Journal of Machine Learning Research 23 (2022) 1-109 Submitted 12/19; Revised 09/21; Published 5/22

An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference

Jeremias Knoblauch J.KNOBLAUCH@WARWICK.AC.UK
The Alan Turing Institute
Dept. of Statistics
University of Warwick
Coventry, CV4 7AL, UK

Jack Jewson J.E.JEWSON@WARWICK.AC.UK
The Alan Turing Institute
Dept. of Statistics
University of Warwick
Coventry, CV4 7AL, UK

Theodoros Damoulas T.DAMOULAS@WARWICK.AC.UK
The Alan Turing Institute
Depts. of Computer Science & Statistics
University of Warwick
Coventry, CV4 7AL, UK

Editor: Frank Wood

Abstract

We advocate an optimization-centric view of Bayesian inference. Our inspiration is the representation of Bayes' rule as infinite-dimensional optimization (Csiszár, 1975; Donsker and Varadhan, 1975; Zellner, 1988). Equipped with this perspective, we study Bayesian inference when one does not have access to (1) well-specified priors, (2) well-specified likelihoods, (3) infinite computing power. While these three assumptions underlie the standard Bayesian paradigm, they are typically inappropriate for modern Machine Learning applications. We propose addressing this through an optimization-centric generalization of Bayesian posteriors that we call the Rule of Three (RoT). The RoT can be justified axiomatically and recovers Bayesian, PAC-Bayesian and VI posteriors as special cases. While the RoT is primarily a conceptual and theoretical device, it also encompasses a novel sub-class of tractable posteriors which we call Generalized Variational Inference (GVI) posteriors. Just as the RoT, GVI posteriors are specified by three arguments: a loss, a divergence and a variational family. They also possess a number of desirable properties, including modularity, Frequentist consistency and an interpretation as approximate ELBO. We explore applications of GVI posteriors, and show that they can be used to improve robustness and posterior marginals on Bayesian Neural Networks and Deep Gaussian Processes.

Keywords: Bayesian Inference, Generalized Bayesian Inference, Variational Inference, Bayesian Neural Networks, Deep Gaussian Proceses

1. Introduction

Though first discovered by the Reverend Thomas Bayes (1763), the version of Bayes' Theorem that a modern audience would be familiar with is much closer to the one in De Laplace (1774). Bayes' rule is one of the most fundamental results in probability theory and states that for a probability measure $\mathbb{P}$ and two events A, B, it holds that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)}.$$

As usual, $\mathbb{P}(A|B)$ denotes the conditional probability of event A given that event B occured. It would take nearly two more centuries for this mathematical result to be used as the basis for an entire school of statistical inference (Fienberg, 2006). More precisely, Fisher (1950) provides the first mention of the term *Bayesian* in accordance with our modern understanding (David, 1998).

Bayesian statistics uses Bayes' Theorem to conduct inference on an unknown and unobservable event A. Specifically, suppose that one can compute for an observable event B the probability $\mathbb{P}(B|A)$ and has a prior belief $\mathbb{P}(A)$ about the event A before observing B. Now, Bayes' rule tells us that we should be able to draw probabilistic inferences on $A|B$ by computing the probability $\mathbb{P}(A|B)$. In practice, the events A quantify the uncertainty about a parameter of interest $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and so are of the form $A \subset \boldsymbol{\Theta}$. The prior beliefs about events A are usually specified by some probability density $\pi : \boldsymbol{\Theta} \to \mathbb{R}_+$ inducing the probability measure $\mathbb{P}(A) = \int_A d\pi(\boldsymbol{\theta})$. This leaves us with the need to specify a probability distribution $\mathbb{P}(B|A)$ that relates the (unobserved) parameter $\boldsymbol{\theta}$ to the (observable) event B. In practice, B will correspond to n observations $x_{1:n}$. The next step is to define a distribution of $B|A$. This amounts to positing a likelihood function $p_n(x_{1:n}|\boldsymbol{\theta})$ and setting $\mathbb{P}(B|A) = p_n(x_{1:n}|\boldsymbol{\theta})$. Put together, this yields the standard *Bayesian posterior* that we denote as $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ throughout the paper and which is given by

$$q^*_{\mathrm{B}}(\boldsymbol{\theta}) = \frac{p_n(x_{1:n}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{Z}.$$

Here, $Z = \int_{\boldsymbol{\Theta}} p_n(x_{1:n}|\boldsymbol{\theta}) d\pi(\boldsymbol{\theta})$ is the normalizing constant—also known as the partition function—whose computation often makes the Bayesian posterior intractable.

Bayesian inference is appealing both conceptually and practically: Unlike Frequentist inference, Bayesian methods allow inferences to be informed by domain expertise in the form of a carefully specified prior belief $\pi(\boldsymbol{\theta})$. Furthermore, Bayesian inference produces belief distributions (rather than point estimates) over the parameter of interest $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. As a consequence, Bayesian inferences automatically quantify uncertainty about $\boldsymbol{\theta}$. This is practically useful in many situations, but especially if one uses $\boldsymbol{\theta}$ predictively: Integrating over $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ avoids being over-confident about the best value of $\boldsymbol{\theta}$, substantially improving predictive performance (see e.g. Aitchison, 1975). Amongst other benefits, it is this enhanced predictive performance that has cast Bayesian inference as one of the predominant paradigms in contemporary large-scale statistical inference and machine learning.

While Bayesian methods automatically quantify the uncertainty about their inferences, this comes at a cost: In the translation of Bayes' rule into the Bayesian posterior $q^*_{\mathrm{B}}(\boldsymbol{\theta})$, we have made three implicit but crucial assumptions. First, we assumed that the modeller has

a prior belief which is worth being taken into account and which the modeller is capable of writing out mathematically in the form of $\pi(\boldsymbol{\theta})$. Second, we specified the likelihood function $p_n(x_{1:n}|\boldsymbol{\theta})$ as a conditional probability. In other words, we assumed that the model is correctly specified, which is to say that $p_n(x_{1:n}|\boldsymbol{\theta}^*) = d\mathbb{P}(x_{1:n})$ for some unknown value of $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$. Third, we assumed the availability of enough computational power to make use of the often intractable posterior $q^*_{\mathrm{B}}(\boldsymbol{\theta})$. In many situations, these three assumptions built into $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ are harmless. For modern large-scale statistical machine learning tasks however, they are frequently violated.

To address this, the current paper takes a step back from the traditional interpretation of the Bayesian posterior $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ as an updating rule—Instead, we adopt an optimization-centric view point. Throughout, we motivate this with the tension between the three main assumptions underlying standard Bayesian inference on the one hand and the requirement of many contemporary statistical applications on the other hand. Aimed at resolving this tension, we define an optimization-centric generalization of Bayesian inference that we call the Rule of Three (RoT). The RoT is specified by an optimization problem over the space of probability measures $\mathcal{P}(\boldsymbol{\Theta})$ on $\boldsymbol{\Theta}$ with three arguments. These arguments are a loss function ℓ, a divergence D measuring the deviation of the posterior from the prior and a space $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$ of feasible solutions. Together, these three ingredients define posterior beliefs of the form

$$q^*(\boldsymbol{\theta}) = \operatorname*{arg\,min}_{q \in \Pi} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \right] + D(q \| \pi) \right\} \stackrel{\text{def}}{=} P(\ell, D, \Pi). \tag{1}$$

While this objective clearly also depends on two additional arguments—data $x_{1:n}$ and a prior π—we consider these fixed throughout and thus notationally suppress this dependence. Posteriors defined via this objective recover previous generalizations of Bayesian inference, including those inspired by Gibbs posteriors (e.g. Ghosh and Basu, 2016; Bissiri et al., 2016; Jewson et al., 2018; Nakagawa and Hashimoto, 2019; Chérief-Abdellatif and Alquier, 2019a), tempered posteriors (e.g. Grünwald, 2011, 2012; Holmes and Walker, 2017; Grünwald and Van Ommen, 2017; Miller and Dunson, 2019), as well as PAC-Bayesian approaches (for a recent overview, see Guedj, 2019). we illustrate this taxonomy in Figure 1. Unlike any of these previous generalizations however, posteriors taking the form $P(\ell, D, \Pi)$ may be non-multiplicative.

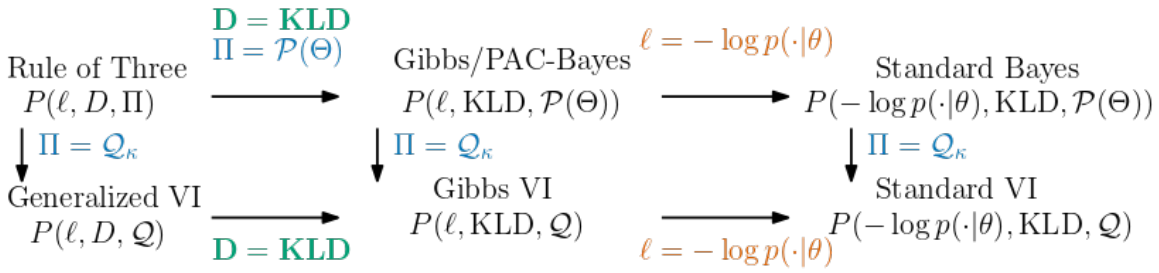


Figure 1: A taxonomy of some important belief distributions as special cases of the RoT.

For example, the RoT recovers Variational Inference (VI) posteriors based on minimizing the Kullback-Leibler Divergence (KLD) to $q^*_{\mathrm{B}}(\boldsymbol{\theta})$. Beyond that, it also gives rise to a new class of distributions we will call Generalized Variational Inference (GVI) posteriors. GVI posteriors are the methodological consequence of our generalization and are defined as the tractable special case for the RoT in which $\Pi = \mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) : \boldsymbol{\kappa} \in \boldsymbol{K}\} \subset \mathcal{P}(\boldsymbol{\Theta})$ is chosen to be a variational family. A number of theoretical and empirical findings lead us to conclude that GVI posteriors are often well-suited to real world inference problems.

The key ideas presented in the current paper are developed in five steps.

Section 2: We recapitulate the standard approach to Bayesian inference and various variational approximation schemes for $q^*_{\mathrm{B}}(\boldsymbol{\theta})$. Unconventionally, we do so through the lens of infinite-dimensional optimization. This view provides a number of interesting insights: For example, it enables a natural breakdown of variational approximation methods. Further, it reveals that relative to the objective characterizing $q^*_{\mathrm{B}}(\boldsymbol{\theta})$, standard VI is the optimal solution in its variational family.

Section 3: We explain why a generalized view on Bayesian inference is useful. To this end, we first recapitulate the three assumptions that justify Bayesian inference: the availability of appropriately specified priors and likelihoods as well as sufficient computational power to address the intractability of $q^*_{\mathrm{B}}(\boldsymbol{\theta})$. We contrast these assumptions with the realities of modern day large-scale inference and explain some problems arising from the severe misalignment between assumptions and reality.

Section 4: We axiomatically derive a generalized representation of Bayesian inference that we call the Rule of Three (RoT). Unlike previous generalizations, the RoT constitutes an optimization-centric outlook on Bayesian inference. We discuss the RoT and explain how it addresses the adverse effects of violating the assumptions underlying standard Bayesian inference. Lastly, we draw connections between the RoT and existing Bayesian methods.

Section 5: The conceptual contribution of the RoT can be translated into a methodological one via a family of methods we call Generalized Variational Inference (GVI). We explain how to use GVI for robust inference and more appropriate marginal variances. We also point to some theoretical findings, including frequentist consistency and an interpretation of GVI as approximate evidence lower bound. Computation of GVI posteriors concludes the section.

Section 6: We demonstrate GVI on two large-scale inference applications: Bayesian Neural Networks (BNNs) and Deep Gaussian Processes (DGPs). In different ways, both model classes are representative for the different ways in which contemporary large-scale inference is often at odds with the assumptions underlying the standard Bayesian posterior. We show that appropriately addressing this misalignment dramatically improves performance.

Throughout, we radically simplify the presentation for improved readability: For example, we do not incorporate latent variables into our notation. Further, we assume that losses are additive, homogeneous and such that the i-th loss term $\ell(\boldsymbol{\theta}, x_i)$ only depends on x_i. Neither

of these assumptions is necessary, and we will explicitly relax them for those parts of the paper where such a relaxation is needed.

2. An optimization-centric view on Bayesian inference

First, we set the stage by introducing an optimization-centric view on (generalized) Bayesian inference. Drawing attention to an isomorphism between the Bayesian posterior and an infinite-dimensional optimization problem, we discuss three implications of this relationship:

> **Section 2.2:** Committing to any exact Bayesian posterior is equivalent to committing to a particular optimization problem over the space of probability measures
>
> **Section 2.3:** With an optimization-centric view on Bayesian inference and a fixed variational family $\mathcal{Q}$, standard Variational Inference (VI) produces *optimal* approximations of the exact Bayesian posterior in $\mathcal{Q}$ by the logic of constrained optimization. Similarly, non-standard VI methods based on alternative objectives are *suboptimal*.
>
> **Section 2.4:** While standard VI posteriors are nominally optimal relative to $q^*_{\mathrm{B}}(\boldsymbol{\theta})$, non-standard approximations often perform much better in practice because they define posteriors through more appropriate objectives.

2.1 Preliminaries

Given a prior belief $\pi(\boldsymbol{\theta})$ and observations $x_{1:n}$ linked to $\boldsymbol{\theta}$ via a likelihood function $p(x_i|\boldsymbol{\theta})$, the **standard Bayesian posterior** belief $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ is computed through a multiplicative updating rule with $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$ as

$$q^*_{\mathrm{B}}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} \exp\{-\ell(\boldsymbol{\theta}, x_i)\}. \tag{2}$$

While this way of writing Bayes rule might seem cumbersome, it reveals that the multiplicative structure is in principle applicable to any loss function. In fact, replacing the negative log likelihood with any loss $\ell : \boldsymbol{\Theta} \times \mathcal{X} \to \mathbb{R}$ yields the **generalized Bayesian posterior**. If the normalizer of eq. (2) exists, these posteriors provide a coherent way for updating beliefs about an arbitrary parameter $\boldsymbol{\theta}$ (Bissiri et al., 2016).

For instance, $\boldsymbol{\theta}$ could be the median for the data generating mechanism that produced $x_{1:n}$. A loss-based Bayesian treatment of this problem would combine a prior belief π about the median with $\ell(\boldsymbol{\theta}, x_i) = |\boldsymbol{\theta} - x_i|_1$. Together, these two ingredients yield a generalized Bayesian posterior belief about the median as above. For some generalized Bayesian posteriors with more interesting loss functions, see Ghosh and Basu (2016); Alquier et al. (2016); Grünwald and Van Ommen (2017); Jewson et al. (2018); Knoblauch et al. (2018); Chérief-Abdellatif and Alquier (2019a); Nakagawa and Hashimoto (2019); Knoblauch and Vomfell (2020).

Throughout this paper, we do not notationally distinguish standard and generalized Bayesian posteriors. Unless we make the distinction explicit, we call both types of belief distributions **Bayesian posterior** and denote any posterior belief computed as in eq. (2) by $q^*_{\mathrm{B}}(\boldsymbol{\theta})$. The asterisk superscript in $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ emphasizes our next observation: *Any* posterior belief distribution is the result of an appropriately specified optimization problem.

2.2 Bayesian inference as infinite-dimensional optimization

While the logic of multiplicative updates inherent in Bayes' rule and eq. (2) is conceptually elegant, there is an independent and completely different path for arriving at $q^*_{\mathrm{B}}(\boldsymbol{\theta})$: dating back at least to Csiszár (1975) and Donsker and Varadhan (1975), it was shown that Bayesian inference can be recast as the solution to an infinite-dimensional optimization problem. This result was rediscovered in statistics by Zellner (1988) and states that for $\mathcal{P}(\boldsymbol{\Theta})$ denoting the space of all probability measures on $\boldsymbol{\Theta}$, the Bayesian posterior is given by

$$q^*_{\mathrm{B}}(\boldsymbol{\theta}) = P(-\log p(\cdot|\boldsymbol{\theta}), \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta})) = \underset{q\in\mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})}\left[-\sum_{i=1}^{n} \log(p(x_i|\boldsymbol{\theta})) \right] + \mathrm{KLD}\left(q||\pi\right) \right\},$$

where KLD is the Kullback-Leibler divergence (Kullback and Leibler, 1951) given by

$$\mathrm{KLD}(q||\pi) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)\right] = \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log q(\boldsymbol{\theta})\right] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log \pi(\boldsymbol{\theta})\right].$$

Similarly, the generalized Bayesian posteriors of Bissiri et al. (2016) solve

$$q^*_{\mathrm{B}}(\boldsymbol{\theta}) = P(\ell, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta})) = \underset{q\in\mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \right] + \mathrm{KLD}\left(q||\pi\right) \right\}. \tag{3}$$

This objective allows a re-interpretation of Bayesian inference as regularized optimization: as in maximum likelihood inference and other empirical risk minimization tasks, one wishes to minimize some loss function over the data. Unlike with frequentist methods however, one wishes to quantify uncertainty and obtain a belief distribution rather than a point estimate. Consequently, one adds the KLD regularization term. In fact, if this KLD term were absent from eq. (3), the solution of the optimization problem would simply be a Dirac mass $\delta_{\widehat{\boldsymbol{\theta}}_n}(\boldsymbol{\theta})$ at the empirical risk minimizer $\widehat{\boldsymbol{\theta}}_n$.

For completeness' sake, we provide a proof of eq. (3) based on the supplementary material of Bissiri et al. (2016). This encompasses the original result of (Csiszár, 1975) and (Donsker and Varadhan, 1975) for $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$.

Theorem 1 *If* $Z = \int_{\boldsymbol{\Theta}} \exp\left\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\right\} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} < \infty$, *then the solution of eq.* (3) *exists and is equivalent to the generalized Bayesian posterior* $q^*_B(\boldsymbol{\theta})$ *as given in eq.* (2).

Proof One may rewrite the objective in eq. (3) as

$$\begin{aligned} q^*(\boldsymbol{\theta}) &= \underset{q\in\mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \int_{\boldsymbol{\Theta}} \left[\log\left(\exp\left\{ \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \right\} \right) + \log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right) \right] q(\boldsymbol{\theta}) d\boldsymbol{\theta} \right\} \\ &= \underset{q\in\mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \int_{\boldsymbol{\Theta}} \log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}) \exp\left\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\right\}} \right) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \right\}. \end{aligned}$$

As one only cares about the minimizer $q^*(\boldsymbol{\theta})$ (and not the objective value), it also holds that for any constant $Z > 0$, the above is equal to

$$\begin{aligned} q^*(\boldsymbol{\theta}) &= \underset{q\in\mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \int_{\boldsymbol{\Theta}} \log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}) \exp\left\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\right\} Z^{-1}} \right) q(\boldsymbol{\theta}) d\boldsymbol{\theta} - \log Z \right\} \\ &= \underset{q\in\mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \mathrm{KLD}\left(q(\boldsymbol{\theta}) \middle\| \pi(\boldsymbol{\theta}) \exp\left\{ -\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \right\} Z^{-1} \right) \right\}. \end{aligned}$$

Lastly, one sets $Z = \int_{\boldsymbol{\theta}} \exp\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\}\, \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ and notes that as the KLD is a statistical divergence, it is minimized uniquely if its two arguments are the same, so $q^*(\boldsymbol{\theta}) = q_B^*(\boldsymbol{\theta})$. ■

The result in Theorem 1 implies an important observation that drives much of the current paper's development: Any commitment to a (standard or generalized) Bayesian posterior is always a commitment to an optimization problem.

Observation 1 *If you conduct inference based upon the Bayesian posterior $q_B^*(\boldsymbol{\theta})$ in eq. (2), you conduct inference by specifying and solving the optimization problem in eq. (3). In other words, $q_B^*(\boldsymbol{\theta})$ is adequate to address a given inference problem* ***if and only if*** *minimizing the objective in eq. (3) reflects the goals of your inference.*

Building on this observation, Section 3, will explain why the usefulness of the standard Bayesian posterior—and thus of the objective in eq. (3)—is at least doubtful for many contemporary machine learning applications.

2.3 Optimality of standard Variational Inference

While $q_{\mathrm{B}}^*(\boldsymbol{\theta})$ is analytically available up to a normalizing constant, this is not immediately useful: As exact computations with $q_{\mathrm{B}}^*(\boldsymbol{\theta})$ are often only possible through sampling methods, using a posterior of this form typically incurs a large computational burden. To alleviate this problem, many approximate Bayesian inference schemes have been proposed. Their principal idea is to force the posterior belief into some parametric form. Specifically, one seeks to approximate $q_{\mathrm{B}}^*(\boldsymbol{\theta}) \approx q_{\mathrm{A}}^*(\boldsymbol{\theta})$, where $q_{\mathrm{A}}^*(\boldsymbol{\theta}) \in \mathcal{Q}$ and

$$\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) : \boldsymbol{\kappa} \in \boldsymbol{K}\} \tag{4}$$

is a family of distributions on $\boldsymbol{\Theta}$ parameterized by $\boldsymbol{\kappa}$. This significantly reduces the computational burden, because it transforms the optimization problem from an infinite-dimensional into a finite-dimensional space.

The literature on such approximations is extensive and has diverse origins. Their development arguably started with Laplace Approximations (see e.g. the seminal papers of Tierney and Kadane, 1986; Shun and McCullagh, 1995; MacKay, 1998), which have recently been refined into Integrated Nested Laplace Approximations (Rue et al., 2009). A second family of approximation methods known as Expectation Propagation (Opper and Winther, 2000; Minka, 2001) was motivated through factor graphs and message passing (Minka, 2005). The third and arguably most successful approach originated by connecting the Expectation-Maximization algorithm (Dempster et al., 1977) and the variational free energy from statistical physics (Neal and Hinton, 1998), culminating in Variational Inference (VI) (Jordan et al., 1999; Beal, 2003). For these methods, $\mathcal{Q}$ is called the variational family.

Two main interpretations of VI prevail: One may derive its objective function as an Evidence Lower Bound (ELBO); And one can show that VI minimizes the KLD between $\mathcal{Q}$ and $q_{\mathrm{B}}^*(\boldsymbol{\theta})$. Here, we introduce a third interpretation of VI: Relative to the objective in eq. (3), it corresponds to the best $\mathcal{Q}$-constrained solution of the Bayesian inference problem.

2.3.1 VI as log evidence bound

One context in which VI was derived is the task of model selection. In Bayesian model selection, the integral $p(x_{1:n}) = \int_{\boldsymbol{\Theta}} \exp\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\}\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$—called *evidence* or *marginal*

likelihood whenever $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$ for some likelihood model p—takes centre stage. Roughly speaking, one selects the model for which this integral is largest. But since $p(x_{1:n})$ is generally intractable, one finds an approximation to it. In particular, one notes that for any $q(\boldsymbol{\theta}) \in \mathcal{Q}$,

$$\begin{aligned}
\log p(x_{1:n}) &= \log\left(\int_{\boldsymbol{\Theta}} \exp\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\right) \\
&= \log\left(\int_{\boldsymbol{\Theta}} \exp\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\}\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}q(\boldsymbol{\theta})d\boldsymbol{\theta}\right) \\
&\overset{\text{Jensen's IE}}{\geq} \int_{\boldsymbol{\Theta}} \log\left(\exp\left\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\right\}\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right) q(\boldsymbol{\theta})d\boldsymbol{\theta}. \qquad (5)
\end{aligned}$$

If the loss function is $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$ for some likelihood model p, then the right hand side of eq. (5) is called the Evidence Lower Bound (ELBO). Rewriting the integral, one now obtains the VI posterior as

$$q^*_{\text{VI}}(\boldsymbol{\theta}) = P(\ell, \text{KLD}, \mathcal{Q}) = \arg\min_{q\in\mathcal{Q}} \left\{\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\right] + \text{KLD}\left(q||\pi\right)\right\}, \qquad (6)$$

where $q^*_{\text{VI}}(\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)$ for some optimal parameter $\boldsymbol{\kappa}^* \in \boldsymbol{K}$.

Taking inspiration from this interpretation, alternative approximations target generalized Evidence Lower Bounds (e.g. Chen et al., 2018; Domke and Sheldon, 2018; Burda et al., 2016). For a given bound $\log p(x_{1:n}) \geq \text{G-ELBO}(q)$, such methods produce posteriors as

$$q^*_{\text{G-ELBO}}(\boldsymbol{\theta}) = \arg\min_{q\in\mathcal{Q}} \left\{-\text{G-ELBO}(q)\right\}.$$

Multi-sample bounds (see e.g. Burda et al., 2016) are a particularly prominent example. As the name implies, these bounds interpret the ELBO term given in eq. (6) by

$$\text{ELBO}(q) = \mathbb{E}_{\boldsymbol{\theta}\sim q(\boldsymbol{\theta})}\left[\log\left(\frac{\exp\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\pi(\boldsymbol{\theta})\}}{q(\boldsymbol{\theta})}\right)\right]$$

as a bound constructed from a single sample of $\boldsymbol{\theta}$ and replace the objective with its K-sample version obtained by

$$\text{MS-ELBO}(q, K) = \mathbb{E}_{\boldsymbol{\theta}_{1:K}\sim\prod_{j=1}^{K} q(\boldsymbol{\theta}_j)}\left[\log\frac{1}{K}\sum_{j=1}^{K}\left(\frac{\exp\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}_j, x_i)\pi(\boldsymbol{\theta})\}}{q(\boldsymbol{\theta}_j)}\right)\right].$$

The rationale for doing so is that $\text{MS-ELBO}(q, 1) = \text{ELBO}(q)$, and that the resulting bound on the (generalized) evidence is tighter than the standard ELBO. More precisely, for any $K \in \mathbb{N}$, $\log p(x_{1:n}) \geq \text{MS-ELBO}(q, K+1) \geq \text{MS-ELBO}(q, K) \geq \text{MS-ELBO}(q, 1) = \text{ELBO}(q)$.

2.3.2 VI as KLD-minimization and Discrepancy VI (

DVI)

A second well-known perspective on standard VI posteriors is motivated by rewriting the objective in eq. (6) in terms of the Kullback-Leibler Divergence (KLD) as follows:

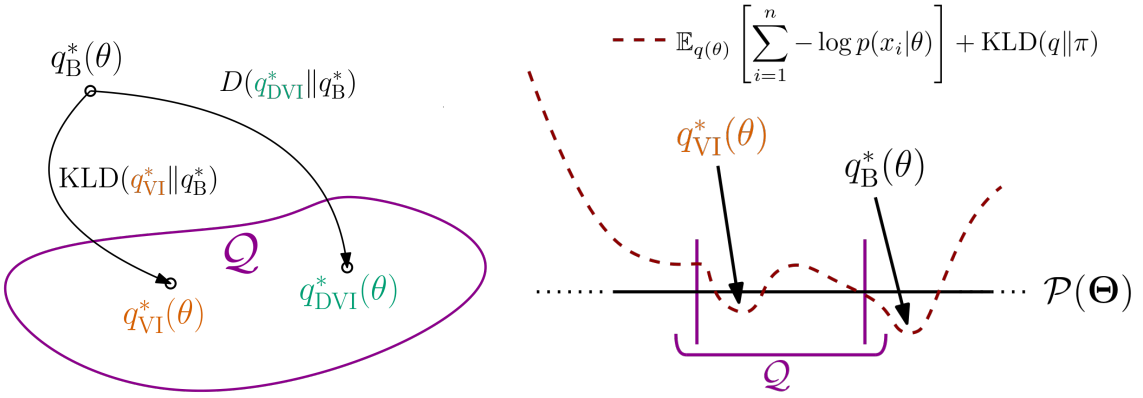
$$q^*_{\text{VI}}(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} \left\{ \text{KLD}\left(q(\boldsymbol{\theta}) \middle\| q^*_{\text{B}}(\boldsymbol{\theta}) \right) \right\}$$

The relevant algebraic arguments are similar to the ones used in the proof of Theorem 1 and show that standard VI finds $q^*_{\text{VI}}(\boldsymbol{\theta}) \in \mathcal{Q}$ closest to $q^*_{\text{B}}(\boldsymbol{\theta})$ in the KLD-sense.

This insight has produced a growing literature seeking to minimize (local or global) discrepancies D between $\mathcal{Q}$ and $q^*_{\text{B}}(\boldsymbol{\theta})$ different from the KLD (e.g. Minka, 2001; Opper and Winther, 2000; Li and Turner, 2016; Dieng et al., 2017; Hernández-Lobato et al., 2016; Yang et al., 2019; Cichocki and Amari, 2010; Ranganath et al., 2016; Wang et al., 2018; Saha et al., 2019). For a disrepancy measure $D : \mathcal{P}(\boldsymbol{\Theta}) \times \mathcal{P}(\boldsymbol{\Theta}) \to \mathbb{R}$, these methods compute

$$q^*_{\text{DVI}}(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} \left\{ D\left(q(\boldsymbol{\theta}) \middle\| q^*_{\text{B}}(\boldsymbol{\theta}) \right) \right\}.$$

In the remainder, we will call such procedures **Discrepancy Variational Inference (DVI)** methods whenever $D \neq \text{KLD}$. We graphically summarize their interpretation in Figure 2a. Note that DVI methods do not fall into our Rule of Three framework: they are generally not recoverable as $\mathcal{P}(\ell, D, \mathcal{Q})$ for any choice of ℓ, D, $\mathcal{Q}$.



(a) **DVI** interpretation of **VI** (b) Interpretation of **VI** as in Theorem 2

Figure 2: Best viewed in color. Depicted is a schematic to clarify the conceptual distinction between two interpretations of **VI**. **DVI** methods interpret **VI** as the KLD-projection of $q^*_{\text{B}}(\boldsymbol{\theta})$ into the variational family $\mathcal{Q}$. New methods are then derived by replacing the KLD with alternative projection operators. In contrast, Theorem 2 interprets **VI** posteriors as best solutions to a constrained optimization problem. Rather than finding the global optimum $q^*_{\text{B}}(\boldsymbol{\theta})$ of the optimization problem in eq. (3), **VI** finds the best solution in the subset $\mathcal{Q} \subset \mathcal{P}(\boldsymbol{\Theta})$.

2.3.3 VI as constrained optimization

While the interpretations of VI as optimizing over an evidence lower bound and as minimizing a discrepancy are well-known, this paper presents a third interpretation: VI posteriors are also the optimal solutions to the $\mathcal{Q}$-constrained version of the optimization problem underlying $q^*_{\mathrm{B}}(\boldsymbol{\theta})$. Specifically, comparing eq. (3) to eq. (6) shows that while $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ is obtained by optimizing over all of $\mathcal{P}(\boldsymbol{\Theta})$, $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ is obtained by optimizing *the same objective*—but only over the finite-dimensional subset $\mathcal{Q} \subset \mathcal{P}(\boldsymbol{\Theta})$. This observation is summarized in the following Theorem.

Theorem 2 (Optimality of standard VI) *Relative to the objective in eq.* (3) *characterizing Bayesian inference, and for any fixed finite-dimensional variational family $\mathcal{Q}$, standard* VI *produces the optimal posterior belief in $\mathcal{Q}$.*

Proof First, notice that the VI posterior belief distribution $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ and the Bayesian posterior belief distribution $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ both seek to minimize

$$\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\right] + \mathrm{KLD}(q\|\pi)$$

over $q(\boldsymbol{\theta})$. Second, notice that $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ is the minimizer of this objective relative to $\mathcal{Q}$ while $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ is the minimizer relative to $\mathcal{P}(\boldsymbol{\Theta})$. Third, note that $\mathcal{Q} \subset \mathcal{P}(\boldsymbol{\Theta})$. ■

This provides another meaningful interpretation of $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ depicted in Figure 2b. Specifically, the result endows standard VI with a special property: in the optimization-centric view on Bayesian inference, we should prefer $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ amongst all possible approximations within $\mathcal{Q}$ *provided* we believe that the optimization objective defining the Bayesian posterior is appropriate for the problem at hand. The following observation explains this in more detail.

Observation 2 *As Observation 1 explained, committing to $q^*_B(\boldsymbol{\theta})$ means committing to the objective function in eq.* (3)*. In other words, if we judge the posterior belief $q^*_B(\boldsymbol{\theta})$ to be desirable, we are also saying that the objective function in eq.* (3) *encodes properties that we want our posterior to adhere to. Accordingly, once we restrict posterior beliefs to a subset $\mathcal{Q} \subset \mathcal{P}(\boldsymbol{\Theta})$, we should want to compute the best possible solution to the* ***same*** *objective in $\mathcal{Q}$. As Theorem 2 shows, this is exactly what* VI *does.*

Relative to the optimization-centric view on Bayesian inference, Theorem 2 also implies the sub-optimality of alternative approximation methods.

Corollary 3 (Suboptimality of alternative methods) *Relative to the objective in eq.* (3) *characterizing Bayesian inference, and for any fixed finite-dimensional variational family $\mathcal{Q}$, methods different from standard* VI *produce sub-optimal posterior beliefs.*

Proof We prove this by contradiction: Suppose we are given a posterior belief $q^*_{\mathrm{A}}(\boldsymbol{\theta})$ that could not have alternatively been produced by standard VI. First, by definition of standard VI, it holds that that for *any* sequence of observations $x_{1:n}$ and for all n,

$$\mathbb{E}_{q^*_{\mathrm{VI}}(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\right] + \mathrm{KLD}\left(q^*_{\mathrm{VI}}||\pi\right) \leq \mathbb{E}_{q^*_{\mathrm{A}}(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\right] + \mathrm{KLD}\left(q^*_{\mathrm{A}}||\pi\right).$$

Since we also assumed that $q^*_{\text{A}}(\boldsymbol{\theta})$ could not have alternatively been produced by standard VI, it also holds that the inequality is strict, i.e.

$$\mathbb{E}_{q^*_{\text{VI}}(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},x_i)\right] + \text{KLD}\left(q^*_{\text{VI}}||\pi\right) < \mathbb{E}_{q^*_{\text{A}}(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},x_i)\right] + \text{KLD}\left(q^*_{\text{A}}||\pi\right).$$

This yields the desired result. ■

Corollary 3 says that for a fixed variational family $\mathcal{Q}$, any alternative approximation $q^*_{\text{A}}(\boldsymbol{\theta}) \in \mathcal{Q}$ that is not equal to $q^*_{\text{VI}}(\boldsymbol{\theta})$ will be sub-optimal under an optimization-centric view of Bayesian inference. This concerns a host of methods, including generalized evidence lower bound formulations, alternative Discrepancy Variational Inference (DVI) methods or Expectation Propagation (EP) approaches. This is significant, as it shows that alternative approximations do *not* provide the optimal posterior relative to eq. (3)—an equation that is endowed with particular meaning under the optimization-centric view on Bayesian inference.

Importantly, the result does *not* imply that these alternative posterior approximations will perform worse than VI *in practice*. In fact, from an optimization-centric standpoint it is quite clear why such alternative approximations can deliver empirical success: if $q^*_{\text{A}}(\boldsymbol{\theta})$ performs better than the standard variational approximation $q^*_{\text{VI}}(\boldsymbol{\theta})$, the objective underlying $q^*_{\text{A}}(\boldsymbol{\theta})$ must implicitly be targeting a more appropriate posterior belief for the problem at hand—an observation we elaborate upon next.

2.4 Reconciling (sub)optimality with empirical evidence

At first glance, the sub-optimality result of Corollary 3 may seem to contradict numerous landmark findings in the area of approximate Bayesian inference: standard VI exhibits various well-known pathologies that hinder its effectiveness in certain situations (see e.g. Turner and Sahani, 2011). For this reason, various alternative approximations have proven successful in practice (e.g. Minka, 2001; Rue et al., 2009) and often produce more desirable posterior inferences.

This paradox resolves itself upon closer examination. Notably, our notion of optimality is *relative to the objective that defined the Bayesian posterior*. In other words, the *practical* relevance of our optimality result hinges on two crucial assumptions that are typically violated in practice. First, one needs to assume that the best possible belief for the task at hand is the minimizer of the original objective in eq. (3)—i.e., the Bayesian posterior $q^*_{\text{B}}(\boldsymbol{\theta})$. Second, one then needs to assume that the variational family $\mathcal{Q}$ is rich enough to make the statement $q^*_{\text{VI}}(\boldsymbol{\theta}) \approx q^*_{\text{B}}(\boldsymbol{\theta})$ not completely vacuous[1]. Conversely, this means that the (nominally sub-optimal) objective underlying $q^*_{\text{A}}(\boldsymbol{\theta})$ often encodes more desirable belief distributions than the (nominally optimal) one underlying $q^*_{\text{VI}}(\boldsymbol{\theta})$ if at least one of the following holds:

(i) The optimization objective in eq. (3) is misspecified and does not reflect the belief distribution we wish to compute. In other words, the *full* posterior $q^*_{\text{B}}(\boldsymbol{\theta})$ (and hence

1. To give a hyperbolic example, consider $q^*_{\text{B}}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; 0, 1)$ and $\mathcal{Q} = \{\mathcal{N}(\boldsymbol{\theta}; -100, 1), \mathcal{N}(\boldsymbol{\theta}; 100, 1)\}$. While the (sub)optimality results of Theorem 2 and Corollary 3 continue to hold, $\mathcal{Q}$ will make the statement $q(\boldsymbol{\theta}) \approx q^*_{\text{B}}(\boldsymbol{\theta})$ meaningless for all $q \in \mathcal{Q}$.

the objective underlying it) are already problematic—and since it is based on the same objective, $q^*_{\text{VI}}(\boldsymbol{\theta})$ will inherit these problems.

(ii) The approximating family $\mathcal{Q}$ makes the statement $q(\boldsymbol{\theta}) \approx q^*_{\text{B}}(\boldsymbol{\theta})$ vacuous for any $q \in \mathcal{Q}$. In other words, thinking about $q^*_{\text{A}}(\boldsymbol{\theta})$ or $q^*_{\text{VI}}(\boldsymbol{\theta})$ as approximations to $q^*_{\text{B}}(\boldsymbol{\theta})$ becomes purely semantic—and we should rather think about $q^*_{\text{A}}(\boldsymbol{\theta})$ and $q^*_{\text{VI}}(\boldsymbol{\theta})$ as belief distributions constructed directly by minimizing some objective function on $\mathcal{Q}$.

Under these conditions, $q^*_{\text{A}}(\boldsymbol{\theta})$ can outperform $q^*_{\text{VI}}(\boldsymbol{\theta})$ whenever the underlying objective implicitly encodes desirable properties for the posterior belief distribution that are not part of the objective in eq. (3).

For example, virtually all posteriors produced within the DVI family (e.g. Li and Turner, 2016; Hernández-Lobato et al., 2016; Dieng et al., 2017; Regli and Silva, 2018) are designed to address (ii): these methods prevent unimodal approximations from focusing too strongly around the empirical risk minimizer of $\boldsymbol{\theta}$. For standard VI, this phenomenon is common whenever $\mathcal{Q}$ is the mean field variational family, which explains why DVI often empirically outperforms standard VI for this choice of $\mathcal{Q}$. Under the optimization-centric view on posterior beliefs, this implies that in spite of being sub-optimal relative to eq. (3), DVI methods pose objectives that are often better-suited to produce belief distributions in $\mathcal{Q}$. This raises an interesting question: Rather than thinking of inference in a subset $\mathcal{Q} \subset \mathcal{P}(\boldsymbol{\Theta})$ as approximate, can we adapt an optimization-centric view from the start and then *directly* design objectives that generate posterior beliefs with desirable properties? Later in this paper, we will operationalize this logic and propose a procedure we call Generalized Variational Inference (GVI).

Beforehand, we step back and ask under which conditions such an optimization-centric design of posteriors would even be desirable. We find the answer in the next Section by re-visiting the original motivations for performing Bayesian inference. Specifically, we explain how the assumptions underpinning the traditional Bayesian paradigm are often misaligned with the reality of contemporary statistical machine learning. This misalignment has three dimensions: The information contained in the prior belief **(P)**, the role of the likelihood model **(L)**, and the availability of computational resources **(C)**.

3. A reality check: Re-examining the traditional Bayesian paradigm

In the following section, we illuminate the misalignment between the assumptions underlying the traditional Bayesian paradigm and the way in which modern statistical machine learning uses (approximate) Bayesian posteriors to conduct inference.

First, **Section 3.1** elaborates on the three crucial assumptions underlying the standard Bayesian posterior: Appropriate specification of prior **(P)** and likelihood **(L)** as well as an infinite computational budget **(C)**.

Next, **Section 3.2** exposes the misalignment of these three assumptions with inferential practices in contemporary statistical machine learning and large-scale inference.

Lastly, **Sections 3.3–3.5** points to the adverse real-world consequences arising from violating these assumptions.

3.1 The traditional Bayesian paradigm

Due to their direct correspondence with the fundamental rules of probability, Bayesian posteriors $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ are desirable objects to be basing inference on. To see why, suppose the following three conditions hold true.

(P) The **P**rior $\pi(\boldsymbol{\theta})$ is correctly specified: It encodes the best available judgement about $\boldsymbol{\theta}$ based on *all* information available to the modeller. Crucially, the distribution $\boldsymbol{\pi}(\boldsymbol{\theta})$ is assumed to reflect this prior belief *exactly*. This implies that $\pi(\boldsymbol{\theta})$ should *completely* reflect all information available to the modeller such as previously observed observations $x_{-m:0}$ of the same phenomenon or domain expertise relating to the problem domain and the statistical model.

(L) There exists an (unknown but fixed) $\boldsymbol{\theta}^*$ making the **L**ikelihood model equivalent to the data generating mechanism of x_i. This is to say that $x_i \sim p(x_i|\boldsymbol{\theta}^*)$.[2]

(C) The budget for **C**omputation is infinite, so the complexity of computing the belief $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ can be ignored.

If **(L)**, **(P)** and **(C)** are satisfied, it immediately follows that the best belief about the best parameter value given the data $\{\boldsymbol{\theta}^* = \boldsymbol{\theta}\}|\{\boldsymbol{x}_{1:n} = x_{1:n}\}$ is simply given by

$$d\mathbb{P}\left(\boldsymbol{\theta}|x_{1:n}\right) \propto d\mathbb{P}\left(\boldsymbol{\theta}\right) \prod_{i=1}^{n} d\mathbb{P}\left(x_i|\boldsymbol{\theta}\right) = \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) \propto q^*_{\mathrm{B}}(\boldsymbol{\theta}). \tag{7}$$

Note that **(P)** and **(L)** lend a meaningful interpretation to Bayes' rule in form of conditional probability updates. Complementing this, **(C)** ensures that it is feasible to compute the often intractable solution $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ of eq. (3). Accordingly, **(C)** generally is interpreted to mean that a Markov Chain Monte Carlo algorithm can be run for long enough to accurately represent $q^*_{\mathrm{B}}(\boldsymbol{\theta})$. In summary, if **(P)**, **(L)** and **(C)** hold, $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ is the only desirable posterior belief distribution.

But how well does reality align with **(P)**, **(L)** and **(C)**? Turning attention to **(C)** first, most traditional scientific disciplines have little need to worry about computational complexity and will resort to sampling schemes for two reasons: First, the models are often relatively simple and thus straightforward to infer. Second, even for more complicated models the experimental setups, the cost of data collection typically outweighs those of computation

2. We note here that to keep the presentation simpler, we are giving conditions that are stricter than what is required for Bayesian analysis. In particular, **(L)** corresponds to an objectivist treatment of the likelihood and can be weakened under the subjectivist paradigm for Bayesian analysis. In this paradigm, the treatment of the likelihood mirrors that of the prior: It now simply corresponds to the modeller's belief about the process that generated the data. While this first sounds like a weaker requirement, it ends up producing the same misspecification problems as **(L)**. Specifically, a subjectivist treatment of the likelihood requires the modeller to express her beliefs about the likelihood function *exactly*. This forces her to make more probability statements than she realistically has time or introspection for (see e.g. Goldstein, 1990; O'Hagan and Oakley, 2004; Goldstein, 2006). The result is that the likelihood function supplied by the modeller is *at best* going to be an approximate description of the modeller's beliefs. This provides the subjectivist interpretation of misspecification. Notice that it directly mirrors the objectivist interpretation of misspecification in **(L)**: The likelihood function supplied is *at best* going to be an approximate description of the true data generating mechanism.

by orders of magnitude. As for **(P)** and **(L)**, neither prior nor likelihood are ever perfect reflections of one's full prior beliefs (see e.g. Goldstein, 1990; O'Hagan and Oakley, 2004; Goldstein, 2006) or the data generating mechanism (see e.g. Bernardo, 2000). In other words, **(P)** and **(L)** are invariably violated when interpreted literally. However and as enshrined in Box's aphorism that *all models are wrong, but some are useful*, this is not a problem so long as these violations are sufficiently small. In traditional statistics, ensuring that these violations are small has typically been enforced through a simple recursion (e.g. Box, 1980; Berger et al., 1994). Specifically, until you are confident that both **(P)** and **(L)** are close enough to the truth, repeat the following: Check if **(L)** or **(P)** are violated severely. If they are, choose a more appropriate likelihood and prior. In order to operationalize this iterative logic, batteries of descriptive statistics, tests and model selection criteria have been developed.

In summary then, ignoring the computational overhead and iteratively refining likelihoods and priors is rightfully the predominant inferential strategy for traditional scientific endeavours. Not only is domain expertise relevant for designing priors and likelihood, but the process of finding an appropriate model often provides valuable insights in itself. Further, the expensive part of the analysis is typically data *collection*. Consequently, it is typically not prohibitive to perform inference even with the most computationally expensive of sampling schemes. In line with this, most methodological contributions in statistical sciences rely to a substantial degree on **(P)**, **(L)** and **(C)**.

3.2 Machine Learning: Challenging the traditional Bayesian paradigm

Contemporary large-scale inference applications have frequently turned the traditional schematic of statistical model design upside down: Rather than carefully designing an appropriate likelihood model $p(\cdot|\boldsymbol{\theta})$ for a specific data domain, statistical machine learning research is typically characterized by the search of a flexible algorithm that can fit *any* data set $x_{1:n}$ well enough to produce useful inferences. The resulting likelihood models are typically not attempting to describe any data generating processes in the sense of **(L)**. Rather, they are highly over-parameterized functions of $\boldsymbol{\theta}$ and typically un-identifiable, meaning that $\boldsymbol{\theta}^*$ is neither interpretable nor unique. Such statistical machine learning models have three major issues under the traditional paradigm of Bayesian inference that are readily identified:

(↯P) Invariably, the **P**rior is misspecified. Two factors compound this issue: Firstly, the large number of parameters over-parameterizing the likelihoods of many statistical machine learning models are no longer interpretable. This often prohibits domain experts to carry out carefully guided prior elicitation. Secondly, priors are typically selected at least in part for their computational feasibility. This fundamentally alters the interpretation of the prior: Rather than the result of an attempt to capture the modeller's knowledge before observing the data, the prior takes the role of a more or less arbitrary reference measure. To make matters worse, the number of parameters is often large relative to n, which means that the priors have a disproportionate effect on inference via $q^*_{\mathrm{B}}(\boldsymbol{\theta})$—a problem we discuss in Example 1 in the context of Bayesian Neural Networks.

(↯L) Clearly, the **L**ikelihood is misspecified. This often has adverse side effects: While using an over-parameterized or off-the-shelf likelihood function can provide a good fit for

the typical behaviour of the data, it will struggle with heterogeneous or untypical data points. We demonstrate this phenomenon on a changepoint problem in Example 4.

(↯C) With increasingly complex statistical models, **(C)** has proven an increasingly infeasible description of reality. Accordingly, this problem has inspired numerous directions of research, including variational methods and Laplace approximations. Example 2 illustrates this for the case of Gaussian Processes.

Under the challenges outlined in **(↯P)**, **(↯L)** and **(↯C)**, standard Bayesian posteriors often do not provide appropriate belief distributions. In the remainder, we will explain how and why this is the case for many parts of modern large-scale inference.

