post-hoc accuracy measure with theoretical guarantees for posterior approximations by variational inference. The only requirement for the theoretical justification of this evaluation tool is the existence of polynomial moments in the exact and approximating posterior distribution.

Current evaluation tools for variational inference, such as the PSIS diagnostic and VSBC as suggested by Yao et al. (2018), have various disadvantages. For instance, the PSIS diagnostic $\hat{k}$ has merely heuristic suggestions in terms of the acceptance range of the value of $\hat{k}$ which do not have any theoretical underlying guarantees. Additionally, the PSIS diagnostic can have a low value $\hat{k}$ even though the posterior approximation is poor. Furthermore, the VSBC measure is not feasible to compute in most cases as it requires the resampling of multiple variational approximations which can easily be very computationally complex. Other metrics for the evaluation of variational inference often face intractable constants or require impractically strong assumptions on the tail behavior of the specified variational distributions. Moreover, variational inference with common objectives, such as KL-divergence or α-divergence, can yield a small value for the discrepency measure while the error of the posterior quantities can be arbitrarily large.

On the other hand, the error bounds on the posterior mean and other uncertainty quantities, such as the covariance or component marginal standard deviation, by Huggins et al. (2019) are computationally efficient as they only require the calculation of the specified bounds. The bounds are simply computed by using results from the VI procedure, analytic calculations and Monte Carlo estimates.

The new general workflow for variational inference by Huggins et al. (2019) proposes the application of VI with the incorporation of the evaluation of the VI approximated posterior quantities by bounds on the 2-divergence $\bar{\delta}_2$ and on the 2-Wasserstein distance $\bar{w}_2$.

Moreover, the workflow provides suggestions for improving the posterior approximations dependent on the error bound values $\bar{\delta}_2$ and $\bar{w}_2$.

1) Select a variational family $\mathcal{Q}$ with sufficiently heavy tails which has a true $k \leq 0.5$.
2) Minimize a discrepancy measure to find a variational approximation $\hat{\pi}$.

   → **KLVI**: maximizing ELBO $\hat{=}$ minimizing KL-divergence

   → **CHIVI**: minimizing CUBO $\hat{=}$ minimizing α-Rényi divergence
3) Compute $\hat{k}$,

→ **if** there is no guarantee that $k \leq 0.5$

→ **if** $\hat{k} > 0.5$,

then refine $\mathcal{Q}$ or reparameterize the model.

4) Compute ELBO($\hat{\pi}$) and $\text{CUBO}_2(\hat{\pi})$.
5) **(optional)** Further optimize the ELBO(ξ).
6) Compute bound on α-divergence $\bar{\delta} \geq \text{D}_2(\pi|\hat{\pi})$.
7) Compute bound on p-Wasserstein distance $\bar{w}_2 \geq \mathcal{W}_2(\pi, \hat{\pi})$.
8) **If** $\bar{\delta}_2 \uparrow$ and $\bar{w}_2 \uparrow$, **then** refine $\mathcal{Q}$ or reparameterize the model.
9) **If** $\bar{\delta}_2 \downarrow$ and $\bar{w}_2 \uparrow$, **then** use IS or PSIS to refine the posterior expectations produced by $\hat{\pi}$.
10) **If** $\bar{\delta}_2 \downarrow$ and $\bar{w}_2 \downarrow$, **then** use $\hat{\pi}$ to approximate π.

Overall, the main advantage of posterior error bounds of variational objectives is the computational efficiency which is usually the reason for choosing variational inference as an approximation method to compute a posterior distribution. However, it should also be mentioned that this method imposes only weak tail restrictions on the selection of variational distributions. Lastly, it focuses on the quantities of interest directly, such as the mean and various uncertainty metrics, and not on the traditional comparison of the approximated and exact posterior distributions by bounding their divergence.

4 Application Study

In the following, we implement the presented diagnostics for a parametric and a non-parametric regression on a real-world data set. Our code is also available on https://github.com/sghalebikesabi/does_VI_work.

4.1 Dataset

For our analysis, we have used a dataset of 442 diabetes patients, first seen in (Efron et al., 2004). The response variable is a standardised quantitative measure of the progression of the disease. The features of the dataset that we use in our modelling approach are age, sex, body-mass index (BMI) and average blood pressure.

