

---

# Generalized Variational Inference

---

**Jeremias Knoblauch**

The Alan Turing Institute  
Dept. of Statistics  
University of Warwick  
j.knoblauch@warwick.ac.uk

**Jack Jewson**

The Alan Turing Institute  
Dept. of Statistics  
University of Warwick  
j.e.jewson@warwick.ac.uk

**Theodoros Damoulas**

The Alan Turing Institute  
Depts. of Computer Science & Statistics  
University of Warwick  
t.damoulas@warwick.ac.uk

## Abstract

This paper introduces a generalized representation of Bayesian inference. It is derived axiomatically, recovering existing Bayesian methods as special cases. We then use it to prove that variational inference (VI) based on the Kullback-Leibler Divergence with a variational family  $\mathcal{Q}$  produces the uniquely optimal  $\mathcal{Q}$ -constrained approximation to the exact Bayesian inference problem. Surprisingly, this implies that standard VI dominates any other  $\mathcal{Q}$ -constrained approximation to the exact Bayesian inference problem. This means that alternative  $\mathcal{Q}$ -constrained approximations such as VI minimizing other divergences [e.g. 49, 4, 66, 19, 5] and Expectation Propagation [e.g. 60, 63, 31] can produce better posteriors than VI only by *implicitly* targeting more appropriate Bayesian inference problems. Inspired by this, we introduce Generalized Variational Inference (GVI), a modular approach for instead solving such alternative inference problems *explicitly*. We explore some applications of GVI, including robustness and better marginals. Lastly, we derive black box GVI and apply it to Bayesian Neural Networks and Deep Gaussian Processes, where GVI can comprehensively outperform competing methods.

## 1 Introduction

Bayesian methods are becoming ever more popular in statistical machine learning because they provide uncertainty quantification as part of their inferences. Their guiding principle is to choose a log-likelihood-based loss  $\ell_n(\boldsymbol{\theta}, \mathbf{x}) = \sum_{i=1}^n -\log(p(x_i|\boldsymbol{\theta}))$  relating  $n$  observations  $\mathbf{x} = (x_1, x_2, \dots, x_n)'$  to a parameter  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$  and a prior belief  $\pi$  over  $\boldsymbol{\Theta}$ . Together,  $\ell_n$  and  $\pi$  define the Bayesian posterior  $q^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^n p(x_i|\boldsymbol{\theta})$ , while on its own  $\ell_n$  defines the parameter  $\boldsymbol{\theta}$  of interest. Accordingly, its minimizer is the in-sample optimal Maximum Likelihood Estimator. Complementing this, the prior belief  $\pi$  over  $\boldsymbol{\theta}$  prevents overfitting and induces uncertainty about the optimum. This division of labour between  $\ell_n$  and  $\pi$  is clearest in Zellner [87], where it is shown that  $q^*$  solves

$$\arg \min_{q \in \mathcal{P}(\boldsymbol{\Theta})} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[ - \sum_{i=1}^n \log(p(x_i|\boldsymbol{\theta})) \right] + \text{KLD}(q||\pi) \right\}, \quad \text{KLD}(q||\pi) = \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right], \quad (1)$$

for  $\mathcal{P}(\boldsymbol{\Theta})$  the set of all probability distributions on  $\boldsymbol{\Theta}$  and KLD the Kullback-Leibler divergence. In the remainder, we take inspiration from eq. (1) to make the following contributions:

- (1) We decompose Bayesian inference methods as  $P(\ell_n, D, \Pi)$  into three clearly interpretable components: The loss  $\ell_n$ , the uncertainty quantifier  $D$  and the admissible posteriors  $\Pi$ ;

- (2) We show that  $\ell_n$ ,  $\Pi$  and  $D$  must relate to each other via eq. (2);
- (3) We prove that standard Variational Inference (**VI**) with a variational family  $\Pi = \mathcal{Q}$  produces the optimal  $\mathcal{Q}$ -constrained Bayesian posterior. Thus, alternative approximations can only outperform standard VI by *implicitly* targeting a different Bayesian inference problem.
- (4) We introduce Generalized VI (**GVI**) to *explicitly* target these inference problems by embedding desirable properties of the posterior directly into  $P(\ell_n, D, \mathcal{Q})$ . To illustrate GVI's modular logic and flexibility, we show how it can be used to improve marginal variances as well as provide robustness to badly specified priors and models.
- (5) We derive black box GVI inference, which contains black box VI [65] as a special case.
- (6) To demonstrate GVI's benefits, we apply it to Bayesian Neural Nets [62] and Deep Gaussian Processes [17], where it substantially outperforms competing methods.

All proofs, definitions for divergences and a full exposition can be found in the Appendix.

## 2 The Bayesian inference problem

In this paper, we axiomatically derive and motivate a generalized representation of Bayesian inference. In particular, we argue that Bayesian inference should take the form  $P(\ell_n, D, \Pi)$  given by

$$q^*(\theta) = \arg \min_{q \in \Pi} \{L(q|\mathbf{x}, \ell_n, D)\}; \quad L(q|\mathbf{x}, \ell_n, D) = \mathbb{E}_{q(\theta)} [\ell_n(\theta, \mathbf{x})] + D(q|\pi), \quad (2)$$

where the constituent parts of the form  $P(\ell_n, D, \Pi)$  are given by

- a **loss**  $\ell_n$  linking a parameter of interest  $\theta$  to the observations  $\mathbf{x} = x_{1:n}$ . Throughout, we will assume additivity, i.e.  $\ell_n(\theta, \mathbf{x}) = \sum_{i=1}^n \ell(\theta, x_i)$  for some  $\ell$ .
- a divergence  $D : \mathcal{P}(\Theta) \times \mathcal{P}(\Theta) \rightarrow \mathbb{R}_+$  regularizing the posterior with respect to the prior  $\pi$ . As  $D$  determines how uncertainty in  $q^*(\theta)$  is quantified, we call it **uncertainty quantifier**.
- a set of **admissible posteriors**  $\Pi \subseteq \mathcal{P}(\Theta)$  the regularized expected loss is minimized over.

Eq. (1) shows that standard Bayesian inference solves  $P(-\sum_{i=1}^n \log(p(x_i|\theta)), \text{KLD}, \mathcal{P}(\Theta))$ . Further, for  $\mathcal{Q}$  a variational family, the objective of  $P(-\sum_{i=1}^n \log(p(x_i|\theta)), \text{KLD}, \mathcal{Q})$  in eq. (2) is the Evidence Lower Bound (ELBO) of VI. Throughout, we refer to any problem of form  $P(\ell_n, D, \mathcal{Q})$  as a **Generalized Variational Inference (GVI)** problem. Section 4 in the Appendix elaborates on different choices for  $D$  in more detail, but we focus on Rényi's  $\alpha$ -divergence as well as the  $\alpha$ -,  $\beta$ - and  $\gamma$ -divergences and refer to them as  $D_{AR}^{(\alpha)}$ ,  $D_A^{(\alpha)}$ ,  $D_B^{(\beta)}$ ,  $D_G^{(\gamma)}$ . We also assume that for a given  $P(\ell_n, D, \Pi)$ , the minimizers  $\hat{\theta}_n = \arg \min_{\theta} \ell_n(\theta, \mathbf{x})$  and  $\theta^* = \arg \min_{\theta} \mathbb{E}_{\mathbf{X}} [\ell(\theta, \mathbf{X})]$  exist.