3.3 Prior misspecification

For most finite-dimensional parameters, even severely misspecified priors can often be harmless. For example, prior misspecification is typically no problem in the asymptotic sense. Specifically, so long as **(L)** holds, it suffices that $\pi(\boldsymbol{\theta}^*) > 0$ for standard Bayesian posteriors to contract around $\boldsymbol{\theta}^*$ at rate $O(n^{-1/2})$ (see e.g. Ghosal, 1998; Ghosal et al., 2000; Shen and Wasserman, 2001; Walker, 2004, and references therein).

Often, these results are used as an apology to neglect the role of prior specification. While it is reassuring that the sequence of standard Bayesian posteriors shrinks to the population-optimum as $n \to \infty$, this does not describe the real world: n is usually fixed and only a single posterior is computed. Further, it is possible to specify arbitrarily bad priors for any n fixed observations. This means that once one departs from assuming that **(P)** is at least approximately correct, the standard Bayesian posterior belief about $\boldsymbol{\theta}^*$ can be made arbitrarily inappropriate. In summary, prior specification is particularly precarious whenever (i) the parameter space is large relative to n or (ii) it is impossible to specify priors in a principled way. As we discuss in the next example, a model invariably affected by both problems is the Bayesian Neural Network (BNN).

Example 1 (Deep Bayesian models as violations of (P)) *Bayesian Neural Networks (BNNs) (MacKay, 1996; Neal, 2012) combine Deep Learning with Bayesian uncertainty quantification. For the parameter vector $\boldsymbol{\theta}$ of network weights, let $F(\boldsymbol{\theta})$ be the function specified by a Neural Network. One way of thinking about BNNs is as an over-parameterized likelihood function with a large number of parameters $d = |\boldsymbol{\Theta}|$. This is to say that one believes that (at least approximately), $x_i \sim p(x_i|F(\boldsymbol{\theta}^*))$ for some $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$. For a prior $\pi(\boldsymbol{\theta})$ about $\boldsymbol{\theta}$, this means that BNNs seek to do inference on the posterior given by*

$$q^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} p(x_i|F(\boldsymbol{\theta})).$$

*This approach is conceptually appealing: One circumvents most issues with **(L)** by making the likelihood function almost arbitrarily flexible, and also quantifies uncertainty in the usual Bayesian manner. While both observations are correct, they mask a severe practical issue with this approach: Specifying $\pi(\boldsymbol{\theta})$ in a principled way and in accordance with **(P)** is generally impossible.*

There are two main reasons for this: Firstly, the vector $\boldsymbol{\theta}$ indexes a black box model and is not interpretable, making domain expertise useless for prior elicitation. Secondly, computational aspects are a major concern for BNN*s, so that one typically is constrained to choose priors that factorize over $\boldsymbol{\theta}$. As a consequence, practitioners often resort to choosing "default priors" that do not even attempt to approximately satisfy* ***(P)****. Specifically, one typically just picks $\pi(\boldsymbol{\theta}) = \prod_{j=1}^{d} \pi_j(\boldsymbol{\theta}_j)$, where $\pi_j(\boldsymbol{\theta}_j)$ is a standard normal distribution for all j. Choosing priors in this ad-hoc fashion violates the principles underlying classical Bayesian modelling (see also Section 5.2.1) and is especially problematic when n is small relative to d (so that the prior has relatively strong influence). At the same time, reliable uncertainty quantification is most important whenever n is small relative to d. In fact, this is a well-known issue and addressed in various contributions by up-weighting the likelihood (down-weighting the* KLD *term in the* ELBO*), see Zhang et al. (2018); Rossi et al. (2019a,b); Sønderby et al. (2016).*

We do not mean to suggest that it is impossible to specify meaningful or useful priors for BNN*s. For example, Toussaint et al. (2006) uses the principles of transformation invariance and maximum entropy, Nalisnick et al. (2020) calibrates priors via their predictive distribution and a 'reference' model, and Matsubara et al. (2020) focuses on the prior's impact on the prediction space (see also Gelman et al., 2017) and in particular its covariance structure to specify more principled priors. While these approaches are all conceptually elegant, they also are computationally cumbersome—thus compounding the issues outlined in* ***(↯C)****. As a result, the fully factorized priors discussed above are the de-facto default choices for most applications.*

For completeness, we note that the current paper does not discuss uninformative and so-called objective priors (see, e.g. Jeffreys, 1961; Zellner, 1977; Bernardo, 1979; Berger and Bernardo, 1992; Jaynes, 2003; Berger, 2006). Such priors are constructed to be as uninformative as possible. In some ways, they would be a natural, principled alternative to ill-informed priors. Generally however, their construction results in improper priors–densities that do not correspond to a finite measure and thus do not integrate to one. While this is not generally prohibitive, it would severely complicate the developments of Section 4 because most divergences are not well-defined for improper priors[3].

3.4 Likelihood Misspecification

While prior misspecification affects inference adversely, the issue for inferential practice is even more serious if **(L)** is violated: Whenever the likelihood model for x_i is severely misspecified, inference outcomes suffer dramatically. Moreover, not even the asymptotic regime offers a remedy: The adverse effects of misspecification persist as $n \to \infty$. The traditional approach to addressing this issue is straightforward: If the likelihood model $p(x_i|\boldsymbol{\theta})$ is misspecified, simply investigate why exactly it fits the data poorly. After residual analysis, intense study of descriptive statistics and consultation with domain experts, redesign it to arrive at a likelihood model $p'(x_i|\boldsymbol{\theta}')$, which provides a better fit to the data and (approximately)

3. The KLD is the exception to this rule: As it depends on the log normalizer of $\pi(\boldsymbol{\theta})$ in an additive fashion, improper priors can still be admissible so long as eq. (3) yields a solution for the unnormalized version of the KLD as given in Cichocki and Amari (2010).

satisfies **(L)**. In other words, the traditional view is that any problem with misspecification is really a problem with careless modelling.

As outlined in Section 3.2, this strategy is neither practiced nor feasible with contemporary large-scale models. The naive interpretation of likelihoods as corresponding to an appropriately good description of the true data generating process in the sense of **(L)** is thus wholly inappropriate. This is especially important as many large-scale models are mainly interested in capturing the *typical* behaviour of the data—rather than *fully* modelling every aspect of a population. While this may appear to be a minor point at first glance, it has serious consequences for inferential practice. To see why, suppose a population contains a small number of outlying observations, local heterogeneities or spiky noise. The naive interpretation of the likelihood as in **(L)** *assumes* that these untypical aspects are encoded in the likelihood function. Hence, if x_i is an outlier, the inference machinery of traditional statistics interprets this as a strong signal: Since the likelihood model is an approximately correct description of the data, the most informative observations are those that do *not* fit the model fitted to the rest of the data. This is why aberrant parts of the data will have a disproportional impact on inference outcomes—leading standard inference methods to break down (see also Jewson et al., 2018).

Moreover, the often-invoked intuition that a sufficiently flexible likelihood family (such as likelihoods parameterized by Neural Networks) will not suffer these problems is dangerously incorrect in at least two ways: firstly, increasing the dimension of the model space for a fixed number of observations amounts to placing more weight on the prior—and so amounts to merely shifting the problem from **(L)** into **(P)**. Secondly, the symmetries and degeneracies of such likelihoods can be shown to induce generalization errors that increase with the number of observations n (see e.g. Watanabe, 2018, Example 19 and Remark 20).

3.5 Computation mismatch

As Theorem 1 shows, the Bayesian posterior $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ is the result of optimizing over the infinite-dimensional space $\mathcal{P}(\boldsymbol{\Theta})$. Generally, this implies that the posterior itself also does not live in a finite-dimensional space. In fact, the only case in which $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ can be represented through a finite-dimensional parameter is when prior and likelihood are conjugate to one another—a fact independently established by Koopman (1936), Pitman (1936), and Darmois (1935) and thus commonly referred to as Koopman-Pitman-Darmois Theorem. This means that inference with $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ is generally a hard problem, which manifests itself through the need to compute the posterior's normalizing constant. To address this problem, Markov Chain Monte Carlo algorithms are typically used. Such algorithms produce an exact representation of $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ if the chain runs indefinitely and collects infinitely many samples. In practice, collecting a finite number of samples from the chain yields can represent $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ almost exactly whenever $d = |\boldsymbol{\Theta}|$ is not too large. For large enough d however, the number of samples required to make the approximation useful is often too large to make samplers computationally viable: For example, in the *best* case scenario, Random Walk Metropolis Hastings scales like $\mathcal{O}(d^2)$ (Roberts et al., 1997), the Metropolis-adjusted Langevin algorithm like $\mathcal{O}(d^{4/3})$ (Roberts and Rosenthal, 1998) and Hamiltonian Monte Carlo like $\mathcal{O}(d^{5/4})$ (Beskos et al., 2013). Note that these results assume independence and Gaussianity—so in practice scaling rates are typically even worse.

Approximation strategies constitute an alternative way to avoid explicit computation of normalizing constants. These methods project $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ into some parameterized subset $\mathcal{Q} \subset \mathcal{P}(\boldsymbol{\Theta})$. Clearly then, they produce approximations $q^*_{\mathrm{A}}(\boldsymbol{\theta})$ of high quality only if the set $\mathcal{Q}$ is chosen to be sufficiently rich. In practice however, most posterior belief distributions $q^*_{\mathrm{A}}(\boldsymbol{\theta})$ computed this way barely deserve to be called approximations of $q^*_{\mathrm{B}}(\boldsymbol{\theta})$. For example, consider the mean field normal variational family given by

$$\mathcal{Q}_{\mathrm{MFN}} = \left\{ \prod_{j=1}^{d} \mathcal{N}(\boldsymbol{\theta}_j | \mu_j, \sigma_j^2) : \mu_j \in \mathbb{R}, \sigma_j^2 \in \mathbb{R}_{>0} \text{ for all } j \right\}. \tag{8}$$

For most interesting non-trivial posterior distributions $q^*_{\mathrm{B}}(\boldsymbol{\theta})$, there will not exist an element $q \in \mathcal{Q}_{\mathrm{MFN}}$ that could be considered a meaningful approximation to $q^*_{\mathrm{B}}(\boldsymbol{\theta})$. This is perhaps unsurprising: After all, $\mathcal{Q}_{\mathrm{MFN}}$ assumes $\mathcal{O}(d^2)$ independence relationships in the approximate posterior belief for $\boldsymbol{\theta}$. Worse still: As approximations are particularly attractive when $|\boldsymbol{\Theta}| = d$ is large, in practice we will resort to such insufficiently expressive "approximations" to $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ *precisely when* the elements in $\mathcal{Q}_{\mathrm{MFN}}$ are structurally most dissimilar from $q^*_{\mathrm{B}}(\boldsymbol{\theta})$. To improve the quality of these approximations, numerous directions of research have proposed ever more flexible variational families in order to make $\mathcal{Q}$ more expressive. Examples include implicit distributions (e.g. Tran et al., 2017; Tiao et al., 2018; Shi et al., 2018; Ma et al., 2019), normalizing flows (e.g. Rezende and Mohamed, 2015), or the variational Gaussian Process (Tran et al., 2016). Generally, there is no free lunch: more expressive families $\mathcal{Q}$ will incur higher computational cost and compound the issues with **(C)**.

In the current paper, we advocate an optimization-centric view of posterior belief computation. As a side-product of this view, we believe that it is often unhelpful to think of $q^*_{\mathrm{A}}(\boldsymbol{\theta})$ as an approximation to $q^*_{\mathrm{B}}(\boldsymbol{\theta})$. Rather, we prefer to think of $q^*_{\mathrm{A}}(\boldsymbol{\theta})$ as defining a new and distinct posterior belief distribution in its own right—which happens to also be an approximation to $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ if $\mathcal{Q}$ is sufficiently expressive.

To highlight the importance that computational considerations as summarized in **(C)** have played in research on statistical machine learning, we end their discussion by pointing to some of the recent research on Bayesian computation for Gaussian Processes.

Example 2 (large-scale Gaussian processes as violations of (C)) *Many Bayesian machine learning models prohibit exact computation. One particularly interesting case are Gaussian Process (*GP*) models: Even in the special cases where they admit closed form posteriors, it may well be impossible to compute them exactly for sufficiently large inference problems. The reason is that for n observations, direct computation of the associated* GP *posterior takes $\mathcal{O}(n^3)$ time. As a consequence, an entire literature is dedicated to bringing down this computational complexity (see for instance Williams and Seeger, 2001; Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006; Titsias, 2009) and developing software or computer-architecture specific methods geared towards inference with* GP*s (e.g. Matthews et al., 2017; Gardner et al., 2018; Balandat et al., 2019; Wang et al., 2019). Furthermore, with deep (i.e., hierarchical) approaches to* GP*s introduced in Damianou and Lawrence (2013) and extended in various directions (e.g. Dai et al., 2016; Hegde et al., 2019), this challenge has only become more important (see e.g. Bui et al., 2016; Cutajar et al., 2017b; Salimbeni and Deisenroth, 2017).*

4. The Rule of Three: Optimization-Centric Bayesian Inference

As the last sections have shown, the traditional view of Bayesian inference as an *update rule* relies on a number of assumptions that are not always a good basis for modern large-scale statistical inference. To address this, we propose a *optimization-centric* view on Bayesian methods. This view encompasses numerous exact Bayesian methods as well as variational approximations, can naturally relax the three assumptions underlying standard Bayesian posteriors, and can be interpreted as a generalization of the Bayesian paradigm. Developing these ideas proceeds in three steps:

> **Section 4.1** sets out axioms that are a minimal requirement for any posterior belief distribution. In accordance with these axioms, we derive the Rule of Three (RoT).
>
> **Sections 4.2 & 4.3** discuss the RoT as a recipe for producing posterior belief distributions and elaborate on its three interpretable ingredients. We also show how the RoT can **directly** address the concerns associated with imposing **(P)**, **(L)** and **(C)**.
>
> **Section 4.4** demonstrates that the axiomatic development is both helpful and useful by comparing the RoT with existing methods that generate belief distributions.

4.1 An axiomatic foundation for Bayesian inference

In this section, we give an axiomatic foundation for Bayesian inference that is flexible enough to relax **(P)**, **(L)** and **(C)**. Before doing so, we set out some preliminaries.

Definition 4 (Loss Function) *Losses are functions $L_n : \boldsymbol{\Theta} \times \mathcal{X}^n \to \mathbb{R}$ which are lower bounded. For observations $x_{1:n} \in \mathcal{X}^n$, their empirical risk minimizers are given by*

$$\widehat{\boldsymbol{\theta}}_n \in \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \{L_n(\boldsymbol{\theta}, x_{1:n})\}.$$

Definition 5 (Statistical Divergence) *Statistical divergences are functions $D : \mathcal{P}(\boldsymbol{\Theta}) \times \mathcal{P}(\boldsymbol{\Theta}) \to \mathbb{R}_{\geq 0}$ so that $D(q\|\pi) \geq 0$ and $D(q\|\pi) = 0 \iff q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$ almost everywhere.*

For simplicity, we will avoid introducing measure-theoretic notation in the remainder. To this end, we will assume that all probability measures of interest have densities with respect to the Lebesgue measure. We also slightly abuse notation in two ways: We write $q \in \mathcal{P}(\boldsymbol{\Theta})$ for probability densities $q(\boldsymbol{\theta})$ on $\boldsymbol{\Theta}$, even though probability densities are not in $\mathcal{P}(\boldsymbol{\Theta})$. However, $q(\boldsymbol{\theta})$ induces a measure $\mu_q \in \mathcal{P}(\boldsymbol{\Theta})$ as $\mu_q(A) = \int_A dq(\boldsymbol{\theta})$ for any measurable set $A \subset \boldsymbol{\Theta}$. Thus, whenever we write $q \in \mathcal{P}(\boldsymbol{\Theta})$, what we mean is that $\mu_q \in \mathcal{P}(\boldsymbol{\Theta})$. Similarly, when we write $q_1 \neq q_2$, we mean that there exists a measurable set $A \in \boldsymbol{\Theta}$ such that $\mu_{q_1^*}(A) \neq \mu_{q_2^*}(A)$.

Moreover, we will generally assume that the loss is additive and given by

$$L_n(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \tag{9}$$

unless explicitly stated otherwise. The additivity assumption unclutters notation: instead of having to specify an infinite sequence $\{L_n\}_{n \in \mathbb{N}}$ of functions specified on the sequence of

spaces $\{\boldsymbol{\Theta} \times \mathcal{X}^n\}_{n\in\mathbb{N}}$, we only need to define the function $\ell : \boldsymbol{\Theta} \times \mathcal{X} \to \mathbb{R}$. Note that while the axiomatic development is presented for ℓ only, the conclusions are unchanged if one uses arbitrary loss sequences $\{L_n\}_{n\in\mathbb{N}}$ instead.

Axiom I (Representation) *The posterior $q^* \in \mathcal{P}(\boldsymbol{\Theta})$ solves an optimization problem over some space $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$. For any finite sample $\{x_i\}_{i=1}^n$, the optimization problem's objective is increasing in two arguments:*

(i) An expected in-sample loss $\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$ taken with respect to $q^*(\boldsymbol{\theta})$.

(ii) The deviation from the prior $\pi(\boldsymbol{\theta})$ as measured by some statistical divergence D.

Reiterating the essence of Observation 1, Axiom I formalizes the optimization-centric view on Bayesian inference. More precisely, it tells us that for a fixed prior π, posteriors are specified through three parts: The loss ℓ, the divergence $D(\cdot\|\pi)$ and the space Π. Making this insight more precise, we can derive the following representation Theorem:

Theorem 6 (Form 1) *Under Axiom I, posterior belief distributions can be written as*

$$q^*(\boldsymbol{\theta}) = \underset{q\in\Pi}{\arg\min} \left\{ f\left(\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right], D(q||\pi) \right) \right\},$$

where $f : \mathbb{R}^2 \to \mathbb{R}$ is some function that may depend on $\pi, \Pi, \ell, \{x_i\}_{i=1}^n$, or D.

Proof This follows directly from Axiom I: Firstly, any posterior belief distribution $q^*(\boldsymbol{\theta})$ is the solution to an optimization problem over Π. Thus, for an appropriately structured objective Obj, one can write

$$q^*(\boldsymbol{\theta}) = \underset{q\in\Pi}{\arg\min} \{\mathrm{Obj}(q)\}.$$

By (i) and (ii) of Axiom I, we also know that the optimization's objective depends only on $D(q||\pi)$ and $\mathbb{E}_{q(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta}, x_i)]$. Clearly then, for some function $f : \mathbb{R}^2 \to \mathbb{R}$,

$$\mathrm{Obj}(q) = f\left(\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right], D(q||\pi) \right),$$

which completes the proof. ■

This result is a first and helpful step, but in itself does not suffice to yield objectives that are useful in practice. Specifically, we need to get a handle on the function f. It is clear that under Axiom I alone, very little can be said about f. Since our explicit target is a generalization of the Bayesian inference problem, we will have to restrict the form of f so that Theorem 6 admits only the Bayesian posterior whenever $D = \mathrm{KLD}$ and $\Pi = \mathcal{P}(\boldsymbol{\Theta})$.

Axiom II (Recovers Bayesian Posteriors) *Function f in Theorem 6 does not depend on $\pi, \Pi, \ell, \{x_i\}_{i=1}^n$, or D. Further, q^* is the posterior q_B^* of eq. (2) if $D = \mathrm{KLD}$, $\Pi = \mathcal{P}(\boldsymbol{\Theta})$.*

The intution of Axiom II is clear: in the case of q_{B}^*, $D = \mathrm{KLD}$, $\Pi = \mathcal{P}(\boldsymbol{\Theta})$; and f does not depend on the data $\{x_i\}_{i=1}^n$, the prior π, the loss ℓ, etc. As we want to recover q_{B}^*, we thus impose the same conditions for other posteriors. Fortunately, this also drastically simplifies the representation of Theorem 6.

Theorem 7 *Suppose the posterior belief $q^* \in \mathcal{P}(\boldsymbol{\Theta})$ satisfies Axioms I and II. Then the objective of Theorem 6 can be identified as $f(x, y) = x + y$ so that*

$$q^*(\boldsymbol{\theta}) = \underset{q \in \Pi}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \right] + D(q \| \pi) \right\} = P(\ell, D, \Pi). \tag{10}$$

Proof This follows by combining Theorem 6 with Axiom II and eq. (3). ■

The last result is of crucial importance for the further development of the paper: Specifically, eq. (10) suggests a flexible recipe for the optimization-centric design of new posterior distributions.

Remark 8 *Theorem 7 demonstrates that in combination with Axiom I, Axiom II enforces an additive relationship between the expected loss and prior regularizer. This additive relationship is desirable for a number of reasons, some of which include*

- ***Invariance to additive, but not multiplicative constants**: adding constants to ℓ will not change the posterior. In other words for, any $C \in \mathbb{R}$, we have $P(\ell, D, \Pi) = P(\ell + C, D, \Pi)$. However, multiplying ℓ by w (or equivalently, D by $\frac{1}{w}$) for some $w \in \mathbb{R}$ changes the posterior that is computed. This means that we recover a well-known feature of other Bayes-like procedures. In fact, exponentiating likelihoods $p(\cdot|\boldsymbol{\theta})^w$—which is equivalent to multiplying $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$ with some $w \in (0, 1)$—is a popular tool in the existing literature on generalized Bayesian methods with $D = \text{KLD}$ (e.g. Grünwald, 2011, 2012; Holmes and Walker, 2017; Grünwald and Van Ommen, 2017; Miller and Dunson, 2019) and serves to up-weight (or down-weight) the information of the data relative to the prior.*

- ***Recovery of (D-approximated) prior without additional Information**: Given no information from the observations (i.e. if $\ell = 0$), the solution of the optimization problem in Theorem 7 is the member of admissible set Π that is closest to the prior $\pi(\theta)$ as measured by D. Put differently, $P(\ell = 0, D, \Pi) = \arg\min_{q \in \Pi} D(q \| \pi)$. Clearly then, if $\pi \in \Pi$ we have that $P(\ell = 0, D, \Pi) = \pi$.*

- ***Generalized (weak) likelihood principle**: Data $x_{1:n}$ favours θ_1 over θ_2 if and only if $\sum_{i=1}^{n} \ell(\boldsymbol{\theta}_1, x_i) < \sum_{i=1}^{n} \ell(\boldsymbol{\theta}_2, x_i)$. This is the natural generalization of the 'weak likelihood principle' outlined by Sober (2008) from $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$ to general loss functions , see also Mayo-Wilson and Saraf (2020).*

These implications of Theorem 7 are a natural requirement for any belief distribution trading off prior against data-driven information.

4.2 The Rule of Three

Following the axiomatic developments of the last section that culminated in Theorem 7, we now discuss the interpretations and theoretical properties of posterior belief distributions generated from objectives as in eq. (10). To simplify the representation throughout the remainder, we first define notation for posteriors of this form.

Definition 9 (Rule of Three (RoT)) *For observations $x_{1:n}$, a prior π, a space $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$, a loss function $\ell : \boldsymbol{\Theta} \times \mathcal{X} \to \mathbb{R}$ and a divergence $D(\cdot\|\pi) : \Pi \to \mathbb{R}_{\geq 0}$, we say that a posterior has been constructed via the Rule of Three (*RoT*) if it can be written as*

$$q^*(\boldsymbol{\theta}) = P(\ell, D, \Pi) = \operatorname*{arg\,min}_{q \in \Pi} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \right] + D(q\|\pi) \right\}.$$

Here, $P(\ell, D, \Pi)$ is a short-hand notation for the RoT *suppressing dependence on $x_{1:n}$ and π.*

In the remainder, we study posteriors taking the form $P(\ell, D, \Pi)$. Just as for standard variational posteriors, it is generally hard to establish existence and uniquenes when Π is a variational family. If $\Pi = \mathcal{P}(\Theta)$ however, $\mathcal{P}(\ell, D, \Pi)$ exists whenever D is convex in its first argument by elementary analysis. If D is strictly convex in its first argument, this minimizer is also guaranteed to be unique. More elaborate arguments can be deployed to prove existence for non-convex divergences by imposing additional assumptions on the loss function and using Prokhorov's Theorem (see Knoblauch (2019a), Lemma 1).

The most practically relevant versions of these posteriors take Π to be a $\boldsymbol{\kappa}$-parameterized family of distributions $\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) : \boldsymbol{\kappa} \in \boldsymbol{K}\}$, a special case we explore in Section 5. Since such parameterized sets $\mathcal{Q}$ are commonly called *variational families*, we call the act of computing posteriors of the form $P(\ell, D, \mathcal{Q})$ Generalized Variational Inference (GVI).

Before we proceed to study GVI, we first give an interpretation of the three components in the RoT. In particular, we show that these components directly address the three violated assumptions **(↯P)**, **(↯L)** and **(↯C)** of standard Bayesian inference via an intuitive modularity result. Beyond that, we recover existing Bayesian methods as special cases of the RoT. Lastly, we discuss the meaning of a Bayesian method (not) being representable via $P(\ell, D, \Pi)$ and use this as a springboard to motivate GVI.

4.3 Modularity of the Rule of Three

Each component of the optimization problem defined by the posterior $P(\ell, D, \Pi)$ serves a specific and separate purpose.

(▶L) A **loss** $\ell : \boldsymbol{\Theta} \times \mathcal{X} \to \mathbb{R}$. The loss defines the parameter of interest $\boldsymbol{\theta}$ by linking it to the observations $x_{1:n}$. To simplify presentation, we will assume that all losses are additive and identical[4], that they depend on a parameter $\boldsymbol{\theta}$ rather than a latent variable[5], and that they are deterministic without dependence on (local or global) latent variables[6].

(▶P) A divergence $D : \mathcal{P}(\boldsymbol{\Theta}) \times \mathcal{P}(\boldsymbol{\Theta}) \to \mathbb{R}_+$ that **regularizes** the posterior by penalizing deviations from the prior $\pi(\boldsymbol{\theta})$. Note that D determines the nature of the uncertainty

4. As pointed out above, losses are not required to be identical. For instance, we could replace $\ell(\boldsymbol{\theta}, x_i)$ by $\ell_i(\boldsymbol{\theta}, x_i)$ and set $\ell_i(\boldsymbol{\theta}, x_i) = \ell(\boldsymbol{\theta}, x_i|x_{1:i-1})$. Here, the x_i-th observation is conditioned on the first $i-1$ observations as is common in time series models. Generally speaking, conditional dependence is easy to incorporate into the RoT at the expense of complicating notation, see also Knoblauch (2019a).
5. This requirement is easily relaxed: For instance, in the experiments on Deep Gaussian Processes in Section 6, all losses are defined relative to latent variables.
6. While latent variable models are not the focus of the current paper, the RoT and GVI are easily extended to the latent variable case, see Knoblauch (2019a).

induced by π. To see this, consider $D = 0$ and the (non-RoT) problem

$$\widehat{q}(\boldsymbol{\theta}) = \operatorname*{arg\,min}_{q \in \mathcal{P}(\boldsymbol{\Theta})} \left\{ \mathbb{E}_q \left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \right] \right\}. \tag{11}$$

Denoting $\widehat{\boldsymbol{\theta}}_n = \operatorname{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \{\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\}$ and $\delta_y(x)$ as the Dirac measure at y, it is clear that $\widehat{q}(\boldsymbol{\theta}) = \delta_{\widehat{\boldsymbol{\theta}}_n}(\boldsymbol{\theta})$, which holds as $\delta_{\widehat{\boldsymbol{\theta}}_n} \in \mathcal{P}(\boldsymbol{\Theta})$. Clearly, the absence of D corresponds to the absence of any uncertainty in the posterior. Similarly, the nature of D determines the nature in which uncertainty about $\boldsymbol{\theta}$ is quantified.

(▶C) A set of **feasible posteriors** $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$: By definition, any $q \in \Pi$ is a feasible solution for the optimization problem associated to the posterior $P(\ell, D, \Pi)$.

Interestingly, each of these three arguments directly addresses one of the problems **(↯P)**, **(↯L)** and **(↯C)** in Section 3: Firstly, the loss ℓ determines the parameter and thus can be used to tackle model misspecification and other violations of **(L)**. Secondly—assuming one has specified the best possible prior—the divergence D can tackle **(P)** by shaping the nature in which priors affect uncertainty quantification. Thirdly, the choice of Π can directly address **(C)**: The more computational power is available, the more complex Π is allowed to become. Formalizing this intuitive modularity, we arrive at the following result:

Theorem 10 (RoT modularity) *Hold $x_{1:n}$, $n, \pi(\boldsymbol{\theta})$ and Π fixed. Let $q_1^*(\boldsymbol{\theta}) \in \Pi = P(\ell, D, \Pi)$. If one wishes to derive an alternative posterior $q_2^*(\boldsymbol{\theta}) \in \Pi$ through the* RoT

(1) which avoids or is robust to model misspecification, this amounts to changing ℓ.

(2) which is robust to prior misspecification without changing the parameter of interest, this amounts to changing D.

(3) which affects quantification of uncertainty without changing the parameter of interest, this amounts to changing D.

Since giving a complete account of the necessary arguments and definitions (such as the definition for robustness) is somewhat laborious, we defer definitions and the proof to Appendix C.

4.4 Connecting the Rule of Three to existing methods

As summarized in Table 1, most existing methods with Bayesian flavour are special cases of $P(\ell, D, \Pi)$. This includes a wide range of approximate Bayesian methods, including **standard Variational Inference (VI)**. In the following paragraphs, we elaborate on some of the most important connections.

4.4.1 Standard Variational Inference (VI)

One of the perhaps most surprising entries in Table 1 are **standard VI** posteriors: the RoT does *not* judge the full Bayesian posterior q_{B}^* to be preferable to **standard VI** posteriors q_{VI}^* *by default*. The reason for this is simple: Unlike the traditional Bayesian paradigm,

Method	$\ell(\boldsymbol{\theta}, x_i)$	D	Π
Standard Bayes	$-\log p(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Power Likelihood Bayes[1]	$-\log p(x_i\|\boldsymbol{\theta})$	$\frac{1}{w}$KLD, $w < 1$	$\mathcal{P}(\boldsymbol{\Theta})$
Composite Likelihood Bayes[2]	$-w_i \log p(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Divergence-based Bayes[3]	divergence-based ℓ	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Gibbs/PAC-Bayes [4]	any ℓ	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
VAE[5,†]	$-\log p_{\boldsymbol{\zeta}}(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{Q}$
β-VAE[6,†]	$-\log p_{\boldsymbol{\zeta}}(x_i\|\boldsymbol{\theta})$	$\beta \cdot$ KLD, $\beta > 1$	$\mathcal{Q}$
Bernoulli-VAE[7,†]	continuous Bernoulli	KLD	$\mathcal{Q}$
Standard VI	$-\log p(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{Q}$
Power VI[8]	$-\log p(x_i\|\boldsymbol{\theta})$	$\frac{1}{w}$KLD, $w < 1$	$\mathcal{Q}$
Utility VI[9]	$-\log p(x_i\|\boldsymbol{\theta}) + \log u(h, x_i)$	KLD	$\mathcal{Q}$
Regularized Bayes[10]	$-\log p(x_i\|\boldsymbol{\theta}) + \phi(\boldsymbol{\theta}, x_i)$	KLD	$\mathcal{Q}$
Gibbs VI[11]	any ℓ	KLD	$\mathcal{Q}$
Generalized VI	any ℓ	any D	$\mathcal{Q}$

Table 1: Relationship of $P(\ell, D, Q)$ to a selection of existing methods. [1](e.g. Grünwald, 2011, 2012; Holmes and Walker, 2017; Grünwald and Van Ommen, 2017; Miller and Dunson, 2019), [2](e.g. Varin et al., 2011; Pauli et al., 2011; Ribatet et al., 2012; Hamelijnck et al., 2019), [3](e.g. Hooker and Vidyashankar, 2014; Ghosh and Basu, 2016; Futami et al., 2018; Jewson et al., 2018; Chérief-Abdellatif and Alquier, 2019a), [4](Bissiri et al., 2016; Germain et al., 2016; Guedj, 2019; Syring and Martin, 2019), [5](Kingma and Welling, 2013), [6](Higgins et al., 2017), [7](Loaiza-Ganem and Cunningham, 2019) [8](e.g. Yang et al., 2017; Huang et al., 2018) [9](e.g. Kuśmierczyk et al., 2019; Lacoste-Julien et al., 2011) [10](Ganchev et al. (2010), but only if the regularizer can be written as $\mathbb{E}_{q(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}, \boldsymbol{x})]$ as in Zhu et al. (2014)), [11](e.g. Alquier et al., 2016) [†]For notational clarification for the VAE entries in the table, see Section 4.4.4.

the RoT explicitly encodes the availability of finite computational resources through the argument Π. Hence, when computational resources are scarce and posterior beliefs can only be computed over a parameterized set $\mathcal{Q} \subset \mathcal{P}(\boldsymbol{\Theta})$, **standard VI** produces the best computationally feasible posterior. In this sense, the RoT respects the optimality result of **standard VI** presented in Theorem 2.

4.4.2 Coherence and the RoT

Unlike previous generalizations such as the Generalized Bayesian update in eq. (2), posteriors generated through the RoT are allowed to break a property referred to as *coherence* or *Bayesian additivity* (e.g. Bissiri et al., 2016; Fong and Holmes, 2019). In a nutshell, coherence says that posteriors have to be generated according to some function $\psi : \mathbb{R}^2 \to \mathbb{R}$ which for

the prior $\pi(\boldsymbol{\theta})$ and loss terms $\ell(\boldsymbol{\theta}, x_1), \ell(\boldsymbol{\theta}, x_2)$ behaves as

$$\psi\left(\ell(\boldsymbol{\theta}, x_2), \psi\left(\ell(\boldsymbol{\theta}, x_1), \pi(\boldsymbol{\theta})\right)\right) = \psi\left(\ell(\boldsymbol{\theta}, x_1) + \ell(\boldsymbol{\theta}, x_2), \pi(\boldsymbol{\theta})\right).$$

Effectively, this property enforces a multiplicative update via exponential additivity as in eq. (2). For the standard Bayesian posterior, the desirability of coherence is a *direct* result of assuming **(P)** and **(C)**. To see this, note that treating the prior belief according to **(P)** and assuming infinite computational power via **(C)** is *exactly* equivalent to setting $D = \text{KLD}$ and $\Pi = \mathcal{P}(\boldsymbol{\Theta})$. Solving eq. (3) with these specifications as in the proof of Theorem 1, one now obtains the coherent exponentially additive update rule in eq. (2). In other words, enforcing coherence is reasonable *only if* **(P)** and **(C)** can be assumed to hold. Conversely, posteriors that violate coherence do not have to rely on **(P)** and **(C)**—which is precisely what we set out to do in the first place.

4.4.3 PAC-Bayes

While PAC-Bayesian results often have intimate links with Bayesian inference (see e.g. Germain et al., 2016; Grünwald and Van Ommen, 2017), their motivations and origins are distinct (see e.g. Shawe-Taylor and Williamson, 1997; Guedj, 2019): Unlike Bayesian inference, PAC-Bayesian results are not constructed based on **(P)** and **(L)**. Rather, their aim is the derivation of generalization bounds for belief distributions $q(\boldsymbol{\theta}) \in \mathcal{P}(\boldsymbol{\Theta})$ defined over some hypothesis space (corresponding to the parameter space $\boldsymbol{\Theta}$) relative to a loss function (corresponding to ℓ). For example, under a prior belief $\pi(\boldsymbol{\theta})$, a loss ℓ and a data generating mechanism for $x_{1:n}$ satisfying appropriate regularity conditions and for any $q(\boldsymbol{\theta}) \in \mathcal{P}(\boldsymbol{\Theta})$ as well as for any fixed value of $\varepsilon > 0$, McAllester's seminal bound (McAllester, 1999a,b) says that with probability at least $1 - \varepsilon$,

$$\mathbb{E}_{q(\boldsymbol{\theta})}\left[\mathbb{E}_{\boldsymbol{x}_{1:n}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right]\right] \leq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i)\right] + \sqrt{\frac{\text{KLD}(q, \pi) + \log\frac{2\sqrt{n}}{\varepsilon}}{2n}}. \quad (12)$$

Minimizing the right hand side of this bound with respect to $q(\boldsymbol{\theta})$ over some set $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$ immediately recovers a special case for the RoT given by $P(\ell, D_{\text{McAllester}}, \Pi)$. Here, $D_{\text{McAllester}}$ is just the last term of the above bound, with a subtracted constant and rescaled by n:

$$D_{\text{McAllester}}(q\|\pi) = \sqrt{n} \cdot \left(\sqrt{\frac{\text{KLD}(q, \pi) + \log\frac{2\sqrt{n}}{\varepsilon}}{2}} - \sqrt{\frac{\log\frac{2\sqrt{n}}{\varepsilon}}{2}}\right).$$

Subtraction of the constant ensures that $D_{\text{McAllester}}(q\|\pi) = 0$ if and only if $\pi(\boldsymbol{\theta}) = q(\boldsymbol{\theta})$ (almost everywhere). The rescaling is necessary as we have to multiply both sides of eq. (12) by n to bring them into the RoT form. Note that neither the addition of the constant nor the rescaling affects the minimizer of the right hand side.

A similar logic can be applied to virtually all PAC-Bayesian bounds, crucially also for bounds based on divergences other than the KLD such as those of Bégin et al. (2016), Alquier and Guedj (2018), or Ohnishi and Honorio (2020).[7] In these settings, GVI—the

7. For more classical bounds based on $D = \text{KLD} \neq D_{\text{McAllester}}$, see Catoni (2007); Zhang (2006).

tractable case when $\Pi = \mathcal{Q}$ is a parameterized subset of $\mathcal{P}(\boldsymbol{\Theta})$—is a promising way forward to scale and operationalize PAC-Bayesian learning. In fact, since the current manuscript was first made publicly available, the work of Letarte et al. (2019) and Alquier (2020) constitute the first steps in this direction. More generally speaking, PAC-Bayesian analysis may prove crucial in deciding which divergence should be used for inference in a given problem: The bounds of Bégin et al. (2016), Alquier and Guedj (2018), and Ohnishi and Honorio (2020) all depend on divergences other than the KLD, and provide generalization guarantees for less restrictive settings than the KLD. For example, the bounds of Alquier and Guedj (2018) depend on f-divergences, and provide generalization guarantees even if the observation sequence exhibits a substantial degree of heterogeneity or temporal dependence. Similarly, unlike bounds based on the KLD, the bounds of Ohnishi and Honorio (2020) provide generalization guarantees even if ℓ is an unbounded loss function.

4.4.4 Latent Variable Models & Variational Autoencoders

While we have thus far stated the entire development in terms of a single *global* latent variable $\boldsymbol{\theta}$, nothing stops us from extending the presented ideas to *local* latent variables. The reason for this is that none of our Axioms prohibit $\boldsymbol{\Theta}$ or Π to depend on n or indeed $x_{1:n}$. In other words, we can seamlessly transfer everything we considered thus far to the context of inference on local latent variables $\boldsymbol{z}_{1:n} \in \mathcal{Z}^n$ by taking $\boldsymbol{\Theta} = \boldsymbol{\Theta}(n) = \mathcal{Z}^n$.

To make this logic more tangible, we will explain how Variational Autoencoders (VAEs) (Kingma and Welling, 2013) can be recast in the RoT form. VAEs use local latent variables, in our notation $\boldsymbol{\theta} = \boldsymbol{\theta}_{1:n}$, to encode lower dimensional representations of observations $x_{1:n}$ via the global parameter $\boldsymbol{\kappa}_g$. Simultaneously, they seek to probabilistically decode the latent variables back to the observation space via the global decoder model with parameters ζ. This involves an optimisation problem over a set of distributions for the latent variables. The corresponding variational family is

$$\Pi_{x_{1:n}} = \left\{ q(\boldsymbol{\theta}|\boldsymbol{\kappa}_g) = \prod_{i=1}^{n} q(\boldsymbol{\theta}_i|\boldsymbol{\kappa}_i) \text{ so that } q(\boldsymbol{\theta}_i|\boldsymbol{\kappa}_i) = \mathcal{N}(\boldsymbol{\theta}_i; \mu(\boldsymbol{\kappa}_g, x_i), \sigma(\boldsymbol{\kappa}_g, x_i) \right\},$$

where the parameters $\boldsymbol{\kappa}_i = (\boldsymbol{\kappa}_g, x_i)$ consist of a fixed local component observation x_i as well as the global parameter $\boldsymbol{\kappa}_g$ that is shared to be optimized over. Here, $\boldsymbol{\kappa}_g$ will define the weights of a neural network indexing a probabilistic model. The optimization problem underlying a VAE is now given by

$$\underset{\zeta, q \in \Pi_{x_{1:n}}}{\arg\min} \left\{ \sum_{i=1}^{n} \mathbb{E}_{q(\boldsymbol{\theta}_i|\boldsymbol{\kappa}_i)} \left[-\log p_\zeta(x_i|\boldsymbol{\theta}_i) \right] + \sum_{i=1}^{n} \mathrm{KLD}\left(q(\boldsymbol{\theta}_i|\boldsymbol{\kappa}_i) \| \pi(\boldsymbol{\theta}_i) \right) \right\}.$$

where $\sum_{i=1}^{n} \mathbb{E}_{q(\boldsymbol{\theta}_i|\boldsymbol{\kappa}_i)} \left[-\log p_\zeta(x_i|\boldsymbol{\theta}_i) \right]$ minimises the expected reconstruction error of decoding the probabilistic encoding and the KLD term regularises this encoding to improve the model's capacity to generate realistic pseudo-observations. Now simply note that for the fully factorized prior $\pi(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(\boldsymbol{\theta}_i)$, one can rewrite the above as

$$\underset{\zeta, q \in \Pi_{x_{1:n}}}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa}_g)} \left[\sum_{i=1}^{n} -\log p_\zeta(x_i|\boldsymbol{\theta}_i) \right] + \mathrm{KLD}\left(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_g) \| \pi(\boldsymbol{\theta}) \right) \right\}, \tag{13}$$

which is a RoT form with an added optimization over the hyperparameter ζ.[8] An important distinction between this example and many of the others in Table 1 is that for VAEs, the variational distributions are introduced in order to regularise the latent space rather than to approximate an underlying Bayesian posterior. As a result, the VAE objective exists solely as a means to generate desirable generative distributions for a particular inference tasks.

4.4.5 Links with Information Theory

One can also draw a close connection between the RoT and another latent variable model: the Predictive Information Bottleneck (PIB) (see Tishby et al., 2000; Bialek et al., 2001). Given a data generating process ϕ so that $\boldsymbol{x}_{1:\infty} \sim \phi$ and a compressed representation $\boldsymbol{\theta}$ of the random variables $\boldsymbol{x}_{1:n}$, the PIB poses the following optimization problem:

$$q^*(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) = \underset{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) \in \Pi_{\text{PIB}}}{\arg\min} \{-I(\boldsymbol{\theta}, \boldsymbol{x}_{n+1:\infty})\} \quad \text{s.t. } I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}) \leq I_0, \tag{14}$$

where all random variables admit densities p with respect to the Lebesgue measure,

$$I(\boldsymbol{Z}, \boldsymbol{Y}) = \text{KLD}\left(p(\boldsymbol{Z}, \boldsymbol{Y}) \| p(\boldsymbol{Z})p(\boldsymbol{Y})\right)$$

denotes the mutual information between random variables $\boldsymbol{Z}$ and $\boldsymbol{Y}$, and

$$\Pi_{\text{PIB}} = \left\{ q \in \mathcal{P}(\boldsymbol{\Theta}|\mathcal{X}^n) : \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) p(\boldsymbol{x}_{1:n}) d\boldsymbol{\theta} = p(\boldsymbol{x}_{1:n}) \right\}$$

is the set of admissible conditional distributions. This shows that the PIB maximizes the mutual information $I(\boldsymbol{\theta}, \boldsymbol{x}_{n+1:\infty})$ between the future $\boldsymbol{x}_{n+1:\infty}$ and the compression (i.e. model) $\boldsymbol{\theta}$ subject to an upper bound I_0 on the mutual information $I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n})$ between said model and the distribution of the training data $\boldsymbol{x}_{1:n}$. The PIB owes its name to the requirement that $I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}) \leq I_0$: in words, this bound prevents the compression from being arbitrarily expressive and forces us to squeeze the information contained in $\boldsymbol{x}_{1:n}$ through a bottleneck.

This PIB form is generally hard to solve, but can be rewritten as a RoT-like objective

$$q^*(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) = \underset{q \in \Pi_{\text{PIB}}}{\arg\min} \left\{ \mathbb{E}_q \left[L_{n,\text{PIB}}(q) \right] + D_{\text{PIB}}(q \| \pi_{\text{PIB}}) \right\}.$$

The nature of $L_{n,\text{PIB}}$ and D_{PIB} as well as mathematical details for arriving at this form are deferred to Appendix D. While the structure of the problem closely resembles that of Definition 9, there are some important differences. Most important among them, the PIB relates to the full distributional characterizations of the random variables $\boldsymbol{x}_{1:n}$ via $p(\boldsymbol{x}_{1:n})$—rather than to any actual observations $x_{1:n}$. As a consequence, the space of feasible solutions Π_{PIB} contains *all* possible conditional distributions $\{q^*(\boldsymbol{\theta}|x_{1:n})\}_{x_{1:n} \in \mathcal{X}^n}$—rather than a single conditional distribution $q^*(\boldsymbol{\theta}|x_{1:n})$ depending on a single realization $x_{1:n}$ of $\boldsymbol{x}_{1:n}$ only.

8. Optimizing over hyperparameters in variational objectives is very common, and our experiments in Section 6 make use of this technique, too. While optimizing over hyperparameters is strictly speaking not part of the RoT definition, we treat and discuss objectives of this kind essentially as members of the RoT.

As shown in Alemi (2019) however, the PIB can also be variationally lower-bounded and approximated with observations $x_{1:n}$ to arrive at the usual data-dependent form of the RoT. Specifically, if we are willing to assume that $\boldsymbol{x}_{1:n}$ are independent, then we may rewrite $p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta}) = \prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{\theta})$. This allows the coarse approximation $\mathbb{E}_{p(\boldsymbol{x}_{1:n})}\left[\log p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta})\right] = \mathbb{E}_{p(\boldsymbol{x}_{1:n})}\left[\sum_{i=1}^n \log p(\boldsymbol{x}_i|\boldsymbol{\theta})\right] \approx \sum_{i=1}^n \log p(x_i|\boldsymbol{\theta})$, which replaces dependence on $p(\boldsymbol{x}_{1:n})$ by dependence on a finite sample $x_{1:n}$. This yields the approximate lower bound

$$\begin{aligned} I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}) &\geq H(\boldsymbol{x}_{1:n}) + \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})}\left[\mathbb{E}_{p(\boldsymbol{x}_{1:n})}\left[\log p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta})\right]\right] \\ &\approx H(\boldsymbol{x}_{1:n}) + \mathbb{E}_{p(\boldsymbol{\theta}|x_{1:n})}\left[\sum_{i=1}^n \log p(x_i|\boldsymbol{\theta})\right] \end{aligned}$$

For any $\pi \in \mathcal{P}(\boldsymbol{\Theta})$, we can also derive another approximate upper bound via

$$\begin{aligned} I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty}) &\leq \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})p(\boldsymbol{x}_{1:n})}\left[\log\left(\frac{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})}{\pi(\boldsymbol{\theta})}\right)\right] \\ &\approx \mathbb{E}_{p(\boldsymbol{\theta}|x_{1:n})}\left[\log\left(\frac{p(\boldsymbol{\theta}|x_{1:n})}{\pi(\boldsymbol{\theta})}\right)\right] \end{aligned}$$

Writing out the resulting bound and minimizing over $p(\boldsymbol{\theta}|x_{1:n}) \in \mathcal{P}(\boldsymbol{\Theta})$, we find that its minimizer is $P(-\log p(\cdot|\boldsymbol{\theta}), \beta\mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$.