4.2 Model Setup

4.2.1 Bayesian Linear Regression

For the analyis, we fit a Bayesian linear regression model to the diabetes dataset, defined as followed by the specification of the prior distributions for the parameters α and β, as well as the standard deviation σ, and the likelihood:

Prior distributions

$$p(\alpha) = \mathcal{N}(\alpha;\ 0,\ 10)$$
$$p(\beta) = \mathcal{N}(\beta;\ 0,\ 1)$$
$$p(\sigma) = Gamma(\sigma;\ 1,\ 1)$$

Likelihood

$$\mu_i = \alpha + \beta_{age} \cdot \text{age}_i + \beta_{sex} \cdot \text{sex}_i + \beta_{bmi} \cdot \text{bmi}_i + \beta_{bp} \cdot \text{bp}_i$$

$$p(\mathbf{y}|\alpha, \beta) = \mathcal{N}(y; \mu, \sigma^2)$$

The variational inference method for the evaluation with the diagnostic tools PSIS and VSBC defined in Section 3.1 and 3.2 is ADVI (Kucukelbir et al., 2017) and with the posterior error bounds is Stochastic Mean-Field Variational Inference (Hoffman et al., 2013).

Furthermore, the chosen variational family $\mathcal{Q}$ is Gaussian for consistency with predefined initializations. It should be noted that Huggins et al. (2019) select a combination of t-distributions with degrees of freedom of 40 or 100 as their variational distribution in order to ensure that the true k is smaller than 0.5 which satisfies the assumption on the tail behaviour of the variational distributions.

Variational Family:

$$q(\alpha) = \mathcal{N}(\alpha;\ \mu_\alpha,\ \sigma_\alpha)$$
$$q(\beta) = \mathcal{N}(\beta;\ \mu_\beta,\ \sigma_\beta)$$
$$q(\sigma) = \mathcal{N}(\sigma;\ \mu_\sigma,\ \sigma_\sigma)$$

Initialisation:

$$q(\alpha) = \mathcal{N}(\alpha;\ 0,\ 1)$$
$$q(\beta) = \mathcal{N}(\beta;\ 0,\ 1)$$
$$q(\sigma) = \mathcal{N}(\sigma;\ 1, 0.05)$$

4.2.2 Gaussian Process Regression

For the non-parametric approach, we fit a Gaussian process regression to the diabetes dataset. The key difference is that it is no longer the aim to fit the variational parameters of the variational distribution put on a parameter of interest. The goal is to fit the mean vector and the covariance matrix of the Gaussian process prior on the function f by specifying a Gaussian variational distribution over f and optimise in regards to its parameters. The model setup is therefore as follows:

Prior distributions and Likelihood:

$$y_i = f(x_i) + \epsilon_i$$

where

$f \sim \mathcal{GP}(0, K)$,

K is the squared exponential kernel,

$\epsilon_i \sim$ Student-t(df)

Variational Inference Method: Variational Gaussian approximation (Opper and Archambeau, 2009)

Variational Family: Gaussians

$$q(f) = N(\mu_f, K_f)$$

We maximise the hyperparameters and the variational parameters iteratively using Adam for the former and natural gradients for the latter.

4.3 Results

Before implementing the parametric and non-parametric models on the diabetes dataset, we implemented VSBC to determine whether the VI approaches in both the parametric and non-parametric cases, perform well on for a response $\mathbf{x}$, sampled from the prior predictive. We are also able to combine the PSIS approach with the VSBC approach, by sampling from the following particle density approximations:

- $\hat{\pi}_{IS}(\theta) = \sum_{i=1}^{S} \tilde{w}_i \delta(\theta^i)$, with $\tilde{w}_i = \frac{w_i}{\sum_{i=1}^{S} w_i}$
- $\hat{\pi}_{PSIS}(\theta) = \sum_{i=1}^{S} \tilde{r}_i \delta(\theta^i)$, with $\tilde{r}_i = \frac{r_i}{\sum_{i=1}^{S} r_i}$

instead of just sampling for the variational approximation to the posterior $\hat{\pi}$. The results in the parametric case, for the parameters β_{age} and σ are presented in figure X. Whilst the variational approximation appears