### 2.1 Observations about $\ell_n$ , $D$ and $\Pi$

**Loss  $\ell_n$ :** In traditional Bayesian inference,  $\ell_n$  is the negative log likelihood. Yet, this is just a special case [12] and  $\ell_n(\theta, \mathbf{x})$  can be *any* loss about whose optimum  $\theta^*$  one learns in a Bayesian manner.

**Uncertainty quantifier  $D$ :** In contrast to Maximum Likelihood inference, Bayesian methods provide uncertainty quantification about  $\theta^*$ . Specifically, uncertainty about  $\theta^*$  is quantified by penalizing how far the posterior  $q$   $D$ -diverges from the prior  $\pi$ . To the best of our knowledge, we are first to consider inference with  $D \neq \text{KLD}$ . This is because *exact* Bayesian inference (i.e.,  $\Pi = \mathcal{P}(\Theta)$ ) is coherent only if  $D = \text{KLD}$  [12]. Yet, once  $\Pi = \mathcal{Q}$  is some constrained subset of  $\mathcal{P}(\Theta)$ , this is no longer a valid reason to restrict attention to  $D = \text{KLD}$ : So long as  $\mathcal{Q}$  is not rich enough to contain the exact posterior form, VI is *never* coherent, regardless of  $D$ . For a thorough explanation of the relationship between coherence and the KLD, see Bissiri et al. [12].

**Admissible posteriors:** If  $\Pi = \mathcal{Q}$  is a variational family, eq. (2) produces the optimal posterior in the constrained set  $\mathcal{Q}$  for given  $\ell_n$  and  $D$ . This means that eq. (2) produces the *optimal*  $\mathcal{Q}$ -constrained posterior relative to the standard Bayesian problem  $P(\ell_n, \text{KLD}, \mathcal{P}(\Theta))$  for standard VI (Thm. 4).

### 2.2 Construction of the form $P(\ell_n, D, \Pi)$

This section axiomatically derives  $P(\ell_n, D, \Pi)$  in eq. (2). For simplicity,  $\mathbf{x} = x_{1:n}$  are treated as  $n$  independent draws, but the presented arguments extend to conditional independence structures.

**Axiom 1** (Representation). Bayesian inference infers posteriors  $q$  on  $\Theta$  by (i) measuring how  $q$  fits a sample  $\mathbf{x}$  via the expectation of a loss  $\ell_n(\theta, \mathbf{x})$ , (ii) quantifying uncertainty about  $\theta^*$  via a divergence  $D$  between prior  $\pi$  and  $q$ , (iii) optimizing  $q$  over a space of probability distributions  $\Pi$  on  $\Theta$ .

This axiom formalizes Bayesian inference inspired by eq. (1), implying that it is representable as a triplet  $P(\ell_n, D, \Pi)$ . Showing that  $P(\ell_n, D, \Pi)$  takes the form in eq. (2) requires three more axioms.

**Axiom 2** (Information difference).  $P(\ell_n, D, \Pi)$  produces different posteriors for  $\mathbf{x} = x_{1:n}$  and  $\mathbf{x}' = x_{1:n+m}$  if there is an information difference of  $\mathbf{x}'$  relative to  $\mathbf{x}$ , i.e. if  $\ell_n(\theta, \mathbf{x}) \neq \ell_{n+m}(\theta, \mathbf{x}')$ .

**Axiom 3** (Prior regularization).  $q$  is regularized against  $\pi$  by penalizing the divergence  $D(q||\pi)$ .

**Axiom 4** (Translation Invariance). For constant  $C$  and  $\ell'_n = \ell_n + C$ ,  $P(\ell'_n, D, \Pi) = P(\ell_n, D, \Pi)$ .

Axiom 2 ensures that different information about  $\theta$  produces different posteriors. Axiom 3 says that  $D(q||\pi)$  acts as penalty. Axiom 4 enforces that adding constants to  $\ell_n$  does not affect inference. Note that we do *not* want invariance to multiplications of  $\ell_n$ , as this contradicts Axiom 2 for additive  $\ell_n$ :

**Theorem 1.** If Axiom 2 holds,  $\ell_n$  is additive and  $C \in \mathbb{N}$ ,  $P(\ell_n, D, \Pi) \neq P(C \cdot \ell_n, D, \Pi)$ .

Finally, we motivate the form of  $P(\ell_n, D, \Pi)$  in eq. (2). Since the additivity of eq. (1) and the ELBO both rely on log-additivity via  $D = \text{KLD}$ , we require more fundamental arguments for general  $D$ .

**Theorem 2** (Form 1). If Axiom 1 holds,  $P(\ell_n, D, \Pi)$  can be written as  $\arg \min_{q \in \Pi} \{L(q|\mathbf{x}, \ell_n, D)\}$  for  $L(q|\mathbf{x}, \ell_n, D) = f(\mathbb{E}_{q(\theta)}[\ell_n(\theta, \mathbf{x})], D(q||\pi))$ , where  $f$  is some function  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ .

**Theorem 3** (Form 2). For  $P(\ell_n, D, \Pi)$  being  $\arg \min_{q \in \Pi} \{L(q|\mathbf{x}, \ell_n, D)\}$  and  $\circ$  an elementary operation on  $\mathbb{R}$ ,  $L(q|\mathbf{x}, \ell_n, D) = \mathbb{E}_{q(\theta)}[\ell_n(\theta, \mathbf{x})] \circ D(q||\pi)$  satisfies Axioms 3 and 4 only if  $\circ = +$ .

## 2.3 Relationships to existing Bayesian methods

Most Bayesian methods are special cases of  $P(\ell_n, D, \Pi)$  with additive losses  $\ell_n$ , see Table 1. Section 3 in the Appendix gives a full exposition. As we derived  $P(\ell_n, D, \Pi)$  for arbitrary losses  $\ell_n$ , we draw attention to Generalized Bayesian Inference [12], which theoretically justifies so-called Gibbs posteriors. Specifically, for additive losses  $\ell_n$ ,  $P(\ell_n, \text{KLD}, \mathcal{P}(\Theta))$  is coherently solved with

$$q^*(\theta) = \frac{\pi(\theta) \exp \left\{ \sum_{i=1}^n -\ell(\theta, x_i) \right\}}{\int_{\Theta} \pi(\theta) \exp \left\{ \sum_{i=1}^n -\ell(\theta, x_i) \right\} d\theta}. \quad (3)$$

In Sections 4.1–4.2 of the Appendix, we show that analogously to VI, GVI for  $D \in \{D_B^{(\beta)}, D_G^{(\gamma)}, D_{AR}^{(\alpha)}\}$  optimizes a lower bound on the evidence corresponding to the generalized posterior of eq. (3).