5. Generalized Variational Inference (GVI)

We now introduce a version of the RoT that is feasible for real-world inference and that we call Generalized Variational Inference (GVI). For $\mathcal{Q}$ a parameterized subset of $\mathcal{P}(\boldsymbol{\Theta})$ (i.e., a variational family), GVI posteriors are given by $P(\ell, D, \mathcal{Q})$. We proceed as follows:

Section 5.1 motivates why GVI generates conceptually appealing posteriors.

Section 5.2 motivates choosing non-standard losses ℓ and divergences D. Particular emphasis is placed on practical advice for making posterior inferences robust to model and prior misspecification.

Section 5.3 discusses two theoretical guarantees for GVI: frequentist consistency and an interpretation as approximate lower bound on the evidence of a (generalized) Bayesian posterior.

Section 5.4 focuses on strategies for inference. We derive closed form objectives for a set of GVI posteriors that are robust to model misspecification, introduce black box inference for GVI, and provide results on closed form divergence expressions.

5.1 Operationalizing the Optimization-Centric View on Bayesian Inference

The driving force of the development thus far has been the idea that undesirable inference outcomes are synonymous with an inappropriately designed optimization objective—an observation we call the optimization-centric view on Bayesian inference. Following this line of reasoning, the most transparent way of improving posteriors is a *direct* adjustment of the optimization problem that generated them. Conveniently, Definition 9 provides a way to specify posteriors precisely this way.

Definition 11 (Generalized Variational Inference (GVI)) *Solving any* RoT *of form* $P(\ell, D, \mathcal{Q})$ *for* $\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) : \boldsymbol{\kappa} \in \boldsymbol{K}\}$ *a parameterized subset of* $\mathcal{P}(\boldsymbol{\Theta})$ *(also called a variational family) constitutes a procedure we call Generalized Variational Inference (*GVI*).*

GVI posteriors have a number of desirable properties. Of particular practical importance is that they inherit the modularity result of Theorem 10. The ramifications are threefold:

(1) GVI can address model misspecification by changing ℓ,

(2) GVI can address prior misspecification by changing D,

(3) GVI can address undesirable uncertainty quantification by changing D.

In the context of potential misspecification problems, this modularity means that GVI posteriors $P(\ell, D, \mathcal{Q})$ are appealing alternatives to $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ or $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$. More precisely, if one can identify whether the assumptions underlying standard Bayesian inference are violated via the likelihood or the prior, GVI can be used to address this in a modular and optimization-centric manner by directly modifying ℓ or D. This means that GVI has an inherently different motivation from other variational methods (such as standard VI or DVI): rather than seeking to approximate $q^*_{\mathrm{B}}(\boldsymbol{\theta})$, GVI designs and computes an inherently different—and hopefully better-suited—posterior belief.

Many Bayesian practitioners may argue that this feature makes GVI *less* desirable than alternative variational methods: why would we prefer these posteriors over approximations to $q^*_{\mathrm{B}}(\boldsymbol{\theta})$? In principle, this is a valid point: in fact, if the assumptions underlying Bayesian inference are at least approximately correct, and if $\mathcal{Q}$ contains qualitatively good approximations to $q^*_{\mathrm{B}}(\boldsymbol{\theta})$, one will *want* to use a method that is motivated as approximation to $q^*_{\mathrm{B}}(\boldsymbol{\theta})$. Yet—even if likelihoods and priors are correctly specified—thinking of variational methods as approximations is often misleading: In many applications, the set $\mathcal{Q}$ does not contain any distributions that can approximate $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ in any meaningful way. In this setting, variational methods seeking to approximate $q^*_{\mathrm{B}}(\boldsymbol{\theta})$ have a clear drawback when compared to GVI posteriors: they are not interpretable as a modularly specified belief distribution—and so their behaviour can have rather undesirable side-effects. We demonstrate this in Example 3 and Figure 3, and will revisit this issue with our experiments in Section 6.1, where we observe its real world consequence on Bayesian Neural Networks.

Example 3 (Label switching and multi-modality) *A recurrent theme in the research on variational approximations* $q^*_A(\boldsymbol{\theta})$ *to* $q^*_B(\boldsymbol{\theta})$ *is the observation that if* $\mathcal{Q}$ *is a mean field normal family,* $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ *will center closely around the maximum likelihood estimate (e.g. Turner and Sahani, 2011). This phenomenon is often referred to as the* ***zero-forcing*** *behaviour of the* KLD *(Minka, 2005). Its effect are undesirably overconfident variational posteriors* $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$*. Moreover, this problem is especially pronounced when the approximated posterior beliefs* $q^*_B(\boldsymbol{\theta})$ *are multi-modal. Popular approaches to address this issue are Expectation Propagation (*EP*) (Minka, 2001; Opper and Winther, 2000) and Divergence Variational Inference (*DVI*) methods as introduced in Section 2.3.2 (e.g. Hernández-Lobato et al., 2016; Li and Turner, 2016; Dieng et al., 2017). All of these approaches seek to (locally or globally) minimize an alternative* ***zero-avoiding*** *divergence D between* $\mathcal{Q}$ *and* $q^*_B(\boldsymbol{\theta})$*. Unlike with* GVI*,* ***changing the divergence in the DVI-sense no longer affects uncertainty***

quantification alone. *In other words, we may accidentally interfere with the loss and warp the way the goodness of a parameter value* $\boldsymbol{\theta}$ *is assessed in undesirable ways.*

*Using Bayesian mixture models (*BMM*s), we show that this is indeed a problem in practice.* BMM*s produce multi-modal posteriors as the likelihood function is invariant to switching parameter labels. In other words,* BMM*s have multiple parameter values that constitute equally good fits to the data. With this in mind, we simulate* $n = 100$ *observations from*

$$p\left(x|\boldsymbol{\theta} = (\mu_1, \mu_2)\right) = 0.5 \cdot \mathcal{N}\left(x|\mu_1, 0.65^2\right) + 0.5 \cdot \mathcal{N}\left(x|\mu_2, 0.65^2\right)$$

for two different parameterizations 1) $\boldsymbol{\theta} = (0, 0.75)$ *and 2)* $\boldsymbol{\theta} = (0, 2)$*. For inference, we use the well-specified prior belief* $\mu_j \sim \mathcal{N}(0, 2^2)$*,* $j = 1, 2$*. Using the correctly specified likelihood function* $\ell(\boldsymbol{\theta}, x_i) = -\log p\left(x_i|\boldsymbol{\theta} = (\mu_1, \mu_2)\right)$*, we compare the standard Bayesian posterior* $q^*_B(\boldsymbol{\theta})$*, the standard* VI *posterior* $q^*_{\text{VI}}(\boldsymbol{\theta})$*, a* DVI *posterior based on Rényi's* α*-divergence (*$D^{(\alpha)}_{AR}$*) as described by Li and Turner (2016), and a* GVI *posterior using* $D = D^{(\alpha)}_{AR}$ *(see eq.* (15) *and Appendix Definition 22). For* $\mathcal{Q}$*, we use the collection of fully-factorized normals on* Θ*.*

Figure 3 shows the results. Because $p\left(x|\boldsymbol{\theta} = (\mu_1, \mu_2)\right) = p\left(x|\boldsymbol{\theta} = (\mu_2, \mu_1)\right)$*, there are two equally good parameter values describing the data—implying that the full posterior* $q^*_B(\boldsymbol{\theta})$ *is bi-modal. By choice of* $\mathcal{Q}$ *however, the variational* DVI *and* GVI *posteriors are unimodal, which endows them with a straightforward interpretation: firstly, the modes of these posteriors should correspond to (one of the two) best parameter values of* $\boldsymbol{\theta} = (\mu_1, \mu_2)$*. Secondly, their variances quantify the uncertainty about this best value. For both settings of the true value for* $\boldsymbol{\theta}$*,* DVI *produces a posterior that reflects a highly undesirable belief: the mode of the* DVI *posterior is located at a (locally) worst value of* $\boldsymbol{\theta}$*. Unsurprisingly and as the bottom right plot shows, this adversely affects predictive performance. This behaviour is entirely attributable to the fact that unlike* GVI *posteriors,* DVI *do not inherit the modularity result of Theorem 10. In this context, Figure 3 serves as a morality tale: In the* GVI *framework, changing the* KLD *to another divergence only changes uncertainty quantification and does* ***not*** *affect the way the best parameter is found. In sharp contrast, the* DVI *framework comes with no such guarantee. Accordingly, posteriors produced with* DVI *may conflate uncertainty quantification and the way the best parameter is found.*

5.2 Choosing ℓ and D: Robustness, better marginals, and beyond

Under an optimization-centric view on Bayesian methods, saying that $q^*_{\text{VI}}(\boldsymbol{\theta})$ produces undesirable inferences amounts to saying that the objective in eq. (6) is inappropriate for the inference task at hand. By virtue of their modularity, it is also clear that GVI posteriors define convenient alternative objectives that can address some of the drawbacks of traditional variational methods. Here, we focus on three situations that often cause problems for standard VI, but can be addressed by GVI: Prior misspecificaton **(↯P)**, model misspecification **(↯L)**, and overly narrow marginal variances.

To conclude this section, we provide some practical advice for using GVI. Since this has been our focus throughout, we focus on choices of ℓ and D that enhance robustness. Additionally, we also give some brief insights into other motivations for changing ℓ and D that go beyond robustness.

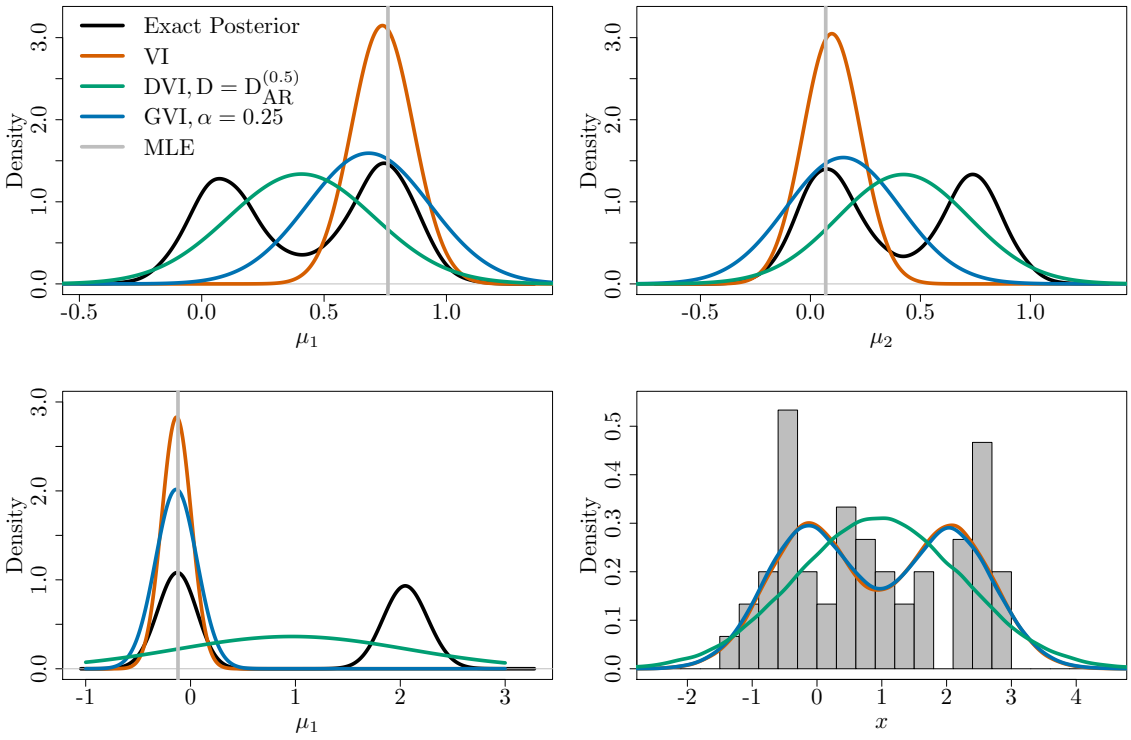


Figure 3: Best viewed in color. Depicted are inference outcomes for a BMM model, namely the (multimodal) **standard Bayesian** posterior, **standard VI** posterior, a **DVI**-approximation based on minimizing $D_{AR}^{(\alpha)}$ between $\mathcal{Q}$ and $q_{\mathrm{B}}^*(\boldsymbol{\theta})$ (Li and Turner, 2016), and a **GVI** posterior taking $D = D_{AR}^{(\alpha)}$. **Top:** Posterior marginals for $\mu_1 = 0, \mu_2 = 0.75$. The mode of the **DVI** posterior is a locally worst value for $\boldsymbol{\theta}$ relative to the **exact Bayesian** posterior. In contrast, **standard VI** and **GVI** respect the loss: They produce a posterior belief centered around one (of the two) values of $\boldsymbol{\theta}$ minimizing the loss. **Bottom left:** Posterior marginal for $\mu_1 = 0, \mu_2 = 2$. The effects of the top row become even stronger as the modes move further apart. **Bottom right:** Posterior predictive for $\mu_1 = 0, \mu_2 = 2$ against the histogram depicting the actual data. **VI**, **GVI** and **exact Bayesian** inference perform well and almost identically. **DVI** performs poorly, failing to capture the mixture components of the BMM.

5.2.1 Robustness to prior misspecification via D

As outlined in Section 3.3, inference outcomes are adversely affected if the prior does not at least approximately reflect the best available judgement about good values of θ before any data is seen. This is a problem whenever the prior is specified according to some (more or less arbitrary) default setting. For example, in the case of Bayesian Neural Networks (BNNs), a typical choice of prior is a multivariate standard normal distribution that factorizes over all network weights. While this may seem harmless or even uninformative, a supposedly uninformative prior specification of this kind actually encompasses a large degree of information, e.g.

(U) The prior belief is *unimodal*. In other words, we believe that there exists a *uniquely most likely* parameterization of the network before observing any data.

(I) The prior belief is that all network weights of a BNN are uncorrelated. In fact, we even believe that all network weights of a BNN are both *pairwise and mutually independent*.[9]

The above implications are in direct and strong contradiction to our best possible judgements about BNNs and thus violate **(P)**:

(↯U) Neural Networks are well-understood to have multiple parameter settings that are equally good (e.g. Choromanska et al., 2015). The unimodality assumption outlined in **(U)** is thus clearly not a reflection of the best judgement available: A prior belief in accordance with **(P)** would encode multimodality.

(↯I) By construction, Neural Networks encode a significant degree of dependence in their parameters: The best values for parameters in the l-th layer will strongly depend on the best values for parameters in the $(l-1)$-th layer (and vice versa). Hence, assuming uncorrelatedness (much less so independence!) directly contradicts our best judgement.

From this, it is obvious that a fully factorized normal distribution is hardly an appropriate default prior for BNNs in the sense of **(P)** in Section 3.1. At the same time, it is often prohibitive or computationally infeasible to construct alternative prior beliefs that reflect our best judgements more accurately. In other words, we are stuck with a sub-optimal prior. Under the standard Bayesian paradigm, this is not an acceptable position. In contrast, the optimization-centric paradigm outlined in Section 4.1 does not require the prior to be flawless. We can thus use our very imperfect prior to design more appropriate posterior beliefs: Simply adapt the argument D which regularizes the posterior belief against the prior. In particular, we want to adapt D such that the resulting posteriors satisfy two criteria: Firstly, they should be more robust to priors which strongly contradict the observed data. Secondly, they should still provide reliable uncertainty quantification.

There is a host of robust alternatives to the KLD that we may hope behave in this way, most of which fall within the families of α-, β-, and γ-divergences. Appendix B studies the way in which these divergences affect prior robustness and uncertainty quantification in great detail. Some of the most important findings are

- D should be unbounded over $\mathcal{Q}$ to prevent the posterior from collapsing to a point mass. This rules out the family of α-divergences as well as the Total Variation Distance, see Appendix B.1. Further and unsurprisingly, the larger the regularizers D, the larger the induced posterior variances, see Appendix B.2
- Using $D = \frac{1}{w}\mathrm{KLD}$ for $w \in (0, 1)$ makes marginal variances larger, but is highly non-robust to misspecified priors. This should not come as a surprise, since all we do is giving more weight to the same regularizer that we were trying to fix in the first place. While $w > 1$ decreases the adversarial effects of misspecified priors, it also rapidly shrinks the posterior's marginal variances. For details, see Appendix B.3.1.

9. For joint normal distributions, variables are uncorrelated if and only if they are independent.

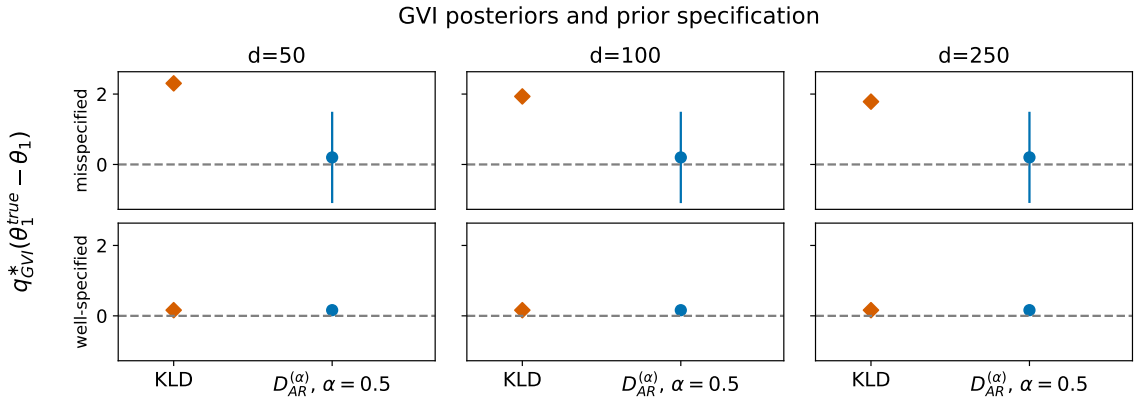


Figure 4: Best viewed in color. Taken from Section 6.2 in Knoblauch (2019a), the plot shows the impact of different prior beliefs on inference in a Bayesian normal mixture model with $n = 50$ observations and mixture components in $\mathbb{R}^d$ for different choices of d. Specifically, the plot compares inference outcomes under a misspecified prior (**Top**) against those under a well-specified prior (**Bottom**). It does so by depicting the average absolute difference between the true parameter values and their MAP estimate on the y-axis. Here, the solid whiskers' length corresponds to one standard deviation of the underlying posterior. For full details, see Appendix E. The plot shows that using the **KLD** as prior regularizer as in **standard Variational Inference (VI)** will produce undesirable uncertainty quantification under misspecified prior beliefs. In contrast, **Generalized Variational Inference (GVI)** with **Rényi's α-divergence** as prior regularizer produces desirable uncertainty quantification in both settings.

- The robust families of β- and γ-divergences induce fairly similar behaviour. While they are robust to misspecified priors for $\beta > 1$ (or $\gamma > 1$), this robustness comes at the price of a smaller marginal variance. For more details, see Appendices B.3.3 and B.3.4.

- Amongst all robust divergences that we examined, Rényi's α-divergence has the most desirable properties. Specifically, it guarantees prior robustness *without* tightening the marginal variances. Thus, it provides the prior robustness of β- and γ-divergences without the associated overconfident uncertainty quantification, see Appendix B.3.2)

In conclusion, we find that Rényi's α-divergence provides prior robustness in the most practically useful way. For values of $\alpha \in (0, 1)$, it generally also provides larger marginal variances than the KLD. Conversely, values of $\alpha > 1$ provide tighter marginal variances than the KLD. As Figures 4 and 10 show, the divergence produces similar posteriors as $D = \text{KLD}$ if the prior is *correctly* specified. The same Figures also show that unlike KLD, choosing Rényi's α-divergence continues to produce desirable uncertainty quantification when the prior is misspecified.

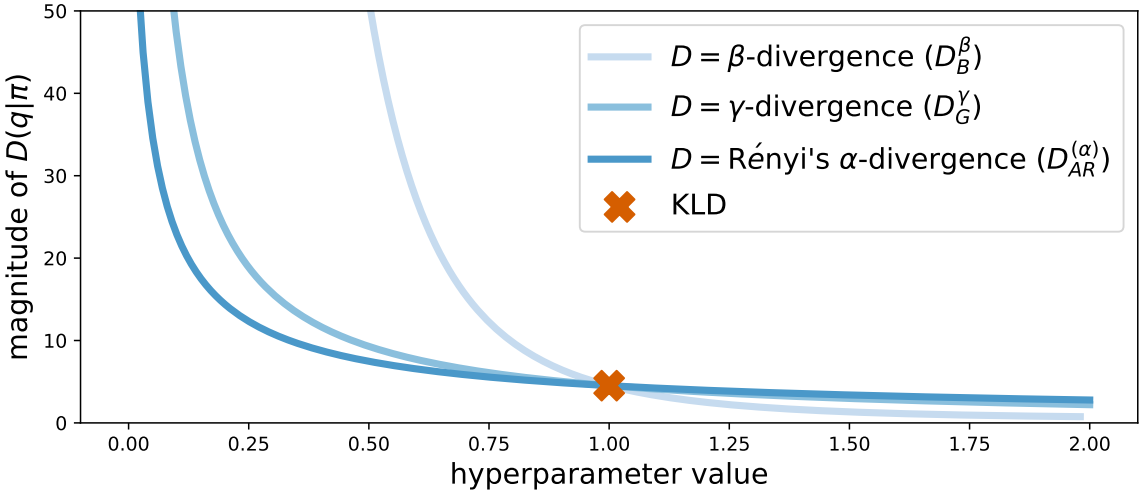


Figure 5: Depicted is the magnitude $D(q\|\pi)$ for different **robust divergences D** and the **KLD** for two Normal Inverse Gamma distributions given by $q(\boldsymbol{\theta}) = \mathcal{NI}^{-1}(\boldsymbol{\theta}; \boldsymbol{\mu}_q, \boldsymbol{V}_q, a_q, b_q)$ and $\pi(\boldsymbol{\theta}) = \mathcal{NI}^{-1}(\boldsymbol{\theta}; \boldsymbol{\mu}_\pi, \boldsymbol{V}_\pi, a_\pi, b_\pi)$ with $\boldsymbol{\mu}_\pi = (0,0)^T$, $\boldsymbol{V}_\pi = 25 \cdot I_2$, $a_\pi = 500$, $b_\pi = 500$ and $\boldsymbol{\mu}_q = (2.5, 2.5)^T$, $\boldsymbol{V}_q = 0.3 \cdot I_2$, $a_q = 512$, $b_q = 543$.

Going into more detail, Rényi's α-divergence—henceforth denoted $D_{AR}^{(\alpha)}$ and introduced by Rényi (1961)—in the parameterization of Cichocki and Amari (2010) is given by

$$D_{AR}^{(\alpha)}(q\|\pi) = \frac{1}{\alpha(\alpha-1)} \log \left(\mathbb{E}_{q(\boldsymbol{\theta})} \left[\left(\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right)^{1-\alpha} \right] \right). \tag{15}$$

Originally, Rényi's α-divergence was motivated as the *geometric mean* information to discriminate between the two hypotheses $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim q(\boldsymbol{\theta})$ of order α, for some $\alpha \in (0,1)$. Similarly, the original motivations for the KLD was its interpretation as the *arithmetic mean* information to discriminate between $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim q(\boldsymbol{\theta})$ (Kullback and Leibler, 1951). Intuitively speaking, geometric means are more robust measures of central tendency than arithmetic means, and so it makes sense that the $D_{AR}^{(\alpha)}$ is generally a more robust discrepancy measure. Conversely, picking smaller values of $\alpha \in (0,1)$ will produce more prior-robust measures of discrepancy than larger values of $\alpha \in (0,1)$. Indeed, $D_{AR}^{(\alpha)}(q\|\pi)$ even recovers the non-robust discrepancy $\mathrm{KLD}(q\|\pi)$ as $\alpha \to 1$. A host of other robust divergences also recover the KLD as their respective hyperparameters appraoch one. This includes α-, β- and γ-divergences as well as their generalizations (see Cichocki and Amari, 2010). To illustrate this phenomenon, we vary their hyperprameters and plot their magnitude in Figure 5. The plot illustrates that in our parameterization, hyperparameter values below (above) unity impose larger (smaller) penalties than the KLD.

While $D_{AR}^{(\alpha)}$ behaves robustly, it has one clear practical drawback relative to other potential regularizers such as the KLD or f-divergences. Specifically, eq. (15) defines it as a log expectation—meaning that standard stochastic inference techniques do not provide unbiased estimates for D. In the experiments of the current paper, we circumvent this issue by

only considering variational families $\mathcal{Q}$ that permit a closed form of $D_{AR}^{(\alpha)}$. Note that this requirement is not particularly restrictive, as Rényi's α-divergence has closed forms for essentially all exponential family members (see Theorem 36).

5.2.2 robustness to model misspecification via ℓ

Section 3.4 explains how and why **(↯L)** can severely impede the usefulness of standard Bayesian posteriors: if $p(\cdot|\boldsymbol{\theta})$ is not an accurate description of the data generating mechanism, inferences are susceptible to outliers, heterogeneity, and other adversarial aspects of the data. Recalling that $q_{\mathrm{B}}^*(\boldsymbol{\theta}) = P(-\log p(\cdot|\boldsymbol{\theta}), \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$, it is also clear that treating the likelihood model as (approximately) correct amounts to using the log score $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$ to assess how well $p(\cdot|\boldsymbol{\theta})$ fits $\{x_i\}_{i=1}^n$. Indeed, this loss processes information about the likelihood model $p(\cdot|\boldsymbol{\theta})$ contained in $x_{1:n}$ optimally within a Bayesian framework *if* this model happens to be correctly specified (Zellner, 1988).

While this implies that robust likelihood-based losses are typically less statistically efficient *under correct specification*, this tradeoff radically reverses even under mild misspecification (see e.g. Basu et al., 1998; Fujisawa and Eguchi, 2008; Hung et al., 2018; Jewson et al., 2018). For notational clarity, we write $L_n(\boldsymbol{\theta}, x_{1:n}) = L_n(p(\cdot|\boldsymbol{\theta}), x_{1:n})$ as a robust loss assessing the fit of likelihood parameter $\boldsymbol{\theta}$ on the sample $x_{1:n}$. The most appealing choices for $L_n : \mathcal{P}(\boldsymbol{\Theta}) \times \mathcal{X}^n \to \mathbb{R}$ are finite-sample estimators of $D(p_x(\cdot)\|p(\cdot|\boldsymbol{\theta}))$ for some robust divergence D. In other words, a natural loss is the estimated divergence between the true data-generating mechanism p_x and the model $p(\cdot|\boldsymbol{\theta})$. A notable advantage of designing losses in this way is the following: even in the unlikely event that $p(\cdot|\boldsymbol{\theta})$ is correctly specified for p_x—so that there is $\boldsymbol{\theta}^*$ for which $p_x = p(\cdot|\boldsymbol{\theta}^*)$—minimizing an unbiased estimate of $D(p_x(\cdot)\|p(\cdot|\boldsymbol{\theta}))$ targets the correct value $\boldsymbol{\theta}^*$ for any statistical divergence D. So even though robust losses are less efficient than the log score under correct misspecification, they nonetheless recover the parameter value if the model is correctly specified. An overview of some robust losses constructed in this way is provided in Table 2. Note that unlike in eq. (9), we have not assumed additivity of L_n since a large class of such robust losses obtained this way are non-additive.

All losses presented in Table 2 guarantee various forms of robustness, and their main limiting factors are often of practical nature. To begin with, all except the Total Variation Distance depend on hyperparameters that are generally difficult to choose. All of the non-additive losses in the table also come with higher computational complexity, since non-additive losses do not admit unbiased estimation by sub-sampling. On top of this, such losses generally come with increased computational overhead. For example, kernel-based discrepancy measures such as the Maximum Mean Discrepancy or Kernel Stein Discrepancy are estimated using double sums. This means that evaluating these losses on a sample of size n has a computational complexity of $\mathcal{O}(n^2)$. Estimating losses based on the α-divergence or the Total Variation Distance is even more computationally demanding, since they require kernel density estimators if $\mathcal{X} = \mathbb{R}^p$.

In summary, computational feasibility makes additive losses much more compelling than their non-additive alternatives. At the time of writing, the only two divergence-based losses that are both robust and additive are those corresponding to the family of β- and γ-divergences (denoted $\mathcal{L}_p^\beta$ and $\mathcal{L}_p^\gamma$ respectively), first introduced by Basu et al. (1998) and

Divergence	Hyperparameters	Additive	References
α-divergence	$\alpha \in (0,1)$	✗	Beran et al. (1977); Tamura and Boos (1986); Simpson (1987); Lindsay et al. (1994); Hooker and Vidyashankar (2014)
β-divergence	$\beta > 1$	✓	Basu et al. (1998); Ghosh and Basu (2016); Futami et al. (2018)
γ-divergence	$\gamma > 1$	✓	Fujisawa and Eguchi (2008); Hung et al. (2018); Nakagawa and Hashimoto (2019)
Maximum Mean Discrepancy	Kernel k_ν and ν	✗	Briol et al. (2019); Chérief-Abdellatif and Alquier (2019b,a)
Kernel Stein Discrepancy	Stein Operator, kernel k_ν and ν	✗	Barp et al. (2019)
Total Variation Distance	—	✗	Yatracos (1985); Devroye and Lugosi (2012); Knoblauch and Vomfell (2020)

Table 2: Overview over robust likelihood-based losses derived from divergences

Hung et al. (2018).

$$\mathcal{L}_p^\beta(\boldsymbol{\theta}, \boldsymbol{y}_i) = -\frac{1}{\beta - 1} p(\boldsymbol{y}_i|\boldsymbol{\theta})^{\beta-1} + \frac{I_{p,\beta}(\boldsymbol{\theta})}{\beta} \tag{16}$$
$$\mathcal{L}_p^\gamma(\boldsymbol{\theta}, \boldsymbol{y}_i) = -\frac{1}{\gamma - 1} p(\boldsymbol{y}_i|\boldsymbol{\theta})^{\gamma-1} \cdot \frac{\gamma}{I_{p,\gamma}(\boldsymbol{\theta})^{\frac{\gamma-1}{\gamma}}}.$$

Here, the integral term $I_{p,c}(\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{\theta}^L)^c d\boldsymbol{y}$ is generally available in closed form for exponential families. These losses are more robust than the log score whenever $\beta > 1$ (or $\gamma > 1$). Relative to other robust losses, they have two additional benefits: firstly and we will explain in Section 5.4.1, they have desirable computational properties. Secondly, the hyperparameter β (or γ) has a clear interpretation since the losses recover the negative log likelihood as $\beta \to 1$ (or $\gamma \to 1$). To see this, one simply notes that $\lim_{x\to 1} \frac{z^{x-1}-1}{x-1} = \log z$ and $I_{p,1}(\boldsymbol{\theta}) = 1$. Thus—unlike the other entries in Table 2 except the α-divergence—the losses $\mathcal{L}_p^\beta$ and $\mathcal{L}_p^\gamma$ can be made arbitrarily close to the standard negative log likelihood. More specifically, choices of $\beta = 1 + \varepsilon$ (or $\gamma = 1 + \varepsilon$) for small $\varepsilon > 0$ will provide a loss function that is both robust *and* nearly as statistically efficient as the negative log likelihood.

Unfortunately, it is generally difficult to pick the optimal degree of robustness ε because its optimal level will depend on the scale of the data $x_{1:n}$. However, in numerous experiments both in the remainder as well as in prior work (e.g. Jewson et al., 2018; Knoblauch et al., 2018; Knoblauch, 2019a) we found that *if the data are standardized*, values for $\varepsilon \in [0.01, 0.1]$ will yield a very favourable trade-off between robustness and efficiency across a very wide range of data sets and models.

Influence functions provide a concise and intuitively appealing way of illustrating this trade-off between robustness and efficiency. In the Bayesian context, influence functions quantify the impact the $(n+1)$-th observation x_{n+1} has on the posterior distribution $q_{\mathrm{B}}^*(\boldsymbol{\theta})$ constructed using the first n observations (Peng and Dey, 1995). This discrepancy is

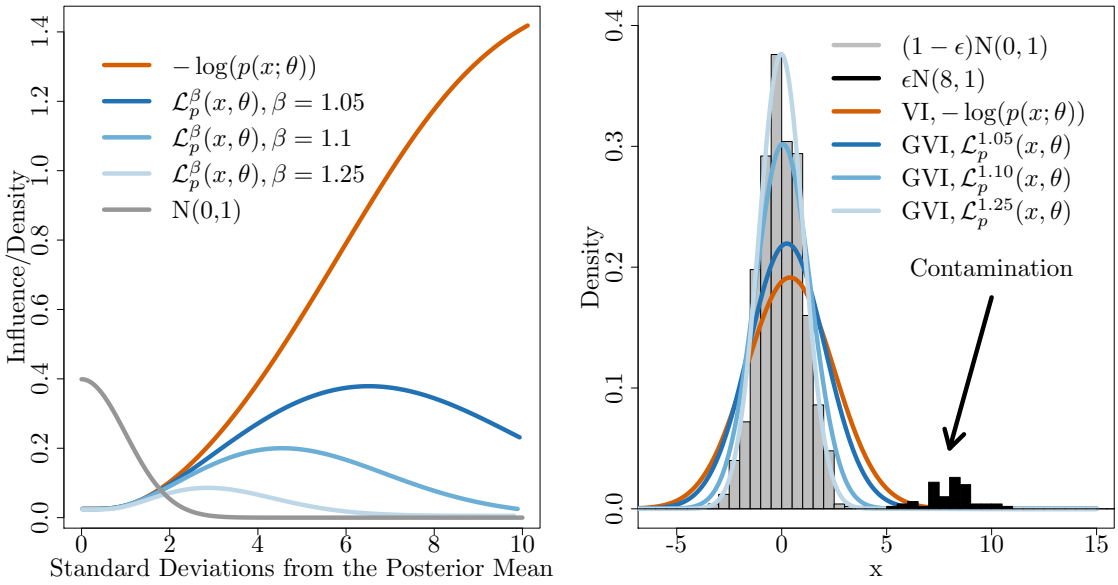


Figure 6: Best viewed in color. The plots compare influence functions (**Left**) and predictive posteriors (**Right**) of a **standard** Bayesian inference against a **GVI** posterior. **Left:** The influence functions of scoring the normal likelihood with a **standard** negative log likelihood against a **robust** scoring rule derived from β-divergences. **Right:** A univariate normal is fitted using all the data depicted, including the outlying contamination. The posterior predictive corresponding to the **robust** scoring rule and $\beta = 1.25$ is able to ignore these outliers. This stands in contrast to the posterior predictive based on **standard Bayesian inference**, which assigns increasingly large influence to outlying observations.

measured by computing a divergence between the posteriors based on $x_{1:n}$ and on $x_{1:(n+1)}$. Using the Fisher-Rao divergence (see Kurtek and Bharath, 2015), Figure 6 compares the influence of a standard Bayesian posterior with that of a posterior belief computed using Generalized Variational Inference (GVI). The left side of the Figure quantifies the lack of robustness for standard Bayesian methods: In this, the influence of x_{n+1} on the posterior belief grows stronger and stronger the more untypical it is relative to previously observed data. Similarly, the right side shows the adverse effect this has on the posterior predictive. To make the implications of influence functions for inferential practice more tangible, we refer to the outlier problem in Example 4.

Example 4 (Outliers as violations of (L)) *Arguably the most useful conceptualization of outliers is that inherent in the ε-contamination model. In this model, most of the data come from the uncontaminated model $p(x_i|\boldsymbol{\theta}^*)$, but a small proportion $\varepsilon \in (0,1)$ comes from an outlier-generating density o:*

$$p_{true}(x_i) = (1-\varepsilon) \cdot p(x_i|\boldsymbol{\theta}^*) + \varepsilon \cdot o(x_i).$$

For data generated like this, an obvious violation of **(L)** *would be to fit the data only to the non-contaminated component* $p(x_i|\boldsymbol{\theta})$ *in order to infer* $\boldsymbol{\theta}^*$*. This type of model misspecification is especially poignant in Bayesian On-line Changepoint Detection (*BOCPD*) (see e.g. Adams and MacKay, 2007; Fearnhead and Liu, 2007; Wilson et al., 2010; Saatçi et al., 2010; Caron et al., 2012; Turner et al., 2013; Knoblauch and Damoulas, 2018; Knoblauch et al., 2018). Through an efficient recursive relationship that updates the Bayesian posterior,* BOCPD *segments a data stream in real time. A canonical application example of* BOCPD *is the well-log data set (O'Ruanaidh, 1994) whose observations are the nuclear responses of rock strata while drilling a well. Generally, different rock strata are clearly distinguishable. However, as rock formation processes are noisy and sometimes interrupted by extraordinary events (e.g., tsunamis, earth quakes, eruptions), the data are surprisingly close to an* ε*-contaminated normal distribution within each of the strata. Figure 7 is taken from Knoblauch et al. (2018) and shows how this phenomenon renders vanilla* BOCPD *an unreliable algorithm. It also shows that this issue can be remedied by constructing alternative posterior belief distributions via a special case of Generalized Variational Inference (*GVI*) based on robust loss functions derived from the* β*-divergence.*

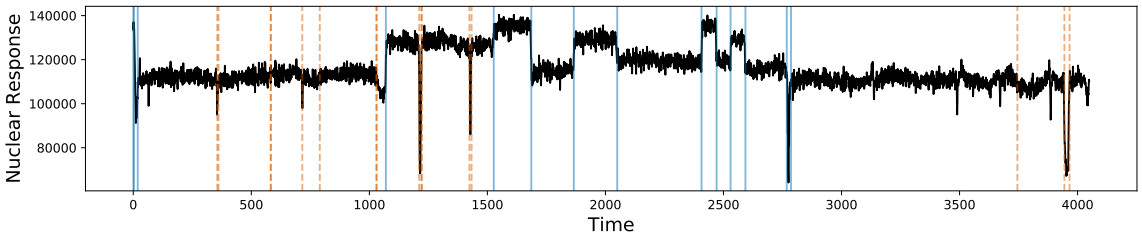


Figure 7: Best viewed in color. Inference outcomes of BOCPD on the well log data set using the **standard Bayesian posterior** and a **GVI posterior** constructed with robust losses based on the β-divergence. Solid vertical lines correspond to Maximum A Posteriori (MAP) segmentation of **GVI posterior**, dashed vertical lines mark incorrect changepoints *additionally* detected under **standard Bayesian inference**.

5.2.3 Beyond Robustness

While we have focused on using robustness throughout the paper, it should be clear that the generalization introduced in Definition 9 is also useful outside this narrow restriction. In fact, the standard VI objective can be inappropriate even in situations where assuming appropriately specified priors **(P)** and likelihood functions **(L)** underlying the traditional Bayesian paradigm are a useful working assumption.

Adjusting marginal variances: the uncertainty quantification of standard VI is often inappropriate when $\mathcal{Q}$ is a mean field variational family factorizing dimension-wide over $\boldsymbol{\theta}$ and the individual entries of $\boldsymbol{\theta}$ exhibit strong dependence (see for instance Example 3). Oftentimes, this phenomenon is referred to as *mode seeking behaviour* (e.g. Minka, 2005). The name itself also reveals that this problem is intimately linked to variational families $\mathcal{Q}$ containing only unimodal distributions. Again, the modularity result of Theorem 10

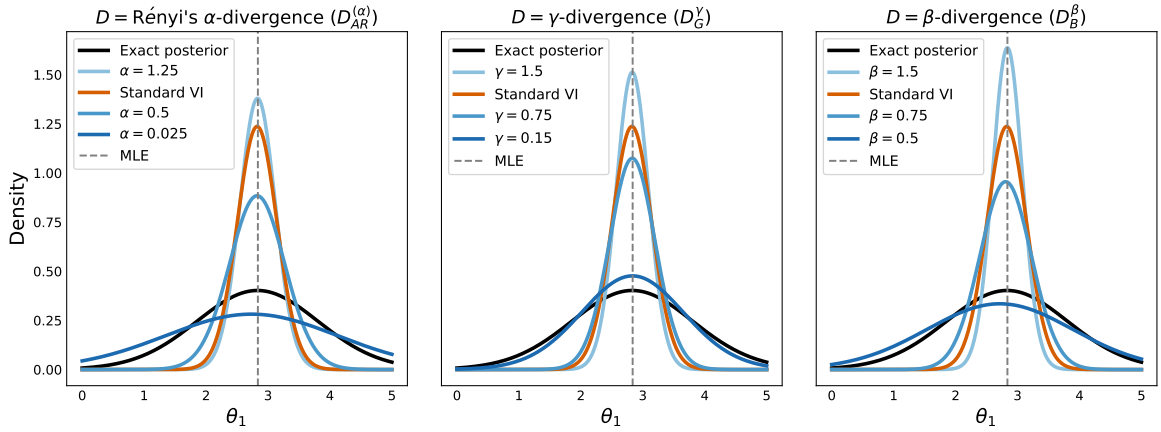


Figure 8: Best viewed in color. Marginal **VI** compared to different **GVI** posteriors for the coefficient θ_1 of data simulated from a Bayesian linear model (see Appendix B for details). For all posteriors, the loss ℓ is the correctly specified negative log likelihood of the true data generating mechanism. Further, for all variational posteriors the belief is constrained to lie inside a mean field normal family $\mathcal{Q}$. Due to high correlation between the coefficients for the **exact posterior**, **standard VI** produces undesirably over-concentrated belief distributions. In contrast, appropriately choosing the hyperparameters of alternative robust divergences $D \neq \text{KLD}$ provides more desirable uncertainty quantification.

can be helpful: Provided that one is limited to choosing the set of possible approximations $\mathcal{Q}$ to contain only unimodal distributions, one can adapt the GVI posterior's uncertainty quantification properties by changing $D = \text{KLD}$ to an alternative divergence. For instance, in Example 3, Rényi's α-divergence ($D_{AR}^{(\alpha)}$) provided a wider marginal variance. As Figure 8 illustrates, most other robust divergences behave similarly.

Inferential machinery for non-standard PAC-Bayes bounds: certain choices of $D \neq \text{KLD}$ would lead to an interpretation of the GVI objective as a PAC-Bayesian generalization bound (see for instance Bégin et al., 2016; Wang et al., 2018; Ohnishi and Honorio, 2020) or a regret bound (see Alquier, 2020). Importantly, these generalization bounds hold under conditions where traditional PAC-Bayesian bounds based on the KLD would fail, such as unbounded losses, heavy tails, or sequential dependence.

Intractability: one can use losses derived from the Fisher divergence or the Kernel Stein Discrepancies (Hyvärinen, 2005; Barp et al., 2019) to perform inference in likelihood models $p(x|\boldsymbol{\theta}) = \hat{p}(x|\boldsymbol{\theta})/Z(\boldsymbol{\theta})$ where the normalization constant $Z(\boldsymbol{\theta})$ is intractable. Specifically, these losses only depend on the derivative of the likelihood model, so that we can ignore the normalization constant since $\nabla_x \log p(x|\boldsymbol{\theta}) = \nabla_x \log \hat{p}(x|\boldsymbol{\theta}) - \nabla_x \log Z(\boldsymbol{\theta}) = \nabla_x \log \hat{p}(x|\boldsymbol{\theta})$. Similarly, we can perform inference for models with implicitly specified, intractable likelihoods by using the Maximum Mean Discrepancy as loss function (see Chérief-Abdellatif and Alquier, 2019a). Unlike the log likelihood, the Maximum Mean Discrepancy can be used as a loss function even if the likelihood function $p(\cdot|\boldsymbol{\theta})$ has no analytic form and can only be sampled from for any given $\boldsymbol{\theta} \in \boldsymbol{\Theta}$—as in simulators. It is also possible to instead adapt D in a

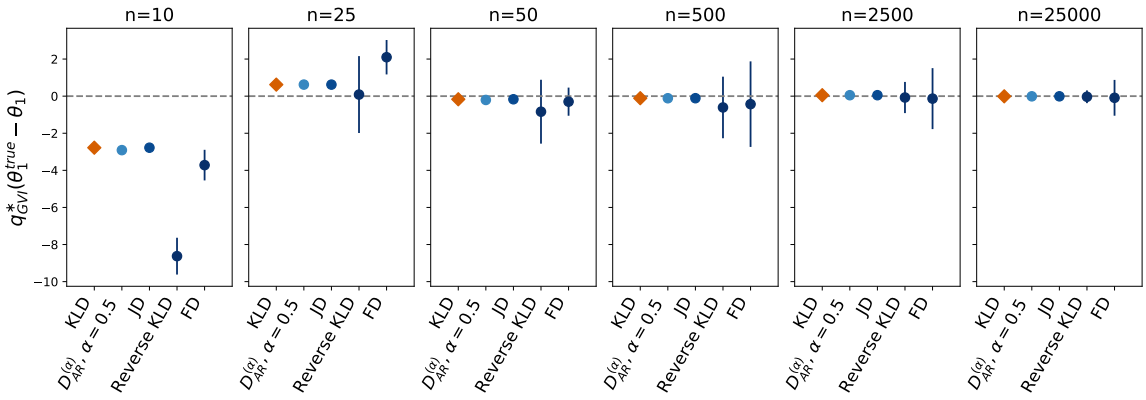


Figure 9: Best viewed in color. Marginal **VI** and different **GVI** posteriors for the first coefficient of a simulated 20-dimensional Bayesian Linear Model based on n observations. The loss ℓ is the correctly specified negative log likelihood of the true data generating mechanism and D is varied along the x-axis. Depicted are the forward and reverse KLD, Rényi's α-divergence ($D^{(\alpha)}_{AR}$), Jeffrey's Divergence (JD) as well as the Fisher Divergence (FD). All posteriors are members of the mean field normal family $\mathcal{Q}$. Because all inferred posterior beliefs are normals, dots are used to mark out the posterior mean and whiskers to denote the posterior standard deviation. All posteriors are re-centered around the true value of the coefficient, so that the y-axis shows how far the posterior belief is from the truth.

similar vain. This would simplify inference with so-called *implicit priors* that are represented by a sample only (see e.g. Tiao et al., 2018).

Simplified optimization: From a more practical point of view, one could choose D (or ℓ) to simplify the corresponding optimization problem. For example, one could make $D = D_{\mathcal{Q}}$ directly defined only for a particular variational family $\mathcal{Q}$ to ensure that $D_{\mathcal{Q}}$ is convex (or even strongly convex) in the variational parameters. In fact, it is surprisingly easy to do this. For example, letting $\pi \in \mathcal{Q}$ with variational parameter $\boldsymbol{\kappa}_\pi$, and $q \in \mathcal{Q}$ with variational parameter $\boldsymbol{\kappa}_q$, the divergence $D_{\mathcal{Q}}(q\|\pi) = \frac{1}{2}\|\boldsymbol{\kappa}_\pi - \boldsymbol{\kappa}_q\|_2^2$ is 1-strongly convex in $\boldsymbol{\kappa}_q$. Though this particular ad-hoc divergence likely would produce strange uncertainty quantification, the crucial point here is that it would result in an easy-to-optimize problem with a unique solution whenever $\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\right]$ is convex in $\boldsymbol{\kappa}_q$.

5.3 Theoretical properties of GVI

The principal appeal of GVI lies in its modularity and the associated subjective choices of ℓ, D and $\mathcal{Q}$. Beyond that, the following section briefly visits two theoretical findings: firstly, we point to novel results showing that GVI posteriors collapse to the population-optimal value of $\boldsymbol{\theta}$ as $n \to \infty$, regardless of D. Secondly, we show that GVI posteriors with certain choices for D have a second interpretation as approximations to Bayesian posteriors with a power likelihood.