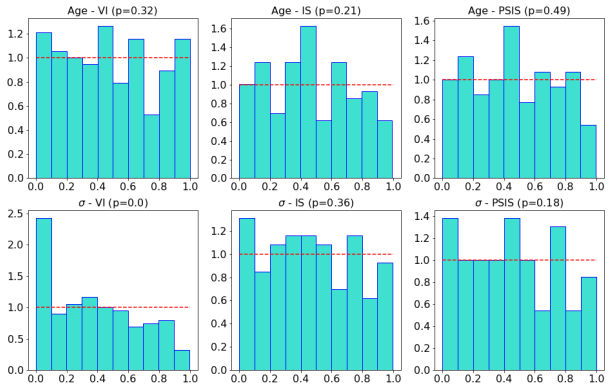

Figure 1: Investigating VI for the parametric model: PSIS with VSBC. P-values for each plot are for the two-sided K-S test for uniformity.

to be reasonable for β_{age}, it is clearly unreasonable for σ, suggesting that the variation approximation does not accurately infer the posterior distribution of this parameter. By sampling from $\hat{\pi}_{IS}(\theta)$ and $\hat{\pi}_{PSIS}(\theta)$ instead of $\hat{\pi}(\theta)$, the accuracy of the approximation of the marginal posterior of σ is clearly improved. For the non-parametric model, we apply VSBC (with an importance sampling adjustment) to the first 3 realisations of the Gaussian process: $f(\mathbf{x}_1), f(\mathbf{x}_2), f(\mathbf{x}_3)$. The results are shown in figure 2. For all 3 points at which the GP is realised, the posterior distribution at these points appears to be reasonably approximated by the variational Gaussian approximation. This analysis restricts to the fit of the first three patients in the data set. A more accurate test should perform VSBC on all patients which would however take several weeks to run on a 2.9GHz Dual-Core Intel Core i5 machine.

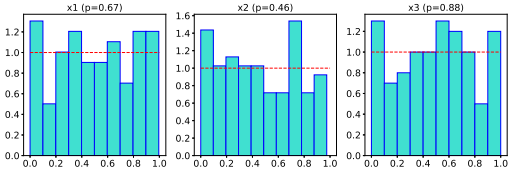

Figure 2: Investigating VI for the non-parametric GP model: VSBC. P-values for each plot are for the two-sided K-S test for uniformity.

After testing the accuracy of VI posteriors under the specified model, we implemented the framework outlined in section 3.3, to implement VI on the diabetes dataset, attaining posterior bounds for variational objectives. We run two separate implementations of the framework:

- **KLVI:** SVI with KL-divergence as variational objective

- **CHIVI:** SVI with 2-Rényi divergence as variational objective

We compare the resulting posteriors to the MCMC method Hamiltonian Monte Carlo (HMC) in combination with the No-U-Turn sampler (Hoffman and Gelman, 2011). The marginal posterior distributions of the parameters are then presented in Figure 3. It should be noted that the HMC approach is treated as the ground truth in this scenario. Hence, the comparison shows that the two VI approximations with the KL-divergence and the 2-Rényi divergence as a variational objectives indicate fairly good fits of the posterior in terms of the location of the posterior mean. However, the standard deviation of the marginals is quite skewed which indicates a poor approximation through VI.

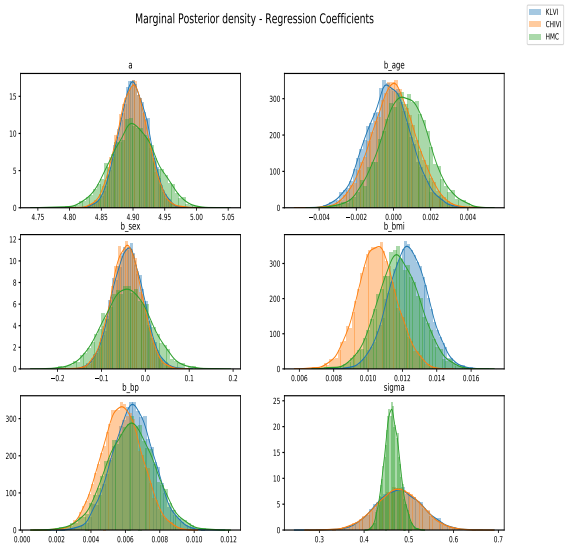

Figure 3: KLVI vs. CHIVI vs. HMC

We also implemented KLVI for the Gaussian process regression. Table 1 hereby shows the error bounds on the mean and standard deviation error through the 2-divergence bound D_2 and the 2-Wasserstein distance bound $\mathcal{W}_2$ for the KLVI and CHIVI implementations for the Bayesian linear regression, and the KLVI for the Gaussian process regression.