While some approximation methods take the form  $P(\ell_n, D, \Pi)$ , Laplace approximations [70], Expectation Propagation (EP) [60, 63, 58, 31] and many variational methods [e.g. 19, 49, 67, 4, 81, 66, 71, 5, 85] do not. Variational methods not of form  $P(\ell_n, D, \Pi)$  construct posteriors via

Method	$\ell(\theta, x_i)$	$D$	$\Pi$
Standard Bayes	$-\log(p(\theta x_i))$	KLD	$\mathcal{P}(\Theta)$
Generalized Bayes <sup>1</sup>	any $\ell$	KLD	$\mathcal{P}(\Theta)$
Power Bayes <sup>2</sup>	$-\log(p(\theta x_i))$	$\frac{1}{w}$ KLD, $w < 1$	$\mathcal{P}(\Theta)$
Divergence Bayes <sup>3</sup>	divergence-based $\ell$	KLD	$\mathcal{P}(\Theta)$
<b>Standard VI</b>	$-\log(p(\theta x_i))$	KLD	$\mathcal{Q}$
Power VI <sup>4</sup>	$-\log(p(\theta x_i))$	$\frac{1}{w}$ KLD, $w < 1$	$\mathcal{Q}$
Regularized Bayes <sup>5</sup>	$-\log(p(\theta x_i)) + \phi(\theta, x_i)$	KLD	$\mathcal{Q}$
$(\beta\text{-})\text{VAE}$ <sup>6</sup>	$-\log(p_{\zeta}(x_i \theta))$	$\beta \cdot \text{KLD}$ , $\beta > 1$	$\mathcal{Q}$
Gibbs VI <sup>7</sup>	any $\ell$	KLD	$\mathcal{Q}$
<b>Generalized VI</b>	any $\ell$	any $D$	$\mathcal{Q}$

Table 1:  $P(\ell_n, D, \mathcal{Q})$  and relation to some existing methods. All losses have the form  $\ell_n(\theta, \mathbf{x}) = \sum_{i=1}^n \ell(\theta, x_i)$  for some  $\ell(\theta, x_i)$ . <sup>1</sup>[12], <sup>2</sup>[e.g. 34, 28, 57], <sup>3</sup>[e.g. 35, 24, 22, 40], <sup>4</sup>[e.g. 86, 36] <sup>5</sup>[23], but only if the regularizer can be written as  $\mathbb{E}_{q(\theta)}[\phi(\theta, \mathbf{x})]$  as in [88], <sup>6</sup>[44, 32], <sup>7</sup>[e.g. 2, 22]

$\tilde{q} = \arg \min_{q \in \mathcal{Q}} \{F(q|q^*)\}$ , where  $q^*$  is the exact posterior and  $F$  a (local or global) discrepancy measure. We call these methods **F-Variational Inference (F-VI)** and note their three disadvantages:

- (1) If  $F \neq \text{KLD}$ , F-VI **violates Axioms 1–4**.
- (2) F-VI with a variational family  $\mathcal{Q}$  constructs **provably suboptimal**  $\mathcal{Q}$ -constrained **approximations** to its exact target  $P(\ell_n, \text{KLD}, \mathcal{P}(\Theta))$  relative to standard VI (Thm. 4).
- (3) F-VI **conflates the effects of  $\ell_n$  and  $D$**  because it induces desirable properties for the posterior through  $F$  rather than through the clearly interpretable modularity of  $P(\ell_n, D, \mathcal{Q})$ .

In contrast, **GVI satisfies Axioms 1–4** and encodes a clear separation of responsibilities. Further, as GVI solves  $P(\ell_n, D, \mathcal{Q})$  directly, it produces the **optimal**  $\mathcal{Q}$ -constrained **posteriors for  $\ell_n$  and  $D$** . Proving the practical implications of these points, the remainder of this paper proceeds as follows: First, we emphasize the special role of standard VI by proving its optimality relative to eq. (1). Next, we introduce model-robust, prior-robust and zero-avoiding GVI to showcase its flexibility.

### 3 Optimality of standard Variational Inference (VI) and consequences

While standard VI was introduced as an approximation technique for the exact Bayesian posterior [e.g., 41, 7], we show its optimality: no other  $\mathcal{Q}$ -constrained approximation can produce a better posterior relative to the exact Bayesian inference problem representable as  $P(\ell_n, \text{KLD}, \mathcal{P}(\Theta))$ .

**Theorem 4** (VI: Uniquely optimal approximation). For exact and coherent Bayesian posteriors solving  $P(\ell_n, \text{KLD}, \mathcal{P}(\Theta))$  and a fixed variational family  $\mathcal{Q}$ , **standard VI** produces the uniquely **optimal  $\mathcal{Q}$ -constrained approximation** to  $P(\ell_n, \text{KLD}, \mathcal{P}(\Theta))$ : Having decided on approximating the Bayesian posterior with some  $q \in \mathcal{Q}$ , VI provides the uniquely optimal solution.

Thm. 4 implies that for a fixed variational family  $\mathcal{Q}$ , standard VI dominates all other  $\mathcal{Q}$ -constrained posteriors. In other words, alternatives like F-VI and EP produce worse approximations to eq. (1). This seems like a paradox: Alternative  $\mathcal{Q}$ -constrained F-VI methods [60, 58, 19, 31, 49, 67] are popular precisely because they produce better posteriors than VI in certain situations. To resolve this paradox, note that VI optimality holds only relative to the exact Bayesian inference problem  $P(\ell_n, \text{KLD}, \mathcal{P}(\Theta))$ . Reversing this logic, alternative  $\mathcal{Q}$ -constrained methods such as **F-VI** must produce more adequate posteriors than standard VI precisely *because they target a different* and more appropriate Bayesian **inference problem**. Yet, they do so *implicitly*, i.e. without specifying a new loss  $\ell'_n$  and uncertainty quantifier  $D'$ . **GVI** instead specifies and solves  $P(\ell'_n, D', \mathcal{Q})$  *explicitly*.

### 4 Generalized Variational Inference (GVI)

Inspired by VI optimality and the form  $P(\ell_n, D, \mathcal{Q})$ , we introduce GVI. GVI is provably modular and incorporates desirable properties of  $\mathcal{Q}$ -constrained posteriors explicitly via  $\ell_n$  and  $D$  (Thm. 5).

**Definition 1** (GVI). Any Bayesian inference method solving  $P(\ell_n, D, \mathcal{Q})$  with admissible choices  $\ell_n$ ,  $D$  and  $\mathcal{Q}$  is a Generalized Variational Inference (GVI) method satisfying Axioms 1 – 4.

**Theorem 5** (GVI modularity). For Bayesian inference via some GVI method  $P(\ell_n, D, \mathcal{Q})$ , making it robust to model misspecification amounts to changing  $\ell_n$ . Conversely, adapting its uncertainty quantification (for fixed  $\mathcal{Q}$ ,  $\pi$ ,  $\theta^*$ ,  $\hat{\theta}_n$ ) amounts to changing  $D$ .

In the remainder, we focus on three applications of GVI illustrated in Fig. 1: robustness to model misspecification (via  $\ell_n$ ), conservative marginal variances (via  $D$ ) and prior robust inference (via  $D$ ).