5.3.1 Frequentist consistency

Knoblauch (2019a) shows that GVI posteriors are consistent in the Frequentist sense. This holds under a wide range of extremely mild regularity conditions on the arguments ℓ, D and $\mathcal{Q}$ so long as $\widehat{\boldsymbol{\theta}}_n$ is unique for all n large enough. Here, we state a simple version of the result for independent data with the mean field normal variational family.

Theorem 12 (Frequentist consistency of GVI) *Suppose that Assumption 1 in Knoblauch (2019a) holds. Choose $\mathcal{Q}$ to be the mean field normal family, let D be lower-semi-continuous in its first argument and suppose that $D(q\|\pi) < \infty$ for all $q \in \mathcal{Q}$. Further, let $\mathbb{P}_{\boldsymbol{x}}$ be the true probability measure of some random variable $\boldsymbol{x}$ and suppose that the observations $x_{1:n}$ are independent and identically distributed draws from $\boldsymbol{x}$. If the prior is not infinitely bad for the population of $\boldsymbol{x}$ (which is to say that $\mathbb{E}_\pi\left[\mathbb{E}_{\mathbb{P}_{\boldsymbol{x}}}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right] < \infty$), then*

$$q^*_{\text{GVI},n}(\boldsymbol{\theta}) \xrightarrow{D} \delta_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}),$$

where $q^*_{\text{GVI},n}(\boldsymbol{\theta})$ is the GVI posterior corresponding to the problem $P(\ell, D, \mathcal{Q})$ based on n observations and $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}}\left[\mathbb{E}_{\mathbb{P}_{\boldsymbol{x}}}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right]$ is the population-optimal parameter value.

Remark 13 *This Theorem is an invocation of Corollary 1 in Knoblauch (2019a). Assumption 1 guarantees a number of conditions that are required to make GVI a well-defined optimization problem. For example, it ensures that the sum of the losses has minimizers for any finite n and in the large data limit and that the loss expected under $\mathbb{P}_{\boldsymbol{x}}$ is finite.*

This finding is illustrated in Figure 9, which is taken from Knoblauch (2019a). As the theory suggests, the posteriors collapse to a point mass under mild regularity conditions on D. Unsurprisingly, speed and nature of the convergence depend on the choice of D.

5.3.2 GVI as a posterior approximation

Although the axiomatic development in Section 4.1 shows that GVI produces a posterior belief distribution that is valid in its own right, one can also interpret certain GVI posteriors as approximations to (generalized) Bayesian posteriors as in eq. (2). In particular, we show that for a range of robust divergences $D^{(\rho)}_{\text{robust}}$ parameterized by some hyperparameter ρ so that $\lim_{\rho\to 1} D^{(\rho)}_{\text{robust}} = \text{KLD}$, the GVI objective constitutes a lower bound on the evidence of generalized Bayesian posterior. Results of this form can be shown to hold for Rényi's α-divergence ($D^{(\alpha)}_{AR}$), the γ-divergence ($D^{(\gamma)}_{G}$) as well as the β-divergence ($D^{(\beta)}_{B}$). As they are structurally similar, we only state the bound corresponding to $D = D^{(\alpha)}_{AR}$ and defer the results for $D^{(\gamma)}_{G}$ and $D^{(\beta)}_{B}$ as well as all proofs to Appendix F.

Theorem 14 (GVI as approximate Evidence Lower bound with $D = D^{(\alpha)}_{AR}$) *The objective of a GVI posterior based on $P(\ell, D^{(\alpha)}_{AR}, \mathcal{Q})$ has an interpretation as lower bound on the $c(\alpha)$-scaled (generalized) evidence lower bound of $P(w(\alpha)\cdot\ell, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$:*

$$\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^n \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right] + D^{(\alpha)}_{AR}(q||\pi) \geq -c(\alpha)\cdot \text{ELBO}^{w(\alpha)\ell}(q) + S_1(\alpha, q, \pi) \tag{17}$$

where $\text{ELBO}^{w(\alpha)\ell}$ *denotes the Evidence Lower Bound associated with standard* VI *relative to the generalized Bayesian posterior given by*

$$q_B^{w(\alpha)\ell}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp\left(-w(\alpha) \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right),$$

where $S_1(\alpha, q, \pi) = \mathbb{1}(0 < \alpha < 1)\left\{D_{AR}^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}) - \text{KLD}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))\right\}$ *is an interpretable slack term,* $c(\alpha) = \min\{1, \alpha^{-1}\}$ *and* $w(\alpha) = \max\{1, \alpha\}$.

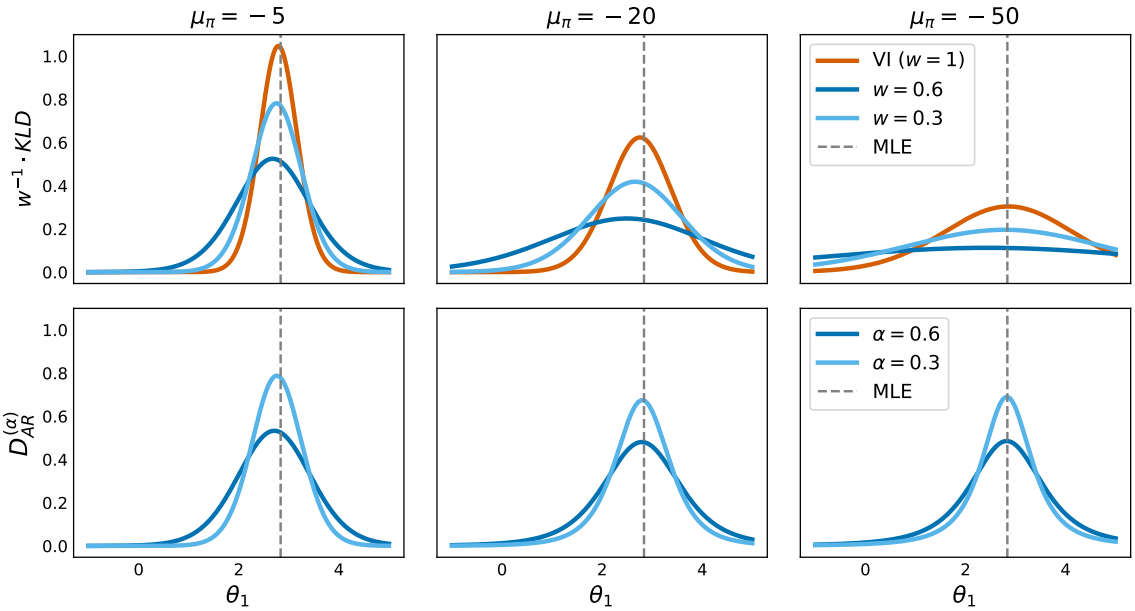


Figure 10: Best viewed in color. Marginal **VI** compared to different **GVI** posteriors for the coefficient θ_1 of data simulated from a d-dimensional Bayesian linear model with different priors (see Appendix B for details). The prior for the coefficients is a Normal Inverse Gamma distribution given by $\boldsymbol{\mu} \sim \mathcal{NI}^{-1}(\mu_\pi \cdot 1_d, v_\pi \cdot I_d, a_\pi, b_\pi)$ with $v_\pi = 4 \cdot I_d$, $a_\pi = 3$, $b_\pi = 5$ and various values for μ_π. For all posteriors, the loss ℓ is the correctly specified negative log likelihood of the true data generating mechanism. Further, all variational posteriors are constrained to lie inside a mean field normal family $\mathcal{Q}$. Notice that the **standard VI** posterior corresponds to the ELBO component on the right hand side of the bound in eq. (17). In contrast, the **GVI** posteriors are obtained by maximizing the left hand side of the same bound.

Remark 15 *Eq.* (17) *shows that the slack term* $S_1(\alpha, q, \pi)$ *introduces the main difference between* $P(\ell, D_{AR}^{(\alpha)}, \mathcal{Q})$ *and* $P(w(\alpha) \cdot \ell, \text{KLD}, \mathcal{Q})$. *It is possible but tedious to make analytically more concise statements about* $S_1(\alpha, q, \pi)$ *(see Appendix F). Doing so reveals that this slack term makes* $P(\ell, D_{AR}^{(\alpha)}, \mathcal{Q})$ *more robust to misspecification of the prior than that of* $P(w(\alpha) \cdot \ell, \text{KLD}, \mathcal{Q})$, *and that this behaviour becomes more pronounced for smaller* α. *This phenomenon is summarized in Figure 10: since* $w(\alpha) = 1$ *for* $\alpha \in (0, 1)$, *if we ignore*

$S_1(\alpha, q, \pi)$ then the bound on the right of eq. (17) is just the ELBO *of the* ***Standard*** **VI** *posterior $P(\ell, \text{KLD}, \mathcal{Q})$) for all $P(\ell, D_{AR}^{(\alpha)}, \mathcal{Q})$ with $\alpha \in (0, 1)$. As the Figure reveals, these two posteriors are quite different—making the slack term rather important in relating $P(\ell, D_{AR}^{(\alpha)}, \mathcal{Q})$ to $P(\ell, \text{KLD}, \mathcal{Q}))$. Since $P(\ell, D_{AR}^{(\alpha)}, \mathcal{Q})$ inflates variance relative to $P(\ell, \text{KLD}, \mathcal{Q}))$, one may expect that up-weighting the* KLD *term with $\frac{1}{\alpha}$ may produce similar posteriors. Thus, Figure 10 additionally compares $P(\ell, D_{AR}^{(\alpha)}, \mathcal{Q})$ with $P(\ell, \frac{1}{\alpha}\text{KLD}, \mathcal{Q}))$. Doing so reveals that while $P(\ell, D_{AR}^{(\alpha)}, \mathcal{Q}) \approx P(\ell, \frac{1}{\alpha}\text{KLD}, \mathcal{Q}))$ for reasonable prior specification, the distributions diverge substantially as the prior becomes more and more misspecified. This clarifies the role of the Slack term $S_1(\alpha, q, \pi)$: while it ensures that $P(\ell, D_{AR}^{(\alpha)}, \mathcal{Q}) \approx P(\ell, \frac{1}{\alpha}\text{KLD}, \mathcal{Q}))$ whenever π is well-specified, it robustifies $P(\ell, D_{AR}^{(\alpha)}, \mathcal{Q})$ (relative to $P(\ell, \frac{1}{\alpha}\text{KLD}, \mathcal{Q}))$) for poor choices of π.*

5.4 Inference with Generalized Variational Inference (GVI)

This section outlines two inference strategies for GVI: quasi-conjugate and fully black box inference. Built on earlier findings in Knoblauch et al. (2018), we show that a class of GVI posteriors based on robust likelihood scoring rules admits closed form variational objectives. This closed form objective emerges when the likelihood is conjugate to the prior, we call the resulting inference procedure quasi-conjugate. For more complicated models, closed form objectives generally are not available. To address this, we introduce a black box inference procedure for arbitrary choices of ℓ and D. Note that while this inference scheme in principle works on any choice of π, $\mathcal{Q}$ and D, the variance of estimated gradients will be much-reduced if $D(q\|\pi)$ is available in closed form for all $q \in \mathcal{Q}$. Conveniently, this generally holds if $\mathcal{Q}$ is a set of exponential family models and $\pi \in \mathcal{Q}$ for all robust divergences studied in the current paper.

5.4.1 Quasi-conjugate inference

An interesting interdependence between loss function and variational family was studied in Knoblauch et al. (2018): When applying the robust scoring rule $\mathcal{L}^\beta$ (see eq. (16)) derived from the β-divergence ($D_B^{(\beta)}$) to a likelihood $p(\cdot|\boldsymbol{\theta})$ associated with a conjgate prior $\pi(\theta|\kappa_0)$, there is advantage in taking $\mathcal{Q}$ to be the family of the conjugate prior: specifically, $\mathcal{L}^\beta(\boldsymbol{\theta}, x_i) \to -\log p(x_i|\boldsymbol{\theta})$ as $\beta \to 1$, so that the (generalized) Bayesian posterior

$$q_{\text{B}}^\beta(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^n \exp\left\{-\mathcal{L}^\beta(\boldsymbol{\theta}, x_i)\right\}$$

becomes conjugate and is contained in $\mathcal{Q}$ as $\beta \to 1$. Thus, so long as $|\beta - 1| < \varepsilon$ for some sufficiently small value $\varepsilon > 0$ and $D = \text{KLD}$, constraining the posterior to be in $\mathcal{Q}$ produces excellent approximations to $q_{\text{B}}^\beta(\boldsymbol{\theta})$. Beyond approximation quality, choosing the quasi-conjugate variational family also offers another advantage: As Theorem 2 in Knoblauch et al. (2018) shows, they make the variational objective of $P(\mathcal{L}^\beta, \text{KLD}, \mathcal{Q})$ available in closed form. Consequently, no stochastic approximation to the objective is required, so that the optimum is usually found within a very small number of iterations.

Proposition 16 extends this quasi-conjugacy to the robust scoring rule $\mathcal{L}^\gamma$ (see eq. (17)) derived from the γ-divergence ($D_G^{(\gamma)}$) of Hung et al. (2018). Similarly to $\mathcal{L}^\beta$, $\mathcal{L}^\gamma(\boldsymbol{\theta}, x_i) \to -\log p(x_i|\boldsymbol{\theta})$ as $\gamma \to 1$, so that the same intuition that applied to $\mathcal{L}^\beta$ also applies here. Note that the conditions for $\mathcal{L}^\gamma$ in Proposition 16 are slightly more restrictive than those

derived for $\mathcal{L}^\beta$ due to the appearance of a multiplicative integral term. While the proof is conceptually straightforward, it is notationally cumbersome and deferred to Appendix G.

Proposition 16 (Closed form GVI objectives with $\mathcal{L}^\gamma$) *Let $\mathcal{L}^\gamma(\boldsymbol{\theta}, \cdot)$ be the γ-divergence based scoring rule for likelihood $p(\cdot|\boldsymbol{\theta})$. Suppose $p(\cdot|\boldsymbol{\theta})$ admits conjugacy relative to the exponential distributions given by $\mathcal{Q}$ and let the conjugate prior $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0) \in \mathcal{Q}$. Writing*

$$\begin{aligned} p(x|\boldsymbol{\theta}) &= h(x) \exp\left\{ g(x)^T T(\boldsymbol{\theta}) - B(x) \right\}, \\ q(\boldsymbol{\theta}|\boldsymbol{\kappa}) &= h(\boldsymbol{\theta}) \exp\left\{ \eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa})) \right\}, \\ \mathcal{N} &= \{ \boldsymbol{\kappa} : \exp\{A(\eta(\boldsymbol{\kappa}))\} < \infty \}, \end{aligned}$$

the objective of $P(\mathcal{L}^\gamma, \text{KLD}, \mathcal{Q})$ has closed form if for observations $x_{1:n}$ and all $q \in \mathcal{Q}$

$$I^{(\gamma)}(\boldsymbol{\theta}) = \int_{\mathcal{X}} p(x|\boldsymbol{\theta})^\gamma dx, \quad F_1(\boldsymbol{\kappa}) = \int_{\boldsymbol{\Theta}} T(\boldsymbol{\theta}) q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\boldsymbol{\theta}, \quad F_2(\boldsymbol{\kappa}) = \int_{\boldsymbol{\Theta}} I^{(\gamma)}(\boldsymbol{\theta})^{\frac{1-\gamma}{\gamma}} q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\boldsymbol{\theta}$$

are closed form functions of $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$ for all x_i such that $(\eta(\boldsymbol{\kappa}) + (\gamma - 1) g(x_i)) \in \mathcal{N}$.

5.4.2 Additional details on Black-Box GVI (BBGVI)

Standard VI is scalable using doubly stochastic, model-agnostic optimization techniques (e.g. Paisley et al., 2012; Hoffman et al., 2013; Titsias and Lázaro-Gredilla, 2014; Salimans and Knowles, 2014; Wu et al., 2019) collectively known as black box VI (Ranganath et al., 2014). We extend these methods to black box GVI (BBGVI), an inference algorithm directly inheriting the modularity of the posteriors defined by $P(\ell, D, \mathcal{Q})$. This makes it easy to build BBGVI into existing software: For example, adapting the Deep Gaussian Process implementation of Salimbeni and Deisenroth (2017) required ¡100 lines of Python code.

Suppose $\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) : \boldsymbol{\kappa} \in K\}$ and that for all $(\boldsymbol{\kappa}, \boldsymbol{\theta}) \in (K, \boldsymbol{\Theta})$, one can sample $\boldsymbol{\theta} \sim q(\boldsymbol{\theta}|\boldsymbol{\kappa})$. Suppose also that the derivatives $\nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}|\boldsymbol{\kappa}))$ and $\nabla_{\boldsymbol{\kappa}} D(q||\pi)$ exist (almost surely relative to the measure on Θ induced by q). For many choices of D, $\mathcal{Q}$ and π, $\nabla_{\boldsymbol{\kappa}} D(q||\pi)$ is available in closed form. In this case, BBGVI is particularly attractive and GVI posteriors can be computed through an unbiased gradient estimate given as

$$\nabla_{\boldsymbol{\kappa}} \hat{L}(q|\ell, D, \mathcal{Q}) = \frac{1}{S} \sum_{s=1}^{S} \left\{ \sum_{i=1}^{n} \ell(\boldsymbol{\theta}^{(s)}, x_i) \cdot \nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa})) \right\} + \nabla_{\boldsymbol{\kappa}} D(q||\pi) \tag{18}$$

and relying on an independent sample $\boldsymbol{\theta}^{(1:S)} \overset{i.i.d}{\sim} q(\boldsymbol{\theta}|\boldsymbol{\kappa})$. Since all models in the experiments of Section 6 admit closed forms for $\nabla_{\boldsymbol{\kappa}} D(q||\pi)$, this is the gradient estimator we use in the current paper. If a closed form for $\nabla_{\boldsymbol{\kappa}} D(q||\pi)$ is not available but $D(q||\pi) = \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[\ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})\right]$ for a function $\ell^D_{\boldsymbol{\kappa},\pi} : \boldsymbol{\Theta} \to \mathbb{R}$, one could use the alternative unbiased gradient estimate

$$\nabla_{\boldsymbol{\kappa}} \hat{L}(q|\ell, D, \mathcal{Q}) = \frac{1}{S} \sum_{s=1}^{S} \left\{ \left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}^{(s)}, x_i) + \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)}) \right] \cdot \nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa})) + \nabla_{\boldsymbol{\kappa}} \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)}) \right\} \tag{19}$$

This can be deployed for most divergences of interest, including the family of f-divergences. In some cases however, divergences will not be linear in q so that one has $D(q\|\pi) =$

$\tau\left(\mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[\ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})\right]\right)$ for some non-linear function $\tau : \mathbb{R} \to \mathbb{R}$. In this case, BBGVI can be performed based on the biased gradient estimate

$$\nabla_{\boldsymbol{\kappa}} \hat{L}(q|\ell, D, \mathcal{Q}) = \frac{1}{S}\sum_{s=1}^{S}\left\{\sum_{i=1}^{n} \ell(\boldsymbol{\theta}^{(s)}, x_i) \cdot \nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa}))\right\} + \tau\left(\frac{1}{S}\sum_{s=1}^{S}\ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)})\right) \cdot \frac{1}{S}\sum_{s=1}^{S}\nabla_{\boldsymbol{\kappa}}\ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)}). \tag{20}$$

Note that the induced bias of this estimator could be eliminated using the work of Grathwohl et al. (2017). Whichever form the gradient takes, one can apply most standard black box variance reduction techniques by introducing some control variate h (e.g. Ranganath et al., 2014; Wu et al., 2019; Grathwohl et al., 2017), see also Appendix H for details. Algorithm 1 summarizes a generic BBGVI procedure.

Proposition 17 clarifies under which conditions closed forms for $\nabla_{\boldsymbol{\kappa}} D(q\|\pi)$ are available for the case of robust divergences.

Proposition 17 (Closed form D) *Let q, π with natural parameters $\boldsymbol{\eta}_q, \boldsymbol{\eta}_\pi$ be in the exponential family $\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\eta}) = h(\boldsymbol{\theta}) \exp\{\boldsymbol{\eta}' T(\boldsymbol{\theta}) - A(\boldsymbol{\eta})\} : \eta \in \mathcal{N}\}$ with natural parameter space $\mathcal{N} = \{\boldsymbol{\eta} : \exp\{A(\boldsymbol{\eta})\} < \infty\}$. Then,*

(1) $D_A^{(\alpha)}(q||\pi)$ and $D_{AR}^{(\alpha)}(q||\pi)$ have a closed form if $\alpha \in (0,1)$ or if $\alpha\boldsymbol{\eta}_q + (1-\alpha)\boldsymbol{\eta}_\pi \in \mathcal{N}$

(2) $D_B^{(\beta)}(q||\pi)$ has a closed form if $h(\boldsymbol{\theta}) = h$ does not depend on $\boldsymbol{\theta}$ and additionally, $(\beta-1)\cdot\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2 \in \mathcal{N}$ for any $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathcal{N}$ (amongst others, this holds for Beta, Gamma, Gaussian, exponential or Laplace distributions)

(3) $D_G^{(\gamma)}(q||\pi)$ has closed form if $D_B^{(\beta)}(q||\pi)$ does for $\beta = \gamma$.

Roughly speaking, the above Proposition holds for all exponential families that are typically of interest in variational inference schemes.

6. Experiments

Having introduced an inference strategy that is generic enough to work on high-dimensional, black box Bayesian models, the remainder of the paper studies GVI on two applications of interest in Bayesian Deep Learning. Before doing so, notice that as indicated in Table 1, previous work constitutes various interesting special cases of GVI with other strong empirical results (e.g., Futami et al., 2018; Knoblauch et al., 2018; Chérief-Abdellatif and Alquier, 2019a; Jankowiak et al., 2019) We add to this body of evidence by deploying GVI on Bayesian Neural Networks (BNNs) and Deep Gaussian Processes (DGPs) to address the particular ways in which these two models challenge the assumptions underlying the standard Bayesian paradigm. All code used for generating the experiments is available from `https://github.com/JeremiasKnoblauch/GVIPublic`.

Algorithm 1 Black box GVI (BBGVI)

Input: $x_{1:n}$, π, D, ℓ, $\mathcal{Q}$, h, StoppingCriterion, $\boldsymbol{\kappa}_0$, K, S, $t = 0$, LearningRate

done $\leftarrow$ False
while not done **do**

// STEP 1: Get a subsample from $x_{1:n}$ of size K
$\rho_{1:K} \leftarrow \text{SampleWithoutReplacement}(1:n, K)$
$x(t)_{1:K} \leftarrow x_{\rho_{1:K}}$

// STEP 2: Sample from $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_t)$ and compute losses
$\boldsymbol{\theta}^{(1:S)} \overset{i.i.d.}{\sim} q(\boldsymbol{\theta}|\boldsymbol{\kappa}_t)$
$\ell_{i,s} \leftarrow \ell(\boldsymbol{\theta}^{(s)}, x(t)_i) \cdot \nabla_{\boldsymbol{\kappa}_t} \log q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa}_t)$ for all $s = 1, 2, \dots S$ and $i = 1, 2, \dots, K$
$\ell_s \leftarrow \frac{n}{K} \sum_{i=1}^{K} \ell_{i,s}$ for all $s = 1, 2, \dots S$

// STEP 3: Compute divergence term
if $D(q\|\pi)$ admits closed form **then**
$\ell_s \leftarrow \ell_s + \nabla_{\boldsymbol{\kappa}} D(q\|\pi)$ for all $s = 1, 2, \dots S$
else if $D(q\|\pi) = \mathbb{E}_q[\ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})]$ **then**
$\ell_s \leftarrow \ell_s + \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)}) \nabla_{\boldsymbol{\kappa}_t} \log q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa}_t) + \nabla_{\boldsymbol{\kappa}_t} \ell^D_{\boldsymbol{\kappa}_t,\pi}(\boldsymbol{\theta}^{(s)})$ for all $s = 1, 2, \dots S$
else if $D(q\|\pi) = \tau\left(\mathbb{E}_q[\ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})]\right)$ **then**
$\ell_s \leftarrow \ell_s + \tau\left(\frac{1}{S}\sum_{s=1}^{S} \ell^D_{\boldsymbol{\kappa}_t,\pi}(\boldsymbol{\theta}^{(s)})\right) \cdot \nabla \boldsymbol{\kappa}_t \ell^D_{\boldsymbol{\kappa}_t,\pi}(\boldsymbol{\theta}^{(s)})$ for all $s = 1, 2, \dots S$

// STEP 4: Apply variance reduction via h if desired
if $h \neq$ None **then**
$h_s \leftarrow h(\boldsymbol{\theta}^{(s)}, \ell_s)$
$\ell_s \leftarrow \ell_s - h_s$ for all for all $s = 1, 2, \dots S$

// STEP 5: Update $\boldsymbol{\kappa}_t$ and stopping criterion
$\rho_t \leftarrow \text{LearningRate}(t)$
$L \leftarrow \frac{1}{S}\sum_{s=1}^{S} \ell_s$
$\boldsymbol{\kappa}_{t+1} \leftarrow \boldsymbol{\kappa}_t + \rho_t \cdot L$
done $\leftarrow \text{StoppingCriterion}(\boldsymbol{\kappa}_{t+1}, \boldsymbol{\kappa}_t, t)$
$t \leftarrow t + 1$

6.1 Bayesian Neural Network Regression

As alluded to in Example 1 and Section 5.2.1, BNN models should be expected to suffer from prior misspecification. Focusing on the regression case, we wish to alleviate this problem using GVI's modularity and thus focus on varying D. Accordingly, we fix the loss function

to the usual negative log likelihood $\ell(\boldsymbol{\theta}, y_i, x_i, \sigma^2) = -\log p_{\mathcal{N}}(y_i|x_i, \sigma^2, F(\boldsymbol{\theta}))$ for

$$p_{\mathcal{N}}(y_i|x_i, \sigma^2, F(\boldsymbol{\theta})) = \mathcal{N}(y_i|F(\boldsymbol{\theta}), x_i, \sigma^2),$$

and choose $\mathcal{Q} = \mathcal{Q}_{\text{MFN}}$ as the normal mean field variational family given in eq. (8). With this in hand, we compare three different constructions of posterior beliefs:

(1) **Standard VI** as described in Section 2.3;

(2) **DVI** methods introduced as approximations to the standard Bayesian posterior $q^*_{\text{B}}(\boldsymbol{\theta})$ that find $q^*_{\text{A}}(\boldsymbol{\theta}) = \arg\min_{q\in\mathcal{Q}} D(q\|q^*_{\text{B}}(\boldsymbol{\theta}))$ with D being the α-divergence (Hernández-Lobato et al., 2016)[10] and Rényi's α-divergence (Li and Turner, 2016);

(3) **GVI** with $D = D^{(\alpha)}_{AR}$.

To make comparisons as fair as possible, our implementation is built on top of that used for the results of Li and Turner (2016) and only changes the objective being optimized. Similarly, all settings and data sets for which the methods are compared are unchanged and taken directly from Li and Turner (2016) and Hernández-Lobato et al. (2016): We use a single-layer network with 50 ReLU nodes on all experiments. Inference is performed via probabilistic back-propagation (Hernández-Lobato and Adams, 2015) and the ADAM optimizer (Kingma and Ba, 2014) with its default settings, 500 epochs and a batch size of 32. Priors and variational posteriors are both fully factorized normal distributions. Further, the results are also evaluated on the same selection of UCI data sets (Lichman, 2013) and in the same way as they were in Li and Turner (2016) and Hernández-Lobato et al. (2016): Using 50 random splits of the relevant data into training (90%) and test (10%) sets, the inferred models are evaluated predictively on the test sets using the average negative log likelihood (NLL) as well as the average root mean square error (RMSE). For each of the 50 splits, predictions are computed based on 100 samples from the variational posterior.

We summarize the two main results of our experiments as follows: First, Figure 11 depicts what appears to be the most typical relationship between **VI**, **DVI** and **GVI** on BNNs. Second, Figure 12 explores a surprising finding about the typical relationship further and connects it back to the modularity result in Theorem 10. The Appendix contains some further results.

6.1.1 Typical patterns (Figure 11)

As Figure 11 demonstrates, several findings form a consistent pattern across a range of data sets. Three findings are most poignant.

(A) **DVI** can often achieve a performance gain for the NLL relative to **standard VI**, but much less so for RMSE. On both metrics, there is no clear pattern of improvement.

(B) Relative to **standard VI**, **GVI** significantly improves performance for both NLL and RMSE if $\alpha > 1$. Conversely, **GVI** worsens performance if $\alpha \in (0, 1)$. In other words, *larger posterior variances adversely affect predictive quality.*

10. We align the parameterization of the $D^{(\alpha)}_A$ with the current paper, meaning $1 - \alpha_{\text{current}} = \alpha_{\text{H.-L. et al. (2016)}}$

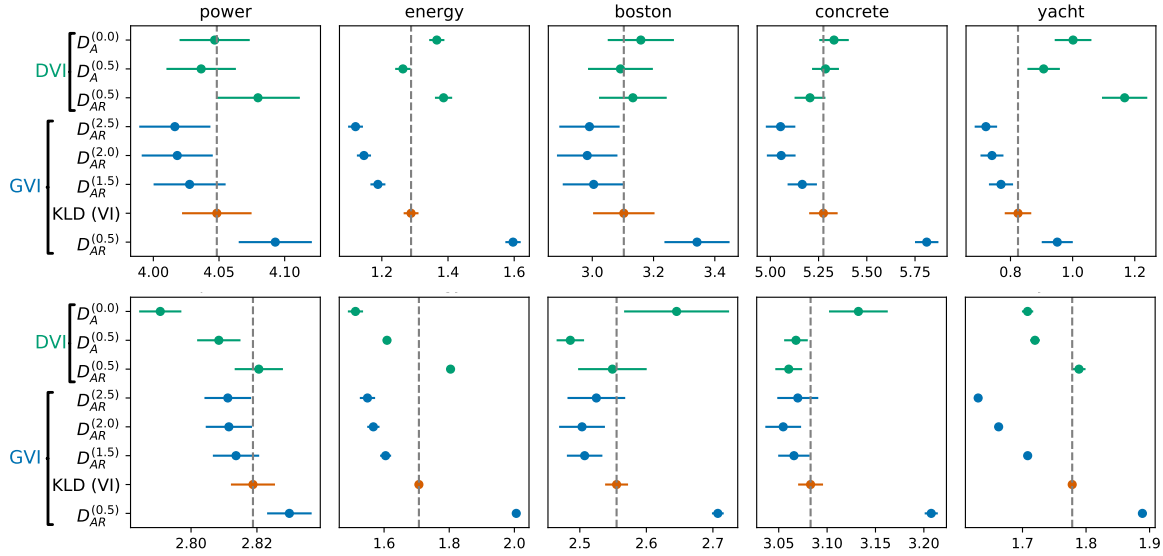


Figure 11: Best viewed in color. Top row depicts RMSE, bottom row the NLL across a range of data sets using BNNs. Dots correspond to means, whiskers to standard errors. The further to the left, the better the predictive performance. For the depicted selection of data sets, a clear common pattern exists for the performance differences between **standard VI**, **DVI** and **GVI**.

(C) **GVI** performance is a clear banana-shaped function of α across all data sets: While predictive performance benefits as α gets larger than one, the improvement flattens out and bends back in a banana shape as α grows too large. In other words, *decreasing uncertainty relative to the standard variational posterior improves predictive performance, but becoming 'overconfident' worsens it.*

Finding (B) has a straightforward interpretation: Since it holds that $D_{AR}^{(\alpha)} \leq \text{KLD}$ for $\alpha > 1$ (see Van Erven and Harremos (2014)[11] and Figure 5), the GVI posteriors associated with $D_{AR}^{(\alpha)}$ for $\alpha > 1$ are *more* concentrated than the standard VI posteriors, a phenomenon also depicted on toy models in Figure 8. In other words: Ignoring more of the poorly specified prior and consequently being closer to a point mass at the empirical risk minimizer is beneficial for predictive performance. As alluded to in Example 1, this is to be expected: it is doubtful if a literal interpretation of the prior as in **(P)** is appropriate for BNNs. As finding (C) shows however, this does not mean that point estimates are preferable to posterior beliefs: Increasing the value of α shrinks the variances too much, eventually impeding predictive performance.

6.1.2 The surprising benefits of modularity (Figure 12)

While findings (B) and (C) should not come as a surprise by themselves, they do raise an interesting question: In particular, GVI for $D_{AR}^{(\alpha)}$ with $\alpha = 0.5$ is the *worst-performing*

11. Note that their result holds for a different parameterization of the $D_{AR}^{(\alpha)}$, but it is easy to show that our parameterization is strictly smaller than theirs for $\alpha > 1$.

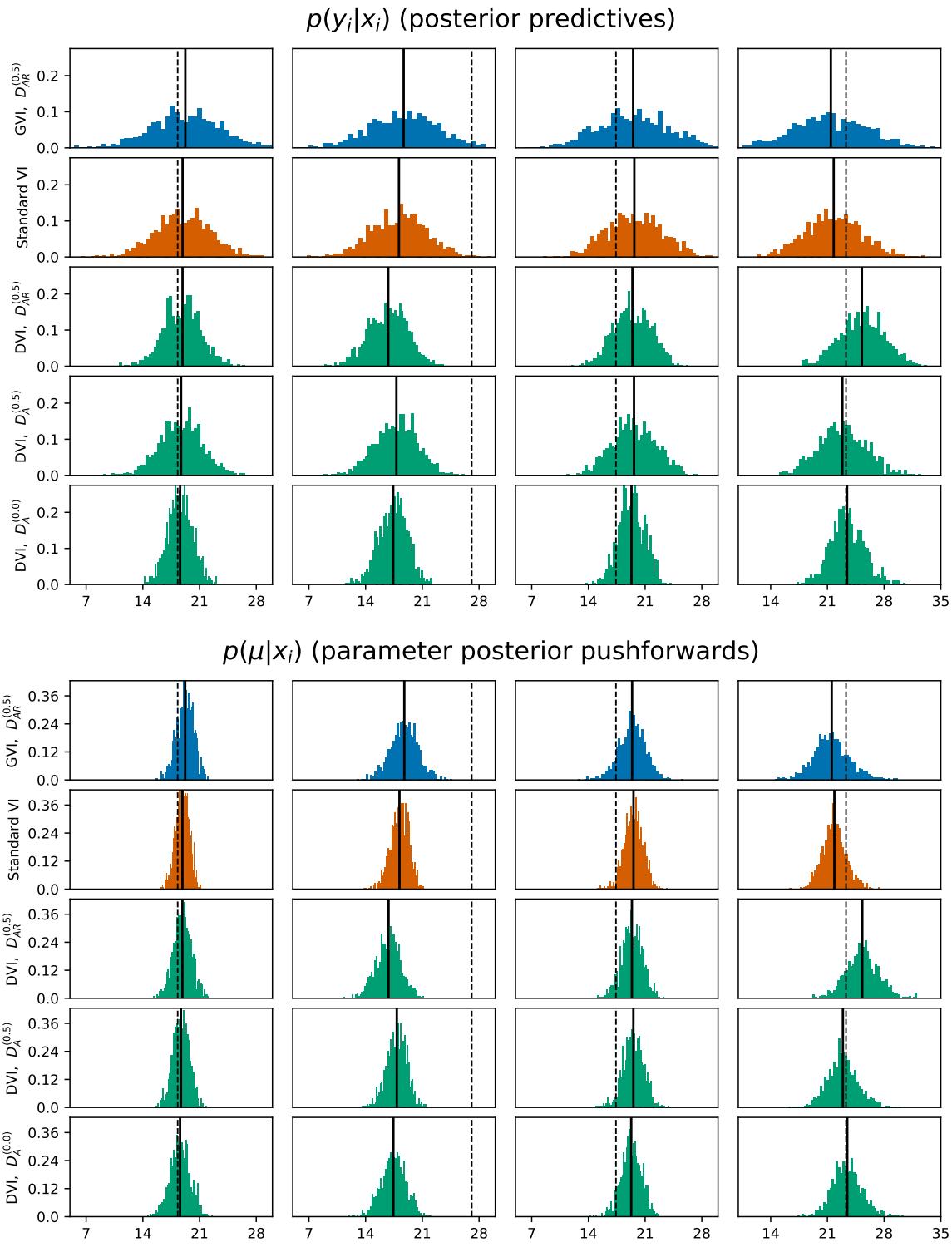


Figure 12: Best viewed in color. Depicted are test set predictions based on posterior predictives (**top panel**) and parameter posterior pushforwards (**bottom panel**) with four observations in the boston data set. Each column shows one observation (dashed line). The predictive distributions (histogram) and their means (solid line) for each row correspond to **standard VI**, **DVI** and **GVI**.

setting across the board. This is remarkable because this setting also constructs the only GVI posteriors in our experiments with *wider variances* than standard VI. At the same time, producing wider variances and more conservative uncertainty quantification is one of the main motivations for Expectation Propagation (EP) and the presented DVI methods, see for example Figure 1(a) in Li and Turner (2016) or Figure 8 in Hernández-Lobato et al. (2016). This is puzzling: Are wider variances for $\boldsymbol{\theta}$ somehow beneficial for DVI posteriors' predictive performance while damaging that of GVI posteriors? As it turns out, this is not the case. Rather, while both GVI with $\alpha = 0.5$ and all DVI methods produce parameter posteriors with larger variances, in the case of DVI this does not translate into predictive uncertainty—as would be expected in standard Bayesian inference.

This phenomenon is depicted in Figure 12, which clearly shows that the additional uncertainty in the DVI parameter posteriors $q^*_{\text{DVI}}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)$ is completely overshadowed by an extreme degree of variance shrinkage in the corresponding posterior predictives. In other words, the increased uncertainty in $\boldsymbol{\theta}$ is outweighed by extremely small values for σ^2. The plot demonstrates this by comparing the push-forward $F\#q^*_{\text{DVI}}(\cdot|\boldsymbol{\kappa}^*)$ with the posterior predictives. Formally, the push-forward is given by

$$p(\mu|x_i) = \left(F\#q^*_{\text{DVI}}(\cdot|\boldsymbol{\kappa}^*)\right)(\mu),$$

where the operation # is simply a formalization of the following two operations: (i) sample $\boldsymbol{\theta} \sim q^*_{\text{DVI}}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)$, (ii) compute $\mu = F(\boldsymbol{\theta})$. The posterior predictive then integrates the push-forward measure $p(\mu|x_i)$ over the likelihood function as

$$p(y_i|x_i) = \int_{\boldsymbol{\Theta}} p_{\mathcal{N}}(y_i|x_i, \sigma^2, F(\boldsymbol{\theta}))q^*_{\text{DVI}}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)d\boldsymbol{\theta} = \int_{\mathbb{R}} p_{\mathcal{N}}(y_i|x_i, \sigma^2, \mu)p(\mu|x_i)d\mu.$$

As Figure 12 shows, the push-forward (i.e. the posterior predictive) behaves as expected for both GVI and DVI. For DVI, the same cannot be said: specifically, the posterior predictive generally has much *less* variance than that of standard VI.

This surprising phenomenon is due to hyperparameter optimization for σ^2 and has an intimate link with the modularity result of Theorem 10. Since variational inference on σ^2 complicates the DVI objectives, both Hernández-Lobato et al. (2016) and Li and Turner (2016) do not infer σ^2 probabilistically. Instead, it is optimized over their objectives. This approach poses an optimization problem which for $D = D_A^{(\alpha)}$ and $D = D_{AR}^{(\alpha)}$ is given by

$$\widehat{\sigma}^2, q^*_{\text{DVI}}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*) = \arg\min_{\sigma^2} \left\{ \arg\min_{q \in \mathcal{Q}} D(q(\boldsymbol{\theta}|\boldsymbol{\kappa})||q^*_{\text{B}}(\boldsymbol{\theta}|\sigma^2, x_{1:n}, y_{1:n})) \right\}. \tag{21}$$

Crucially, the inner part of this objective conditions on the exact Bayesian posterior for a *fixed* value of σ^2 and then seeks to approximate the posterior belief given by

$$q^*_{\text{B}}(\boldsymbol{\theta}|\sigma^2, x_{1:n}, y_{1:n}) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} p_{\mathcal{N}}(y_i|x_i, \sigma^2, F(\boldsymbol{\theta})).$$

At the same time however, the outer part of the objective seeks to find a value for σ^2 which makes the posterior $q^*_{\text{B}}(\boldsymbol{\theta}|\sigma^2, x_{1:n}, y_{1:n})$ as easily approximable as possible. In other words, an objective which is explicitly motivated as a projection of $q^*_{\text{B}}(\boldsymbol{\theta}|\sigma^2, x_{1:n}, y_{1:n})$ into $\mathcal{Q}$ also changes the very point from which to project into $\mathcal{Q}$.

Though it would be computationally easy to perform probabilistic inference on σ^2 within GVI, we also optimize σ^2 as a hyperparameter for comparability. Thus, we pose the alternative optimization problem

$$\widehat{\sigma}^2, q^*_{\text{GVI}}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*) = \arg\min_{\sigma^2} \left\{ \arg\min_{q\in\mathcal{Q}} \left\{ \mathbb{E}_q \left[\sum_{i=1}^{n} -\log p_{\mathcal{N}}(y_i|x_i, \sigma^2, F(\boldsymbol{\theta})) \right] + D^{(\alpha)}_{AR}(q||\pi) \right\} \right\} \quad (22)$$

As Figure 12 shows, the outcomes are drastically different: Unlike in the DVI case, the predictive uncertainty for the GVI posteriors move in the same direction as parameter uncertainty as α varies. The modularity of GVI makes it obvious what the optimization over σ^2 corresponds to in eq. (22): Rather than choosing a posterior $q^*_{\text{B}}(\boldsymbol{\theta}|\sigma^2, x_{1:n}, y_{1:n})$ which minimizes the cost of projecting into $\mathcal{Q}$ via $D^{(\alpha)}_{AR}$, the optimization problem simply seeks to find the best possible loss $\ell_{\sigma^2}(y_i|x_i, F(\boldsymbol{\theta})) = -\log p(y_i|x_i, \sigma^2, F(\boldsymbol{\theta}))$ over all $\sigma^2 \in \mathbb{R}_+$.

6.2 Deep Gaussian Processes

Deep Gaussian Processes (DGPs) were introduced by Damianou and Lawrence (2013) and extend the logic of deep learning to the nonparametric Bayesian setting. The principal idea is to construct a hierarchy of Gaussian Process (GP) priors over latent spaces. Unlike with BNNs, the priors in DGPs are usually refined at run-time by using various hyperparameter optimization schemes. This is in fact crucial for DGPs as it ensures that the inputs $\boldsymbol{X}$ are mapped into latent spaces which are informative for the outputs $\boldsymbol{Y}$. As a consequence, and unlike with BNNs, we expect there to be comparatively little merit in varying D for DGPs—a suspicion we experimentally confirm in Appendix J.2.3. Accordingly, we instead focus on experiments that vary the loss ℓ. More specifically, we consider replacing the negative log score with a robust scoring rule for the likelihood which is derived from the γ-divergence (Hung et al., 2018), which drastically improves predictive performance.

In the remainder, we first introduce DGPs (Section 6.2.1). Next, we provide a brief overview of the doubly stochastic inference procedure in Salimbeni and Deisenroth (2017) (Section 6.2.2) and show how to adapt DGPs to GVI (Section 6.2.3). Lastly, we present numerical experiments and their results (Section 6.2.4). These findings are also summarized with a higher level of detail in a separate technical report (Knoblauch, 2019b).

6.2.1 Preliminaries for DGPs

Given observations $(\boldsymbol{X}, \boldsymbol{Y})$ where $\boldsymbol{X} \in \mathbb{R}^{n\times D_x}$ and $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$, a DGP of L layers introduces L latent functions $\{\boldsymbol{F}^l\}_{l=1}^L$. Here, $\boldsymbol{F}^l$ is matrix-valued and of dimension $D^l \times D^{l+1}$. Setting $\boldsymbol{F}^0 = \boldsymbol{X}$, $D^0 = D_x$ and $D^{l+1} = p$, one can write the DGP construction as

$$\begin{aligned}
\boldsymbol{Y} \mid \boldsymbol{F}^L &\sim p\left(\boldsymbol{Y} \middle| \boldsymbol{F}^L\right) \\
\boldsymbol{F}^L|\boldsymbol{F}^{L-1} &= f^L(\boldsymbol{F}^{L-1}) \sim \text{GP}\left(\mu^L(\boldsymbol{F}^{L-1}), \text{K}^L(\boldsymbol{F}^{L-1}, \boldsymbol{F}^{L-1})\right) \\
&\cdots \\
\boldsymbol{F}^1 \mid \boldsymbol{F}^0 &= f^1(\boldsymbol{F}^0) \sim \text{GP}\left(\mu^1(\boldsymbol{F}^0), \text{K}^1(\boldsymbol{F}^0, \boldsymbol{F}^0)\right),
\end{aligned}$$

where the mean and covariance functions are $\mu^l : \mathbb{R}^{D^l} \to \mathbb{R}^{D^{l+1}}$ and $\text{K}^l : \mathbb{R}^{D^l \times D^l} \to \mathbb{R}^{D^{l+1}\times D^{l+1}}$. Scalable inference strategies for this model generally rely on VI (Damianou and

Lawrence, 2013; Dai et al., 2016; Salimbeni and Deisenroth, 2017; Hensman and Lawrence, 2014), Monte Carlo methods (Vafa, 2016; Wang et al., 2016) or more specialized approaches (Cutajar et al., 2017a). In the remainder, we discuss the implications of Generalized Variational Inference (GVI) in relation to the arguably most popular VI approach of Salimbeni and Deisenroth (2017). Unlike previous VI methods, it encodes conditional dependence into the variational family $\mathcal{Q}$ and outperformed Expectation Propagation (EP) based alternatives (Bui et al., 2016).

6.2.2 Doubly stochastic VI in DGPs

First, define m inducing points $\boldsymbol{Z}^l = (\boldsymbol{z}_1^l, \boldsymbol{z}_2^l, \dots, \boldsymbol{z}_m^l)^T$ and their function values $\boldsymbol{U}^l = (f^l(\boldsymbol{z}_1^l), f^l(\boldsymbol{z}_2^l), \dots, f^l(\boldsymbol{z}_m^l))^T$ (for details on inducing points, see Snelson and Ghahramani, 2006; Titsias, 2009; Bonilla et al., 2019; Matthews et al., 2016). For improved readability, we drop $\boldsymbol{X}$ and $\boldsymbol{Z}^l$ from the conditioning sets and denote the i-th row of $\boldsymbol{F}^l$ as $\boldsymbol{f}_i^L$. With this, the joint distribution of the DGP is

$$p\left(\boldsymbol{Y}, \{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L\right) = \underbrace{\prod_{i=1}^n p(\boldsymbol{y}_i|\boldsymbol{f}_i^L)}_{\text{likelihood}} \times \underbrace{\prod_{l=1}^L p\left(\boldsymbol{F}^l \middle| \boldsymbol{U}^l, \boldsymbol{F}^{l-1}, \boldsymbol{Z}^{l-1}\right) p\left(\boldsymbol{U}^l \middle| \boldsymbol{Z}^{l-1}\right)}_{\text{(DGP) prior}}.$$

The posteriors $p\left(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L\right)$ and $p\left(\{\boldsymbol{F}^l\}_{l=1}^L\right)$ are intractable. The VI method proposed in Salimbeni and Deisenroth (2017) overcomes this with the variational family given by

$$q\left(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L\right) = \prod_{l=1}^L p\left(\boldsymbol{F}^l \middle| \boldsymbol{U}^l, \boldsymbol{F}^{l-1}, \boldsymbol{Z}^{l-1}\right) q\left(\boldsymbol{U}^l\right); \quad q\left(\boldsymbol{U}^l\right) = \mathcal{N}\left(\boldsymbol{U}^l \middle| \boldsymbol{m}^l, \boldsymbol{S}_l\right).$$

This allows for exact integration over the inducing points $\{\boldsymbol{U}^l\}_{l=1}^L$, yielding

$$q\left(\{\boldsymbol{F}^l\}_{l=1}^L\right) = \prod_{l=1}^L \mathcal{N}\left(\boldsymbol{F}^l \middle| \boldsymbol{\mu}^l, \boldsymbol{\Sigma}_l\right).$$

As shown in Salimbeni and Deisenroth (2017), this enables a doubly stochastic minimization of the negative Evidence Lower Bound (ELBO) given by

$$\begin{aligned} &\mathbb{E}_{q(\boldsymbol{F}^L)}\left[\sum_{i=1}^n -\log p(\boldsymbol{y}_i|\boldsymbol{F}^L)\right] + \mathrm{KLD}\left(q(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L)\middle\| p(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L)\right) \\ &= -\sum_{i=1}^n \mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\log p(\boldsymbol{y}_i|\boldsymbol{f}_i^L)\right] + \sum_{l=1}^L \mathrm{KLD}(q(\boldsymbol{U}^l)||p(\boldsymbol{U}^l)). \end{aligned} \tag{23}$$

For optimization, the samples for $\boldsymbol{F}^l$ are drawn using the variational posteriors from the previous layers so that approximating the expectations over $q(\boldsymbol{f}_i^L)$ induces the first layer of stochasticity. The second layer is due to drawing mini-batches from $\boldsymbol{X} = \boldsymbol{F}^0$ and $\boldsymbol{Y}$. Because of this large degree of stochasticity, it is appealing if $\mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\log p(\boldsymbol{y}_i|\boldsymbol{f}_i^L)\right]$ is available in closed form, which is for instance the case if $p = p_{\mathcal{N}}$ is a normal likelihood.