We note that the W2 bound and the mean error are the same across all models. This is supposed to result from the theoretical guarantees provided by the Wasserstein bounds in Huggins et al. (2019). Conclusively, the VI approximations are quite poor as is indicated by the high values for the D_2 bound and the $\mathcal{W}_2$ bound for KLVI and CHIVI. In this scenario, step 8) of the variational workflow in Section 3.3 needs to be applied

	KLVI	CHIVI	GP KLVI
D_2 bound	9.84	11.50	0.07
$\mathcal{W}_2$ bound	1.64	2.48	4.25
mean error	1.64	2.48	4.25
std error	3.16	4.79	8.22

Table 1: Error Bounds on Posterior Quantities

and either the variational family $\mathcal{Q}$ should be refined or the model should be reparameterized. On the other hand, the GP KLVI has a value lower than 0.7 for the D_2 bound with 0.07 as shown in Table 1. However, the $\mathcal{W}_2$ has no universal scale and hence the level of desired accuracy is dependent on the data. The KLVI and CHIVI values for the 2-Wasserstein bound are 1.64 and 2.48 respectively whereas the GP KLVI method has a value of 4.25 for this bound which concludes that step 9) of the workflow should be applied and the posterior expectations should be refined by either importance sampling or PSIS.

5 Discussion

In this paper, we presented diagnostics for variational inference that were introduced in the papers by Yao et al. (2018) and Huggins et al. (2019) and furthermore assessed their performance in a parametric and non-parametric setting on the so-called diabetes dataset in the form of a real-world application.

Whilst PSIS and the posterior error bounds framework assess the accuracy of a VI approximation to the true posterior for a particular observed dataset $\mathbf{x}$, VSBC assesses the average performance of VI on average over numerous implementations of VI on datasets $\mathbf{x^i}$ generated from the prior predictive distribution, for a specified Bayesian model. When using a particular VI algorithm, applying VSBC is a useful tool for validating the functionality of your algorithm before applying the VI approximation to your dataset. However, VSBC requires repeated generation from the specified Bayesian model, and repeated application of VI to approximate the posterior. As such, the approach does not scale well to high-dimensional, complex problems. Further, the VSBC approach would not indicate that VI would perform well for a particular modelling problem, if model chosen for that problem was misspecified. That is, if it were highly unlikely that the observed data $\mathbf{x}$ could possibly have been generated from the specified Bayesian model.

The diagnostic $\hat{k}$ given by PSIS allows for statements to be made regarding the quality of a VI approximation to the true posterior for a particular modelling problem, based on a simple statistic. However, the decision bounds seem to be chosen arbitrarily. It is ques-

tionable whether a VI approach with $\hat{k} = 0.69$ should be followed, while another with a $\hat{k} = 0.71$ is considered not acceptable and as a poor approximation for a posterior distribution.

While $\hat{k}$ only provides an approximation of a lower bound of α for which the Rényi divergence $D_\alpha(\pi, \hat{\pi})$ is finite, the bounds computed by Huggins et al. (2019) offer a tool to evaluate the posterior quantities of a VI approximation directly by calculating their error bounds instead of comparing the discrepancy between the exact and approximating posterior. At the same time, the provided posterior error bounds are computationally more efficient than VSBC as they only require quantities from the VI approximation. Furthermore, they have few requirements and only enforce weak tail restrictions on the variational family in comparison the other existing diagnostic measures with theoretical gurantees on the error bounds.

All in all, we recommend to apply the proposed variational workflow by Huggins et al. (2019) for assessing the performance of VI models as it only requires the calculation of error bounds and does not rely on heuristics for making decisions on the quality of variational approximation.

References

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112:855–877.

Cook, S., Rubin, D., and Gelman, A. (2006). Validation of software for bayesian models, using posterior quantiles. *Journal of Computational and Graphical Statistics*, 3:675–692.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Hoffman, M., Blei, D., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Hoffman, M. D. and Gelman, A. (2011). The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo.

Huggins, J. H., Kasprzak, M., Campbell, T., and Broderick, T. (2019). Practical posterior error bounds from variational objectives. *arXiv preprint arXiv:1910.04102*.

Kucukelbir, A., Tran, D., Ranganath, R., et al. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18:1–45.

Opper, M. and Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792.

Rényi, A. (1961). On measures of entropy and information. *Berkeley Symposium on Mathematical Statistics and Probability*, 1:547–561.

Vehtari, A., Gelman, A., and Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Yes, but did it work?: Evaluating variational inference. *arXiv preprint arXiv:1802.02538*.