#### 4.1 Robustness to model misspecification and outliers

Model misspecification and outliers occur relative to  $\ell_n$ . Likelihood-free robust losses are well-understood [38, 37] and useful for Bayesian classification [2] or regression [28]. E.g.,  $\ell_n(\theta, \mathbf{x}) = \sum_{i=1}^n |x_i - \theta|$  measures the central tendency  $\theta$  of  $\mathbf{X}$  more robustly than  $\ell_n(\theta, \mathbf{x}) = \sum_{i=1}^n (x_i - \theta)^2$ . Model-free robustness is attractive, but Bayesian methods typically infer full probabilistic models. Thus, robustness is often encoded in the model – e.g., by replacing normal with heavy-tailed errors.

More recently, likelihood-based but *model-agnostic* robustification strategies have been proposed [35, 24, 40, 46, 22, 39, 61]. First, note that for  $X \sim g$ , minimizing  $-\log(p(x_i|\theta))$  with respect

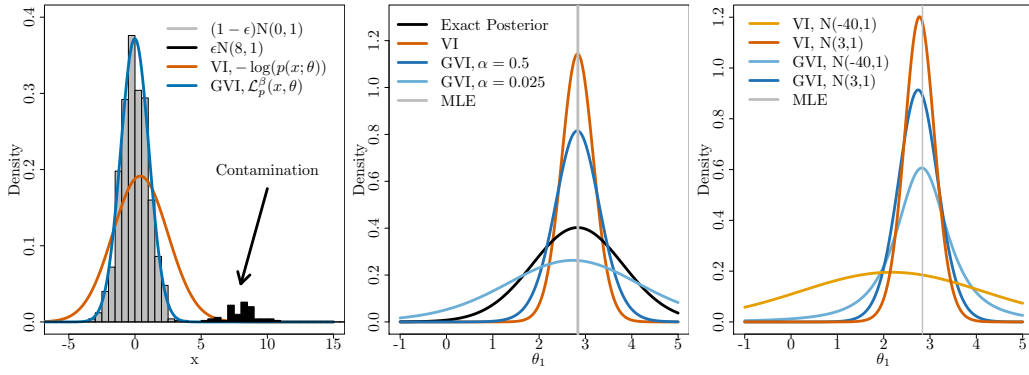


Figure 1: Illustration of three use cases for GVI: robustness to misspecification, improved marginal variances and prior robustness. We compare **standard VI** to **GVI** posteriors with Gaussian likelihoods and mean-field Gaussian approximations. **Left:** Transforming the loss provides robustness against model misspecification. Depicted are posterior predictives under  $\varepsilon = 5\%$  outlier contamination using **VI** and  $P(\sum_{i=1}^n \mathcal{L}_p^\beta(\theta, x_i), \text{KLD}, \mathcal{Q})$ ,  $\mathcal{L}_p^\beta(\theta, x_i)$  as in eq. (4) for  $\beta = 1.5$ . **Center:** Changing  $D$  improves marginal variances. Depicted are exact and approximate marginals. The exact posterior is correlated, causing **VI** to over-concentrate. **GVI** has the flexibility to avoid this by solving  $P(\ell_n, D_{AR}^{(\alpha)}, \mathcal{Q})$ , where  $D_{AR}^{(\alpha)}$  is Rényi’s  $\alpha$ -divergence. **Right:** Changing  $D$  provides prior robustness. Depicted are approximate marginals for two different priors  $\pi \in \{N(-40, 1), N(3, 1)\}$ . **VI** is sensitive to the badly specified prior. **GVI** avoids this by solving  $P(\ell_n, D_{AR}^{(\alpha)}, \mathcal{Q})$  with  $\alpha = 0.5$ .

to  $\theta$  minimizes the KLD between  $p(\cdot|\theta)$  and  $g$ :  $\frac{1}{n} \sum_{i=1}^n -\log(p(x_i|\theta)) \approx \mathbb{E}_X[-\log(p(x|\theta))] = \text{KLD}(g||p(\cdot|\theta)) - \mathbb{E}_X[\log(g(x))]$ . However, the KLD is not a robust measure of discrepancy, see Figs. 1, 2. To address this, model-based losses minimizing more robust divergences can be derived [40]. Popular choices for this are  $\gamma$ - and  $\beta$ -divergences [see e.g. 6, 21, 39], whose losses are

$$\mathcal{L}_p^\beta(\theta, x_i) = -\frac{1}{\beta-1} p(x_i|\theta)^{\beta-1} + \frac{I_{p,\beta}(\theta)}{\beta}, \quad \mathcal{L}_p^\gamma(\theta, x_i) = -\frac{1}{\gamma-1} p(x_i|\theta)^{\gamma-1} \frac{\gamma}{I_{p,\gamma}(\theta)^{\frac{\gamma-1}{\gamma}}} \quad (4)$$

where  $I_{p,c}(\theta) = \int p(z|\theta)^c dz$ . The losses  $\mathcal{L}_p^\beta(\theta, x_i)$ ,  $\mathcal{L}_p^\gamma(\theta, x_i)$  are both based on a probabilistic model as well as naturally robust to outliers and misspecification [see 22, 46, 24, 46]. Fig. 2 illustrates how such losses guard against outliers and model misspecification via influence functions [48].

## 4.2 Better approximate marginals and zero-avoiding behaviour

Methods using  $D = \text{KLD}$  can suffer from over-concentration, a property also called *zero-forcing* [e.g. in 59, 60, 58, 49, 19]. This is particularly pronounced for mean-field VI [79], but it also adversely affects *exact* Bayesian inference under model misspecification [e.g. 34, 28, 57, 82].

Numerous F-VI procedures address this with *zero-avoiding* divergences [e.g. 59, 60, 58, 31, 49, 19, 81]. GVI instead avoids over-concentration using an uncertainty quantifier  $D$  that is more conservative than the KLD. E.g., one can use  $\frac{1}{w}$  KLD for some  $0 < w < 1$ , which is equivalent to using a  $w$ -power likelihood [see also 86]. Other candidates are Rényi’s  $\alpha$ -divergence ( $D_{AR}^{(\alpha)}$ ), the  $\beta$ -divergence ( $D_B^{(\beta)}$ ) or the  $\gamma$ -divergence ( $D_G^{(\gamma)}$ ) with their hyperparameters in  $(0, 1)$ . Note that these divergences approach the KLD as their hyperparameters go to 1, see Fig. 2. Unlike  $\frac{1}{w}$  KLD, they provide not only more conservative marginal variances, but also prior robustness, see Fig. 1 and Section 4.4 in the Appendix.

## 4.3 Robustness to the prior

While the literature on VI has primarily focused on robustness to model misspecification [67, 22] and zero-avoiding methods [63, 60, 58, 31, 49], robustness to the prior [see 9, 10] has been addressed only very recently [27]. For GVI, it is clear that robustness to the prior  $\pi$  should enter via the uncertainty quantifier  $D$ , see Fig. 1. There is no shortage of divergences for robust inference [16], and we identify  $D_{AR}^{(\alpha)}$ ,  $D_B^{(\beta)}$  and  $D_G^{(\gamma)}$  as especially suitable for prior-robust inference, see Section 4.4 of the Appendix.