6.2.3 Adaption to GVI

The objective in eq. (23) suggests itself naturally to a GVI variant. This raises two questions:

(***D***) Do we still penalize the deviation from the *global prior* if we simply replace the KLD-terms layer-wise?

(***ℓ***) Can one derive closed forms for the expectations when the log scoring rule is replaced by robust alternatives $\mathcal{L}^\beta$ or $\mathcal{L}^\gamma$ derived from the β- and γ-divergence?

As shown next, we can give a positive answer to both these questions.

(*D*) Conveniently and as shown in Salimbeni and Deisenroth (2017),

$$\text{KLD}\left(q(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L)\middle\| p(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L)\right) = \sum_{l=1}^L \text{KLD}(q(\boldsymbol{U}^l)||p(\boldsymbol{U}^l)). \tag{24}$$

A natural question is whether one can reverse-engineer this finding: If we simply pick a collection of other divergences $D^l(q(\boldsymbol{U}^l)||p(\boldsymbol{U}^l))$ for each layer l and combine them additively, does the result define a valid divergence between $q(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L)$ and $p(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L)$? As the next Corollary shows, one can prove that reverse-engineering prior regularizers inspired by eq. (24) is feasible so long as the layer-specific divergences D^l are f-divergences or monotonic transformations of f-divergences. The proof relies on a technical Lemma and is given in Appendix J.2.1

Corollary 18 *In the* DGP *construction of eq.* (23)*, replacing the sum of* KLD*-terms by*

$$\sum_{l=1}^L D^l(q(\boldsymbol{U}^l)||p(\boldsymbol{U}^l))$$

defines a valid divergence between $q(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L)$ *and* $p(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L)$ *so long as* D^l *is an* f*-divergence or a divergence obtained as a monotonic transform* g *of an* f*-divergence for all* $l = 1, 2, \dots L$*.*

(*ℓ*) Next, we turn attention to modifying the loss terms in eq. (23). First, note that

$$\mathbb{E}_{q(\boldsymbol{F}^L)}\left[\sum_{i=1}^n -\log p(\boldsymbol{y}_i|\boldsymbol{F}^L)\right] = -\sum_{i=1}^n \mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\log p(\boldsymbol{y}_i|\boldsymbol{f}_i^L)\right].$$

This identity still holds if one replaces the negative log with other scoring rules. As the next Proposition shows, we even retain closed forms for the expectations over $q(\boldsymbol{f}_i^L)$ in the regression case and for the scoring rules given by

$$\mathcal{L}_p^\beta(\boldsymbol{f}_i^L, \boldsymbol{y}_i) = -\frac{1}{\beta-1} p(\boldsymbol{y}_i|\boldsymbol{f}_i^L)^{\beta-1} + \frac{I_{p,\beta}(\boldsymbol{f}_i^L)}{\beta}$$
$$\mathcal{L}_p^\gamma(\boldsymbol{f}_i^L, \boldsymbol{y}_i) = -\frac{1}{\gamma-1} p(\boldsymbol{y}_i|\boldsymbol{f}_i^L)^{\gamma-1} \cdot \frac{\gamma}{I_{p,\gamma}(\boldsymbol{f}_i^L)^{\frac{\gamma-1}{\gamma}}}.$$

Crucially, the integral term $I_{p,c}(\boldsymbol{f}_i^L) = \int p(\boldsymbol{y}|\boldsymbol{f}_i^L)^c d\boldsymbol{y}$ is generally available in closed form for exponential families. As the notation suggests, $\mathcal{L}_p^\beta$ is linked to the β-divergence in the same way we linked the log score to the KLD in Section 5.2.2, see also Basu et al. (1998). Similarly, $\mathcal{L}_p^\gamma$ is derived from the γ-divergence as explained in Hung et al. (2018). As also alluded to in Section 5.4.1, $\mathcal{L}_p^\gamma$ ($\mathcal{L}_p^\beta$) recovers the log score as $\gamma \to 1$ ($\beta \to 1$) and produces robust inferences for $\gamma > 1$ ($\beta > 1$). Figure 6 depicts this for $\mathcal{L}_p^\beta$, and the behaviour is very similar for $\mathcal{L}_p^\gamma$.

Proposition 19 (Closed forms for robust DGP regression) *If it holds that* $\boldsymbol{y}_i \in \mathbb{R}^d$,

$$p(\boldsymbol{y}_i|\boldsymbol{f}_i^L) = \mathcal{N}\left(\boldsymbol{y}_i; \boldsymbol{f}_i^L, \sigma^2 I_d\right); \quad q(\boldsymbol{f}_i^L) = \mathcal{N}(\boldsymbol{f}_i^L; \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

then for the quantities given by

$$\widetilde{\boldsymbol{\Sigma}}^{-1} = \left(\frac{c}{\sigma^s}\boldsymbol{I}_d + \boldsymbol{\Sigma}^{-1}\right); \quad \widetilde{\boldsymbol{\mu}} = \left(\frac{c}{\sigma^2}\boldsymbol{y}_i + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right); \quad I(c) = (2\pi\sigma^2)^{-0.5dc} c^{-0.5d}$$

and for

$$E(c) = \frac{1}{c}\left(2\pi\sigma^2\right)^{-0.5dc} \frac{|\widetilde{\boldsymbol{\Sigma}}|^{0.5}}{|\boldsymbol{\Sigma}|^{0.5}} \exp\left\{-\frac{1}{2}\left(\frac{c}{\sigma^2}\boldsymbol{y}_i^T\boldsymbol{y}_i + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}^T\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\mu}}\right)\right\}$$

the following expectations are available in closed form:

$$\mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\mathcal{L}_p^\beta(\boldsymbol{f}_i^L, \boldsymbol{y}_i)\right] = -E(\beta - 1) + \frac{I(\beta)}{\beta}$$
$$\mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\mathcal{L}_p^\beta(\boldsymbol{f}_i^L, \boldsymbol{y}_i)\right] = -E(\gamma - 1) \cdot \frac{\gamma}{I(\gamma)^{\frac{\gamma-1}{\gamma}}}$$

As shown in Appendix J.2.2, it is easy but tedious to derive this result. While the results of using $\mathcal{L}_p^\beta$ and $\mathcal{L}_p^\gamma$ are often virtually identical (see for instance Figure 22 in Appendix J.1), our experiments on DGPs will exclusively use $\mathcal{L}_p^\gamma$. This is done because unlike for $\mathcal{L}_p^\beta$, computations with $\mathcal{L}_p^\gamma$ can be performed in its numerically more stable log form.

6.2.4 Results

As with the experiments on BNNs in the previous section, we make comparisons as fair as possible by using the `gpflow` (Matthews et al., 2017) implementation of Salimbeni and Deisenroth (2017). Further, we use the same settings, meaning that all experiments use 20,000 iterations of the ADAM optimizer (Kingma and Ba, 2014) with a learning rate of 0.01 and default settings for all other hyperparameters. We perform inference for each of the UCI data sets (Lichman, 2013) after normalization using the RBF kernel with dimension-wise lengthscales, 100 inducing points, with batch sizes of $\min(1000, n)$ and $D^l = \min(D_x, 30)$. As before, we use 50 random splits with 90% training and 10% test data to assess predictive performance in terms of negative log likelihood (NLL) and root mean square error (RMSE). With this, we compare two inference schemes:

(1) The state of the art **standard VI** techniques of Salimbeni and Deisenroth (2017);

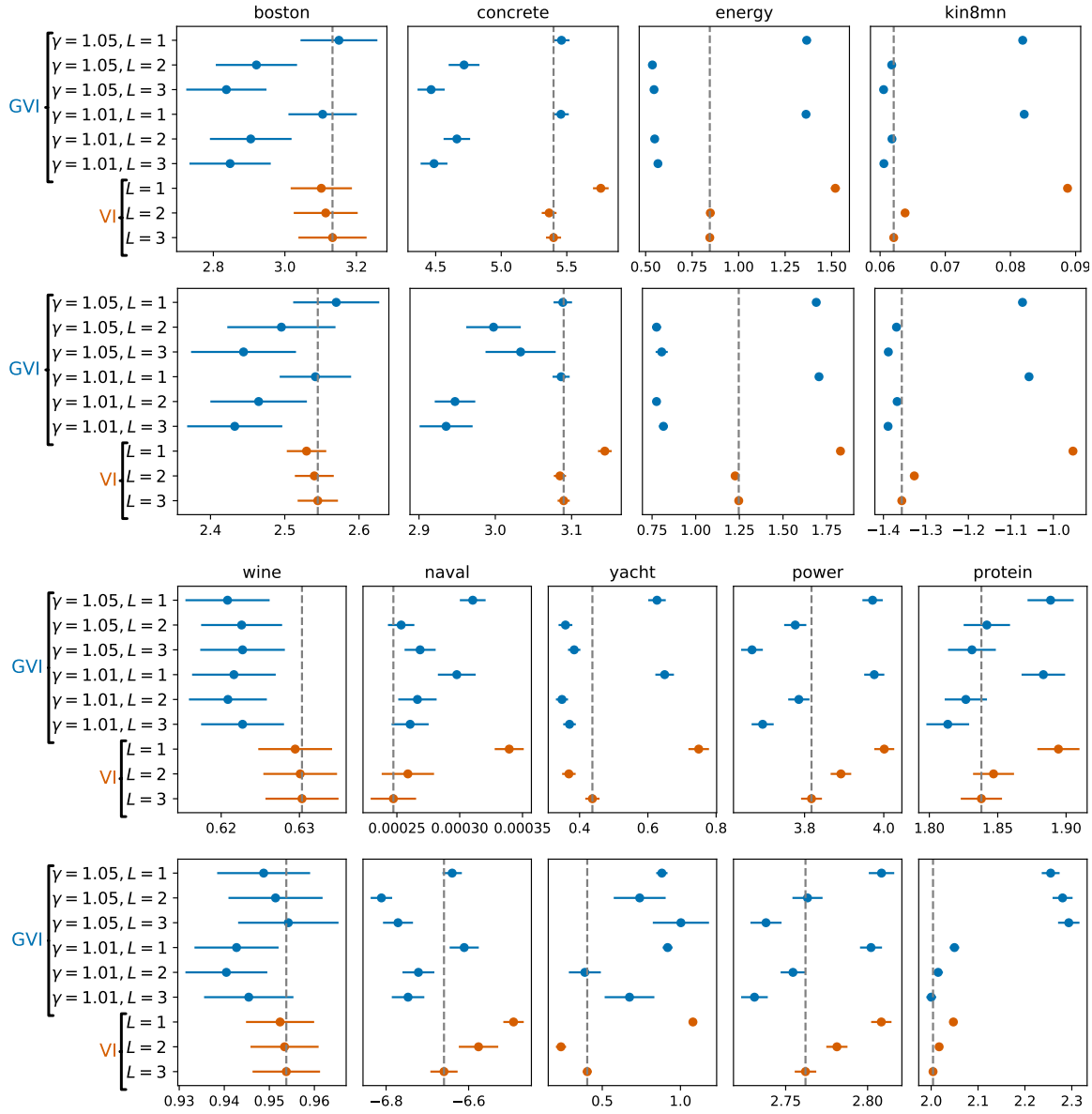


Figure 13: Best viewed in color. Top rows depict RMSE, bottom rows the NLL across a range of data sets using DGPs. Dots correspond to means, whiskers to standard errors. The further to the left, the better the predictive performance. For the depicted selection of data sets, **GVI** comprehensively outperforms **standard VI**.

(2) A **GVI** variant of the same inference method which replaces the log score with the robust γ-divergence based scoring rule $\mathcal{L}_p^\gamma$.

For choosing γ, we note that inferences are robust for $\gamma > 1$ and that $\mathcal{L}_p^\gamma$ recovers the log score as $\gamma \to 1$. At the same time, the scoring rule will grow increasingly happy to ignore

virtually all of the data as $\gamma \to \infty$. Accordingly, one will typically want to pick

$$\gamma = 1 + \varepsilon$$

for a small $\varepsilon > 0$. Choosing γ in this way encodes the intuition that a good scoring rule will behave like the log score for all but the most extreme outliers. We thus pick $\varepsilon \in \{0.01, 0.05\}$. We note that hyperparameter optimization might appear to be the natural choice for picking γ, but will not perform well in practice: Rather than producing robust inferences, this will select for a value of γ generally producing the smallest GVI objective values across $\mathcal{Q}$[12].

The results are depicted in Figure 13 and confirm our two main intuitions about robustness: Firstly, the robust scoring rule provides a significant performance improvement. Secondly, the smaller value of γ (which will be closer to the log score) generally outperforms the larger value of γ, though both choices are equally good in many data sets[13]. We believe that the performance gains of the robust scoring rule is due to large parts of the latent spaces being non-informative. This implies that it is beneficial to implicitly down-weight the influence of these non-informative parts of the latent space. It is clear that robust scoring rules do exactly that (see for instance Figure 6), which explains their superior performance in the DGP experiments. This intuition is further bolstered by the following observation: Generally, performance *improves* with a larger number of layers L under the robust score $\mathcal{L}_p^\gamma$, but *worsens* under the log score. In other words: The more dispersed the prior over the latent space (i.e., the DGP) becomes, the more inferential outcomes benefit from implicitly ignoring its non-informative regions. In Appendix J.2 we provide a small batch of additional results showing that as expected, modifying D is less beneficial for DGPs than it is for BNNs. Most likely, this is due to hyperparameter optimization for the kernels of the DGP: together with the fact that Gaussian Processes are far more informative priors than fully factorized normals, careful selection of the hyperparameters ensures that unlike in the BNN case, the prior is informative.

7. Discussion & Conclusion

In this work, we re-examined the working assumptions that have proven powerful and useful in spreading Bayesian inference into virtually all domains of scientific endeavour. Studying the challenges of contemporary inference, we concluded that the traditional assumptions underlying Bayesian statistics are misaligned with the realities of modern large-scale problems. At the same time, we adopted an optimization-centric view on Bayesian inference that endows standard Vatiational Inference (VI) with a particular form of optimality relative to other approximations. In spite of this, belief distributions computed as alternative approximations to the Bayesian posterior often perform better in practice. This is because standard VI is optimal *only* relative to a particular objective function—an objective function whose form is based on the very assumptions that are misaligned with reality. Inspired by this insight, we proceed to derive a new class of optimization-centric posterior belief distributions that do not

12. In practice, this means that hyperparameter optimization pushes $\gamma \to 1$ or $\gamma \to \infty$, depending on the magnitudes of $\{p(\boldsymbol{y}_i|\boldsymbol{f}_i^L)\}_{i=1}^n$.
13. We expect this second finding about γ to generalize to new settings so long as the inputs are normalized and the outputs are not high-dimensional (see also Figure 22 for some empirical evidence of this on BNNs), which would make $\gamma = 1.01$ a decent default choice in such scenarios.

rely on these assumptions. To do so, we first set out a new axiomatic foundation for Bayesian inference culminating in the Rule of Three (RoT). The RoT is an optimization problem with three arguments, each of which can address one of the shortcomings of the standard Bayesian assumptions. Building on this mostly theoretical device, we introduce Generalized Variational Inference (GVI) as the sub-class of tractable posteriors derived from the RoT. We show that GVI satisfies a number of desirable theoretical properties: most notably, it is modular (in the sense of Theorem 10) and consistent in the frequentist sense. On the practical side, we show how GVI can be used to adjust posterior variances and produce inferences that are robust to model and prior misspecification. Lastly, we demonstrate the benefits of GVI posteriors on two model classes that encapsulate the misalignment between the assumptions underlying the traditional Bayesian paradigm and the realities of modern large-scale Bayesian inference: Bayesian Neural Networks (BNNs) and Deep Gaussian Processes (DGPs).

The current work makes two major contributions. The first of these is conceptual: We propose an optimization-centric generalization of Bayesian inference through the Rule of Three (RoT). This aspect of our work stands in the tradition of previous generalizations of Bayesian inference such as the one in Bissiri et al. (2016) and Jewson et al. (2018). Unlike previous work however, we take the first step in the development of Bayesian inference procedures that generalize beyond multiplicative belief updates. As explained in Sections 2 and 4, an immediate consequence of this generalization is a taxonomy of various variational inference procedures: Unlike most other variational approximations to the Bayesian posterior, VI is a special case of the RoT. More specifically, VI is the most conceptually appealing of all approximations under an optimization-centric view (see Theorem 2).

The second contribution is methodological and consists in making the RoT useful for real world inference problems via Generalized Variational Inference (GVI). We show that GVI modularly addresses the three shortcomings associated with traditional Bayesian inference. As Section 6 shows with two applications on large-scale inference problems, GVI posteriors can yield significant predictive performance improvements in modern statistical machine learning models.

With the provision of a new optimization-centric generalization on Bayesian inference, the current paper is only the first step on a long road to designing posteriors that conform with the demands of contemporary models and inferential problems. In the wake of this, several important questions have been left unanswered. For example, it is unclear how to choose hyperparameters occurring in the loss or prior regularizer beyond simplistic (albeit well-working) rules of thumb. GVI also has an obvious intimate connections with PAC-Bayesian approaches that we will be exploring in the near future. Moreover, the flexibility in choosing different prior regularizers D brings about another interesting question: Given that frequentist consistency holds, what impact does D have on the contraction rate? And do certain special cases of D endow GVI with compelling geometric interpretations?

In summary, the current work is but the start of an investigation into the theoretical, methodological and applied consequences of the RoT and GVI. It is clear that the ideas introduced in the current paper—while barely scratching the surface of the possible—have produced valuable insights and shown much promise in all three of these regards. Consequently, it is with much excitement that we look forward to future contributions on questions of theory, methodology and practice surrounding the RoT and GVI.

Acknowledgments

We would like to cordially thank Edwin Fong, Benjamin Guedj, Chris Holmes, David Dunson, Mark van der Wilk, Giles Hooker and Alex Alemi for fruitful discussions, insights, comments and pointers that were invaluable for improving the paper. JK and JJ are funded by EPSRC grant EP/L016710/1 as part of the Oxford-Warwick Statistics Programme (OxWaSP). JK is additionally funded by the Facebook Fellowship Programme and the London Air Quality project at the Alan Turing Institute for Data Science and AI. TD is funded by the UKRI Turing AI Fellowship EP/V02678X/1, EPSRC grant EP/T004134/1 and the Lloyd's Register Foundation programme on Data Centric Engineering at The Alan Turing Institute. This work was furthermore supported by The Alan Turing Institute for Data Science and AI under EPSRC grant EP/N510129/1 in collaboration with the Greater London Authority.

Appendix A. Definitions for robust divergences

The following is an overview of definitions for the most important divergences that are used throughout the paper.

Definition 20 (The $\alpha\beta\gamma$-divergence $D_G^{(\alpha,\beta,r)}$ (Cichocki and Amari, 2010)) *The $\alpha\beta\gamma$-divergence $D_G^{(\alpha,\beta,r)}$ Cichocki and Amari (2010) takes the form*

$$D_G^{(\alpha,\beta,r)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\alpha(\beta-1)(\alpha+\beta-1)r}\left[\left(\tilde{D}_G^{(\alpha,\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))+1\right)^r - 1\right]$$

where $r>0$, $\alpha\neq 0$, $\beta\neq 1$ and

$$\tilde{D}_G^{(\alpha,\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \int\left(\alpha q(\boldsymbol{\theta})^{\alpha+\beta-1} + (\beta-1)\pi(\boldsymbol{\theta})^{\alpha+\beta-1} - (\alpha+\beta-1)q(\boldsymbol{\theta})^{\alpha}\pi(\boldsymbol{\theta})^{\beta-1}\right)d\boldsymbol{\theta}$$

Below we list some well-known special cases of the family of divergences defined by $D_G^{(\alpha,\beta,r)}$ that we use in the main paper. This exposition is a summary of the review conducted in Cichocki and Amari (2010).

Definition 21 (The α-divergence ($D_A^{(\alpha)}$) (Chernoff, 1952; Amari, 2012)) *The α-divergence is defined as*

$$D_A^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\alpha(1-\alpha)}\left\{1 - \int q(\boldsymbol{\theta})^{\alpha}\pi(\boldsymbol{\theta})^{1-\alpha}d\boldsymbol{\theta}\right\},$$

where $\alpha\in\mathbb{R}\setminus\{0,1\}$. Note that $D_A^{(\alpha)}$ is recovered from $D_G^{(\alpha,\beta,r)}$ when $r=1$ and $\beta=2-\alpha$. $D_A^{(\alpha)}$ is also a member of the f-divergence family.

Definition 22 (Rényi's α-divergence ($D_{AR}^{(\alpha)}$) (Rényi, 1961)) *Rényi's α-divergence is defined as*

$$D_{AR}^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\alpha(\alpha-1)}\log\left(\int q(\boldsymbol{\theta})^{\alpha}\pi(\boldsymbol{\theta})^{1-\alpha}d\boldsymbol{\theta}\right),$$

where $\alpha\in\mathbb{R}\setminus\{0,1\}$. $D_{AR}^{(\alpha)}$ is recovered from $D_G^{(\alpha,\beta,r)}$ in the limit as $r\to 0$ and $\beta=2-\alpha$. Note that we use the rescaled version proposed by Liese and Vajda (1987); Cichocki and Amari (2010) rather than the original parameterization of Rényi (1961) because it links the divergence more closely to other robust alternatives of the KLD.

Definition 23 (The β-divergence ($D_B^{(\beta)}$) (Basu et al., 1998; Mihoko and Eguchi, 2002)) *The β-divergence (Mihoko and Eguchi, 2002) was originally introduced under the name "density power divergence" and is defined as*

$$D_B^{(\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\beta(\beta-1)}\int q(\boldsymbol{\theta})^{\beta}d\boldsymbol{\theta} + \frac{1}{\beta}\int\pi(\boldsymbol{\theta})^{\beta}d\boldsymbol{\theta} - \frac{1}{\beta-1}\int q(\boldsymbol{\theta})\pi(\boldsymbol{\theta})^{\beta-1}d\boldsymbol{\theta},$$

where $\beta\in\mathbb{R}\setminus\{0,1\}$. $D_B^{(\beta)}$ is recovered from $D_G^{(\alpha,\beta,r)}$ when $r=\alpha=1$. $D_B^{(\beta)}$ is a member of the Bregman-divergence family.

Definition 24 (The γ-divergence ($D_G^{(\gamma)}$) (Fujisawa and Eguchi, 2008)) *The γ-divergence (Fujisawa and Eguchi, 2008) is defined as*

$$D_G^{(\gamma)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\gamma(\gamma-1)} \log \frac{\left(\int q(\boldsymbol{\theta})^\gamma d\boldsymbol{\theta}\right)\left(\int \pi(\boldsymbol{\theta})^\gamma d\boldsymbol{\theta}\right)^{\gamma-1}}{\left(\int q(\boldsymbol{\theta})\pi(\boldsymbol{\theta})^\gamma d\boldsymbol{\theta}\right)^{\gamma}},$$

where $\gamma \in \mathbb{R} \setminus \{0,1\}$. $D_G^{(\gamma)}$ is recovered from $D_G^{(\alpha,\beta,r)}$ in the limit as $r \to 0$, $\alpha = 1$ and $\beta = \gamma$. The $D_G^{(\gamma)}$ can be shown to be generated from the $D_B^{(\beta)}$ applying the following transformation

$$c_0 \int g(x)^{c_1} f(x)^{c_2} dx \to c_0 \log \int g(x)^{c_1} f(x)^{c_2} dx$$

to all three of the $D_B^{(\beta)}$ terms. This is of interest because the $D_{AR}^{(\alpha)}$ is generated by the $D_A^{(\alpha)}$ by applying the same transformation of its two terms.

Remark 25 (Recovering the KLD) *The $D_A^{(\alpha)}$, $D_{AR}^{(\alpha)}$, $D_B^{(\beta)}$ and $D_G^{(\gamma)}$ all recover the* KLD *in the limit as $\alpha = \beta = \gamma \to 1$. This can be shown using the replica trick:*

$$\lim_{x \to 0} \frac{Z^x - 1}{x} = \log(Z).$$

Appendix B. Comparing robust divergences as prior regularizer

In order to understand the impact the choice of divergence used for regularization and its hyperparameter have on the inference, this section studies variations in the argument D. This investigation is conducted on a simple Bayesian linear regression example with two highly correlated predictors given by

$$\begin{aligned} \sigma^2 &\sim \mathcal{IG}(a_0, b_0) \\ \boldsymbol{\theta}|\sigma^2 &\sim \mathcal{N}_2\left(\boldsymbol{\mu}_0, \sigma^2 V_0\right) \\ y_i|\boldsymbol{\theta}, \sigma^2 &\sim \mathcal{N}\left(X_i\boldsymbol{\theta}, \sigma^2\right). \end{aligned} \tag{25}$$

We choose this example because it provides a closed form exact Bayesian posteriors and closed form objectives for the variational objectives of VI and GVI. Consequently, no sampling is required—neither for calculating the exact posterior nor for the optimization of the GVI and VI posteriors—so that numerical errors and uncertainties are kept to a minimum.

Studying the exact closed form Bayesian (normal) posterior for $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$, one observes that if the two predictors are correlated, then the posterior covariance of $\boldsymbol{\theta}$ will inherit this correlation. As we wish to investigate the underestimation of marginal variances for standard VI as well as the way in which GVI can address this, our numerical studies leverage this finding. In particular, we simulate the highly correlated predictors

$$(x_1, x_2)^T \sim \mathcal{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.9\\0.9 & 1\end{pmatrix}\right)$$

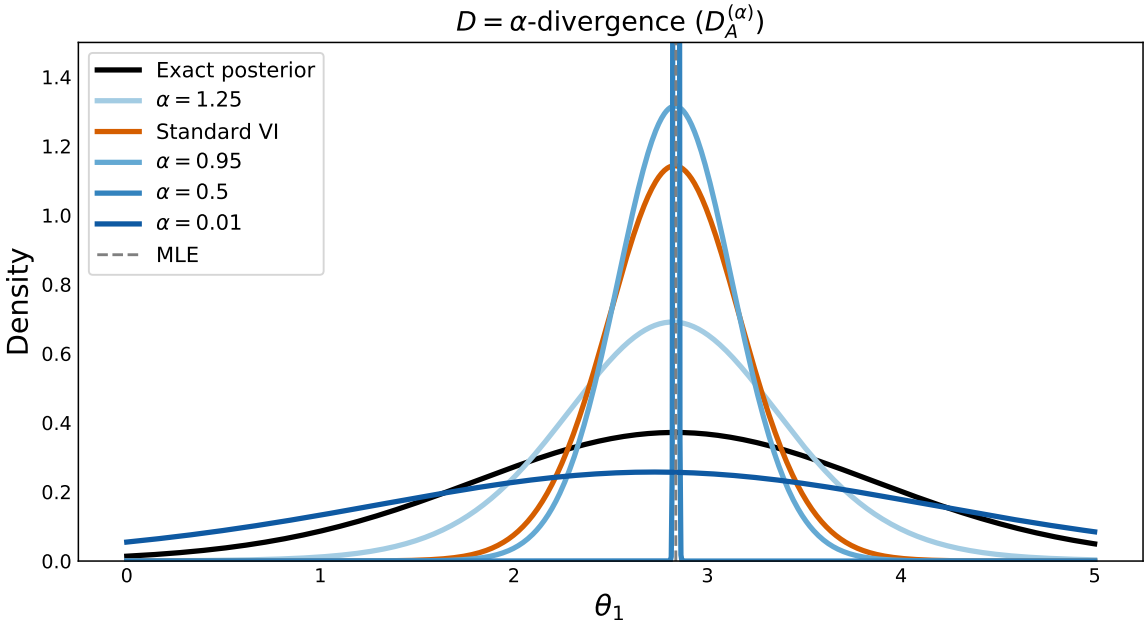


Figure 14: Best viewed in color. Marginal **VI** and **GVI** posterior for the $\boldsymbol{\theta}_1$ coefficient of a Bayesian linear model under the $D_A^{(\alpha)}$ prior regularizer for different values of α. The boundedness of the $D_A^{(\alpha)}$ causes **GVI** posteriors to severely over-concentrate if α is not carefully specified. Prior Specification: $\sigma^2 \sim \mathcal{IG}(20, 50)$, $\theta_1|\sigma^2 \sim \mathcal{N}(0, 25\sigma^2)$ and $\theta_2|\sigma^2 \sim \mathcal{N}(0, 25\sigma^2)$.

and compare the performance of the different **GVI** and **VI** posteriors on the resulting Bayesian linear regression. All posteriors are based on the the mean field normal variational family

$$\begin{aligned}
\mathcal{Q} = \{q(\theta_1|\sigma^2, \boldsymbol{\kappa}_n)q(\theta_2|\sigma^2, \boldsymbol{\kappa}_n)q(\sigma^2|\boldsymbol{\kappa}_n)\} \text{ so that } \boldsymbol{\kappa}_n = (a_n, b_n, \mu_{1,n}, \mu_{2,n}, v_{1,n}, v_{2,n})^T \\
\text{with } a_n, b_n, v_{1,n}, v_{2,n} > 0 \text{ and } \mu_{1,n}, \mu_{2,n} \in \mathbb{R} \\
q(\sigma^2|\boldsymbol{\kappa}_n) = \mathcal{IG}(\sigma^2|a_n, b_n) \\
q(\theta_1|\sigma^2, \boldsymbol{\kappa}_n) = \mathcal{N}\left(\theta_1|\mu_{1,n}, \sigma^2 v_{1,n}\right) \\
q(\theta_2|\sigma^2, \boldsymbol{\kappa}_n) = \mathcal{N}\left(\theta_2|\mu_{2,n}, \sigma^2 v_{2,n}\right).
\end{aligned}$$

For all experiments, $n = 25$ observations are simulated from eq. (26) with $\boldsymbol{\theta} = (2, 3)$ and $\sigma^2 = 4$. We use the negative log-likelihood of the correctly specified model as given in eq. (26) as loss function. To investigate GVI's behaviour across different prior regularizers, we vary its choice as $D \in \left\{D_A^{(\alpha)}, D_B^{(\beta)}, D_{AR}^{(\alpha)}, D_G^{(\gamma)}\right\}$. The results are depicted in Figs. 14 and 16-20. We summarize the most interesting results from these plots in the following three subsections.

B.1 A cautionary tale: The boundedness of the α-divergence ($D_A^{(\alpha)}$)

Of the alternative divergences to the KLD contained within the $D_G^{(\alpha,\beta,r)}$ family defined in Appendix A, $D_A^{(\alpha)}$ is arguably the most well known. Our results in Figure 14 show that in spite of its popularity in other contexts, the $D_A^{(\alpha)}$ is not a reliable prior regularizer within

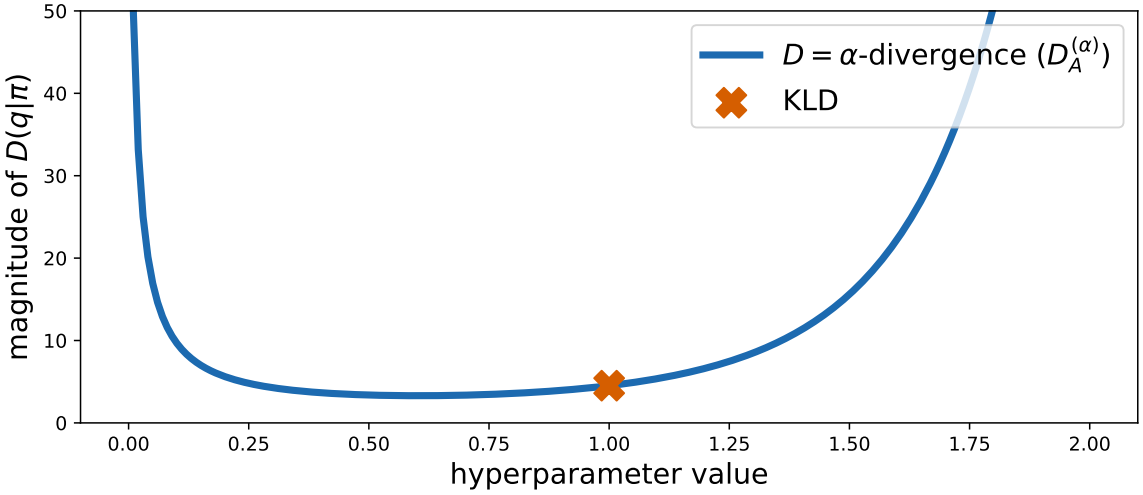


Figure 15: A comparison of the size of $D_A^{(\alpha)}$ for various values of α between two bivariate Normal Inverse Gamma distributions with $a_n = 512$, $b_n = 543$, $\boldsymbol{\mu}_n = (2.5, 2.5)$, $\mathbf{V}_n = \text{diag}(0.3, 2)$ and $a_0 = 500$, $b_0 = 500$, $\mu_0 = (0, 0)$, $V_0 = \text{diag}(25, 2)$.

GVI, at least for $\alpha \in (0, 1)$. In particular, the plot shows that the solutions to $P(\ell, D_A^{(\alpha)}, \mathcal{Q})$ can produce essentially degenerate posteriors if $\alpha \in (0, 1)$. Note also that this happens in spite of the relatively small sample size of $n = 25$. For example, when $\alpha = 0.5$, $P(\ell, D_A^{(\alpha)}, \mathcal{Q})$ is visually indistinguishable from a point mass at the maximum likelihood estimate. This is a consequence of the boundedness of $D_A^{(\alpha)}$ for $\alpha \in (0, 1)$: Specifically, it holds that $D_A^{(\alpha)} \leq (\alpha(1-\alpha))^{-1}$ for $\alpha \in (0, 1)$. As α decreases from 1, this upper-bound initially also decreases until reaching its minimum for $\alpha = 0.5$. As a result, decreasing α from unity to 0.5 significantly decreases the maximal penalty for posterior beliefs far from the prior. In turn, this forces the posterior to focus mostly on minimising the in-sample loss.

This phenomenon is depicted in Figure 15, which also shows that the divergence magnitude increases again as α approaches zero or if $\alpha > 1$. Comparing the plot with that in Figure 5, it is clear why hyperparameter selection for the other members of the $D_G^{(\alpha,\beta,r)}$ family of divergences is a less complicated endeavour than for the α-divergence. This does not mean that the $D_A^{(\alpha)}$ cannot be used for producing GVI posteriors: For example, in Figure 14, the $D_A^{(\alpha)}$ is able to achieve marginal variances that more closely correspond to the exact posterior for $\alpha = 1.25$ and $\alpha = 0.01$. Generally speaking, for values of α close to zero or above unity, it is possible to achieve more conservative uncertainty quantification. Yet, the $D_A^{(\alpha)}$ also functions primarily as a cautionary tale: Without understanding the properties of the prior regularizer D sufficiently well, GVI may well yield unsatisfactory posteriors.

B.2 Larger divergences produce larger marginal variances

In this section, we summarize the impact that a selection of robust divergences have on the marginal variances of the solution to $P(\ell, D, \mathcal{Q})$, again using the Bayesian Linear regression model from before. For a range of robust divergences, Figure 16 illustrates the impact that

changes in D have on the marginal variances of the resulting posteriors. As one should expect from re-examining Figure 5, the plot shows that $D_B^{(\beta)}$, $D_{AR}^{(\alpha)}$ and $D_G^{(\gamma)}$ are able to produce larger posterior variances for $\beta, \alpha, \gamma < 1$ and smaller posterior variances for $\beta, \alpha, \gamma > 1$. This is a manifestation of the posterior being penalized more heavily ($\beta, \alpha, \gamma < 1$) or less heavily

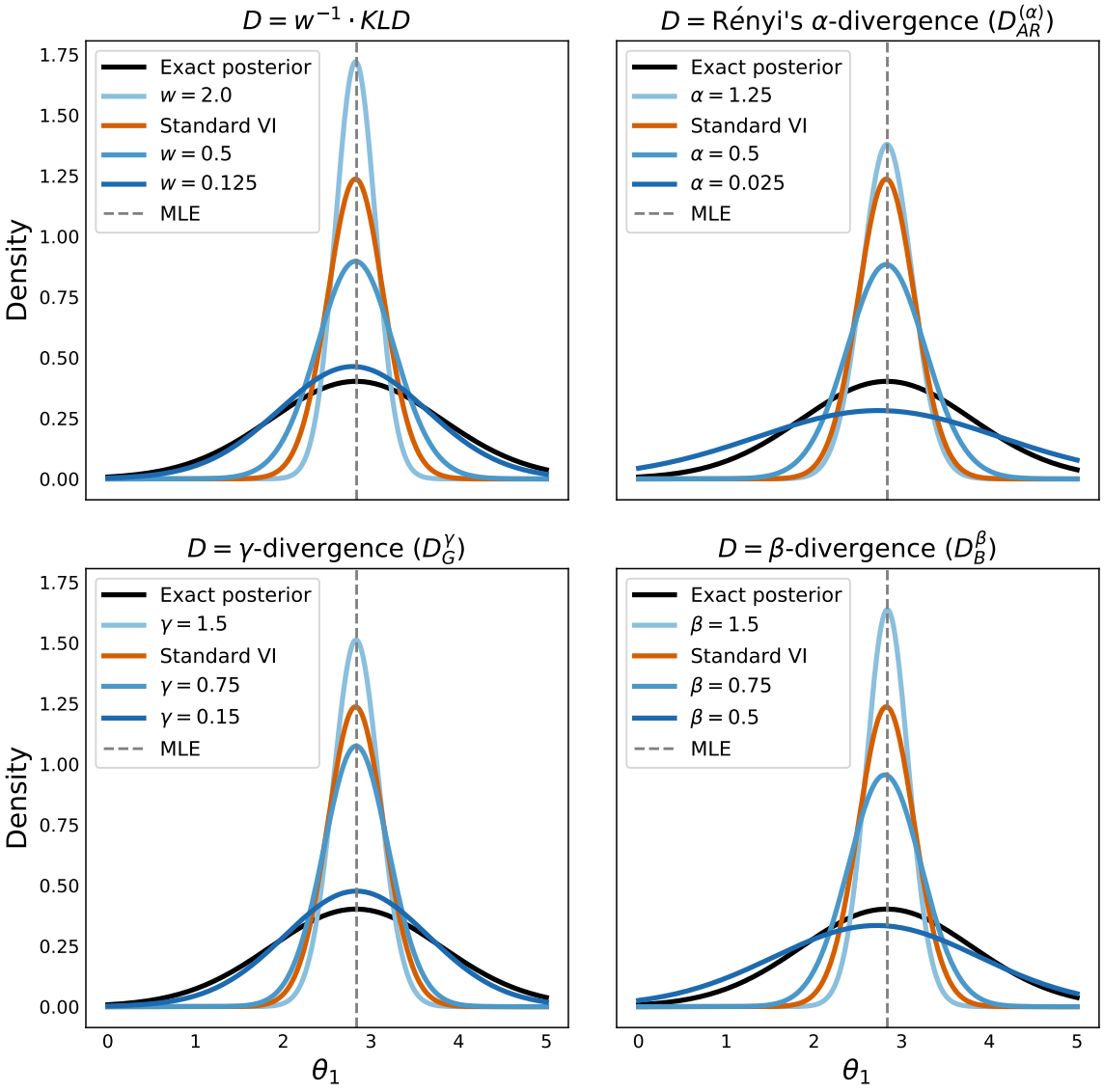


Figure 16: Best viewed in color. Marginal **VI** and **GVI** posterior for the first coefficient of a Bayesian linear model under the $D_{AR}^{(\alpha)}$, $D_B^{(\beta)}$, $D_G^{(\gamma)}$ and $\frac{1}{w}$KLD prior regularizers. Correlated covariates cause dependency in the **exact Bayesian posterior** of the coefficients and as a result **VI** underestimates marginal variances. **GVI** has the flexibility to produce wider marginal variances. Prior Specification: $\sigma^2 \sim \mathcal{IG}(20, 50)$, $\theta_1 \sim \mathcal{N}(0, 5^2)$ and $\theta_2 \sim \mathcal{N}(0, 5^2)$.

$(\beta, \alpha, \gamma > 1)$ for deviating from the prior than under the traditional VI. It follows that by choosing the divergence appropriately, GVI can allow greater control over the uncertainty quantification characteristics of the resulting posterior than what is possible under standard VI. Note that Figure 16 also compares the robust divergences against the re-weighted KLD. While the re-weighted KLD can prove a successful alternative for producing desirable variational posteriors with larger variances robust divergences if the prior is well-specified, this is no longer the case if the information contained in the prior cannot be relied upon. We study this and related findings surrounding robustness to the prior in the next section.

B.3 Robustness to the prior

Next, we compare the impact of changing the prior regularizer on the posterior's sensitivity to appropriate specification of the prior. Specifically, we consider and compare $D_B^{(\beta)}$, $D_{AR}^{(\alpha)}$, $D_G^{(\gamma)}$ and $\frac{1}{w}$KLD. When comparing $\frac{1}{w}$KLD with $D_{AR}^{(\alpha)}$ and $D_G^{(\gamma)}$, we fixed $\alpha = \gamma = w$. Setting the values of these various hyperparameters to be the same is intuitively appealing for comparison due to GVI's interpretation as approximate Evidence Lower Bound (ELBO), see Theorems 14 and 30. For the $D_B^{(\beta)}$, different values of β had to be selected to ensure its availability in a closed form.

B.3.1 weighted KLD ($\frac{1}{w}$KLD)

Figure 17 examines how changing the weight w affects the posteriors $P(\ell, \frac{1}{w}\text{KLD}, \mathcal{Q})$. Notice that this is equivalent to changing the negative log likelihood to a power likelihood with power w. Further, it should be clear that choosing $w < 1$ leads to posteriors that encourage larger variances, making them amenable to conservative uncertainty quantification. Unfortunately and again unsurprisingly, this comes at the price of making posteriors *more* sensitive to the prior: After all, one up-weights the term penalizing deviations from the prior. Conversely, $w > 1$ will result in posteriors that are less sensitive to the prior than standard VI. At the same time, they will also be more concentrated around the Maximum Likelihood Estimator. This makes $D = \frac{1}{w}$KLD less attractive than it could be: Setting w to smaller values will yield larger posteriors variances (at the expense of not being robust to the prior), while setting w to larger values will make the posterior more robust to misspecified priors (but at the expense of far more concentrated posteriors). As we shall see, this undesirable trade-off is *not* shared by the other (robust) divergences considered in this section. Unlike the $\frac{1}{w}$KLD, they often provide a way to have your cake and eat it, too.

B.3.2 Rényi's α-divergence ($D_{AR}^{(\alpha)}$)

Figure 18 demonstrates the sensitivity of $P(\ell, D_{AR}^{(\alpha)}, \mathcal{Q})$ to prior specification. For $0 < \alpha < 1$, the posterior exhibits the kind of behaviour that is difficult to attain with standard VI: It both produces larger marginal variances *and* is robust to badly specified priors. This is no longer true if $\alpha > 1$: For $\alpha > 1$, $D_{AR}^{(\alpha)} \leq$ KLD, so that it is more sensitive to the prior than the KLD. This flip in robustness as α crosses from $(0, 1)$ into values larger than one may seem strange, but can be understood by investigating the form of the $D_{AR}^{(\alpha)}$:

$$D_{AR}^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\alpha(\alpha-1)} \log \int q(\boldsymbol{\theta})^{\alpha} \pi(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta} = \frac{1}{\alpha(\alpha-1)} \log \int \frac{q(\boldsymbol{\theta})^{\alpha}}{\pi(\boldsymbol{\theta})^{\alpha-1}} d\boldsymbol{\theta}.$$

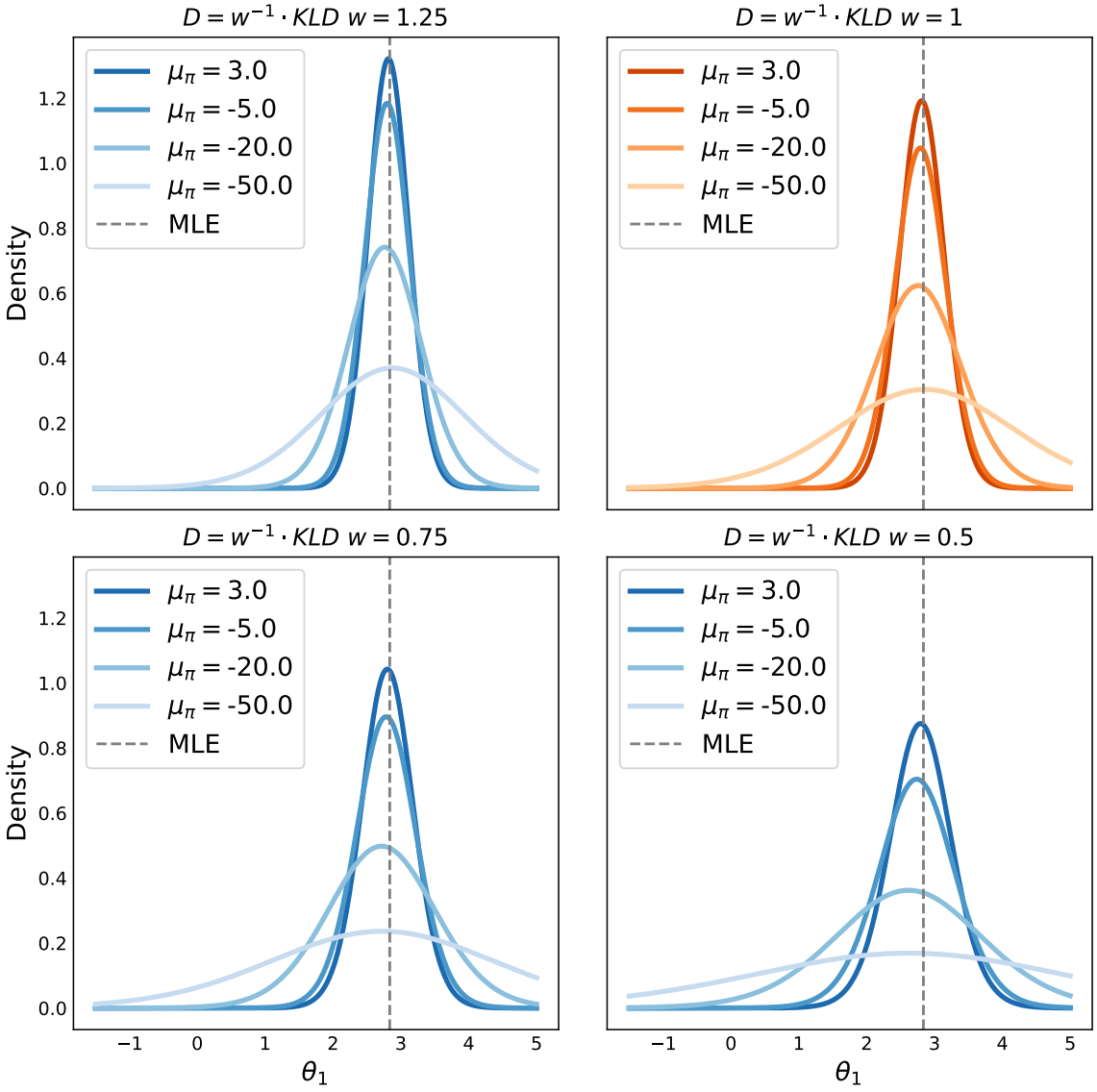


Figure 17: Best viewed in color. Marginal **VI** and **GVI** posterior for the coefficient of a Bayesian linear model under different priors using $D = \frac{1}{w}\text{KLD}$ as prior regularizer ($\frac{1}{w}\text{KLD}$ recovers KLD for $w = 1$). The prior specification is given by $\theta_1|\sigma^2 \sim \mathcal{N}(\mu_\pi, \sigma^2)$ with $\sigma^2 \sim \mathcal{IG}(3, 5)$.

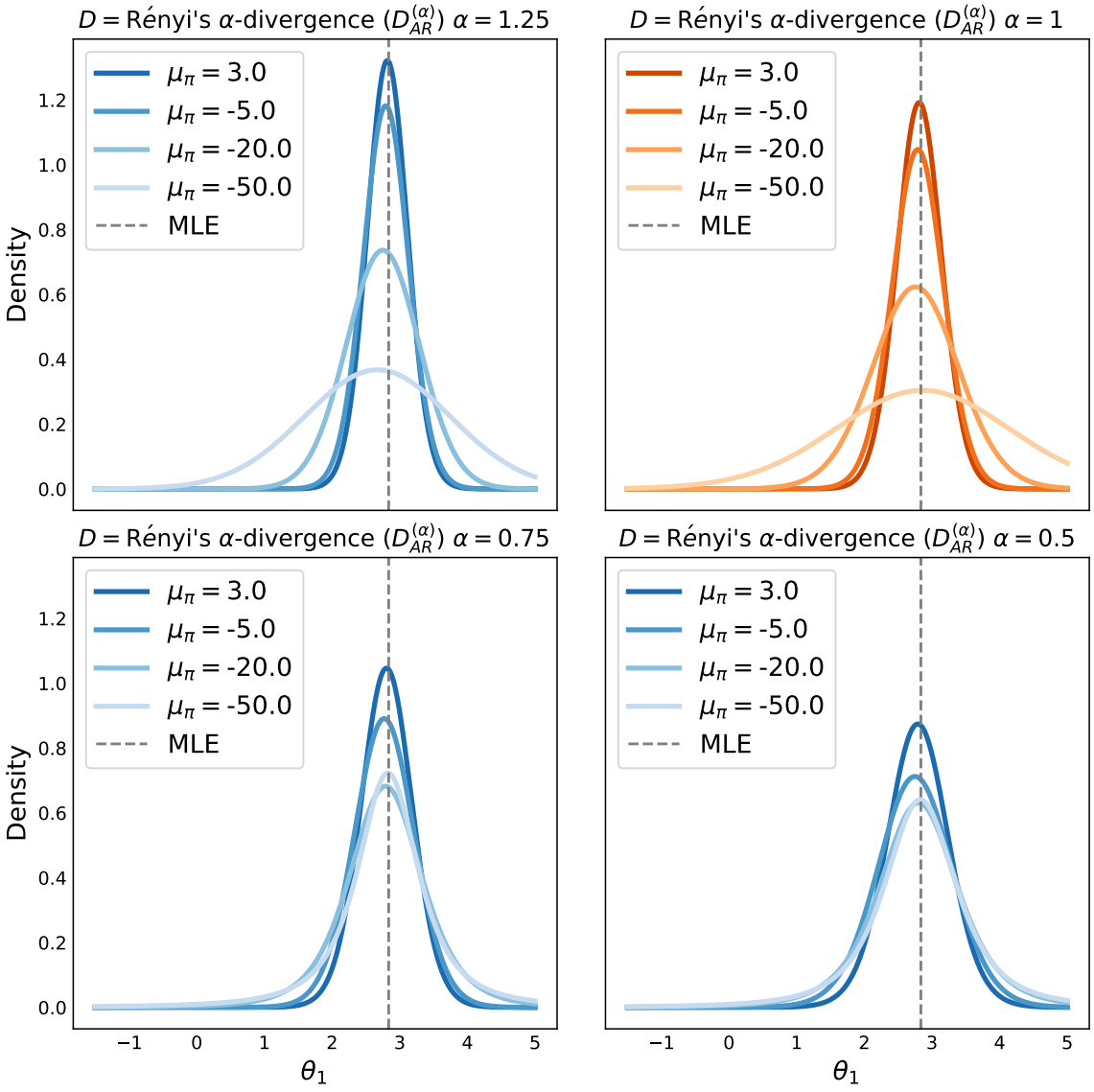


Figure 18: Best viewed in color. Marginal **VI** and **GVI** posterior for the coefficient of a Bayesian linear model under different priors using $D = D_{AR}^{(\alpha)}$ as prior regularizer ($D_{AR}^{(\alpha)}$ recovers KLD as $\alpha \to 1$). The prior specification is given by $\theta_1|\sigma^2 \sim \mathcal{N}(\mu_\pi, \sigma^2)$ with $\sigma^2 \sim \mathcal{IG}(3, 5)$.

It is clear that the magnitude of the divergence is determined by a ratio of two densities. Glancing closer, for $\alpha > 1$ this means that if $q(\boldsymbol{\theta})$ is large in an area where $\pi(\boldsymbol{\theta})$ is not, then a severe penalty is incurred. This limits how far the $q(\boldsymbol{\theta})$ can move from the prior and thus results in lack of prior robustness. Conversely, if $\alpha \in (0, 1)$, then $\pi(\boldsymbol{\theta})^{\alpha-1} > \pi(\boldsymbol{\theta})$ for regions where $\pi(\boldsymbol{\theta}) < 1$, which allows the posterior to spread its mass in a less concentrated way than for $\alpha > 1$. In fact, this very finding is also implicitly stated in Theorem 14.

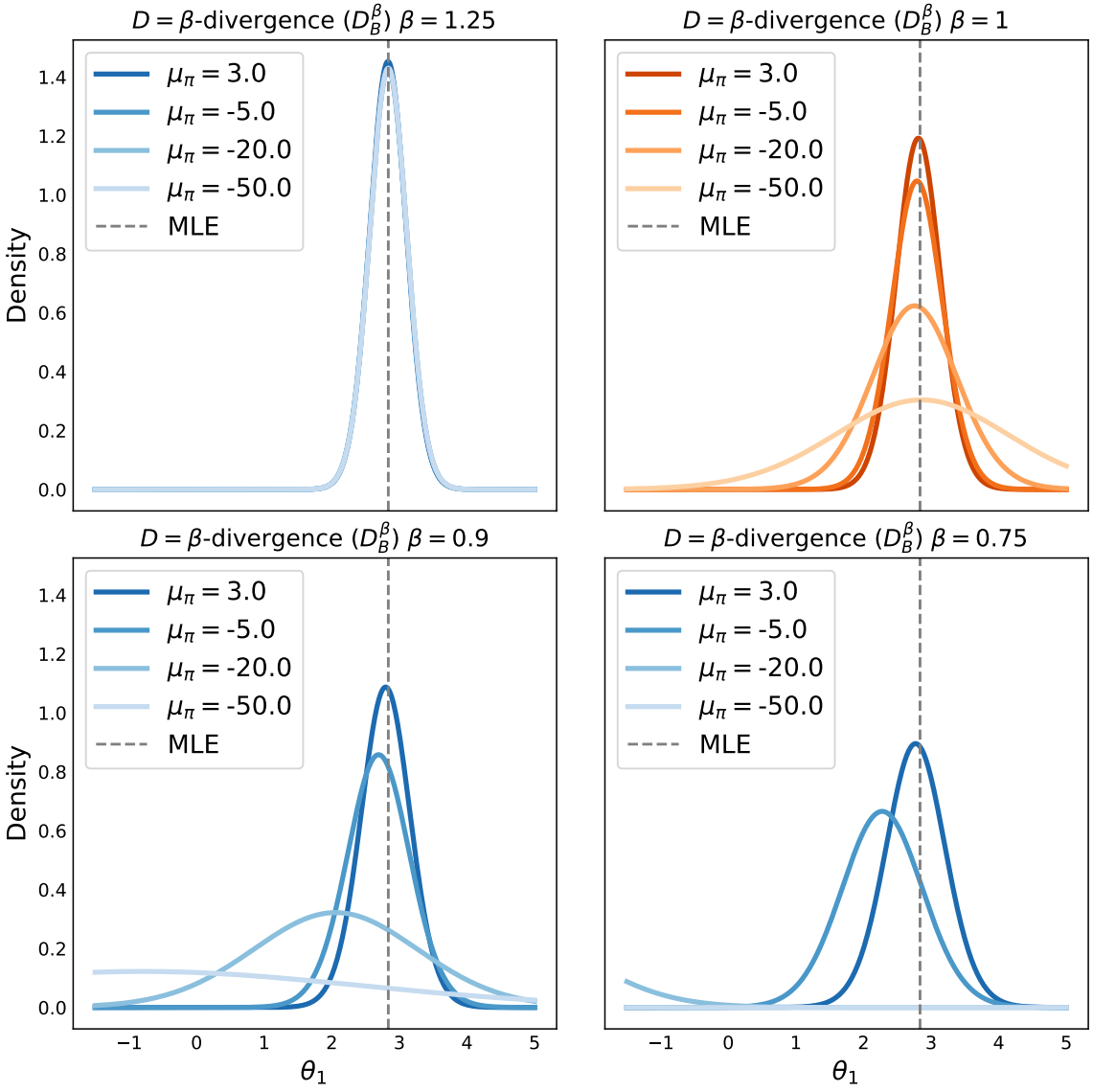


Figure 19: Best viewed in color. Marginal **VI** and **GVI** posterior for the coefficient of a Bayesian linear model under different priors using $D = D_B^{(\beta)}$ as prior regularizer ($D_B^{(\beta)}$ recovers KLD as $\beta \to 1$). The prior specification is given by $\theta_1|\sigma^2 \sim \mathcal{N}(\mu_\pi, \sigma^2)$ with $\sigma^2 \sim \mathcal{IG}(3, 5)$.

B.3.3 β-divergence ($D_B^{(\beta)}$)

Figure 19 demonstrates the sensitivity of $P(\ell, D_B^{(\beta)}, \mathcal{Q})$ to prior specification. The plot shows that $\beta > 1$ is able to achieve extreme robustness to the prior, while $\beta < 1$ causes extreme sensitivity to the prior. This phenomenon is a result of the fact that the $D_B^{(\beta)}$ decomposes into three integrals, one containing just the prior, one containing just $q(\boldsymbol{\theta})$ and one containing an interaction between them.

$$D_B^{(\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\beta}\int \pi(\boldsymbol{\theta})^\beta d\boldsymbol{\theta} - \frac{1}{\beta-1}\int \pi(\boldsymbol{\theta})^{\beta-1} q(\boldsymbol{\theta}) d\boldsymbol{\theta} + \frac{1}{\beta(\beta-1)}\int q(\boldsymbol{\theta})^\beta d\boldsymbol{\theta} \quad (26)$$

The integral depending only on the prior does not depend $q(\boldsymbol{\theta})$, so we can ignore it (since the prior is fixed across the different values of β).

For $\beta \in (0,1)$ the signs of both of the remaining terms flip and it is instructive to rewrite the middle term as $+\frac{1}{1-\beta}\int \frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})^{1-\beta}} d\boldsymbol{\theta}$ with $1-\beta > 0$. This shows that the prior appears as a denominator. The consequences of this are similar to the behaviour of the $D_{AR}^{(\alpha)}$ for $\alpha > 1$, if $q(\boldsymbol{\theta})$ has density where the prior has little density, then we divide a not-so-small number by a very small number and a huge penalty is incurred for this. As a result, the corresponding posterior will be very close to the prior. (In fact, notice that that two of the four posteriors for $\beta = 0.75$ in Figure 19 favour the prior so much that the density around the maximum likelihood estimate is virtually zero.) For $\beta > 1$ the opposite effect is observed. The prior no longer appears as a denominator and therefore deviations from the prior are punished in a milder manner by the middle term. This allows the third term, which depends on $q(\boldsymbol{\theta})$ independently of the prior, to have greater influence on how uncertainty is quantified. This third integral will become very large if the variance of $q(\boldsymbol{\theta})$ gets very small, which prevents it from quickly converging to a point mass at the maximum likelihood estimate. As a consequence, the $D_B^{(\beta)}$ is able to provide virtually prior-invariant uncertainty quantification for $\beta > 1$.

B.3.4 γ-divergence ($D_G^{(\gamma)}$)

Lastly, Fig. 20 demonstrates the sensitivity of $P(\ell, D_G^{(\gamma)}, \mathcal{Q})$ to prior specification. For $\gamma < 1$ it appears as though the $D_G^{(\gamma)}$ reacts similarly to the $\frac{1}{w}$KLD for $w < 1$. The $D_G^{(\gamma)}$ with $\gamma > 1$ produces greater robustness to the prior than the $\frac{1}{w}$KLD prior regularizer with $w > 1$, but this robustness is less extreme as it was for $D = D_B^{(\beta)}$. The reason for this is that although the $D_G^{(\gamma)}$ consists of the same three integral terms as the $D_B^{(\beta)}$, these terms are now transformed into the logarithmic scale. This means that the three integrals are combined multiplicatively (in the $D_G^{(\gamma)}$) rather than additively (in the $D_B^{(\beta)}$), which makes the variation across γ much smoother than across β: Unlike for the $D_B^{(\beta)}$, minimising the $D_G^{(\gamma)}$ no longer disregards any one term in order to minimise the others.

Appendix C. Proof of Theorem 10

Before we can prove Theorem 10, we first formally define the notion of robustness to model misspecification. Our understanding of robustness to model misspecification is aligned with Hampel et al. (2011) and Tukey (1960). In the words of the latter, robustness stands for

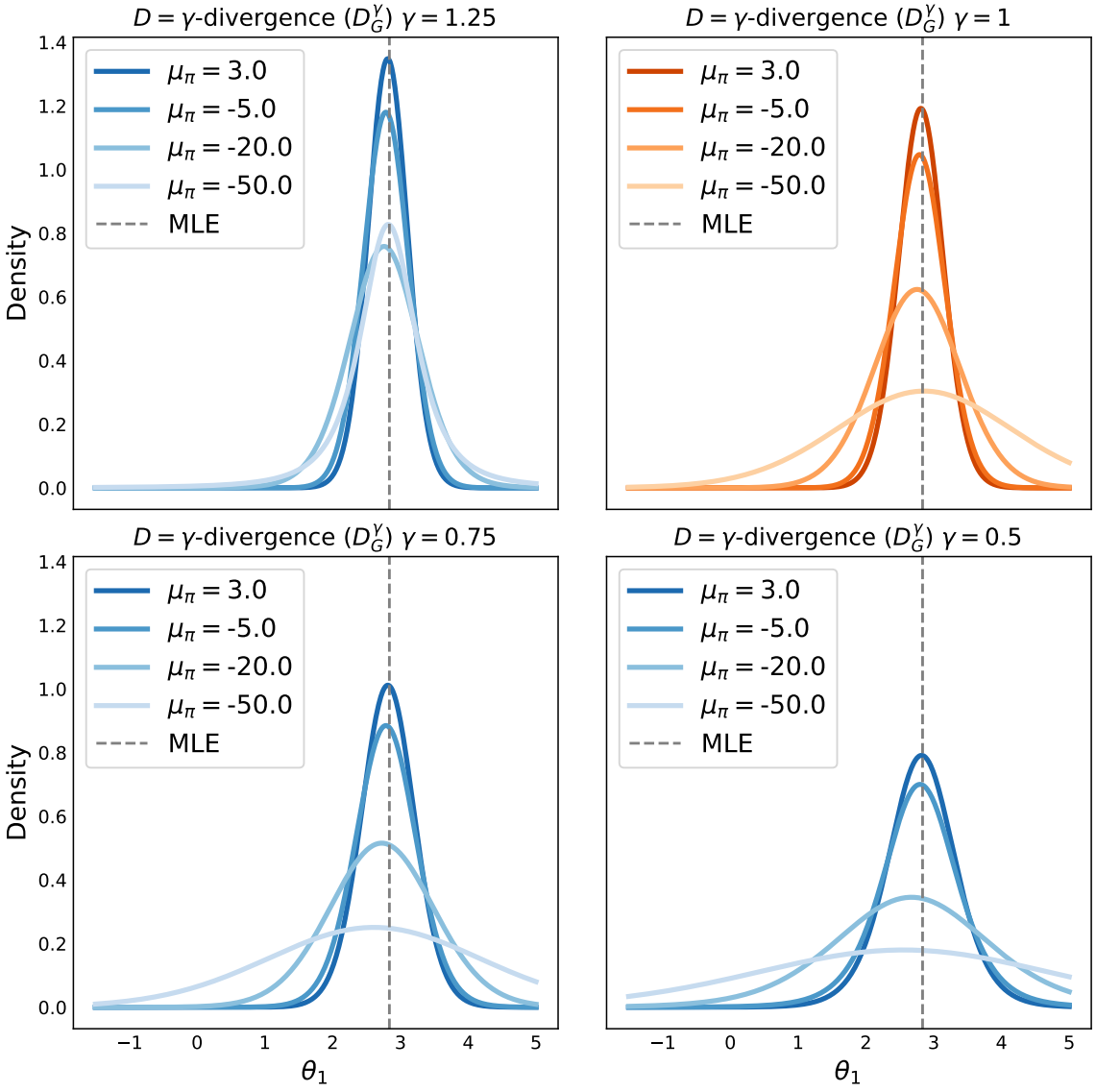


Figure 20: Best viewed in color. Marginal **VI** and **GVI** posterior for the coefficient of a Bayesian linear model under different priors using $D = D_G^{(\gamma)}$ as prior regularizer ($D_G^{(\gamma)}$ recovers KLD as $\gamma \to 1$). The prior specification is given by $\theta_1|\sigma^2 \sim \mathcal{N}(\mu_\pi, \sigma^2)$ with $\sigma^2 \sim \mathcal{IG}(3, 5)$.

> *a tacit hope in ignoring deviations from ideal models was that they would not matter; that statistical procedures which are optimal under the strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians.*

Formalizing this, we arrive at the following definition.

Definition 26 (Robustness) *Let $M_j = P(D_j, \ell_j, \Pi)$ with $\boldsymbol{\theta}_j^* = \arg\min_{\boldsymbol{\theta}} \{\mathbb{E}_{\boldsymbol{X}} [\ell_j(\boldsymbol{\theta}, \boldsymbol{X})]\}$ for $j = 1, 2$. Then, M_1 is more robust for $\boldsymbol{\theta}$ than M_2 relative to the (implicit) assumptions A on the data generating mechanism of $\boldsymbol{X}$ if (i) $\boldsymbol{\theta}_1^*$ is a better result than $\boldsymbol{\theta}_2^*$ if A is untrue and (ii) $\boldsymbol{\theta}_1^* = \boldsymbol{\theta}_2^*$ if A is true.*

Remark 27 *It is hard to say what a better result means, but we note that regardless of its precise meaning, this definition requires that robust inference directly affects $\boldsymbol{\theta}^*$, i.e. that $\boldsymbol{\theta}_1^* \neq \boldsymbol{\theta}_2^*$ unless A is true. While one could substantially strengthen this definition by formalizing what exactly a better result means, this would necessarily be context-dependent, complicate matters substantially and obfuscate the point of robustness.*

Proof First, we prove claim (i) about **robustness** to model misspecification: By Definition 26, robustness implies a change in $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \{\mathbb{E}_{\boldsymbol{X}} [\ell(\boldsymbol{\theta}, \boldsymbol{X})]\}$ if distributional assumptions about $\boldsymbol{X}$ are incorrect. Notice that $\boldsymbol{\theta}^*$ is not affected by D or Π, but is affected by ℓ. Next, we turn to the claims (ii) and (iii) about **uncertainty quantification** and **prior robustness**: First, note that Π and π are not allowed to change by assumption and so cannot affect uncertainty quantification. Next, while ℓ is allowed to change, the parameter of interest it not allowed to change. In other words, ℓ may only be changed in a ways that leave $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*$ unaffected. Notice that changing ℓ to ℓ' will affect $\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} \left\{\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\right\}$ and $\boldsymbol{\theta}^* = \mathbb{E}_{\boldsymbol{X}} [\ell(\boldsymbol{\theta}, \boldsymbol{x})]$ unless $\ell' = C + w \cdot \ell$ for some constants C and $w > 0$. Since $P(\ell, D, \Pi) = P(\ell + C, D, \Pi)$ for any C, we can disregard C and turn to w. Indeed, the uncertainty quantification of $P(\ell, D, \Pi)$ will be different from that of $P(w \cdot \ell, D, \Pi)$ for any constant $w \neq 1$. However, dividing by w in eq. (10) shows that $P(w \cdot \ell, D, \Pi) = P(\ell, \frac{1}{w} D, \Pi)$. Hence, any change in the loss that does not affect $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*$ can be rewritten as a change in D. It follows that changing the uncertainty quantification or making the RoT robust to the prior belief amounts to changing D. ■

Appendix D. Link to the Predictive Information Bottleneck

First, one can rewrite eq. (14) as an unconstrained optimization problem by a well-known argument. For a scalar $\beta = \beta(I_0, \boldsymbol{x}_{1:n})$ derived as in Theorem 1 of Tishby et al. (2000), we have that

$$q^*(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) = \underset{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) \in \Pi_{\text{PIB}}}{\arg\min} \{-I(\boldsymbol{\theta}, \boldsymbol{x}_{n+1:\infty}) + (1-\beta) I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n})\}.$$

But we can do even better: by noting that any distribution on $\boldsymbol{\theta}$ is obtained by compressing (i.e. training on) $\boldsymbol{x}_{1:n}$ only, we also know that $\boldsymbol{\theta}$ and $\boldsymbol{x}_{n+1:\infty}$ are independent once

conditioned on $\boldsymbol{x}_{1:n}$. This means that $p(\boldsymbol{\theta}, \boldsymbol{x}_{n+1:\infty}|\boldsymbol{x}_{1:n}) = p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})p(\boldsymbol{x}_{n+1:\infty}|\boldsymbol{x}_{1:n})$, so that $I(\boldsymbol{\theta}, \boldsymbol{x}_{n+1:\infty}|\boldsymbol{x}_{1:n}) = 0$. By elementary operations (see Alemi, 2019), this implies that we can rewrite

$$I(\boldsymbol{\theta}, \boldsymbol{x}_{n+1:\infty}) = I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}) - I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty}),$$

which we can plug into the unconstrained form to find that

$$q^*(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) = \underset{p(\boldsymbol{\theta}|x_{1:n})\in\Pi_{\text{PIB}}}{\arg\min} \left\{ I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty}) - \beta I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}) \right\}. \tag{27}$$

Though this may not be immediately obvious, eq. (27) has a close relationship with the RoT. To see how this conclusion can be reached, first note that

$$\begin{aligned}
\beta I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}) &= \beta \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})p(\boldsymbol{x}_{1:n})} \left[\log \left(\frac{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) \cancel{p(\boldsymbol{x}_{1:n})}}{p(\boldsymbol{\theta}) \cancel{p(\boldsymbol{x}_{1:n})}} \right) \right] \\
&= \beta \mathbb{E}_{p(\boldsymbol{x}_{1:n})} \left[\text{KLD} \left(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) \| p(\boldsymbol{\theta}) \right) \right]. \\
&=: D_{\text{PIB}}(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) \| \pi_{\text{PIB}}(\boldsymbol{\theta})),
\end{aligned}$$

where we have defined the marginal $\pi_{\text{PIB}}(\boldsymbol{\theta}) = \int_{\mathcal{X}^n} p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})p(\boldsymbol{x}_{1:n})d\boldsymbol{x}_{1:n}$. Clearly, $D_{\text{PIB}}(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) \| \pi_{\text{PIB}}(\boldsymbol{\theta})) \geq 0$ and $D_{\text{PIB}}(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) \| \pi_{\text{PIB}}(\boldsymbol{\theta})) = 0 \iff p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) = \pi_{\text{PIB}}(\boldsymbol{\theta})$. Notice that unlike in the Bayesian paradigm, the prior π_{PIB} here is *not* a free variable. Instead, it gives the distribution over $\boldsymbol{\theta}$ which is obtained over all possible configurations of $\boldsymbol{x}_{1:n}$, which makes this prior conceptually close to a bootstrap distribution.

Similarly, we can rewrite the first term as a loss function by noting that

$$\begin{aligned}
& I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty}) \\
&= \mathbb{E}_{p(\boldsymbol{x}_{n+1:\infty})} \left[\text{KLD} \left(p(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty}) \| p(\boldsymbol{\theta}|\boldsymbol{x}_{n+1:\infty}) p(\boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty}) \right) \right] \\
&= \mathbb{E}_{p(\boldsymbol{x}_{n+1:\infty})} \left[\text{KLD} \left(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) \cancel{p(\boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty})} \| p(\boldsymbol{\theta}|\boldsymbol{x}_{n+1:\infty}) \cancel{p(\boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty})} \right) \right] \\
&= \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})} \Bigg[\underbrace{\mathbb{E}_{p(\boldsymbol{x}_{1:n})} \left[\log \left(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) \right) \right] - \mathbb{E}_{p(\boldsymbol{x}_{n+1:\infty})} \left[\log p(\boldsymbol{\theta}|\boldsymbol{x}_{n+1:\infty}) \right]}_{= L_{n,\text{PIB}}(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}))} \Bigg].
\end{aligned}$$

While this loss is not computable in practice, it has a clear interpretation. Specifically, it jointly minimizes (i) the information that $\boldsymbol{\theta}$ loses on future data $\boldsymbol{x}_{n+1:\infty}$ and (ii) the difference between the information that $\boldsymbol{\theta}$ loses on $\boldsymbol{x}_{1:n}$ versus $\boldsymbol{x}_{n+1:\infty}$. The loss $L_{n,\text{PIB}}$ has two properties that set it apart from the losses we have considered thus far: first of, $L_{n,\text{PIB}}$ does not depend on a sample $x_{1:n}$ (but the distributions of the underlying random variables $\boldsymbol{x}_{1:n}, \boldsymbol{x}_{n+1:\infty}$). Second, $L_{n,\text{PIB}}$ is not summable. Neither of these properties affect the axiomatic development in Section 4.1, since any empty sample is a finite sample and because summability was imposed for presentational purposes only.

Putting everything together, we can rewrite the PIB as

$$q^*(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) = \underset{q\in\Pi_{\text{PIB}}}{\arg\min} \left\{ \mathbb{E}_q \left[L_{n,\text{PIB}}(q) \right] + D_{\text{PIB}}(q \| \pi_{\text{PIB}}) \right\}.$$

Appendix E. Experimental Details for Figure 4

For Figure 4, n observations are generated from the d-dimensional Bayesian Mixture Model with two equally likely normal mixture components $z = 1, 2$ with dimension-wise unit variance and mean given by

$$\boldsymbol{\mu}^z = (\mu_1^z, \mu_2^z, \dots \mu_d^z)^T = \begin{cases} 2 \cdot e_d & \text{if } z = 1 \\ -2 \cdot e_d & \text{if } z = 2 \end{cases},$$

where $e_d = (1, 1, \dots 1)^T$ is the d-dimensional column vector of ones. The n observations x_i^o are drawn with equal probability from two mixture, meaning that

$$\begin{aligned} \boldsymbol{z}_i &\overset{i.i.d.}{\sim} \text{Bernoulli}(0.5) \\ x_i^o | \{\boldsymbol{z}_i = z_i\} &\overset{i.i.d.}{\sim} \mathcal{N}(x_i^o | \boldsymbol{\mu}^{z_i}, \boldsymbol{I}_d). \end{aligned} \tag{28}$$

Notice in particular that this generates n latent variables $\boldsymbol{z}_{1:n}$ that indicate mixture memberships for $x_{1:n}^o$, but are unobserved. With this, inference is conducted on $\boldsymbol{\mu}^c$ for $c = 1, 2$ via the negative log likelihood loss of the correct model. For $\boldsymbol{\theta} = (\boldsymbol{\mu}^1, \boldsymbol{\mu}^2)$, this is given by

$$\ell(\boldsymbol{\theta}, x_i^o, \boldsymbol{z}_i) = -\log p_{\mathcal{N}}(x_i^o | \boldsymbol{\mu}^{\boldsymbol{z}_i}, \boldsymbol{I}_d).$$

The benefits of alternative choices of D are explored for the fixed number of observations $n = 50$. To this end, $B = 100$ artificial data sets are generated according to the above description.

If the prior is poorly specified, $D = \text{KLD}$ will produce posterior beliefs that place the same weight on the prior as they do on the data. In contrast, robust alternatives to the KLD do not suffer this problem: They can produce posterior beliefs that take the prior into account, but are robust to prior misspecification, see also Knoblauch et al. (2019). To illustrate the phenomenon empirically, the next experiment compares the KLD with Rényi's α-divergence ($D_{AR}^{(\alpha)}$) for $\alpha = 0.5$ under two settings: A well-specified prior $\pi_1(\boldsymbol{\theta})$ and a misspecified prior $\pi_2(\boldsymbol{\theta})$, which are given as normal distributions

$$\begin{aligned} \pi_1(\boldsymbol{\theta}) &= \mathcal{N}\left(\boldsymbol{\theta}; 0_d, \sqrt{10}\boldsymbol{I}_d\right) \\ \pi_2(\boldsymbol{\theta}) &= \mathcal{N}\left(\boldsymbol{\theta}; -10 \cdot e_d, \sqrt{0.1}\boldsymbol{I}_d\right) \end{aligned}$$

To evaluate the experiments, 100 data sets are generated with $n = 50$ observations each. Across these, Figure 4 reports the average posterior computed as

$$\mathcal{N}\left(\bar{m}, \bar{s}\right), \quad \bar{m} = \frac{1}{100} \sum_{j=1}^{2d} \sum_{b=1}^{B} m_{b,j}, \quad \bar{s} = \frac{1}{100} \sum_{j=1}^{2d} \sum_{b=1}^{B} s_{b,j}.$$

Here, $s_{b,j}$ corresponds to the standard deviation computed for the j-th dimension of the mean field normal posterior on the b-th artificial data sets. Similarly, $m_{b,j}$ corresponds to the mean of the same parameter posterior, albeit re-centered around the true value of the inferred parameter.

As Figure 4 shows, $D_{AR}^{(\alpha)}$ is an interesting alternative to the KLD in finite samples: If the prior is misspecified (top row), the KLD produces belief distributions that take the prior too strongly into account and are far from the truth. In contrast, the $D_{AR}^{(\alpha)}$ provides both prior robustness as well as better uncertainty quantification under misspecification. At the same time, $D_{AR}^{(\alpha)}$ has no tangible disadvantage relative to the KLD if the prior is well-specified (bottom row).

Appendix F. Proof of Theorem 14 and additional lower bounds

This section of the Appendix provides proofs for the lower bound interpretation of certain GVI objectives. First, we prove the result stated in the main paper. Next, we show equivalent results for the case of the β-divergence ($D_B^{(\beta)}$) and γ-divergence ($D_G^{(\gamma)}$). While the following results and corresponding proofs are somewhat tedious to read, they are conceptually simple: In fact, all that is needed to derive the results is some basic algebra, Jensen's inequality and a further inequality involving the logarithm, see Lemma 28.

F.1 Proof for $D_{AR}^{(\alpha)}$ (Theorem 14)

Firstly, we recall explicit forms for the function quoted in Theorem 14

$$S_1(\alpha, q, \pi) = \begin{cases} D_{AR}^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}) - \mathrm{KLD}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) & \text{if } 0 < \alpha < 1 \\ 0 & \text{if } \alpha > 1. \end{cases} \tag{29}$$

Next we provide a proof of the Theorem

Proof For this proof we have to consider two cases for α as the positivity and negativity of $\alpha - 1$ affect the results that can be used.

Case 1) $\alpha > 1$: Jensen's inequality and the concavity of the natural logarithm give us that

$$\begin{aligned} &\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right] + \frac{1}{\alpha(\alpha-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})}\left[\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)^{\alpha-1}\right] \\ &\geq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right] + \frac{1}{\alpha}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)\right] \\ &= \frac{1}{\alpha}\mathrm{KLD}(q(\boldsymbol{\theta})||\pi^{\alpha\ell}(\boldsymbol{\theta}|\mathbf{x})) - \frac{1}{\alpha}\log \int \pi(\boldsymbol{\theta}) \exp\left(-\alpha \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right) d\boldsymbol{\theta}. \end{aligned}$$

Case 2) $0 < \alpha < 1$: Here the negativity of $\frac{1}{\alpha(\alpha-1)}$ means we cannot apply Jensen's inequality in the above way. Instead, we can write

$$\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i)\right] + D_{AR}^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) \\
&= \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\left(\frac{\pi(\boldsymbol{\theta})\exp(-\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i))}{\int\pi(\boldsymbol{\theta})\exp(-\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i))d\boldsymbol{\theta}}\right)\right] \\
&\quad - \log\int\pi(\boldsymbol{\theta})\exp(-\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i))d\boldsymbol{\theta} + D_{AR}^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) \\
&= \mathrm{KLD}(q(\boldsymbol{\theta})||\pi^{\ell}(\boldsymbol{\theta}|x)) - \log\int\pi(\boldsymbol{\theta})\exp(-\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i))d\boldsymbol{\theta} \\
&\quad + D_{AR}^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) - \mathrm{KLD}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})).
\end{aligned}$$

Combined these two cases provides the term in Eq. (17) and (29) ■

Next we state, prove and interpret equivalent results for the $D_B^{(\beta)}$ and $D_G^{(\gamma)}$ prior regularisers. But before we do so we need the following lemma

Lemma 28 (A Taylor series bound for the natural logarithm) *The natural logarithm of a positive real number Z can be bounded as follows*

$$\begin{cases}\log(Z) \leq \frac{Z^x-1}{x} & \text{if } x > 0 \\ \log(Z) \geq \frac{Z^x-1}{x} & \text{if } x < 0.\end{cases}$$

Proof Using the series expansion of $\exp(x)$ and the Lagrange form of the remainder we see that

$$\begin{aligned}
\frac{Z^x-1}{x} &= \frac{\exp\left(x\log Z\right)-1}{x} = \frac{\left(x\log Z\right) + \frac{1}{2!}\left(x\log Z\right)^2 + \frac{1}{3!}\left(x\log Z\right)^3 + \ldots}{x} \\
&= \frac{\left(x\log Z\right) + \frac{1}{2}\exp(c)\left(x\log Z\right)^2}{x} = \log Z + \frac{\frac{1}{2!}\exp(c)\left(x\log Z\right)^2}{x}
\end{aligned}$$

where $c \in [0, x\log(Z)]$. Now the numerator of the remainder term $\frac{\frac{1}{2!}\exp(c)(x\log Z)^2}{x}$ is always positive and therefore the sign of x determines whether this remainder term forms an upper or lower bound for $\log(Z)$. ■

F.2 The $D_B^{(\beta)}$ prior regulariser

Theorem 29 (GVI as approximate Evidence Lower bound with $D = D_B^{(\beta)}$) *The objective of a* GVI *posterior based on $P(\ell, D_B^{(\beta)}, \mathcal{Q})$ has an interpretation as lower bound on the $c(\beta)$-scaled (generalized) evidence lower bound of $P(w(\beta)\cdot\ell, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$:*

$$\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i)\right] + D_B^{(\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) \geq -c(\beta)\mathrm{ELBO}^{w(\beta)\ell}(q) + S_1(\beta,q,\pi) \qquad (30)$$

where $\text{ELBO}^{w(\beta)\ell}$ *denotes the Evidence Lower Bound associated with standard* VI *relative to the generalized Bayesian posterior given by*

$$q_B^{w(\beta)\ell}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp\left(-w(\beta)\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right),$$

where $c(\beta) = \min\{1, \beta^{-1}\}$, $w(\beta) = \max\{1, \beta\}$ *and where* $S_1(\beta, q, \pi)$ *is a closed form slack term with*

$$S_1(\beta, q, \pi) = \begin{cases} \frac{1}{\beta(\beta-1)}\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\beta-1}\right] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log q(\boldsymbol{\theta})\right] - \frac{1}{\beta-1} & \text{if } 0 < \beta < 1 \\ \frac{1}{\beta}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log \pi(\boldsymbol{\theta})\right] - \frac{1}{\beta-1}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right] - \frac{1}{\beta(\beta-1)} & \text{if } \beta > 1. \end{cases} \tag{31}$$

Proof Firstly we note that the objective function associated with the RoT $P(D_B^{(\beta)}, \ell_n, Q)$ can be simplified by removing the terms in the $D_B^{(\beta)}$ that don't depend on $q(\boldsymbol{\theta})$

$$\begin{aligned} &\arg\min_{q\in\mathcal{Q}} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right] + D_B^{(\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) \right\} \\ =&\arg\min_{q\in\mathcal{Q}} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right] + \frac{1}{\beta(\beta-1)}\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\beta-1}\right] - \frac{1}{(\beta-1)}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right] \right\}. \end{aligned}$$

We have to consider two cases for β as the positivity and negativity of $\beta - 1$ affect which part of Lemma 28 we use.

Case 1) $0 < \beta < 1$**:** Lemma 28 gives us that for $\beta - 1 < 0$, $\frac{Z^{\beta-1}}{\beta-1} \leq \log(Z) + \frac{1}{\beta-1}$ therefore

$$\begin{aligned} &= \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right] + \frac{1}{\beta(\beta-1)}\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\beta-1}\right] - \frac{1}{(\beta-1)}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right] \\ &\geq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right] + \frac{1}{\beta(\beta-1)}\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\beta-1}\right] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right] - \frac{1}{\beta-1} \\ &= \text{KLD}(q(\boldsymbol{\theta})||\pi^{\ell}(\boldsymbol{\theta}|x)) - \log \int \exp(-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i))\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &\quad + \frac{1}{\beta(\beta-1)}\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\beta-1}\right] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(q(\boldsymbol{\theta}))\right] - \frac{1}{\beta-1}. \end{aligned}$$

Case 2) $\beta > 1$**:** Lemma 28 gives us that for $\beta - 1 > 0$, $\frac{Z^{\beta-1}}{\beta-1} \geq \log(Z) + \frac{1}{\beta-1}$ therefore

$$\begin{aligned} &= \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right] + \frac{1}{\beta(\beta-1)}\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\beta-1}\right] - \frac{1}{(\beta-1)}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right] \\ &\geq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right] + \frac{1}{\beta}\left(\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\left(q(\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)\right] + \frac{1}{\beta-1}\right) - \frac{1}{(\beta-1)}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right] \\ &= \frac{1}{\beta}\text{KLD}(q(\boldsymbol{\theta})||\pi^{\beta\ell}(\boldsymbol{\theta}|\mathbf{x})) - \frac{1}{\beta}\log \int \pi(\boldsymbol{\theta}) \exp(-\beta\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i))d\boldsymbol{\theta} \\ &+ \frac{1}{\beta}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right] - \frac{1}{(\beta-1)}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right] + \frac{1}{\beta(\beta-1)}. \end{aligned}$$

Combined these two cases provides the term in Eq. (30) and (31) ■

F.3 The $D_G^{(\gamma)}$ prior regulariser

Theorem 30 (GVI as approximate Evidence Lower bound with $D = D_G^{(\gamma)}$) *The objective of a* GVI *posterior based on $P(\ell, D_G^{(\gamma)}, \mathcal{Q})$ has an interpretation as lower bound on the $c(\gamma)$-scaled (generalized) evidence lower bound of $P(w(\gamma)\cdot\ell, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$:*

$$\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i)\right] + D_G^{(\gamma)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = -c(\gamma)\mathrm{ELBO}^{w(\gamma)\ell}(q) + S(\gamma,q,\pi) \tag{32}$$

where $\mathrm{ELBO}^{w(\gamma)\ell}$ *denotes the Evidence Lower Bound associated with standard* VI *relative to the generalized Bayesian posterior given by*

$$q_B^{w(\gamma)\ell}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta})\exp\left(-w(\gamma)\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i)\right),$$

where $c(\gamma) = \min\{1,\gamma^{-1}\}$, $w(\gamma) = \max\{1,\gamma\}$ and where $S_1(\gamma,q,\pi)$ is a closed form slack term with

$$S_1(\gamma,q,\pi) = \begin{cases} \frac{1}{\gamma(\gamma-1)}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\gamma-1}\right] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log q(\boldsymbol{\theta})\right] & \text{if } 0<\gamma<1 \\ \frac{1}{\gamma}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\pi(\boldsymbol{\theta})\right] - \frac{1}{\gamma-1}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\gamma-1}\right] & \text{if } \gamma>1. \end{cases} \tag{33}$$

Proof Firstly we note that the objective function associated with $P(D_G^{(\gamma)}, \ell_n, Q)$ can be simplified by removing the terms in the $D_G^{(\gamma)}$ that don't depend on $q(\boldsymbol{\theta})$

$$\arg\min_{q\in\mathcal{Q}}\left\{\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i)\right] + D_G^{(\gamma)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))\right\} =$$

$$\arg\min_{q\in\mathcal{Q}}\left\{\mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i)\right] + \frac{1}{\gamma(\gamma-1)}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\gamma-1}\right] - \frac{1}{(\gamma-1)}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\gamma-1}\right]\right\}.$$

We have to consider two cases for γ as the positivity and negativity of $\gamma-1$ affect the results we can use.

Case 1) $0<\gamma<1$: Jensen's inequality and the concavity of the natural logarithm applied to $\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\gamma-1}\right]$ provides

$$\begin{aligned}
&= \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i)\right] + \frac{1}{\gamma(\gamma-1)}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\gamma-1}\right] - \frac{1}{(\gamma-1)}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\gamma-1}\right] \\
&\geq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i)\right] + \frac{1}{\gamma(\gamma-1)}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\gamma-1}\right] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\pi(\boldsymbol{\theta})\right] \\
&= \mathrm{KLD}(q(\boldsymbol{\theta})||\pi^{\ell}(\boldsymbol{\theta}|x)) - \log\int\pi(\boldsymbol{\theta})\exp\left(-\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i)\right)d\boldsymbol{\theta} \\
&\quad + \frac{1}{\gamma(\gamma-1)}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\gamma-1}\right] + -\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log q(\boldsymbol{\theta})\right].
\end{aligned}$$

Case 2) $\gamma > 1$: Jensen's inequality and the concavity of the natural logarithm applied to $\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\gamma-1}\frac{\pi(\boldsymbol{\theta})^{\gamma-1}}{\pi(\boldsymbol{\theta})^{\gamma-1}}\right]$ provides

$$
\begin{aligned}
&= \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i)\right] + \frac{1}{\gamma(\gamma-1)}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\gamma-1}\right] - \frac{1}{(\gamma-1)}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\gamma-1}\right] \\
&\geq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i)\right] + \frac{1}{\gamma}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\frac{q(\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right] - \frac{1}{(\gamma-1)}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\gamma-1}\right] \\
&= \frac{1}{\gamma}\mathrm{KLD}(q(\boldsymbol{\theta})||\pi^{\gamma\ell}(\boldsymbol{\theta}|x)) - \frac{1}{\gamma}\int\pi(\boldsymbol{\theta})\exp\left(-\gamma\sum_{i=1}^{n}\ell(\boldsymbol{\theta},\boldsymbol{x}_i)\right)d\boldsymbol{\theta} \\
&\quad + \frac{1}{\gamma}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\pi(\boldsymbol{\theta})\right] - \frac{1}{(\gamma-1)}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\gamma-1}\right].
\end{aligned}
$$

Combined these two cases provides the term in Eq. (32) and (33) ■

F.3.1 Interpretation

Theorems 29 and 30 provide a lower bound on an objective function that is to be minimised so that interpreting this lower bound provides some insight into the behaviour of the GVI posterior. First, we investigate the case where the hyperparameters β and γ are in $(0, 1)$. As expected, the form of the GVI objective leads us to conclude that the variance will be larger than that for standard VI within this range of values. Next, we investigate the case where the hyperparameters β and γ are > 1. Again unsurprisingly, this leads to a shrinkage of the posterior variance relative to standard VI.

Case 1: $0 < \beta = \gamma < 1$**.** For $0 < \beta = \gamma < 1$ the terms $c(\beta) = c(\gamma)$ and $w(\beta) = w(\gamma)$ produce an objective equivalent to standard VI. This suggests that GVI continues to minimise the KLD between the variational and standard Bayesian posterior. Unlike standard VI however, GVI with $D = D_B^{(\beta)}$ or $D = D_G^{(\gamma)}$ additionally minimises the slack terms $S_1(\beta, q, \pi)$ or $S_1(\gamma, q, \pi)$. It is easy to show that the these adjustment terms encourage the solution to $P(D_B^{(\beta)}, \ell, Q)$ with $0 < \beta < 1$ and $P(D_G^{(\gamma)}, \ell, Q)$ with $0 < \gamma < 1$ to have greater variance than the standard VI posterior given by $P(\mathrm{KLD}, \ell_n, Q)$. For the $D_B^{(\beta)}$, we can see this by rewriting

$$S^{(0,1)}(\beta, q, \pi) = -\frac{1}{\beta}h_T^{(\beta)}(q(\boldsymbol{\theta})) + h_{\mathrm{KLD}}(q(\boldsymbol{\theta})) + \frac{1-\beta}{\beta}.$$

Here, $h_{\mathrm{KLD}}(q(\boldsymbol{\theta}))$ is the Shannon entropy of $q(\boldsymbol{\theta})$ and $h_T^{(\beta)}(q(\boldsymbol{\theta}))$ is the Tsallis entropy of $q(\boldsymbol{\theta})$ with parameter β. Again applying Lemma 28, we find that for $0 < \beta < 1$, $h_T^{(\beta)}(q(\boldsymbol{\theta})) > h_{\mathrm{KLD}}(q(\boldsymbol{\theta}))$. It immediately follows that minimising $-\frac{1}{\beta}h_T^{(\beta)}(q(\boldsymbol{\theta})) + h_{\mathrm{KLD}}(q(\boldsymbol{\theta}))$ for $0 < \beta < 1$ will make $h_T^{(\beta)}(q(\boldsymbol{\theta}))$ large—an effect that is achieved by increasing the variance of $q(\boldsymbol{\theta})$.

Applying the same type of logic to the $D_G^{(\gamma)}$, one can rewrite

$$S^{(0,1)}(\gamma, q, \pi) = -\frac{1}{\gamma}h_R^{(\gamma)}(q(\boldsymbol{\theta})) + h_{\mathrm{KLD}}(q(\boldsymbol{\theta})).$$

As before, $h_{\mathrm{KLD}}(q(\boldsymbol{\theta}))$ is the Shannon entropy of $q(\boldsymbol{\theta})$, but unlike before $h_R^{(\gamma)}(q(\boldsymbol{\theta}))$ is now the Rényi entropy of $q(\boldsymbol{\theta})$ with parameter γ. With this, one can extend Theorem 3 in Van Erven and Harremos (2014) to show that $h_R^{(\gamma)}(q(\boldsymbol{\theta}))$ is decreasing in γ. Since it is also well-known that $\lim_{\gamma\to 1} h_R^{(\gamma)}(q(\boldsymbol{\theta})) = h_{\mathrm{KLD}}(q(\boldsymbol{\theta}))$, it follows that minimising $-\frac{1}{\gamma}h_R^{(\gamma)}(q(\boldsymbol{\theta})) + h_{\mathrm{KLD}}(q(\boldsymbol{\theta}))$ for $0 < \gamma < 1$ will make $h_R^{(\gamma)}(q(\boldsymbol{\theta}))$ large—an effect that is again achieved by increasing the variance of $q(\boldsymbol{\theta})$.