## 4.4 Architecture-specific GVI

GVI is adaptable to many challenges in inference. Excellent work abounds on changing  $\ell_n$  [12, 40, 6, 24, 22, 39] or  $\Pi$  [69, 43, 8, 77, 73, 71]. Thm. 6 illustrates GVI’s additional flexibility due to  $D$ .

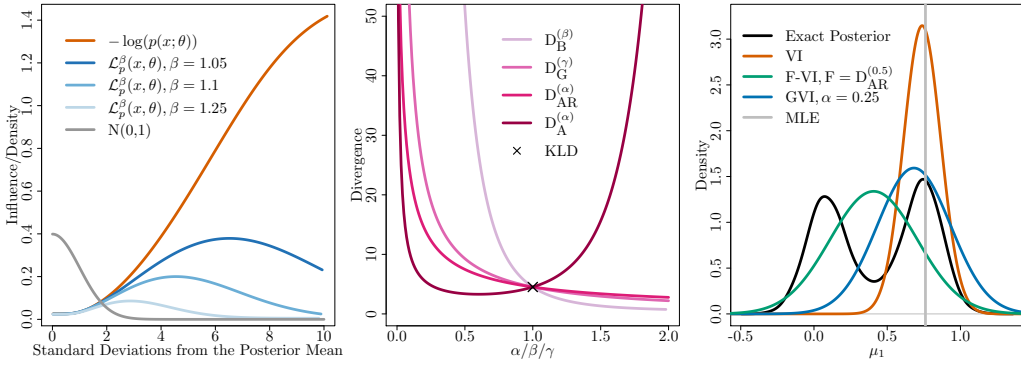


Figure 2: Illustrations of different model-based losses (Left), uncertainty quantifiers (Center) and behaviours of **GVI**, **F-VI** and **VI** (Right). **Left:** From [46]. Depicted is the influence [see 48] of  $x_i$  on exact posteriors for robust and non-robust losses. Higher influence is assigned for  $-\log(p(x_i, \theta))$  the more unlikely  $x_i$  is under the current model. In contrast,  $\mathcal{L}_p^\beta(\theta, x_i)$  guards against assigning the highest influence to outliers. **Center:** Magnitude of the penalty incurred by  $D(q||\pi)$  for different uncertainty quantifiers  $D$  and fixed densities  $\pi, q$ . **Right:** Exact, **VI**, **F-VI** ( $F = D_{AR}^{(0.5)}$ ) and  $P(\ell_n, D_{AR}^{(\alpha)}, \mathcal{Q})$  based **GVI** marginals of the location in a 2 component mixture model. Respecting  $\ell_n$ , **VI** and **GVI** provide uncertainty quantification around the most likely value  $\hat{\theta}_n$  via  $D$ . In contrast, **F-VI** implicitly changes the loss and has a mode at the locally most *unlikely* value of  $\theta$ .

**Theorem 6** (Divergence recombination). Let  $D_l$  be divergences and  $c_l \geq 0$  scalars for  $l = 1, 2, \dots, L$ . The following are divergences between probability densities  $q, \pi$ : **(i)**  $\sum_{l=1}^L c_l D_l(q||\pi)$ ; **(ii)**  $\sum_{l=1}^L c_l D_l(q_l||\pi_l)$  if  $q = \prod_{l=1}^L q_l(\theta_l)$  and  $\pi = \prod_{l=1}^L \pi_l(\theta_l)$ ; **(iii)**  $D^{\theta_{-(1:L)}}(q||\pi) = \sum_{l=1}^L c_l D_l(q_l||\pi_l)$  if  $q = \prod_{l=1}^L q_l(\theta_l|\theta_{-l})$  and  $\pi = \prod_{l=1}^L \pi_l(\theta_l|\theta_{-l})$  and if  $D^{\theta_{-(1:L)}}(q||\pi) = D^{\theta'_{-(1:L)}}(q||\pi)$  for all  $\theta_{-(1:L)}, \theta'_{-(1:L)}$  for  $\theta_{-(1:L)}, \theta'_{-(1:L)}$  the conditioning sets and if a Hammersley-Clifford holds; **(vi)**  $f(D_l(q||\pi))$ , if  $f(x) \geq 0$  for any  $x \in \mathbb{R}$  with  $f(x) = 0$  if and only if  $x = 0$ .

Thm. 6 enables architecture-specific variational inference: E.g., a Bayesian network with  $L$  layers and posterior  $q = \prod_{l=1}^L q_l$  can have layer-specific uncertainty quantifiers  $D_l$ . The same holds for variational approximations of Deep Gaussian Processes with conditional dependency [e.g., 74]. Section 2.6 of the Appendix also shows how to use GVI for mixture models.

## 5 Black Box Generalized Variational Inference (GVI)

VI is scalable using doubly stochastic, model-agnostic optimization techniques [33, 64, 76, 72, 84] known as black box VI [65]. We extend this to black box GVI (BBGVI), an algorithm inheriting the modularity of  $P(\ell_n, D, \Pi)$ . Its modularity makes it easy to build BBGVI into existing software: E.g., adapting the Deep Gaussian Process implementation of [74] required <100 lines of Python code.

Suppose  $\mathcal{Q} = \{q(\theta|\kappa) : \kappa \in K\}$  and that for all  $(\kappa, \theta) \in (K, \Theta)$ , one can sample  $\theta \sim q(\theta|\kappa)$  and the derivatives  $\nabla_\kappa \log(q(\theta|\kappa))$  and  $\nabla_\kappa D(q||\pi)$  exist. For many  $D, \mathcal{Q}$  and  $\pi$ ,  $\nabla_\kappa D(q||\pi)$  is available in closed form (see Thm. 7). In this case, BBGVI can use the unbiased gradient estimate

$$\nabla_\kappa \hat{L}(q|\ell_n, D) = \frac{1}{S} \sum_{s=1}^S \left\{ \ell_n(\theta^{(s)}, x) \cdot \nabla_\kappa \log(q(\theta^{(s)}|\kappa)) \right\} + \nabla_\kappa D(q||\pi) \quad (5)$$

for an independent sample  $\theta^{(1:S)} \sim q(\theta|\kappa)$ . If a closed form for  $\nabla_\kappa D(q||\pi)$  is not available but  $D(q||\pi) = \mathbb{E}_{q(\theta|\kappa)} [\ell_{\kappa, \pi}^D(\theta)]$  for a function  $\ell_{\kappa, \pi}^D : \Theta \rightarrow \mathbb{R}$ , one can also use

$$\nabla_\kappa \hat{L}(q|\ell_n, D) = \frac{1}{S} \sum_{s=1}^S \left\{ \left[ \ell_n(\theta^{(s)}, x) + \ell_{\kappa, \pi}^D(\theta^{(s)}) \right] \cdot \nabla_\kappa \log(q(\theta^{(s)}|\kappa)) + \nabla_\kappa \ell_{\kappa, \pi}^D(\theta^{(s)}) \right\}. \quad (6)$$

Either form of BBGVI admits black box variance reduction [84, 65], see Section 5 of the Appendix. We note in passing that while sufficiency of first order optimization depends on the interplay of  $\ell_n, D$  and  $\mathcal{Q}$ , GVI objectives are typically convex with respect to  $q$  (see Section 2.5 in the Appendix).