Case 2: $\beta = \gamma = k > 1$**.** For $k = \gamma = \beta > 1$, $c(k) = \frac{1}{k}$ and $w(k) = k$. Minimising $\mathrm{KLD}(q||q_k^*)$ for $k > 1$ will encourage $P(D_B^{(\beta)}, \ell, Q)$ or $P(D_G^{(\gamma)}, \ell, Q)$ to be more concentrated around the empirical risk minimizer $\hat{\boldsymbol{\theta}}_n$ of ℓ than the standard VI posterior given by $P(\mathrm{KLD}, \ell, Q)$. Additionally, one can show that minimising the adjustment term also favours shrinking the variance of $q(\boldsymbol{\theta})$. To see this for the case of $D_B^{(\beta)}$, rewrite

$$S^{(1,\infty)}(\beta, q, \pi) = \frac{1}{\beta}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right] - \frac{1}{\beta - 1}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1} - 1\right] - \frac{1}{\beta}. \tag{34}$$

Applying Lemma 28 then shows that for $\beta > 1$,

$$\frac{1}{\beta - 1}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1} - 1\right] \geq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right] \geq \frac{1}{\beta}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right].$$

From this, it follows that minimising Eq. (34) will make $\frac{1}{\beta-1}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right]$ large. Fixing $\pi(\boldsymbol{\theta})$, maximising $\frac{1}{\beta-1}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right]$ plus $\frac{1}{\beta}\times$ the Tsallis entropy of $q(\boldsymbol{\theta})$ is equivalent to minimising $D_B^{(\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))$. Because $D_B^{(\beta)}$ is a divergence, this maximization would naturally seek to choose $q(\boldsymbol{\theta})$ close to $\pi(\boldsymbol{\theta})$. The Tsallis entropy term in this formulation would have acted to increase the variance of $q(\boldsymbol{\theta})$. But since we maximize only $\frac{1}{\beta-1}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right]$—i.e. without adding the Tsallis entropy of $q(\boldsymbol{\theta})$—choices of $\beta > 1$ will lead to shrinking the variance of $q(\boldsymbol{\theta})$ relative to standard VI.

For the $D_G^{(\gamma)}$, Jensen's inequality shows that for $\gamma > 1$,

$$\frac{1}{\gamma - 1}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\gamma-1}\right] \geq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right] \geq \frac{1}{\gamma}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right].$$

As a result, minimising $S^{(1,\infty)}(\gamma, q, \pi)$ will seek to make $\frac{1}{\gamma-1}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\gamma-1}\right]$ large. Fixing again $\pi(\boldsymbol{\theta})$, maximising $\frac{1}{\gamma-1}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right]$ plus $\frac{1}{\gamma}\times$ the Rényi entropy of $q(\boldsymbol{\theta})$ is equivalent to minimising $D_G^{(\gamma)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))$, and thus seeks $q(\boldsymbol{\theta})$ close to $\pi(\boldsymbol{\theta})$. The Rényi entropy term would have acted to increase the variance of $q(\boldsymbol{\theta})$. Therefore and similarly to the case of $D_B^{(\beta)}$, maximising $\frac{1}{\gamma-1}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\gamma-1}\right]$ without adding the Rényi entropy will lead to shrinkage of the variance of $q(\boldsymbol{\theta})$.

Appendix G. Proof of Proposition 16

Proof Proposition 16 considers the following forms of the prior and likelihood

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0) &= h(\boldsymbol{\theta})\exp\left\{\eta(\boldsymbol{\kappa}_0)^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}_0))\right\} \\
q(\boldsymbol{\theta}|\boldsymbol{\kappa}) &= h(\boldsymbol{\theta})\exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}))\right\} \\
p(\boldsymbol{x}|\boldsymbol{\theta}) &= h(\boldsymbol{\theta})\exp(g(\boldsymbol{x})^T T(\boldsymbol{\theta}) - B(\boldsymbol{x})),
\end{aligned}$$

where $A(\eta(\boldsymbol{\kappa})) = \log \int h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta})\right\} d\boldsymbol{\theta}$ and $h(\boldsymbol{\theta}) = \frac{1}{\int \exp(g(\boldsymbol{x})^T T(\boldsymbol{\theta}) - B(\boldsymbol{x})) d\boldsymbol{x}}$.

The GVI objective function in this scenario, which we term an ELBO as we use the KLD prior regulariser is

$$\begin{aligned}\text{ELBO}(\boldsymbol{\kappa}) &= \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[\sum_{i=1}^{n} \ell_G^{(\gamma)}(\boldsymbol{\theta}, \boldsymbol{x}_i)\right] + \text{KLD}(q(\boldsymbol{\theta}|\boldsymbol{\kappa})||q(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)) \\ &= \sum_{i=1}^{n} \underbrace{\int \underbrace{\ell_G^{(\gamma)}(\boldsymbol{\theta}, \boldsymbol{x}_i)}_{C_1(\boldsymbol{\kappa}, \boldsymbol{\theta}, x_i)} q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\theta}_{C_2(\boldsymbol{\kappa}, \boldsymbol{x}_i)} + \underbrace{\text{KLD}(q(\boldsymbol{\theta}|\boldsymbol{\kappa})||\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0))}_{C_3(\boldsymbol{\kappa}, \boldsymbol{\kappa}_0)}.\end{aligned}$$

We have decomposed this into three terms that we need to check are closed forms of $\boldsymbol{\kappa}$. Firstly

$$C_1(\boldsymbol{\kappa}, \boldsymbol{\theta}, \boldsymbol{x}_i) = \ell_G^{(\gamma)}(\boldsymbol{x}_i, \boldsymbol{\theta}) = -\frac{1}{\gamma - 1} p(\boldsymbol{x}_i; \boldsymbol{\theta})^{\gamma - 1} \frac{\gamma}{\left[\int p(\boldsymbol{z}; \boldsymbol{\theta})^\gamma d\boldsymbol{z}\right]^{\frac{\gamma-1}{\gamma}}},$$

and in order for this to be a closed form function of $\boldsymbol{\kappa}$, $\boldsymbol{\theta}$, and $\boldsymbol{x}_i$ requires that

$$I^{(\gamma)}(\boldsymbol{\theta}) = \int p(\boldsymbol{z}|\boldsymbol{\theta})^\gamma dz = \int h(\boldsymbol{\theta})^\gamma \exp(\gamma g(\boldsymbol{z})^T T(\boldsymbol{\theta}) - \gamma B(\boldsymbol{z})) dz,$$

where the theorem statement ensures that $I^{(\gamma)}(\boldsymbol{\theta})$ is a closed form function of $\boldsymbol{\theta}$. Next

$$\begin{aligned}&C_2(\boldsymbol{\kappa}, \boldsymbol{x}_i) \\ =& -\frac{\gamma}{\gamma - 1} \int h(\boldsymbol{\theta})^{\gamma - 1} \exp((\gamma - 1) g(\boldsymbol{x}_i)^T T(\boldsymbol{\theta}) - (\gamma - 1) B(\boldsymbol{x}_i)) \frac{1}{\left[h(\boldsymbol{\theta})^\gamma I^{(\gamma)}(\boldsymbol{\theta})\right]^{\frac{\gamma-1}{\gamma}}} q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\boldsymbol{\theta} \\ =& -\frac{\gamma}{\gamma - 1} \frac{\exp\left((1 - \gamma) B(\boldsymbol{x}_i) + A\left(\eta(\boldsymbol{\kappa}) + (\gamma - 1) g(\boldsymbol{x}_i)\right)\right)}{\exp\left(A(\eta(\boldsymbol{\kappa}))\right)} \mathbb{E}_{q(\boldsymbol{\theta}|(\eta(\boldsymbol{\kappa}) + (\gamma - 1) g(\boldsymbol{x}_i)))}\left[I^{(\gamma)}(\boldsymbol{\theta})^{\frac{1-\gamma}{\gamma}}\right],\end{aligned}$$

where the theorem statement ensures that $(\eta(\boldsymbol{\kappa}_n) + (\gamma - 1) g(\boldsymbol{x}_i)) \in \mathcal{N}$ for all x_i and that $F_2(\boldsymbol{\kappa}^*) = \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)}\left[I^{(\gamma)}(\boldsymbol{\theta})^{\frac{1-\gamma}{\gamma}}\right]$ is closed form function of $\boldsymbol{\kappa}^*$ for all $\boldsymbol{\kappa}^* \in \mathcal{N}$. Lastly

$$\begin{aligned}C_3(\boldsymbol{\kappa}, \kappa_0) =& \int h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}))\right\} \log \frac{h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}))\right\}}{h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa}_0)^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}_0))\right\}} d\boldsymbol{\theta} \\ =& A(\eta(\boldsymbol{\kappa}_0)) - A(\eta(\boldsymbol{\kappa})) + (\eta(\boldsymbol{\kappa}) - \eta(\boldsymbol{\kappa}_0))^T \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[T(\boldsymbol{\theta})\right],\end{aligned}$$

where the theorem statement ensures that $F_1(\kappa^*) = \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)}\left[T(\boldsymbol{\theta})\right]$ is a closed form function of $\boldsymbol{\kappa}^*$ for all $\kappa^* \in \mathcal{N}$. ■

Appendix H. Black Box GVI (BBGVI)

The following sections first recall the (implicit and explicit) assumptions one typically makes for black box VI. They are then compared to assumptions that are reasonable for black box GVI (BBGVI). The corresponding methods, their special cases and the relevant black box

variance reduction techniques are then derived and elaborated upon. While there are many black box VI strategies, we center attention on the framework provided for by Ranganath et al. (2014). Throughout, we denote $q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\kappa})$ as a posterior distribution in a set of variational families $\mathcal{Q}$ and parameterized by some parameter $\boldsymbol{\kappa} \in \boldsymbol{K}$.

H.1 Preliminaries and assumptions

The variance reduction techniques of Ranganath et al. (2014) crucially rely on three implicit assumptions that are reasonable for many applications of standard VI.

(A1) Structured mean-field variational inference is used, which means that we can factorize the variational family as $\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) = \prod_{j=1}^{k} q_j(\boldsymbol{\theta}_j|\boldsymbol{\kappa}_j) : \boldsymbol{\kappa}_j \in K_j \text{ for all } j\}$.

(A2) For all factors $\boldsymbol{\theta}_j$, we have a Markov blanket $\boldsymbol{\theta}_{(j)}$ for which we can additively decompose $\ell(\boldsymbol{\theta}, x_i) = \ell^{(j)}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{(j)}, x_i) + \ell^{(-j)}(\boldsymbol{\theta}_{-j}, x_i)$. Here, $\ell^{(j)}$ is an additive component of the loss ℓ that only depends on the j-th factor and its Markov blanket, while $\ell^{(-j)}$ is an additive component of the loss that may depend on all of $\boldsymbol{\theta}$ except for its j-th factor. Note that such additivity holds for standard VI for which the likelihood and the prior are such that the components $\boldsymbol{\theta}_j$ are conditionally independent. In this case, the conditioning set is the Markov blanket.

(A3) $D = \frac{1}{w} \cdot \mathrm{KLD}$ (with $w = 1$ for standard VI).

Note that (A1) is always satisfied for both standard VI and GVI, because any variational family factorizes into at least a single factor. In contrast, note that (A2) does not even necessarily hold for standard VI unless one imposes some conditional independence structure on the $\boldsymbol{\theta}_j$. For GVI, both (A2) and (A3) do not necessarily hold. If they do however, they can greatly simplify BBGVI or improve its numerical performance. In the remainder of this section, we discuss different constellations of assumptions and their consequences for BBGVI.

H.2 Standard black box VI with (A2) and (A3)

If the regularizer used is still a rescaled version of the KLD, one recovers an internally rescaled version of the objective in (Ranganath et al., 2014). Namely, the gradient is given by

$$\mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\Big[\nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}|\boldsymbol{\kappa}))\Big(- \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) - w \log \pi(\boldsymbol{\theta}) - w \log(q(\boldsymbol{\theta}|\boldsymbol{\kappa}))\Big)\Big].$$

and can be approximated in a smart way by sampling from $q(\boldsymbol{\theta}|\boldsymbol{\kappa})$, see for instance Ranganath et al. (2014) for details and the viable strategies for variance reduction. Next, we turn attention to the cases that are more interesting: If (A3) does not hold (so that $D \neq \mathrm{KLD}$) and when the losses are not necessarily negative log likelihoods, meaning that (A2) requires more careful consideration.

H.3 BBGVI under (A2)

If the losses are decomposable along the factors, two cases need to be distinguished:

(D1) $\nabla_{\boldsymbol{\kappa}} D(q\|\pi)$ has closed form for all $q \in \mathcal{Q}$;

(D2) $D(q\|\pi) = \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[\ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})\right]$ for some function $\ell^D_{\boldsymbol{\kappa},\pi} : \boldsymbol{\Theta} \to \mathbb{R}$.

Under each condition, we find a different solution using as much of the available information as possible to improve inference outcomes. For simplicity, we first explain how the derivation works without using the additional information that (A2). In a second step, we shall see how this additional information can be used for variance reductions in the Rao-Blackwellization spirit also used by Ranganath et al. (2014).

H.3.1 Gradients if (D1) holds, not using (A2)

In this case, we can obtain the objective given in the main paper. Define $L(q)$ to be the GVI objective function of $q(\boldsymbol{\theta}|\boldsymbol{\kappa})$. It holds that

$$\begin{aligned}
\nabla_{\boldsymbol{\kappa}} L(q) &= \nabla_{\boldsymbol{\kappa}} \left[\int_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\boldsymbol{\theta} + D(q||\pi) \right] \\
&= \int_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}, \boldsymbol{x}) \nabla_{\boldsymbol{\kappa}} q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\boldsymbol{\theta} + \nabla_{\boldsymbol{\kappa}} D(q||\pi) \\
&= \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})} \left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}|\boldsymbol{\kappa})) \right] + \nabla_{\boldsymbol{\kappa}} D(q||\pi).
\end{aligned}$$

Correspondingly, the gradient can then be estimated without bias and computing the corresponding sample average $\frac{1}{S}\sum_{s=1}^{S} G(\boldsymbol{\theta}^{(s)})$, where the individual terms are given by

$$G(\boldsymbol{\theta}^{(s)}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}^{(s)}, x_i) \nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}^{(s)})) + \nabla_{\boldsymbol{\kappa}} D(q||\pi)$$

H.3.2 Gradients if (D2) holds, not using (A2)

If the prior regularizer is not available in closed form, one instead can rely on

$$\begin{aligned}
\nabla_{\boldsymbol{\kappa}} L(q) &= \nabla_{\boldsymbol{\kappa}} \left[\int_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) + \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}) \right] q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\boldsymbol{\theta} \right] \\
&= \int_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) + \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}) \right] \nabla_{\boldsymbol{\kappa}} q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\boldsymbol{\theta} + \int_{\boldsymbol{\theta}} \left[\nabla_{\boldsymbol{\kappa}} \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}) \right] q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\boldsymbol{\theta} \\
&= \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})} \left[\left(\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) + \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}) \right) \nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}|\boldsymbol{\kappa})) \right] + \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})} \left[\nabla_{\boldsymbol{\kappa}} \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}) \right].
\end{aligned}$$

This derivation is a more general case of the one given in Ranganath et al. (2014), but further simplifies to the one therein if $D = \text{KLD}$. The gradient is estimated without bias by sampling $\boldsymbol{\theta}^{(1:S)}$ from $q(\boldsymbol{\theta}|\boldsymbol{\kappa})$ and again computing $\frac{1}{S}\sum_{s=1}^{S} G(\boldsymbol{\theta}^{(s)})$ for the slightly different

$$G(\boldsymbol{\theta}^{(s)}) = \left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}^{(s)}, x_i) + \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)}) \right] \nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa})) + \nabla_{\boldsymbol{\kappa}} \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)}).$$

H.3.3 Rao-Blackwellization for variance reduction, using (A2)

If the losses define a markov blanket over the factors $\boldsymbol{\theta}_j$, one can employ Rao-Blackwellization for variance reduction. This is done by rewriting for $q_{-j}(\boldsymbol{\theta}_{-j}|\boldsymbol{\kappa}_{-j}) = \prod_{l=1,l\neq j}^{k} q_l(\boldsymbol{\theta}_l|\boldsymbol{\kappa}_l)$ the partial derivatives as

$$\nabla_{\boldsymbol{\kappa}_j} L(q) = \nabla_{\boldsymbol{\kappa}_j} \mathbb{E}_{q_j(\boldsymbol{\theta}_j|\boldsymbol{\kappa}_j)} \left[\mathbb{E}_{q_{-j}(\boldsymbol{\theta}_{-j}|\boldsymbol{\kappa}_{-j})} \left[L(q) | \boldsymbol{\theta}_j \right] \right].$$

The hope is then to get around computing as many of the inner expectations over $q_{-j}(\boldsymbol{\theta}_{-j}|\boldsymbol{\kappa}_{-j})$ as possible. Assume for the moment that at least (D2) holds. Further, denote $q_{-j}(\boldsymbol{\theta}_{-j}|\boldsymbol{\kappa}_{-j}) = q_{-j}$, $q_j(\boldsymbol{\theta}_j|\boldsymbol{\kappa}_j) = q_j$, and in similar fashion the distributions $q_{(j)}$, $q_{-(j)}$, q. Moreover, denote $\ell_i = \ell(\boldsymbol{\theta}, x_i)$, $\ell^D = \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})$ and in a similar fashion $\ell_i^{(j)}$, $\ell_i^{-(j)}$. Now, assuming that (A2) holds relative to the factors $\boldsymbol{\theta}_j$ of the variational family $\mathcal{Q}$, one finds

$$\nabla_{\boldsymbol{\kappa}_j} L(q) = \mathbb{E}_{q_j} \left[\nabla_{\boldsymbol{\kappa}_j} \log(q_j) \left(\mathbb{E}_{q_{-j}} \left[\sum_{i=1}^{n} \ell_i^{(j)} \right] + \mathbb{E}_{q_{-j}}[\ell_n^{-j}] + \mathbb{E}_{q_{-j}}[\ell^D] \right) \right] + \mathbb{E}_{q_{-j}}[\nabla_{\boldsymbol{\kappa}_j} \ell^D].$$

Observing that $\mathbb{E}_{q_j}[\nabla_{\boldsymbol{\kappa}_j} \log(q_j)] = 0$ and that $\mathbb{E}_{q_{-j}}[\ell^{-(j)}]$ is constant in $\boldsymbol{\theta}_j$ by (A2), this drastically simplifies to

$$\nabla_{\boldsymbol{\kappa}_j} L(q) = \mathbb{E}_{q_j} \left[\nabla_{\boldsymbol{\kappa}_j} \log(q_j) \mathbb{E}_{q_{-j}} \left[\sum_{i=1}^{n} \ell_i^{(j)} \right] + \mathbb{E}_{q_{-j}} \left[\ell^D + \nabla_{\boldsymbol{\kappa}_j} \ell^D \right] \right].$$

Next, observe that by virtue of how $\ell^{(j)}$ was constructed, it holds that we can also simplify

$$\mathbb{E}_{q_j} \left[\nabla_{\boldsymbol{\kappa}_j} \log(q_j) \mathbb{E}_{q_{-j}} \left[\sum_{i=1}^{n} \ell_i^{(j)} \right] \right] = \mathbb{E}_{q_{(j)}} \left[\sum_{i=1}^{n} \ell_i^{(j)} \right].$$

Putting the above together, we finally arrive at

$$\begin{aligned} \nabla_{\boldsymbol{\kappa}_j} L(q) &= \mathbb{E}_{q_j} \left[\nabla_{\boldsymbol{\kappa}_j} \log(q_j) \left(\mathbb{E}_{q_{-j}} \left[\sum_{i=1}^{n} \ell_i^{(j)} \right] + \mathbb{E}_{q_{-j}}[\ell^D] \right) + \mathbb{E}_{q_{-j}}[\nabla_{\boldsymbol{\kappa}_j} \ell^D] \right] \\ &= \mathbb{E}_{q_{(j)}} \left[\nabla_{\boldsymbol{\kappa}_j} \log(q_j) \sum_{i=1}^{n} \ell_i^{(j)} \right] + \mathbb{E}_q \left[\nabla_{\boldsymbol{\kappa}_j} \log(q_j) \ell^D + \nabla_{\boldsymbol{\kappa}_j} \ell^D \right]. \end{aligned}$$

which is the final form under (D1). Should (D1) to hold, one can instead use the lower variance estimate

$$\nabla_{\boldsymbol{\kappa}_j} L(q) = \mathbb{E}_{q_{(j)}} \left[\nabla_{\boldsymbol{\kappa}_j} \log(q_j) \sum_{i=1}^{n} \ell_i^{(j)} \right] + \nabla_{\boldsymbol{\kappa}_j} D(q\|\pi).$$

These derivations are very similar to the ones in the supplement of Ranganath et al. (2014), but importantly the former are restricted to negative log likelihood losses. The more general version presented here holds for arbitrary decomposable losses. The J terms $\nabla_{\boldsymbol{\kappa}_j} L(q)$ can be combined into a global gradient estimate simply by setting

$$\nabla_{\boldsymbol{\kappa}} L(q) = \left(\nabla_{\boldsymbol{\kappa}_1} L(q), \nabla_{\boldsymbol{\kappa}_2} L(q), \ldots \nabla_{\boldsymbol{\kappa}_J} L(q) \right)^T.$$

To make the meaning of (A2) more tangible for the case of general losses, we next provide a short example in the context of multivariate regression.

Example 5 (Markov blankets without conditional independence) *Suppose each* $x_i = (x_{i,1}, x_{i,2}, x_{i,3})'$ *consists of three measurements that we wish to relate to some other observables* y_i *through*

$$\begin{aligned} x_{i,1} &= a + y_i b + \xi_1 \\ x_{i,2} &= b + y_i c + \xi_2 \\ x_{i,3} &= d + \xi_3 \end{aligned}$$

where ξ_j *are unknown slack variables (or errors), the parameters of interest are* $\boldsymbol{\theta} = (a, b, c, d, e)$ *and we wish to produce a belief distribution over* $\boldsymbol{\theta}$ *that is informative about good values of* $\boldsymbol{\theta}$ *relative to some prediction loss*

$$\ell(\boldsymbol{\theta}, x_i) = \|f_1^1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{(1)}, y_i) - x_{i,1}\|_p^p + \|f_2^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{(1)}, y_i) - x_{i,2}\|_p^p + \|f_2^3(\boldsymbol{\theta}_2, \boldsymbol{\theta}_{(2)}, y_i) - x_{i,3}\|_p^p,$$

where $\|\cdot\|_p^p$ *denotes some p-norm for* $p \geq 1$ *and* f_l^j *seeks to predict only the l-th dimension of* x_i *by means of the l-th factor of* $\boldsymbol{\theta}$ *and its blanket. Suppose that* f_l^j *will correspond to the l-th row written down in the above model for* x_i *(excluding of course the error term), which means that*

$$\begin{aligned} f_1^1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{(1)}) &= a + y_i b \\ f_2^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{(1)}, y_i) &= b + y_i c \\ f_2^3(\boldsymbol{\theta}_2, \boldsymbol{\theta}_{(2)}, y_i) &= d \end{aligned}$$

In this case, the two factors of $\boldsymbol{\theta}$ *will clearly be given by*

$$\boldsymbol{\theta}_1 = (a, b, c)^T, \qquad \boldsymbol{\theta}_2 = (d).$$

As before, one will in practice need to approximate the gradients with a sample $\boldsymbol{\theta}^{(1:S)}$ drawn from $q(\boldsymbol{\theta}|\boldsymbol{\kappa})$. For one of the fixed samples $\boldsymbol{\theta}^{(s)}$, the relevant terms are computed as

$$G_j(\boldsymbol{\theta}^{(s)}) = \nabla_{\boldsymbol{\kappa}_j} \log(q_j(\boldsymbol{\theta}_j^{(s)}|\boldsymbol{\kappa}_j)) \sum_{i=1}^{n} \ell^{(j)}(\boldsymbol{\theta}_j^{(s)}, \boldsymbol{\theta}_{(j)}^{(s)}, x_i) + \widetilde{D}(s, j)$$

for some function $\widetilde{D}(s, j)$. If (D2) holds and there is no closed form for the prior regularizer, this function is given by

$$\widetilde{D}(s, j) = \nabla_{\boldsymbol{\kappa}_j} \log(q_j(\boldsymbol{\theta}_j^{(s)}|\boldsymbol{\kappa}_j)) \ell_{\pi,\boldsymbol{\kappa}}^D(\boldsymbol{\theta}^{(s)}) + \nabla_{\boldsymbol{\kappa}_j} \ell_{\pi,\boldsymbol{\kappa}}^D(\boldsymbol{\theta}^{(s)})$$

and in case the stricter requirement (D1) holds, it is simply given by the closed form

$$\widetilde{D}(s, j) = \nabla_{\boldsymbol{\kappa}_j} D(q\|\pi).$$

H.4 BBGVI if neither (A2) nor (A3) hold

It is of course possible that neither (A2) nor (A3) hold. Alternatively, it may simply be convenient to build an implementation that can work reliably without imposing any assumptions. In this case, one will have to use the naive version of BBGVI that is given in the main paper and only depends on the distinction between (D2) and (D1). However—even though we do not do so in our experiments–there still are valid black box variance reduction techniques for this case. The next section presents these techniques, again by adapting notation and logic from Ranganath et al. (2014).

H.5 Generically applicable variance reduction

While the Rao-Blackwellization variance reduction will generally be more effective, some variance reduction techniques can work in circumstances where Rao-Blackwellization does not. Conversely, this means that if the Rao-Blackwellization is applicable, one can actually deploy two variance reduction schemes at once to substantially speed up convergence. The control variate we use is simply

$$h(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\kappa}} \log q(\boldsymbol{\theta}|\boldsymbol{\kappa})$$

with an optimal scaling parameter that can be estimated as

$$\hat{a}^* = \frac{\sum_{s=1}^{S} \widehat{\mathrm{Cov}}(L(\boldsymbol{\theta}^{(s)}), h(\boldsymbol{\theta}^{(s)}))}{\sum_{s=1}^{S} \widehat{\mathrm{Var}}(h(\boldsymbol{\theta}^{(s)}))}.$$

Based on this, one may now compute the variance reduced term $G_{\mathrm{VR}}(\boldsymbol{\theta}^{(s)})$ from $G(\boldsymbol{\theta}^{(s)})$ as

$$G_{\mathrm{VR}}(\boldsymbol{\theta}^{(s)}) = G(\boldsymbol{\theta}^{(s)}) - \hat{a}^* \cdot h(\boldsymbol{\theta}^{(s)}).$$

Of course, the exact same logic can be applied to the Rao-Blackwellized terms $G_j(\boldsymbol{\theta}^{(s)})$ to reduce the variance a second time.

Appendix I. Closed forms for divergences & proof of Proposition 17

This section proves various closed forms for the prior regularizers in the GVI problem. We do so by proving conditions for closed forms of the $\alpha\beta\gamma$-divergence ($D_G^{(\alpha,\beta,r)}$) introduced in Appendix A. Note that the special case of these results for the $D_{AR}^{(\alpha)}$ has been derived before (see Gil et al., 2013; Gil, 2011; Liese and Vajda, 1987). Unlike previous work, our results apply to a range of other divergences, too. This is convenient because all other robust divergences we discuss throughout the paper are special cases of $D_G^{(\alpha,\beta,r)}$.

I.1 High-level overview of results and preliminaries

Summarizing some of the most important findings of this section, we find that if both $q(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ are in the same exponential variational family $\mathcal{Q}$,

- $D_{AR}^{(\alpha)}(q||\pi)$ and $D_A^{(\alpha)}(q|\pi)$ are always available in closed form if $\alpha \in (0,1)$ (see Corollary 35)
- $D_{AR}^{(\alpha)}(q||\pi)$ and $D_A^{(\alpha)}(q|\pi)$ are available in closed form if $\alpha > 1$ for most exponential families (see again Corollary 35)
- $D_B^{(\beta)}(q||\pi)$ and $D_G^{(\gamma)}(q||\pi)$ are available in closed form for $\beta > 1$ and $\gamma > 1$ for most exponential families (See Corollary 41).

We note that these findings are interesting because closed forms for the divergence term drastically reduce the variance of black box GVI, see also Appendix H. The remainder of this section is devoted to tedious but rigorous derivations of these findings. Before stating any results, it is useful to state the definition of an exponential family and its natural parameter space upon which the proofs rely.

Definition 31 (Exponential families) *Object $\theta \in \Theta \subset \mathbb{R}^d$, $d \geq 1$ has an exponential family distribution with parameters $\kappa \in \boldsymbol{K} \subset \mathbb{R}^{p'}$, $p' \geq 1$ if there exist functions $\eta : \boldsymbol{K} \to \mathcal{N} \subset \mathbb{R}^p$, $p \geq 1$, $T : \Theta \to \mathcal{T} \subset \mathbb{R}^p$, $h : \Theta \to \mathbb{R}_{\geq 0}$ and $A : \mathcal{N} \to \mathbb{R}$ such that*

$$p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa})) = h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}))\right\},$$

where $A(\eta(\boldsymbol{\kappa})) = -\log\left(\int h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta})\right\} d\theta\right)$. The set $\mathcal{N}$ is called the natural parameter space and is defined to ensure $p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa}))$ is a normalised probability density, $\mathcal{N} = \{\eta(\boldsymbol{\kappa}) : A(\eta(\boldsymbol{\kappa})) < \infty\}$.

Throughout the rest of this section, we assume that the following condition holds for both the prior and the variational family $\mathcal{Q}$.

Condition 1 (The prior and variational families) *It holds that*

i) the variational family $\mathcal{Q} = \{q(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa}))\}$ is an exponential family of the form given by Definition 31

ii) the prior $\pi(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa}_0))$ is a member of that variational family.

Amongst other things, this implies that the log-normalising constant is a closed form function of the natural parameters and that we can derive generic conditions for closed forms by using the canonical representation of exponential families.

To showcase the implications of the derived results, we use the Mulitvariate Gaussian (MVN) to provide examples along the way.

Definition 32 (The MVN exponential family) *The density of the MVN exponential family for vector $\boldsymbol{\theta}$ of dimension d is $p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa})) = h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}))\right\}$ where*

$$\eta(\boldsymbol{\kappa}) = \begin{pmatrix} \boldsymbol{V}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{V}^{-1} \end{pmatrix} \qquad T(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\theta}\boldsymbol{\theta}^T \end{pmatrix}$$
$$h(\boldsymbol{\theta}) = (2\pi)^{-d/2} \qquad A(\eta(\boldsymbol{\kappa})) = \left[\frac{1}{2}\log|\boldsymbol{V}| + \frac{1}{2}\boldsymbol{\mu}\boldsymbol{V}^{-1}\boldsymbol{\mu}\right]$$

and the natural parameter space requires that $\boldsymbol{\mu}$ is a real valued vector of the same dimension as $\boldsymbol{\theta}$ and $\boldsymbol{V}$ is a $d \times d$ symmetric semi-positive definite matrix.

I.2 Results, proofs & examples

The remainder of this section is structued as follows: First, we give the main result for the $\alpha\beta\gamma$-divergence ($D_G^{(\alpha,\beta,r)}$) in Proposition 33. This "master result" is then applied to various special cases for $D_G^{(\alpha,\beta,r)}$ that are of practical interest, namely the α-divergence ($D_A^{(\alpha)}$), Rényi's α-divergence ($D_{AR}^{(\alpha)}$), the β-divergence ($D_B^{(\beta)}$) as well the γ-divergence ($D_G^{(\gamma)}$).

I.2.1 Master result for $D_G^{(\alpha,\beta,r)}$

While the following result and corresponding proof are somewhat tedious to read, they are conceptually simple: In fact, all that is needed to derive the results is some basic algebra and the canonical form of the exponential family.

Proposition 33 (Closed form $D_G^{(\alpha,\beta,r)}$ between exponential families) *The $D_G^{(\alpha,\beta,r)}$ between a variational posterior $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)$ and prior $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)$ is available in closed form under the following conditions*

i) $\eta(\boldsymbol{\kappa}_0), \eta(\boldsymbol{\kappa}_n) \in \mathcal{N} \Rightarrow (\alpha\eta(\boldsymbol{\kappa}_0) + (\beta-1)\eta(\boldsymbol{\kappa}_n)) \in \mathcal{N}$*;*

ii) $\mathbb{E}_{p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa}))}\left[h(\boldsymbol{\theta})^{\alpha+\beta-2}\right]$ *is a closed form function of* $\eta(\boldsymbol{\kappa}) \in \mathcal{N}$*.*

If these conditions hold the $D_G^{(\alpha,\beta,r)}$ can be written as

$$\begin{aligned}&\tilde{D}_G^{(\alpha,\beta)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)||\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0))\\ &= \alpha B(\boldsymbol{\kappa}_n, (\alpha+\beta-1))E(\boldsymbol{\kappa}_n, (\alpha+\beta-1)) + (\beta-1)B(\boldsymbol{\kappa}_0, (\alpha+\beta-1))E(\boldsymbol{\kappa}_0, (\alpha+\beta-1))\\ &\quad - (\alpha+\beta-1)C(\boldsymbol{\kappa}_n, \boldsymbol{\kappa}_0, \alpha, (\beta-1))\tilde{E}(\boldsymbol{\kappa}_n, \boldsymbol{\kappa}_0, \alpha, (\beta-1))\end{aligned}$$

where

$$B(\boldsymbol{\kappa}, \delta) = \frac{\exp\{A(\delta\eta(\boldsymbol{\kappa}))\}}{\exp\{A(\eta(\boldsymbol{\kappa}))\}^{\delta}}, \qquad C(\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2, \delta_1, \delta_2) = \frac{\exp\{A\left(\delta_1\eta(\boldsymbol{\kappa}_1) + \delta_2)\eta(\boldsymbol{\kappa}_2)\right)\}}{\exp\{A(\eta(\boldsymbol{\kappa}_1))\}^{\delta_1}\exp\{A(\eta(\boldsymbol{\kappa}_2))\}^{\delta_2}}$$

$$E(\boldsymbol{\kappa}, \delta) = \mathbb{E}_{p(\theta|\delta\eta(\boldsymbol{\kappa}))}\left[h(\boldsymbol{\theta})^{\delta-1}\right], \qquad \tilde{E}(\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2, \delta_1, \delta_2) = \mathbb{E}_{p(\theta|\delta_1\eta(\boldsymbol{\kappa}_1)+\delta_2\eta(\boldsymbol{\kappa}_2))}\left[h(\boldsymbol{\theta})^{\delta_1+\delta_2-1}\right]$$

we suppress the dependence of these functions on $A(\cdot)$ and $h(\cdot)$ as these derive form the definition of the exponential family (Definition 31).

Proof The $D_G^{(\alpha,\beta,r)}$ is a closed form function of $\tilde{D}_G^{(\alpha,\beta)}$ given in Definition 20. Hence if $\tilde{D}_G^{(\alpha,\beta)}$ is available in closed form, then so is $D_G^{(\alpha,\beta,r)}$. In order to ensure that $\tilde{D}_G^{(\alpha,\beta)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)||\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0))$ has closed form, we need to make sure the three integrals below are available in closed form for the exponential family.

$$G_1 := \int q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)^{\alpha+\beta-1}d\boldsymbol{\theta}, \quad G_2 := \int \pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)^{\alpha+\beta-1}d\boldsymbol{\theta},$$
$$G_3 := \int q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)^{\alpha}\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)^{\beta-1}d\boldsymbol{\theta}.$$

First we tackle G_1.

$$\begin{aligned}G_1 &= \int h(\boldsymbol{\theta})^{\alpha+\beta-1}\exp\left\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)^T T(\boldsymbol{\theta}) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_n))\right\}d\boldsymbol{\theta}\\ &= \exp\{A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_n))\}\,\mathbb{E}_{p(\boldsymbol{\theta}|(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)}\left[h(\boldsymbol{\theta})^{\alpha+\beta-2}\right],\end{aligned}$$

where condition (i) with $\eta(\boldsymbol{\kappa}_0) = \eta(\boldsymbol{\kappa}_n)$ ensures that

$$A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)) = \int h(\boldsymbol{\theta})\exp\left\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)^T T(\boldsymbol{\theta})\right\}d\boldsymbol{\theta} < \infty,$$

which in turn ensures that $p(\boldsymbol{\theta}|(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)$ is a normalised probability density and that $\mathbb{E}_{p(\boldsymbol{\theta}|(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)}\left[h(\boldsymbol{\theta})^{\alpha+\beta-2}\right]$ is a valid expectation. Now, condition (ii) guarantees this is a closed form function of $\eta(\boldsymbol{\kappa}_n)$. Similarly for G_2,

$$\begin{aligned}G_2 &= \int h(\boldsymbol{\theta})^{\alpha+\beta-1}\exp\left\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0)^T T(\boldsymbol{\theta}) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_0))\right\}d\boldsymbol{\theta}\\ &= \exp\{A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0)) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_0))\}\,\mathbb{E}_{p(\boldsymbol{\theta}|(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0)}\left[h(\boldsymbol{\theta})^{\alpha+\beta-2}\right],\end{aligned}$$

where in analogy to G_1, conditions (i) and (ii) with $\eta(\boldsymbol{\kappa}_k) = \eta(\boldsymbol{\kappa}_0)$ ensure this has a closed form. Lastly for G_3,

$$\begin{aligned} G_3 = &\int h(\boldsymbol{\theta})^{\alpha} \exp\left\{\alpha\eta(\boldsymbol{\kappa}_n)^T T(\boldsymbol{\theta}) - \alpha A(\eta(\boldsymbol{\kappa}_n))\right\} \\ &\cdot h(\boldsymbol{\theta})^{\beta-1} \exp\left\{(\beta-1)\eta(\boldsymbol{\kappa}_0)^T T(\boldsymbol{\theta}) - (\beta-1)A(\eta(\boldsymbol{\kappa}_0))\right\} d\boldsymbol{\theta} \\ = &\exp\left\{A\left(\alpha\eta(\boldsymbol{\kappa}_n) + (\beta-1)\eta(\boldsymbol{\kappa}_0)\right) - \alpha A(\eta(\boldsymbol{\kappa}_n)) - (\beta-1)A(\eta(\boldsymbol{\kappa}_0))\right\} \\ &\cdot \mathbb{E}_{p(\boldsymbol{\theta}|(\alpha\eta(\boldsymbol{\kappa}_n)+(\beta-1)\eta(\boldsymbol{\kappa}_0))}\left[h(\boldsymbol{\theta})^{\alpha+\beta-2}\right], \end{aligned}$$

where once again in analogy to G_1 and G_2, conditions (i) and (ii) ensure this is a closed form function of $\eta(\boldsymbol{\kappa}_n)$ and $\eta(\boldsymbol{\kappa}_0)$.

Therefore, provided conditions (i) and(ii) hold, the integrals G_1, G_2 and G_3 are available in closed form, implying that the same holds for $D_G^{(\alpha,\beta,r)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)||\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0))$.

■

Remark 34 (Conditions of Proposition 33 for the MVN exponential family) *In order to illuminate the meaning and generality of the conditions of Theorem 33, we apply them to the* MVN *exponential family described in Definition 32. In this case the two conditions become:*

i) For $\boldsymbol{\mu}^* := \left\{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 - \left(\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1} + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\right)^{-1}\left(\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1}\boldsymbol{\mu}_2 + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\boldsymbol{\mu}_1\right)\right\}$ *we require that*

$$\begin{aligned} \alpha\begin{pmatrix}\boldsymbol{V}_1^{-1}\boldsymbol{\mu}_1 \\ -\frac{1}{2}\boldsymbol{V}_1^{-1}\end{pmatrix} + (\beta-1)\begin{pmatrix}\boldsymbol{V}_2^{-1}\boldsymbol{\mu}_2 \\ -\frac{1}{2}\boldsymbol{V}_2^{-1}\end{pmatrix} &= \begin{pmatrix}\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1}\boldsymbol{\mu}_1 + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\boldsymbol{\mu}_2 \\ -\frac{1}{2}\left\{\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1} + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\right\}\end{pmatrix} \\ &= \begin{pmatrix}\left\{\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1} + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\right\}\boldsymbol{\mu}^* \\ -\frac{1}{2}\left\{\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1} + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\right\}\end{pmatrix} \in \mathcal{N} \end{aligned}$$

ii) $\mathbb{E}_{p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa}))}\left[(2\pi)^{-d/2(\alpha+\beta+2)}\right] = (2\pi)^{-d/2(\alpha+\beta+2)} = f(\eta(\boldsymbol{\kappa}))$ *where* f *is a closed form function.*

Part ii) shows that the second condition is trivially satisfied for the MVN *exponential family. Part i) shows that for the* MVN *exponential family, the first condition is satisfied provided* $(V^*)^{-1} = \left\{\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1} + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\right\}$ *is a positive definite matrix. This condition is enough to ensure that* V^* *is invertible and thus that* μ^* *is well-defined. We elaborate further on what this means for certain parametrisations below.*

I.2.2 Corollary: The special cases of $D_A^{(\alpha)}$, $D_{AR}^{(\alpha)}$

Next, we consider the $D_A^{(\alpha)}$ and $D_{AR}^{(\alpha)}$ special cases of the $D_G^{(\alpha,\beta,r)}$ family. Definitions 21 and 22 can be used to show that the $D_{AR}^{(\alpha)}$ is available as the following closed form function of the $D_A^{(\alpha)}$. In particular, it holds that

$$D_{AR}^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\alpha(\alpha-1)} \log\left\{1 + \alpha(1-\alpha) D_A^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))\right\}. \tag{35}$$

Thus, as demonstrated in Corollary 36 below, the $D_A^{(\alpha)}$ being available in closed form immediately provides the $D_{AR}^{(\alpha)}$ in closed form. Before stating these results, we note that Gil et al. (2013); Gil (2011); Liese and Vajda (1987) have shown our closed form results for the $D_{AR}^{(\alpha)}$ (and thus implicitly the $D_A^{(\alpha)}$) before. We nevertheless think there is merit in stating them, since our results refer to the $D_G^{(\alpha,\beta,r)}$ and thus are more general, recovering both the $D_A^{(\alpha)}$ and $D_{AR}^{(\alpha)}$ only as a special case.

Corollary 35 (Closed form $D_A^{(\alpha)}$ for exponential families) *The $D_A^{(\alpha)}$ between a variational posterior $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)$ and prior $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)$ is available in closed form under the following conditions*

i) $(\alpha\eta(\boldsymbol{\kappa}_n) + (1-\alpha)\eta(\boldsymbol{\kappa}_0)) \in \mathcal{N}$

and in this case the $D_A^{(\alpha)}$ can be written as

$$D_A^{(\alpha)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)||\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0) = \frac{1}{\alpha(1-\alpha)}\left[1 - C(\boldsymbol{\kappa}_n, \boldsymbol{\kappa}_0, \alpha, (1-\alpha))\right],$$

where $C(\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2, \delta_1, \delta_2)$ was defined in Proposition 33 .

Proof Following Cichocki and Amari (2010) the single-parameter $D_A^{(\alpha)}$ is recovered as a member of the $D_G^{(\alpha,\beta,r)}$ family when $r = 1$ and $\beta = 2 - \alpha$. In this situation, Condition (ii) of Theorem 33 holds automatically and we are left with Condition (i). Substituting $\beta = 2 - \alpha$ provides Condition (i) of the Theorem above.

If $\alpha \in (0, 1)$ then the convexity of the natural parameter space ensures that providing $\eta(\boldsymbol{\kappa}_n) \in \mathcal{N}$ and $\eta(\boldsymbol{\kappa}_0) \in \mathcal{N}$ then $\alpha\eta(\boldsymbol{\kappa}_n) + (1-\alpha)\eta(\boldsymbol{\kappa}_0) \in \mathcal{N}$. If $\alpha < 0$ or $\alpha > 1$, then this can no longer be guaranteed. ■

Corollary 36 is then an immediate consequence of Corollary 35.

Corollary 36 (Closed form $D_{AR}^{(\alpha)}$ for exponential families) *The $D_{AR}^{(\alpha)}$ between a variational posterior $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)$ and prior $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)$ will have closed form providing the $D_A^{(\alpha)}$ between the same two densities for the same value of α has closed form.*

Proof The proof of this follows immediately from the fact that the $D_{AR}^{(\alpha)}$ can be recovered using the closed form function of the $D_A^{(\alpha)}$ shown in eq. (35) ■

Remark 37 (Conditions for Corollary 35 for the MVN exponential family) *The condition that $\alpha\eta(\boldsymbol{\kappa}_n) + (1-\alpha)\eta(\boldsymbol{\kappa}_0) \in \mathcal{N}$ can only be guaranteed for $\alpha \in (0, 1)$. However we*

can see from Remark 34 that provided $\boldsymbol{V}^* = \left(\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1} + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1} \right)^{-1}$ *is a symmetric semi-positive definite (*SPD*) matrix for* $\beta = 2 - \alpha$ *then this condition will be satisfied. For* $\alpha > 1$ *or* $\alpha < 0$ *we cannot guarantee that* $\boldsymbol{V}^*$ *is* SPD*. However, we implement the* $D_{AR}^{(\alpha)}$ *to quantify uncertainty for* $\alpha > 1$ *in the main paper. Corollary 35 demonstrates that these parameters will still produce a closed form divergence provided the prior has sufficiently large variance, which can always be guaranteed to hold in practice.*

I.2.3 Corollary: The special cases of $D_B^{(\beta)}$, $D_G^{(\gamma)}$

Next, we turn attention to the β- and γ-divergence families. Definition 24 shows that the $D_G^{(\gamma)}$ can be recovered as a closed form function of the terms of the $D_B^{(\beta)}$ and thus, as demonstrated in Corollary 39 below, the $D_B^{(\beta)}$ being available in closed form immediately provides that the $D_G^{(\gamma)}$ is available in closed form While the conditions for these are slightly more restrictive than they were for the $D_A^{(\alpha)}$ and $D_{AR}^{(\alpha)}$, one can still obtain closed form prior regularizers for a large range of settings.

Corollary 38 (Closed form $D_B^{(\beta)}$ for exponential families) *The* $D_B^{(\beta)}$ *between a variational posterior* $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)$ *and prior* $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)$ *is available in closed form under the following conditions*

i) $\eta(\boldsymbol{\kappa}_1), \eta(\boldsymbol{\kappa}_2) \in \mathcal{N} \Rightarrow ((\beta-1)\eta(\boldsymbol{\kappa}_1) + \eta(\boldsymbol{\kappa}_2)) \in \mathcal{N}$

ii) $\mathbb{E}_{p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa}))}\left[h(\boldsymbol{\theta})^{\beta-1}\right]$ *is a closed form function of* $\eta(\boldsymbol{\kappa}) \in \mathcal{N}$*.*

and in this case the $D_B^{(\beta)}$ *can be written as*

$$D_B^{(\beta)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)||\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)) = \frac{1}{\beta(\beta-1)}B(\boldsymbol{\kappa}_n,\beta)E(\boldsymbol{\kappa}_n,\beta) + \frac{1}{\beta}B(\boldsymbol{\kappa}_0,\beta)E(\boldsymbol{\kappa}_0,\beta)$$
$$-\frac{1}{(\beta-1)}C(\boldsymbol{\kappa}_n,\boldsymbol{\kappa}_0,1,(\beta-1))\tilde{E}(\boldsymbol{\kappa}_n,\boldsymbol{\kappa}_0,1,(\beta-1)),$$

where the functions $B(\boldsymbol{\kappa},\delta)$*,* $C(\boldsymbol{\kappa}_1,\boldsymbol{\kappa}_2,\delta_1,\delta_2)$*,* $E(\boldsymbol{\kappa},\delta)$ *and* $\tilde{E}(\boldsymbol{\kappa}_1,\boldsymbol{\kappa}_2,\delta_1,\delta_2)$ *are defined in Proposition 33.*

Proof Following Cichocki and Amari (2010), the single-parameter $D_B^{(\beta)}$ is recovered as a member of the $D_G^{(\alpha,\beta,r)}$ family when $r = 1$ and $\alpha = 1$. In this situation, Condition (i)-(ii) of Theorem 33 become (i)-(ii) above. ■

Corollary 39 is then an immediate consequence of Corollary 38.

Corollary 39 (Closed form $D_G^{(\gamma)}$ for exponential families) *The* $D_G^{(\gamma)}$ *between a variational posterior* $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)$ *and prior* $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)$ *will have closed form providing the* $D_B^{(\beta)}$ *between the same two densities with* $\beta = \gamma$ *has closed form.*

Proof The proof of this follows immediately from the fact that the $D_G^{(\gamma)}$ can be recovered from the $D_B^{(\beta)}$ using closed form function as outlined in Definition 24. ■

Remark 40 (Conditions for Corollary 38 under the MVN exponential family) *Following Remark 34, Corollary 38 is satisfied providing* $\boldsymbol{V}^* = \left((\boldsymbol{V}_n)^{-1} + \left(\frac{1}{\beta - 1} \boldsymbol{V}_0 \right)^{-1} \right)^{-1}$ *is a symmetric* SPD *matrix. The sum of two symmetric* SPD *matrices is symmetric* SPD *and additionally the inverse of a symmetric* SPD *matrix is also* SPD*. Therefore provided* $\beta > 1$ *we can be sure that Condition iii) will be satisfied. Similarly to Remark 37, when* $\beta < 1$ *closed forms will require that the prior has a sufficiently large variance.*

In fact Remark 40 can be extended to many other exponential families if we constrain $\beta = \gamma > 1$, this is formalised in Corollary 41.

Corollary 41 (Closed form $D_B^{(\beta)}$ and $D_G^{(\gamma)}$ for exponential families when $\beta = \gamma > 1$) *When* $\beta = \gamma > 1$*, the conditions for Corollary 38 are satisfied by any exponential family whose* $h(\boldsymbol{\theta})$ *is a constant function of* $\boldsymbol{\theta}$ *and whose natural parameter space is closed under addition and scalar multiplication. This includes the Beta, Gamma, Gaussian, exponential and Laplace families.*

Proof The proof of Corollary 41 follows straight from that of Corollary 38. ■

Appendix J. Experiments

While the most interesting findings of our numerical studies can be found in the main paper, here we give a brief overview over additional results. More importantly, we state the proofs for the theoretical groundwork necessary to deploy GVI on DGPs.

J.1 Bayesian Neural Networks (BNNs)

We provide two more sets of experiments for further insights into BNNs. The first set consists in three more data sets with the same settings as used in the main paper. While these findings do not change the overall picture, they do require more careful analysis and dissemination. The second set of results investigates the interaction between robustifying inference relative to the loss with robustifying it relative to the prior. The results suggest a clear relationship for predictive performance as measured by the root mean square error: If robust losses are used, the KLD generally performs better. Moreover, the combination of robust loss and $D = \text{KLD}$ outperforms VI and the investigated DVI methods on all data sets studied. The relationship is less clear for the predictive negative log likelihood, both between loss and prior regularizer as well as between the performance to be expected under GVI, VI and DVI.

J.1.1 First set of additional experiments (Figure 21)

Figure 21 provides the predictive outcomes on three more data sets using the exact same settings and experimental setup as described in the main paper. The findings generally reinforce the findings of the main paper. First, while the GVI methods with $\alpha > 1$ still perform as good as or better than standard VI on the `kin8mn` data set, DVI methods show a clear performance gain relative to either of the two. Crucially, it is not clear what

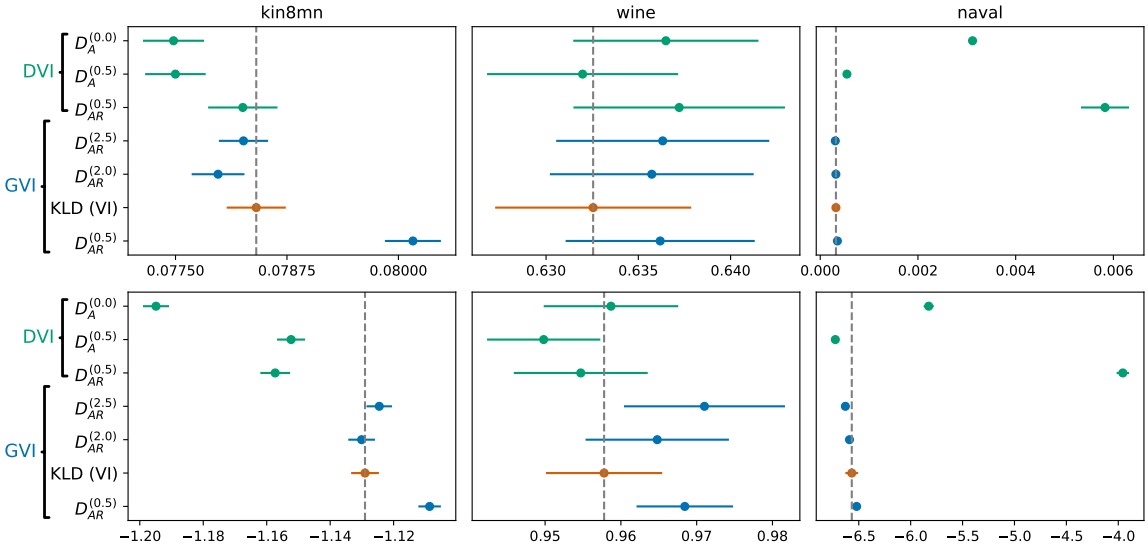


Figure 21: Top row depicts RMSE, bottom row the NLL across a range of data sets using BNNs. Dots correspond to means, whiskers the standard errors. The further to the left, the better the predictive performance. For the depicted selection of data sets, no common pattern exists for the performance differences between **standard VI**, **DVI** and **GVI**.

leads to this improvement gain, though the fact that the best-performing DVI method is the one recovering EP ($D_A^{(\alpha)}$ for $\alpha = 0$) suggests that there is tangible merit in producing mass-covering approximations to the posterior of $\boldsymbol{\theta}$ on this data set. While the deployment of DVI methods looks tempting on the `kin8mn` data set, the results on the `naval` data set are a reminder that the behaviour of these methods is in many ways unpredictable. Moreover, it shows that the risks we identified in Example 3 readily translate into real world applications: By using DVI methods, we may accidentally conflate the role of the loss and the role of uncertainty quantification. If the loss is well-suited for the data at hand—as the RMSE panel suggests it is in the `naval` case—the mass-covering behaviour of DVI methods can be extremely detrimental. Lastly, the `wine` data set provides a very similar picture to the results in Figure 11: Varying α introduces a banana-shaped curve for the GVI methods. As it so happens, the ideal choice of α on the `wine` data set appears to be around $\alpha = 1$ (i.e., standard VI). Taking into account the predictive uncertainty in form of the whiskers, it is doubtful if any of the methods is dominating another one on `wine`. Presumably, the reason for this is that the true posterior is relatively well approximated with the mean field normal family, yielding very similar results across all settings.

J.1.2 Second set of additional experiments (Figure 22)

In a second set of additional experiments, we varied the loss function to be a robust scoring rule. Specifically, we used scoring rules based on the β-divergence and the γ-divergence. See Section 6.2.3 for the definition and more detail on these robust scoring rules. As for the DGP examples, we choose values of the scoring rule that are close to the log score, but

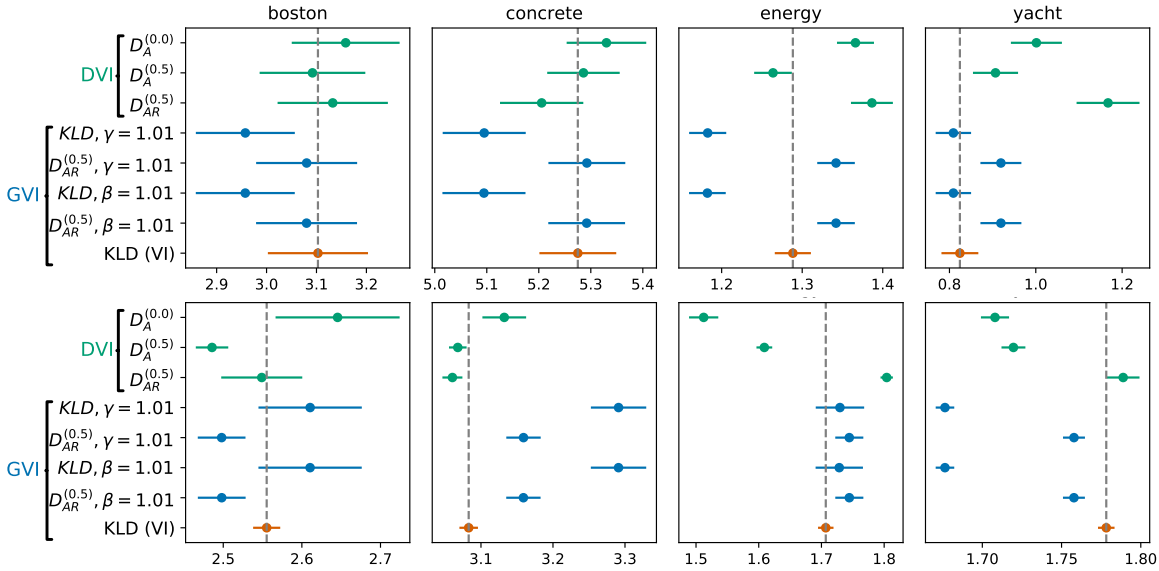


Figure 22: Top row depicts RMSE, bottom row the NLL across a range of data sets using BNNs. Dots correspond to means, whiskers the standard errors. The further to the left, the better the predictive performance. For the depicted selection of data sets, patterns exists for the interplay between the loss and prior regularizer for **GVI**.

sufficiently far to induce robust behaviour. All settings for optimization, initialization as well as the code are the same as for the results provided in the main paper. Figure 22 shows the results: For the RMSE, the results are unambiguous: Combining a robust scoring rule with the standard prior regularizer $D = \text{KLD}$ appears to be the winning combination across all four data sets. The picture is less clear for the NLL: Relative to both VI and DVI, the performance gains depend on the data set. Even within the class of GVI posteriors, it is data-set dependent which prior regularizer should be chosen: For example, it is clearly beneficial to choose the $D_{AR}^{(\alpha)}$ as prior regularizer in the `boston` and `concrete` data sets, but the opposite is true on the `yacht` data set. Above all other things, this highlights the need for a good selection strategy of GVI hyperparameters: Oftentimes, intuitions about the correct prior regularizer or the appropriate loss may be incorrect.

J.2 Deep Gaussian Processes (DGPs)

Unlike BNNs, DGPs require some theoretical groundwork before they are amenable to changes in the loss and prior regularizer. Specifically, we need to show that it is valid to define new divergences layer-wise. Moreover, while not required it is beneficial if one can obtain closed forms for the robustified likelihood terms. The following sections proceed to do both. Thereafter, we also show an additional short example to illustrate the effect of changing the prior regularizer in DGPs.

J.2.1 Proof of Corollary 18

We first prove a Lemma that plays a key role in the proof of Corollary 18.

Lemma 42 (Divergence recombination) *Let D_l be divergences and $c_l > 0$ scalars for $l = 1, 2, \ldots, L$. Further, denote $\boldsymbol{\theta}_{-l} = \boldsymbol{\theta}_{1:l-1,l+t:L}$ and let $q_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})$ and $\pi_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})$ be the conditional distributions of $\boldsymbol{\theta}_l$ for $q(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ conditioned on $\boldsymbol{\theta}_{-l} = \boldsymbol{\theta}'_{-l}$. Then, $D^{\boldsymbol{\theta}'}(q||\pi) = \sum_{l=1}^{L} c_l D_l\left(q_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})||\pi_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})\right)$ is a divergence between $q(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ if (i) $D^{\boldsymbol{\theta}^\circ}(q||\pi) = D^{\boldsymbol{\theta}'}(q||\pi)$ for all conditioning sets $\boldsymbol{\theta}^\circ$, $\boldsymbol{\theta}'$ and (ii) a Hammersley-Clifford Theorem holds for the collection of conditionals $\pi_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})$ and $q_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})$.*

Proof First, observe by definition of a divergence, $D_l(q_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})||\pi_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})) = 0$ for all l and over all potential conditioning sets $\boldsymbol{\theta}'$ holds if and only if $q_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l}) = \pi_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})$. Next, note that we have assumed that $D^{\boldsymbol{\theta}'}(q||\pi) = D^{\boldsymbol{\theta}^\circ}(q||\pi)$ for all conditioning sets $\boldsymbol{\theta}'$, $\boldsymbol{\theta}^\circ$. In other words, if $D^{\boldsymbol{\theta}'}(q||\pi) = 0$ for some $\boldsymbol{\theta}'$, then it will also be 0 for *any* conditioning set $\boldsymbol{\theta}^\circ$. This immediately entails that for arbitrary $\boldsymbol{\theta}'$, $D^{\boldsymbol{\theta}'}(q||\pi) = 0$ if and only if $q_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l}) = \pi_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})$ for *all* l and for *any* choice of $\boldsymbol{\theta}'_{-l}$. In other words, the conditionals are the same. Since the positivity condition holds, we can then apply the Hammersley-Clifford Theorem to conclude that the conditionals fully specify the joint. This finally yields the desired result: $D^{\boldsymbol{\theta}'}(q||\pi) = 0$ if and only if $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$. ■

With this technical result in hand, one can now prove Corollary 18, which shows that reverse-engineering prior regularizers inspired by eq. (24) is feasible so long as the layer-specific divergences D^l are f-divergences or monotonic transformations of f-divergences.

Proof Suppressing again $\boldsymbol{Z}^l$ and $\boldsymbol{X}$ for readability, first recall that

$$q(\{\boldsymbol{U}^l\}_{l=1}^L, \{\boldsymbol{F}^l\}_{l=1}^L) = \prod_{l=1}^{L} p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}^{l-1}) q(\boldsymbol{U}^l)$$
$$p(\{\boldsymbol{U}^l\}_{l=1}^L, \{\boldsymbol{F}^l\}_{l=1}^L) = \prod_{l=1}^{L} p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}^{l-1}) p(\boldsymbol{U}^l)$$

and write for a *fixed* conditioning set $\{\boldsymbol{F}_\circ^l\}_{l=1}^L$ the new divergence

$$\begin{aligned} &D^{\{\boldsymbol{F}_\circ^l\}_{l=1}^L}\left(q(\{\boldsymbol{U}_l\}_{l=1}^L, \{\boldsymbol{F}_l\}_{l=1}^L) \| p(\{\boldsymbol{U}_l\}_{l=1}^L, \{\boldsymbol{F}_l\}_{l=1}^L)\right) \\ &= \sum_{l=1}^{L} D^l\left(p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) q(\boldsymbol{U}^l) \| p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) p(\boldsymbol{U}^l)\right) = \sum_{l=1}^{L} D^l\left(q(\boldsymbol{U}^l) \| p(\boldsymbol{U}^l)\right) \end{aligned}$$

The first equality is simply the definition of the new divergence. The second equality follows by virtue of D^l being a monotonic function of an f-divergences or an f-divergence for all l, which ensures that the l-th term is given by

$$\begin{aligned} &D^l\left(p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) q(\boldsymbol{U}^l) \| p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) p(\boldsymbol{U}^l)\right) \\ &= g\left(\mathbb{E}_{p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) p(\boldsymbol{U}^l)}\left[f\left(\frac{p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) q(\boldsymbol{U}^l)}{p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) p(\boldsymbol{U}^l)}\right)\right]\right). \\ &= g\left(\mathbb{E}_{p(\boldsymbol{U}^l)}\left[f\left(\frac{q(\boldsymbol{U}^l)}{p(\boldsymbol{U}^l)}\right)\right]\right) = D^l\left(q(\boldsymbol{U}^l) \| p(\boldsymbol{U}^l)\right). \end{aligned} \tag{36}$$

Now note that we can invoke Lemma 42: The first condition is satisfied because the derivation was independent of the chosen $\{\boldsymbol{F}_\circ^l\}_{l=1}^L$. The second condition is satisfied as both conditionals satisfy the positivity condition required for the Hammersley-Clifford Theorem to hold. ■

J.2.2 Proof of Proposition 19

Proof The likelihood is Gaussian with a fixed variance parameter σ^2, i.e. for $\boldsymbol{y}_i \in \mathbb{R}^d$ with $i = 1, 2, \ldots, n$

$$p(\boldsymbol{y}_i|\boldsymbol{f}_i^L) = (2\pi\sigma^2)^{-0.5d} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{f}_i^L)^T(\boldsymbol{y}_i - \boldsymbol{f}_i^L)\right\}$$

With this, note that integrating out the normal density yields

$$I_{p,c}(\boldsymbol{f}_i^L) = (2\pi\sigma^2)^{-0.5dc} c^{-0.5d}. \tag{37}$$

Note in particular that this is a constant and does not depend on $\boldsymbol{f}$, which makes computing the expectation over $q(\boldsymbol{f}_i^L)$ depend only on the power likelihood. Next, we show that the power likelihood is also available in closed form. This is laborious but not difficult and relies on the same algebraic tricks in the Appendix of Knoblauch et al. (2018). To simplify notation, we write $\boldsymbol{f} = \boldsymbol{f}_i^L$. Note also that the variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are (stochastic) functions of the draws of $\boldsymbol{f}_i^{1:L-1}$ from the previous layers, but we suppress this dependency, again for readability.

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\frac{1}{c}p(\boldsymbol{y}_i|\boldsymbol{f})^c\right] &= \frac{1}{c}(2\pi\sigma^2)^{-0.5dc} \cdot \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\exp\left\{-\frac{c}{2\sigma^2}(\boldsymbol{y}_i^T\boldsymbol{y}_i + \boldsymbol{f}^T\boldsymbol{f} - 2\boldsymbol{f}^T\boldsymbol{y}_i)\right\}\right] \\
&= \frac{1}{c}(2\pi\sigma^2)^{-0.5dc} \exp\left\{-\frac{c}{2\sigma^2}\boldsymbol{y}_i^T\boldsymbol{y}_i\right\} \cdot \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\exp\left\{-\frac{c}{2\sigma^2}(\boldsymbol{f}^T\boldsymbol{f} - 2\boldsymbol{f}^T\boldsymbol{y}_i)\right\}\right] \\
&= \frac{1}{c}(2\pi\sigma^2)^{-0.5dc}(2\pi\sigma^2)^{-0.5d}|\boldsymbol{\Sigma}|^{-0.5} \exp\left\{-\frac{c}{2\sigma^2}\boldsymbol{y}_i^T\boldsymbol{y}_i\right\} \times \\
&\quad \int \exp\left\{-\frac{1}{2}\left(\frac{c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{f} - \frac{2c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{y}_i + (\boldsymbol{f}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{f}-\boldsymbol{\mu})\right)\right\} d\boldsymbol{f} \\
&= \frac{1}{c}(2\pi\sigma^2)^{-0.5dc}(2\pi)^{-0.5d}|\boldsymbol{\Sigma}|^{-0.5} \exp\left\{-\frac{1}{2}\left(\frac{c}{\sigma^2}\boldsymbol{y}_i^T\boldsymbol{y}_i + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right\} \times \\
&\quad \int \exp\left\{-\frac{1}{2}\left(\frac{c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{f} - \frac{2c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{y}_i + \boldsymbol{f}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{f} - 2\boldsymbol{f}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right\} d\boldsymbol{f}
\end{aligned}$$

The integral suggests one can obtain a closed form through the Gaussian integral by completing the squares. Defining $\widetilde{\boldsymbol{\Sigma}}^{-1} = \left(\frac{c}{\sigma^s}\boldsymbol{I}_d + \boldsymbol{\Sigma}^{-1}\right)$, $\widetilde{\boldsymbol{\mu}} = \left(\frac{c}{\sigma^2}\boldsymbol{y}_i + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$, $\widehat{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\mu}}$, one indeed has

$$\begin{aligned}
\frac{c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{f} - \frac{2c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{y}_i + \boldsymbol{f}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{f} - 2\boldsymbol{f}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \boldsymbol{f}^T\left(\boldsymbol{I}_d\frac{c}{\sigma^2} + \boldsymbol{\Sigma}^{-1}\right)\boldsymbol{f} - 2\boldsymbol{f}^T\left(\frac{c}{\sigma^2}\boldsymbol{y}_i + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\
&= (\boldsymbol{f}-\widehat{\boldsymbol{\mu}})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{f}-\widehat{\boldsymbol{\mu}}) - \widetilde{\boldsymbol{\mu}}^T\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\mu}},
\end{aligned}$$

which allows us to finally rewrite the integral as

$$\begin{aligned}
&\int \exp\left\{-\frac{1}{2}\left(\frac{c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{f} - \frac{2c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{y}_i + \boldsymbol{f}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{f} - 2\boldsymbol{f}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right\} d\boldsymbol{f} \\
&= \exp\left\{-\frac{1}{2}\widetilde{\boldsymbol{\mu}}^T\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\mu}}\right\} \int \exp\left\{-\frac{1}{2}(\boldsymbol{f}-\widehat{\boldsymbol{\mu}})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{f}-\widehat{\boldsymbol{\mu}})\right\} d\boldsymbol{f} = \exp\left\{\frac{1}{2}\widetilde{\boldsymbol{\mu}}^T\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\mu}}\right\}(2\pi)^{0.5d}|\widetilde{\boldsymbol{\Sigma}}|^{0.5}.
\end{aligned}$$

Putting everthing together and simplifying expressions, this means that

$$\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\frac{1}{c}p(\boldsymbol{y}_i|\boldsymbol{f})^c\right] = \frac{1}{c}\left(2\pi\sigma^2\right)^{-0.5dc}\frac{|\widetilde{\boldsymbol{\Sigma}}|^{0.5}}{|\boldsymbol{\Sigma}|^{0.5}}\exp\left\{-\frac{1}{2}\left(\frac{c}{\sigma^2}\boldsymbol{y}_i^T\boldsymbol{y}_i + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}^T\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\mu}}\right)\right\}$$

Depending on whether one uses the β- or γ-divergence for robustifying the loss, one thus obtains the closed form expressions

$$\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[-\frac{1}{\beta-1}p(\boldsymbol{y}_i|\boldsymbol{f})^{\beta-1} + \frac{I_{p,\beta}(\boldsymbol{f})}{\beta}\right] = \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[-\frac{1}{\beta-1}p(\boldsymbol{y}_i|\boldsymbol{f})^{\beta-1}\right] + \frac{I_{p,\beta}(\boldsymbol{f})}{\beta}$$

$$\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[-\frac{1}{\gamma-1}p(\boldsymbol{y}_i|\boldsymbol{f})^{\gamma-1}\cdot\frac{\gamma}{I_{p,\gamma}(\boldsymbol{f})^{\frac{\gamma-1}{\gamma}}}\right] = \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[-\frac{1}{\gamma-1}p(\boldsymbol{y}_i|\boldsymbol{f})^{\gamma-1}\right]\cdot\frac{\gamma}{I_{p,\gamma}(\boldsymbol{f})^{\frac{\gamma-1}{\gamma}}},$$

with the expectation over $q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ as in and the integrals $I_{p,\beta}(\boldsymbol{f})$, $I_{p,\gamma}(\boldsymbol{f})$ as defined above. Note that we have derived the general case for $\boldsymbol{y}_i \in \mathbb{R}^d$, where $\boldsymbol{\Sigma}$, $\boldsymbol{f}$ and $\boldsymbol{\mu}$ are matrix- and vector-valued. ■

In fact, we can simplify everything even further in the univariate case. We summarize this in the next part.

Remark 43 *Since the derivation of Salimbeni and Deisenroth (2017) shows that one in fact only needs to integrate over the marginals $\boldsymbol{f}_i^L$, if $d = 1$ (as in all experiments in both this paper and (Salimbeni and Deisenroth, 2017)), the computation corresponding to the expression above simplifies considerably as no matrix inverses and determinants are needed. In particular, denoting the uni-variate mean and variance parameters as μ, Σ and defining $\widetilde{\Sigma} = \frac{1}{\frac{c}{\sigma^2}+\frac{1}{\Sigma}}$ and $\widetilde{\mu} = \left(\frac{cy_i}{\sigma^2} + \frac{\mu}{\Sigma}\right)$, the expectation term over the posterior q simplifies to*

$$\mathbb{E}_{q(f|\mu,\Sigma)}\left[\frac{1}{c}p(y_i|f)^c\right] = \frac{1}{c}s\left(2\pi\sigma^2\right)^{-0.5c}\sqrt{\frac{\widetilde{\Sigma}}{\Sigma}}\cdot\exp\left\{-\frac{1}{2}\left(\frac{cy_i^2}{\sigma^2} + \frac{\mu^2}{\Sigma} - \widetilde{\mu}^2\widetilde{\Sigma}\right)\right\}.$$

J.2.3 Additional experiments varying D (Figure 23)

While we showed that DGPs allow for the variation of both losses and prior regularizers, the main paper did not use the flexibility afforded by varying D. The main reason for this is that much like for the BNNs, the results when jointly varying loss and prior regularizer are less intuitively interpretable. We showcase this in Figure 23, which compares a number of different GVI posteriors for DGPs with $L = 3$ layers. The loss is either the robust loss $\mathcal{L}_p^\gamma$ for $\gamma \in \{1.01, 1.05\}$ (top 8 entries in each row) or the standard log score (bottom 4 entries in each row). We also compare $D = \frac{1}{w}\text{KLD}$ for $w = 2.0, 1.0, 0.5$ as well as the composite layer-wise divergence

$$D(q\|\pi) = \sum_{l=1}^{3} D_l(q_l\|\pi_l), \quad D_1 = D_2 = \text{KLD}, D_3 = D_{AR}^{(\alpha)} \text{ for } \alpha = 0.5.$$

Aligned with the intuition that the priors in DGPs are rather informative due to various hyperparameter optimization schemes, changing the prior regularizer from the KLD to the

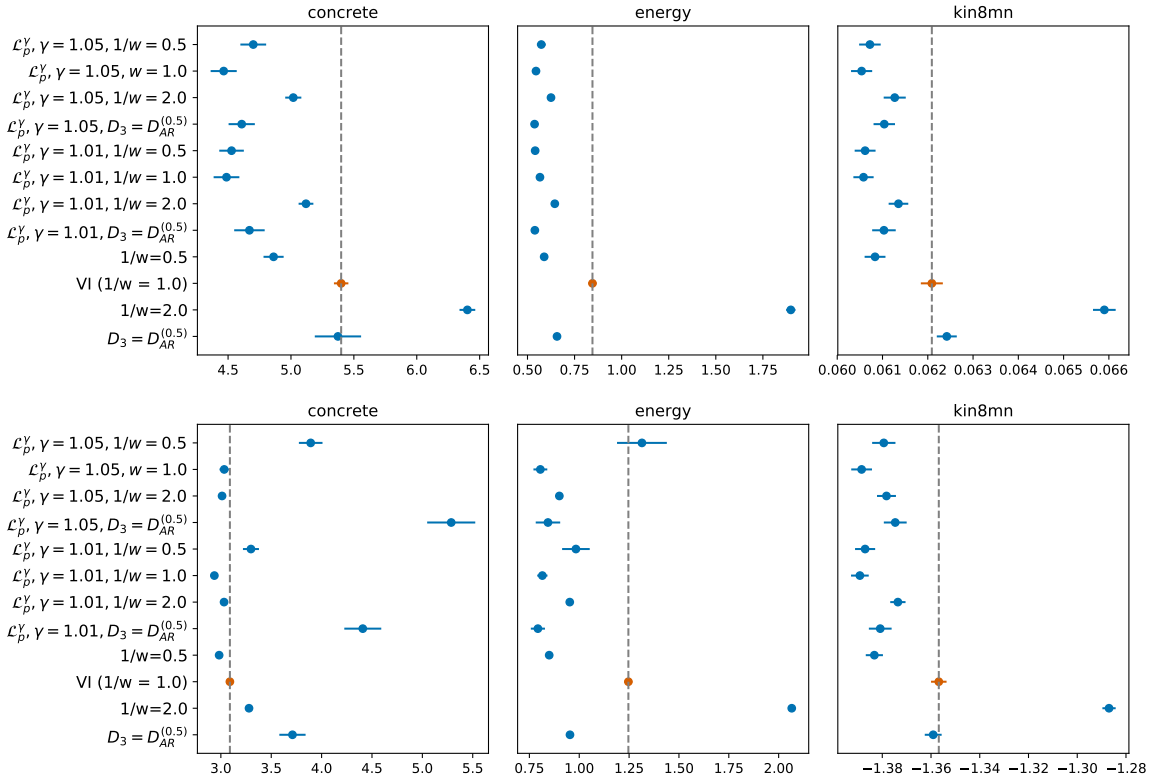


Figure 23: Best viewed in color. Top row depicts RMSE, bottom row the NLL across a range of data sets using DGPs with $L = 3$ layers. Dots correspond to means, whiskers the standard errors. The further to the left, the better the predictive performance.

$D_{AR}^{(\alpha)}$ generally typically has either fairly little or even adverse impact. Similarly, up- or down-weighting the KLD seems not to be beneficial across the board and will depend on the loss function. For the case of the log score however, we find a consistent improvement for down-weighting the KLD: Predictively, it improves the predictions on both metrics and across all data sets relative to standard VI. Similarly, up-weighting the KLD term is counterproductive under the log score and yields a performance deterioration across all data sets. This indicates that despite best efforts to the contrary, DGPs are probably violating **(P)** so that their predictive performance can be enhanced by ignoring more prior information, ensuring posteriors that are concentrated around the empirical risk minimizer.

References

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

James Aitchison. Goodness of prediction fit. *Biometrika*, 62(3):547–554, 1975.

Alexander A. Alemi. Variational predictive information bottleneck. In *Workshop on Information Theory, Advances in Neural Information Processing Systems*, 2019.

Pierre Alquier. Non-exponentially weighted aggregation: regret bounds for unbounded loss functions. *arXiv preprint arXiv:2009.03017*, 2020.

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1): 8374–8414, 2016.

Shun-ichi Amari. *Differential-geometrical methods in statistics*, volume 28. Springer Science & Business Media, 2012.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: Programmable Bayesian optimization in PyTorch. *arXiv preprint arXiv:1910.06403*, 2019.

A. Barp, F.-X. Briol, A. B. Duncan, M. Girolami, and L. Mackey. Minimum Stein discrepancy estimators. In *Neural Information Processing Systems*, pages 12964–12976, 2019.

Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

Matthew James Beal. *Variational algorithms for approximate Bayesian inference*. University College London, 2003.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian bounds based on the Rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444, 2016.

Rudolf Beran et al. Minimum hellinger distance estimates for parametric models. *The annals of Statistics*, 5(3):445–463, 1977.

James O. Berger. The case for objective Bayesian analysis. *Bayesian analysis*, 1(3):385–402, 2006.

James O Berger and José M Bernardo. On the development of the reference prior method. *Bayesian statistics*, 4(4):35–60, 1992.

James O. Berger, Elias Moreno, Luis Raul Pericchi, M. Jesus Bayarri, Jose M. Bernardo, Juan A. Cano, Julian De la Horra, Jacinto Martin, David Rios-Insua, and Bruno Betro. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.

Jose M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128, 1979.

José M. Bernardo. Bayesian theory. *Wiley Series in Probability and Statistics. 23 cm. 586 p.*, 2000.

Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesus-Maria Sanz-Serna, and Andrew Stuart. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013.

William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural computation*, 13(11):2409–2463, 2001.

Pier Giovanni Bissiri, Chris Holmes, and Stephen Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

Edwin V. Bonilla, Karl Krauth, and Amir Dezfouli. Generic inference in latent Gaussian process models. *Journal of Machine Learning Research*, 20(117):1–63, 2019.

George E. P. Box. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, pages 383–430, 1980.

F-X. Briol, A. Barp, A. B. Duncan, and M. Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv:1906.05944*, 2019.

Thang Bui, Daniel Hernández-Lobato, Jose Hernandez-Lobato, Yingzhen Li, and Richard Turner. Deep Gaussian processes for regression using approximate expectation propagation. In *International Conference on Machine Learning*, pages 1472–1481, 2016.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

François Caron, Arnaud Doucet, and Raphael Gottardo. On-line changepoint detection and parameter estimation with application to genomic data. *Statistics and Computing*, 22(2): 579–595, 2012.

Olivier Catoni. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

Liqun Chen, Chenyang Tao, Ruiyi Zhang, Ricardo Henao, and Lawrence Carin Duke. Variational inference and model selection with generalized evidence bounds. In *International Conference on Machine Learning*, pages 892–901, 2018.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. *arXiv preprint arXiv:1909.13339*, 2019a.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Finite sample properties of parametric mmd estimation: robustness to misspecification and dependence. *arXiv preprint arXiv:1912.05737*, 2019b.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015.

Andrzej Cichocki and Shun-ichi Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158, 1975.

Kurt Cutajar, Edwin V Bonilla, Pietro Michiardi, and Maurizio Filippone. Random feature expansions for deep Gaussian processes. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 884–893. JMLR, 2017a.

Kurt Cutajar, Edwin V Bonilla, Pietro Michiardi, and Maurizio Filippone. Random feature expansions for deep Gaussian processes. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 884–893. JMLR. org, 2017b.

Z. Dai, A. Damianou, J. Gonzalez, and N. Lawrence. Variational auto-encoded deep Gaussian processes. In *International Conference on Learning Representations*, 2016.

Andreas Damianou and Neil Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

G Darmois. Sur les lois de probabilité à estimation exhaustive. *Comptes Rendus de l'Académie des Sciences*, 200:1265–1266, 1935.

Herbert Aron David. First (?) occurrence of common terms in probability and statistics—a second list, with corrections. *The American Statistician*, 52(1):36–40, 1998.

Pierre-Simon De Laplace. Mémoire sur la probabilité des causes par les événements. *Mém. de math. et phys. présentés à l'Acad. roy. des sci*, 6:621–656, 1774.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, New York, 2012.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pages 2732–2741, 2017.

Justin Domke and Daniel R. Sheldon. Importance weighting and variational inference. In *Advances in neural information processing systems*, pages 4470–4479, 2018.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Paul Fearnhead and Zhen Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

Stephen E Fienberg. When did Bayesian inference become "Bayesian"? *Bayesian analysis*, 1(1):1–40, 2006.

Ronald Aylmer Fisher. Contributions to mathematical statistics. 1950.

Edwin Fong and Chris Holmes. On the marginal likelihood and cross-validation. *arXiv preprint arXiv:1905.08737*, 2019.

Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 813–822. PMLR, 2018.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049, 2010.

Jacob Gardner, Geoff Pleiss, Kilian Q. Weinberger, David Bindel, and Andrew G. Wilson. Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems*, pages 7576–7586, 2018.

Andrew Gelman, Daniel Simpson, and Michael Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555, 2017.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016.

Subhashis Ghosal. A review of consistency and convergence rates of posterior distributions. In *Proc. Varanasi Symp. on Bayesian Inference*, 1998.

Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

Abhik Ghosh and Ayanendranath Basu. Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, 2016.

Manuel Gil. *On Rényi divergence measures for continuous alphabet sources*. PhD thesis, 2011.

Manuel Gil, Fady Alajaji, and Tamas Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131, 2013.

M Goldstein. Influence and belief adjustment. *Influence Diagrams, Belief Nets and Decision Analysis*, pages 143–174, 1990.

Michael Goldstein. Subjective Bayesian analysis: principles and practice. *Bayesian Analysis*, 1(3):403–420, 2006.

Will Grathwohl, Dami Choi, Yuhuai Wu, Geoffrey Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. *arXiv preprint arXiv:1711.00123*, 2017.

Peter Grünwald. Safe learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 397–420, 2011.

Peter Grünwald. The safe Bayesian. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer, 2012.

Peter Grünwald and Thijs Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

Benjamin Guedj. A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

Oliver Hamelijnck, Theodoros Damoulas, Kangrui Wang, and Mark Girolami. Multi-resolution multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, 2019.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

Pashupati Hegde, Markus Heinonen, Harri Lähdesmäki, and Samuel Kaski. Deep learning with differential gaussian process flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1812–1821, 2019.

James Hensman and Neil D. Lawrence. Nested variational compression in deep Gaussian processes. *stat*, 1050:3, 2014.

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

José Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Daniel Hernández-Lobato, Thang D Bui, and Richard E Turner. Black-box α-divergence minimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1511–1520, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3, 2017.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Chris Holmes and Stephen Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503, 2017.

Giles Hooker and Anand N Vidyashankar. Bayesian model robustness via disparities. *Test*, 23(3):556–584, 2014.

Chin-Wei Huang, Shawn Tan, Alexandre Lacoste, and Aaron C. Courville. Improving explorability in variational inference with annealed variational objectives. In *Advances in Neural Information Processing Systems*, pages 9724–9734, 2018.

Hung Hung, Zhi-Yu Jou, and Su-Yun Huang. Robust mislabel logistic regression without modeling mislabel probabilities. *Biometrics*, 74(1):145–154, 2018.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–708, 2005.

Martin Jankowiak, Geoff Pleiss, and Jacob R Gardner. Sparse gaussian process regression beyond variational inference. *arXiv preprint arXiv:1910.07123*, 2019.

Edwin T. Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

H. Jeffreys. Theory of probability: Oxford Univ. *Press (earlier editions 1939, 1948)*, 1961.

Jack Jewson, Jim Smith, and Chris Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2013.

Jeremias Knoblauch. Frequentist consistency of generalized variational inference. *arXiv preprint arXiv:1912.04946*, 2019a.

Jeremias Knoblauch. Robust deep Gaussian processes. *arXiv preprint arXiv:1904.02303*, 2019b.

Jeremias Knoblauch and Theodoros Damoulas. Spatio-temporal Bayesian on-line changepoint detection with model selection. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2018.

Jeremias Knoblauch and Lara Vomfell. Robust bayesian inference for discrete outcomes with the total variation distance. *ArXiv*, abs/2010.13456, 2020.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Doubly robust Bayesian inference for non-stationary streaming data using β-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 64–75, 2018.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference. *arXiv preprint arXiv:1904.02063*, 2019.

Bernard Osgood Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409, 1936.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Sebastian Kurtek and Karthik Bharath. Bayesian sensitivity analysis with the Fisher–Rao metric. *Biometrika*, 102(3):601–616, 2015.

Tomasz Kuśmierczyk, Joseph Sakaya, and Arto Klami. Variational Bayesian decision-making for continuous utilities. In *Advances in Neural Information Processing Systems*, 2019.

Simon Lacoste-Julien, Ferenc Huszár, and Zoubin Ghahramani. Approximate inference for the loss-calibrated Bayesian. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 416–424, 2011.

Gaël Letarte, Pascal Germain, Benjamin Guedj, and François Laviolette. Dichotomize and generalize: Pac-bayesian binary activated deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

Moshe Lichman. UCI machine learning repository, 2013.

F Liese and I Vajda. Convex statistical distances, volume 95 of teubner texts in mathematics. *BSB BG Teubner Verlagsgesellschaft, Leipzig*, 1987.

Bruce G Lindsay et al. Efficiency versus robustness: the case for minimum hellinger distance and related methods. *The annals of statistics*, 22(2):1081–1114, 1994.

Gabriel Loaiza-Ganem and John P. Cunningham. The continuous Bernoulli: fixing a pervasive error in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2019.

Chao Ma, Yingzhen Li, and José Miguel Hernández-Lobato. Variational implicit processes. In *International Conference on Machine Learning*, pages 4222–4233. PMLR, 2019.

David J. C. MacKay. Bayesian methods for backpropagation networks. In *Models of neural networks III*, pages 211–254. Springer, 1996.

David J. C. MacKay. Choice of basis for Laplace approximation. *Machine learning*, 33(1): 77–86, 1998.

Takuo Matsubara, Chris J Oates, and François-Xavier Briol. The ridgelet prior: A covariance function approach to prior specification for bayesian neural networks. *arXiv preprint arXiv:2010.08488*, 2020.

Alexander G. de G. Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. *Journal of Machine Learning Research*, 51:231–239, 2016.

Alexander G. de G. Matthews, Mark Van Der Wilk, Tom Nickson, Keisuke Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. Gpflow: A Gaussian process library using tensorflow. *The Journal of Machine Learning Research*, 18 (1):1299–1304, 2017.

Conor Mayo-Wilson and Aditya Saraf. Qualitative robust bayesianism and the likelihood principle. *arXiv preprint arXiv:2009.03879*, 2020.

David A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999a.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170. ACM, 1999b.

Minami Mihoko and Shinto Eguchi. Robust blind source separation by beta divergence. *Neural computation*, 14(8):1859–1886, 2002.

Jeffrey W. Miller and David B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

Thomas Minka. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

Thomas P Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

Tomoyuki Nakagawa and Shintaro Hashimoto. Robust Bayesian inference via γ-divergence. *Communications in Statistics-Theory and Methods*, pages 1–18, 2019.

Eric Nalisnick, Jonathan Gordon, and José Miguel Hernández-Lobato. Predictive complexity priors. *arXiv preprint arXiv:2006.10801*, 2020.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Radford M Neal and Geoffrey E Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

Anthony O'Hagan and Jeremy E Oakley. Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering & System Safety*, 85(1):239–248, 2004.

Y. Ohnishi and J. Honorio. Novel change of measure inequalities with applications to pac-bayesian bounds and monte carlo estimation. *arXiv: Learning*, 2020.

Manfred Opper and Ole Winther. Gaussian processes for classification: Mean-field algorithms. *Neural computation*, 12(11):2655–2684, 2000.

Joseph J. K. O'Ruanaidh. *Numerical Bayesian methods applied to signal processing*. PhD thesis, University of Cambridge, 1994.

John Paisley, David M. Blei, and Michael I. Jordan. Variational Bayesian inference with stochastic search. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1363–1370, 2012.

Francesco Pauli, Walter Racugno, and Laura Ventura. Bayesian composite marginal likelihoods. *Statistica Sinica*, pages 149–164, 2011.

Fengchun Peng and Dipak K Dey. Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics*, 23(2):199–213, 1995.

E.J.G. Pitman. Sufficient statistics and intrinsic accuracy. *Proceedings of the Cambridge Philosophical Society*, 32, 1936.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(Dec): 1939–1959, 2005.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, pages 496–504, 2016.

Jean-Baptiste Regli and Ricardo Silva. Alpha-beta divergence for variational inference. *arXiv preprint arXiv:1805.01045*, 2018.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 1530–1538, 2015.

Mathieu Ribatet, Daniel Cooley, and Anthony C Davison. Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, pages 813–845, 2012.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

Gareth O Roberts, Andrew Gelman, and Walter R Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The annals of applied probability*, 7(1): 110–120, 1997.

Simone Rossi, Sebastien Marmin, and Maurizio Filippone. Walsh-Hadamard variational inference for Bayesian deep learning. *arXiv preprint arXiv:1905.11248*, 2019a.

Simone Rossi, Pietro Michiardi, and Maurizio Filippone. Good initializations of variational Bayes for deep models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5487–5497, 2019b.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

Yunus Saatçi, Ryan D. Turner, and Carl E. Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning*, pages 927–934, 2010.

Abhijoy Saha, Karthik Bharath, and Sebastian Kurtek. A geometric variational approach to bayesian inference. *Journal of the American Statistical Association*, pages 1–25, 2019.

Tim Salimans and David A Knowles. On using control variates with stochastic approximation for variational Bayes and its connection to stochastic linear regression. *arXiv preprint arXiv:1401.1022*, 2014.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

John Shawe-Taylor and Robert C Williamson. A PAC analysis of a Bayesian estimator. In *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*, volume 6, pages 2–9, 1997.

Xiaotong Shen and Larry Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. Kernel implicit variational inference. In *International Conference on Learning Representations*, 2018.

Zhenming Shun and Peter McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(4):749–760, 1995.

Douglas G Simpson. Minimum hellinger distance estimation for the analysis of count data. *Journal of the American statistical Association*, 82(399):802–807, 1987.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264, 2006.

Elliott Sober. *Evidence and evolution: The logic behind the science*. Cambridge University Press, 2008.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in neural information processing systems*, pages 3738–3746, 2016.

Nicholas Syring and Ryan Martin. Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486, 2019.

Roy N Tamura and Dennis D Boos. Minimum hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81(393):223–229, 1986.

Louis C Tiao, Edwin V Bonilla, and Fabio Ramos. Cycle-consistent adversarial learning as approximate bayesian inference. *arXiv preprint arXiv:1806.01771*, 2018.

Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86, 1986.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*, pages 1971–1979, 2014.

Udo v Toussaint, Silvio Gori, and Volker Dose. Invariance priors for bayesian feed-forward neural networks. *Neural Networks*, 19(10):1550–1557, 2006.

Dustin Tran, Rajesh Ranganath, and David M Blei. The variational gaussian process. In *4th International Conference on Learning Representations, ICLR 2016*, 2016.

Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533, 2017.

John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. In *Bayesian time series models*. Cambridge University Press, 2011.

Ryan D. Turner, Steven Bottone, and Clay J. Stanek. Online variational approximations to non-exponential family change point models: with application to radar tracking. In *Advances in Neural Information Processing Systems*, pages 306–314, 2013.

Keyon Vafa. Training deep Gaussian processes with sampling. In *NIPS 2016 Workshop on Advances in Approximate Bayesian Inference*, 2016.

Tim Van Erven and Peter Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42, 2011.

Stephen Walker. New approaches to Bayesian consistency. *The Annals of Statistics*, 32(5): 2028–2043, 2004.

Dilin Wang, Hao Liu, and Qiang Liu. Variational inference with tail-adaptive f-divergence. In *Advances in Neural Information Processing Systems*, pages 5742–5752, 2018.

Ke Alexander Wang, Geoff Pleiss, Jacob R Gardner, Stephen Tyree, Kilian Q Weinberger, and Andrew Gordon Wilson. Exact Gaussian processes on a million data points. *arXiv preprint arXiv:1903.08114*, 2019.

Yali Wang, Marcus Brubaker, Brahim Chaib-Draa, and Raquel Urtasun. Sequential inference for deep Gaussian process. In *Artificial Intelligence and Statistics*, pages 694–703, 2016.

Sumio Watanabe. *Mathematical Theory of Bayesian Statistics*. CRC Press, 2018.

Christopher KI Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.

Robert C Wilson, Matthew R Nassar, and Joshua I Gold. Bayesian online learning of the hazard rate in change-point problems. *Neural computation*, 22(9):2452–2476, 2010.

Mike Wu, Noah Goodman, and Stefano Ermon. Differentiable antithetic sampling for variance reduction in stochastic variational inference. In *Proceedings of Machine Learning Research*, volume 89, pages 2877–2886, 2019.

Yue Yang, Ryan Martin, and Howard Bondell. Variational approximations using Fisher divergence. *arXiv preprint arXiv:1905.05284*, 2019.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *arXiv preprint arXiv:1710.03266*, 2017.

Yannis G Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, pages 768–774, 1985.

Arnold Zellner. Maximal data information prior distributions. *New developments in the applications of Bayesian methods*, pages 211–232, 1977.

Arnold Zellner. Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):278–280, 1988.

Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. Noisy natural gradient as variational inference. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5852–5861, 2018.

Tong Zhang. From ϵ-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

Jun Zhu, Ning Chen, and Eric P Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1): 1799–1847, 2014.