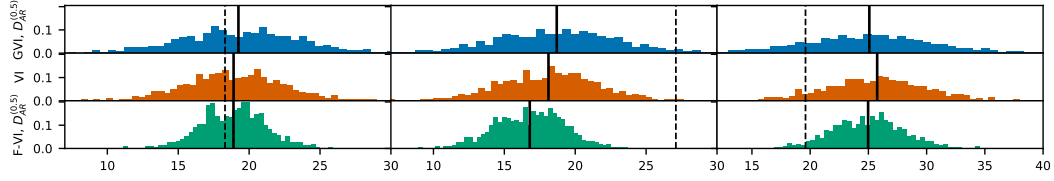


Figure 3: Posterior predictives for three test points on the BNN for boston: **GVI** with  $D = D_{AR}^{(0.5)}$  (top row), **VI** (middle row) and **F-VI** with  $F = D_{AR}^{(0.5)}$  (bottom row) with predictive means (solid) and actual observations (dashed). Note that F-VI/GVI predictives are the most/least concentrated.

**Theorem 7** (Closed form  $D$ ). Let  $q, \pi$  with parameters  $\eta_q, \eta_\pi$  be in the exponential family  $f(\theta|\eta) = h(\theta) \exp\{\eta'T(\theta) - A(\eta)\}$  with natural parameter space  $\mathcal{N} = \{\eta : A(\eta) < \infty\}$ . Then,

- (1)  $D_A^{(\alpha)}(q||\pi)$  and  $D_{AR}^{(\alpha)}(q||\pi)$  have closed form if  $\alpha \in (0, 1)$  or if  $(\alpha\eta_q + (1 - \alpha)\eta_\pi) \in \mathcal{N}$ ;
- (2)  $D_B^{(\beta)}(q||\pi)$  has closed form if  $h(\theta)$  does not depend on  $\theta$  and  $(\beta - 1) \cdot \eta_1 + \eta_2 \in \mathcal{N}$  for  $\eta_1, \eta_2 \in \mathcal{N}$  (which is the case for e.g. Beta, Gamma, Gaussian, exponential and Laplace families);
- (3)  $D_G^{(\gamma)}(q||\pi)$  has closed form if  $D_B^{(\beta)}(q||\pi)$  does for  $\beta = \gamma$ .

## 6 Experiments

We use GVI on Bayesian Neural Networks (BNNs) [62] and Deep Gaussian Processes (DGPs) [17] on data from the UCI repository [50], with typical benchmark settings [as in 31, 49, 74, 14]. We compare test likelihoods and RMSE on 50 splits with 90% training and 10% test data. More details, results and derivations are in Sections 6 and 7 of the Appendix. For readers with strong interests in the DGP application, [45] provides a summary. All code will be made public upon publication<sup>1</sup>.

**BNNs:** We compare F-VI methods using  $F = D_A^{(\alpha)}$  [31]<sup>2</sup> and  $F = D_{AR}^{(\alpha)}$  [49] to GVI using  $D = D_{AR}^{(\alpha)}$ . These GVI methods provide prior-robust uncertainty quantification via Rényi’s  $\alpha$ -divergence  $D_{AR}^{(\alpha)}$ , which equals the KLD for  $\alpha = 1$ . For  $D = D_{AR}^{(\alpha)}$  with  $\alpha \in (0, 1)$ , GVI is less concentrated than VI. Conversely,  $D = D_{AR}^{(\alpha)}$  provides more concentrated posteriors if  $\alpha > 1$ . Fig. 4 shows that GVI’s test performance is a banana shaped curve in  $\alpha$ : Over-concentration relative to VI is an advantage, but too much affects performance adversely. This suggests that for over-parameterized models (such as BNNs), one should select  $D$  to concentrate slightly *more* than standard VI, not less: The posteriors produced by  $P(\ell_n, D_{AR}^{(\alpha)}, \mathcal{Q})$  for all settings of  $\alpha > 1$  uniformly beat VI. Fig. 3 shows that F-VI tends to outperform VI for the same reason: They *shrink* the posterior predictives. This is surprising: The F-VI methods are based on zero-avoiding divergences and would be expected to do the opposite. In Section 7.1.3 of the Appendix, we show that the unintended over-concentration results from the fact that F-VI conflates uncertainty quantification and the loss. As it clearly separates  $D$  and  $\ell_n$ , GVI does not have this problem. Further, Fig. 4 shows that F-VI implicitly changes the loss  $\ell_n$  to its detriment: the GVI settings with  $\alpha > 1$  outperform all F-VI methods on predictive RMSE.

**DGPs:** We compare GVI with VI using the variational families of Salimbeni and Deisenroth [74] that outperformed competing F-VI methods [14]. Our GVI methods provide robustness to model misspecification by using  $\ell_n(\theta, \mathbf{x}) = \sum_{i=1}^n \mathcal{L}_p^\gamma(\theta, x_i)$ . The results are shown in Fig. 5 and are nearly identical for  $\mathcal{L}_p^\beta(\theta, x_i)$ , see Fig. 14 in the Appendix. For either loss, GVI outperforms VI, regardless of the number of layers  $L$ . We use  $D = \text{KLD}$  in Fig. 5, but Sections 6.1.2, 7.3 of the Appendix explain how Thm. 6 allows any  $D^l \in \{\text{KLD}, D_{AR}^{(\alpha)}\}$  for DGP layer  $l$  and show results for varying  $D^l$ .

## 7 Conclusion

We have axiomatically derived a generalized representation  $P(\ell_n, D, \Pi)$  of Bayesian inference (Thms. 1, 2 and 3). The arguments of this triplet are interpretable and serve different functions:  $\Pi$  gives the space of probability distributions over which the posterior is constructed,  $\ell_n$  defines the parameter of interest and  $D$  regulates uncertainty quantification. This representation encompasses a surprisingly wide range of Bayesian methods, e.g. exact inference, Variational Autoencoders and Regularized Bayes (see Table 1 and Section 3 in the Appendix). Moreover, this representation allows the proof of

<sup>1</sup>to be found at <https://github.com/JeremiasKnoblauch/GVIPublic>

<sup>2</sup>We report results with our parameterization of  $D_A^{(\alpha)}$  which is equivalent to using the one in [31] for  $1 - \alpha$

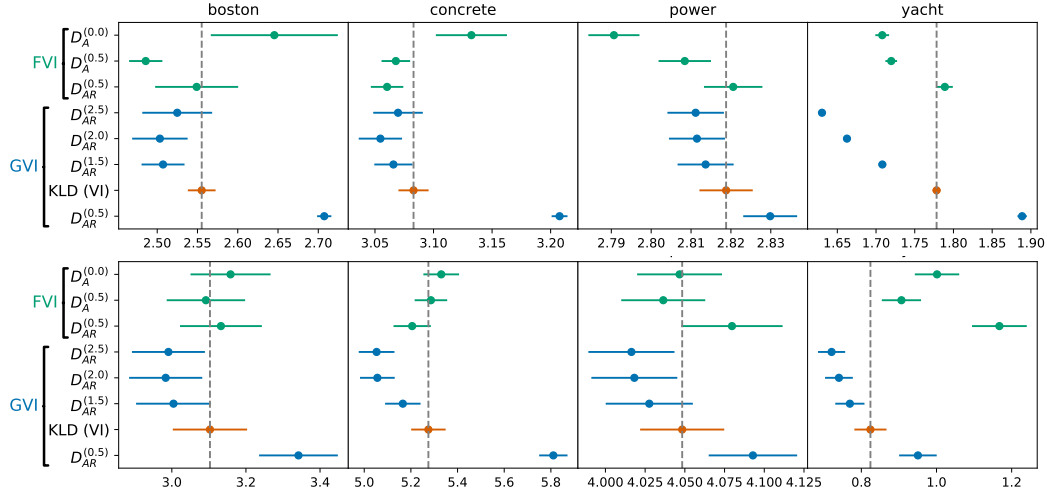


Figure 4: Comparing test set performance on the BNN between **F-VI**, **GVI** with alternative choices for  $D$ , and **VI**. **Top row**: Negative test log likelihoods. **Bottom row**: Test RMSE. The lower the better.

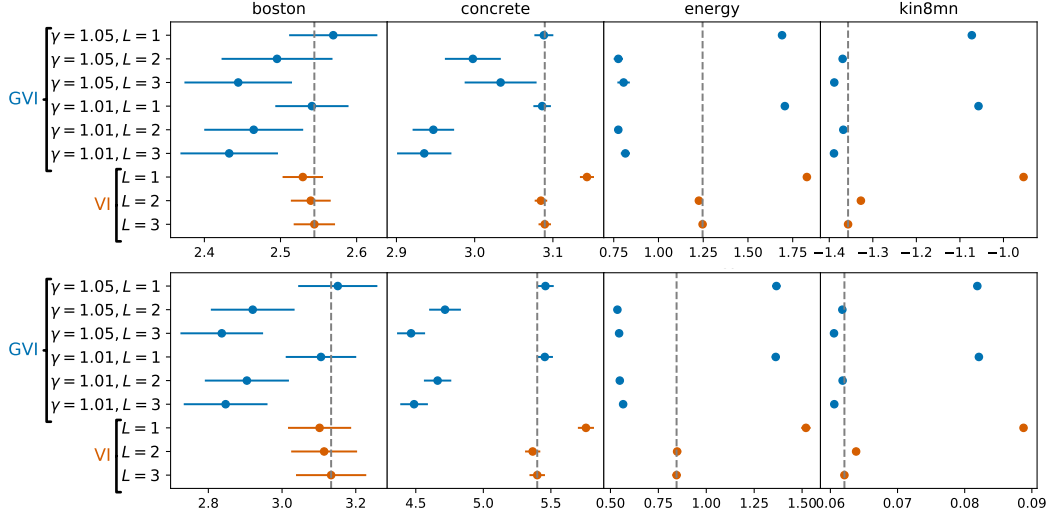


Figure 5: Comparing performance in DGPs with  $L$  layers for **GVI** with  $\ell_n(\theta, x) = \sum_{i=1}^n \mathcal{L}_p^\gamma(\theta, x_i)$  and **VI**. **Top row**: Negative test log likelihoods. **Bottom row**: Test RMSE. The lower the better.

an important optimality result for standard Variational Inference (VI) relative to the exact Bayesian posterior (Thm. 4). The implications of this result lead to the introduction of Generalized Variational Inference (GVI), the natural methodological consequence of the representation  $P(\ell_n, D, \Pi)$ . GVI inherits the modularity of  $P(\ell_n, D, \Pi)$  (see Thms. 5, 6), leading to interpretable alternatives to VI. Just like standard VI, GVI can be interpreted as maximizing a lower bound on the (generalized) Bayesian posteriors (see Appendix, Sections 4.1–4.2). We explore some of GVI’s applications, including more conservative marginals as well as robustness to misspecification and badly specified priors. Due to its modularity, GVI admits black box inference algorithms that can be built into existing black box VI implementations in a matter of minutes. Most importantly, GVI performs very well in practice and leads to substantial performance gains: For example, we show how GVI improves test set metrics in Bayesian Neural Networks and Deep Gaussian Processes (see Figs. 4, 5 and Section 7 in the Appendix). In the future, we would like to work on automatic forms of GVI based on reliable on-line optimization methods for potential hyperparameters of  $D$  and  $\ell_n$ .



## **Acknowledgements**

We would like to cordially thank Chris Holmes for fruitful discussions and helpful remarks regarding the axiomatic foundations of GVI. We would also like to thank David Dunson and Giles Hooker for insightful comments that helped improve 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. TD is funded by the Lloyds Register Foundation programme on Data Centric Engineering through the London Air Quality project. 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

---

## Contents

<b>1</b>	<b>Divergences and Loss functions</b>	<b>12</b>
1.1	Statistical Divergences . . . . .	12
1.2	Loss functions . . . . .	13
<b>2</b>	<b>Theorems &amp; proofs of the main paper</b>	<b>15</b>
2.1	Axiomatic derivation of $P(\ell_n, D, \Pi)$ . . . . .	15
2.1.1	Preliminaries & Axioms . . . . .	15
2.1.2	Proofs . . . . .	15
2.2	Optimality of standard Variational Inference . . . . .	16
2.2.1	Preliminaries . . . . .	16
2.2.2	Proof . . . . .	16
2.3	GVI modularity . . . . .	16
2.3.1	Preliminaries . . . . .	16
2.3.2	Proofs . . . . .	17
2.4	Divergence recombination . . . . .	17
2.5	Convexity of GVI (in $q$ ) . . . . .	18
2.6	GVI for mixtures . . . . .	18
2.7	Closed form $\alpha\beta\gamma$ -divergence ( $D_G^{(\alpha,\beta,r)}$ ) for exponential families . . . . .	19
2.7.1	Preliminaries . . . . .	19
2.7.2	Proofs, results & examples . . . . .	19
<b>3</b>	<b>Examples of existing methods encompassed by <math>P(\ell_n, D, \Pi)</math></b>	<b>23</b>
3.1	Standard Bayesian inference . . . . .	23
3.2	Generalized Bayesian Inference (GBI) [12] . . . . .	23
3.3	Power-likelihood inference . . . . .	24
3.4	Divergence-minimizing Bayesian inference . . . . .	24
3.5	Regularized Bayesian Inference (RegBayes) [23] . . . . .	25
3.6	Variational Inference (for generalized posteriors) . . . . .	25
3.7	( $\beta$ -)Variational Autoencoders . . . . .	26
<b>4</b>	<b>A comparison of different divergences for uncertainty quantification</b>	<b>26</b>
4.1	Notation . . . . .	26
4.2	High-level overview . . . . .	26
4.2.1	Kullback-Leibler Divergence/standard VI ( $D = \text{KLD}$ ) . . . . .	27

4.2.2	Rényi's $\alpha$ -divergence ( $D = D_{AR}^{(\alpha)}$ ) with $\alpha \in (0, 1)$ . . . . .	28
4.2.3	Rényi's $\alpha$ -divergence ( $D = D_{AR}^{(\alpha)}$ ) with $\alpha > 1$ . . . . .	29
4.2.4	$\beta$ -divergence ( $D = D_B^{(\beta)}$ ) with $\beta \in (0, 1)$ . . . . .	29
4.2.5	$\beta$ -divergence ( $D = D_B^{(\beta)}$ ) with $\beta > 1$ . . . . .	29
4.2.6	$\gamma$ -divergence ( $D = D_G^{(\gamma)}$ ) with $\gamma \in (0, 1)$ . . . . .	30
4.2.7	$\gamma$ -divergence ( $D = D_G^{(\gamma)}$ ) with $\gamma > 1$ . . . . .	30
4.3	Theorems and Proofs . . . . .	31
4.3.1	Bounds for Rényi's $\alpha$ -divergence ( $D_{AR}^{(\alpha)}$ ) . . . . .	31
4.3.2	Bounds for the $\beta$ -divergence ( $D_B^{(\beta)}$ ) . . . . .	33
4.3.3	Bounds for the $\gamma$ -divergence ( $D_G^{(\gamma)}$ ) . . . . .	34
4.4	Demonstrations . . . . .	36
4.4.1	The boundedness of the $\alpha$ -divergence ( $D_A^{(\alpha)}$ ) . . . . .	37
4.4.2	Increasing the magnitude of the divergence results in posteriors with large variances . . . . .	37
4.4.3	Robustness to the prior . . . . .	39
4.5	Comparing GVI and F-VI for a multimodal posterior . . . . .	41
<b>5</b>	<b>Variance Reduction in Black box GVI</b>	<b>42</b>
5.1	Preliminaries and assumptions . . . . .	43
5.2	Families of variance-reduced black box GVI methods . . . . .	43
5.2.1	VRBBVI . . . . .	43
5.2.2	BBVI . . . . .	44
5.2.3	BBGVI . . . . .	44
5.2.4	VRBBGVI . . . . .	45
<b>6</b>	<b>Derivations for experiments</b>	<b>47</b>
6.1	Robust Deep Gaussian Processes with alternative uncertainty quantifiers . . . . .	47
6.1.1	Changing the uncertainty quantifier . . . . .	47
6.1.2	Changing the loss . . . . .	49
6.2	Robust Bayesian Neural Nets with alternative uncertainty quantifiers . . . . .	50
<b>7</b>	<b>Experiments</b>	<b>50</b>
7.1	Bayesian Neural Nets . . . . .	50
7.1.1	Details . . . . .	50
7.1.2	Additional results . . . . .	50
7.1.3	Posterior predictives: F-VI vs GVI methods . . . . .	51
7.2	Interpreting performance improvements with over-concentration (relative to VI) . . . . .	55
7.3	Deep Gaussian Processes . . . . .	55
7.3.1	Details . . . . .	55
7.3.2	Additional results . . . . .	55

# 1 Divergences and Loss functions

## 1.1 Statistical Divergences

Axioms 1 and 3 in the main paper states that the principled Bayesian decision problem should regularise how far  $q$  can move from  $\pi$ . In order to do this, we need to be able to quantify the discrepancy between prior  $\pi(\theta)$  and  $q(\theta)$ . Discrepancy measures of this kind are called statistical divergences [see e.g. 3].

**Definition 1** (Statistical Divergences [20]). A statistical divergence  $D(g||f)$  is a measure of discrepancy between two probability densities  $f$  and  $g$  with the following two properties

1.  $D(g||f) \geq 0, \forall f, g$
2.  $D(g||f) = 0$  if and only if  $g = f$ .

We note this definition is not sufficient for  $D(g||f)$  to be a metric. Divergences are often asymmetrical and do not necessarily satisfy the triangle-inequality. Arguably the most famous divergence is the Kullback-Leibler Divergence (KLD). For densities with respect to the Lebesgue measure, it is given by

$$\text{KLD}(g||f) = \int g(x) \log \frac{g(x)}{f(x)} dx. \quad (7)$$

Originally introduced in the seminal paper of Kullback and Leibler [47], it has since appeared throughout many strands of literature, including statistics, information theory and signal processing. It is however by no means the only divergence in use and there are many different families of divergences. Two well-known families, both containing the KLD, are the  $f$ - (or  $\phi$ -) and the Bregman divergences. In this paper we focus on a third particular general divergence family [16] containing cases of  $f$ -, Bregman, and KLD divergences.

**Definition 2** (The  $\alpha\beta\gamma$ -divergence  $D_G^{(\alpha,\beta,r)}$  [16]). The  $\alpha\beta\gamma$ -divergence  $D_G^{(\alpha,\beta,r)}$  [16] takes the form

$$\begin{aligned} & D_G^{(\alpha,\beta,r)}(q(\theta)||\pi(\theta)) \\ &= \frac{1}{\alpha(\beta-1)(\alpha+\beta-1)r} \left[ \left( \tilde{D}_G^{(\alpha,\beta)}(q(\theta)||\pi(\theta)) + 1 \right)^r - 1 \right] \end{aligned} \quad (8)$$

where  $r > 0$  and

$$\begin{aligned} & \tilde{D}_G^{(\alpha,\beta)}(q(\theta)||\pi(\theta)) \\ &= \int (\alpha q(\theta)^{\alpha+\beta-1} + (\beta-1)\pi(\theta)^{\alpha+\beta-1} - (\alpha+\beta-1)q(\theta)^\alpha \pi(\theta)^{\beta-1}) d\theta \end{aligned} \quad (9)$$

with  $\alpha \neq 0, \beta \neq 1$ .

Below we list some well known special cases of the  $D_G^{(\alpha,\beta,r)}$  family that we use in the main paper. This exposition is a summary of the excellent review conducted in [16]. We note that the parametrizations of these divergences may vary throughout the literature.

**Definition 3** (The  $\alpha$ -divergence  $(D_A^{(\alpha)})$  [15, 51, 3]). The  $\alpha$ -divergence is defined as

$$D_A^{(\alpha)}(q(\theta)||\pi(\theta)) = \frac{1}{\alpha(1-\alpha)} \left\{ 1 - \int q(\theta)^\alpha \pi(\theta)^{1-\alpha} d\theta \right\}, \quad (10)$$

where  $\alpha \in \mathbb{R} \setminus \{0, 1\}$ .  $D_A^{(\alpha)}$  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 4** (The Rényi  $\alpha$ -divergence  $(D_{AR}^{(\alpha)})$  [68]). The Rényi [68]  $\alpha$ -divergence is defined as

$$D_{AR}^{(\alpha)}(q(\theta)||\pi(\theta)) = \frac{1}{\alpha(\alpha-1)} \log \left( \int q(\theta)^\alpha \pi(\theta)^{1-\alpha} d\theta \right), \quad (11)$$

where  $\alpha \in \mathbb{R} \setminus \{0, 1\}$ .  $D_A^{(\alpha)}$  is recovered from  $D_G^{(\alpha,\beta,r)}$  in the limit as  $r \rightarrow 0$  and  $\beta = 2 - \alpha$ . Note that we use the scaled version proposed by Liese and Vajda [51] and frequently used in the literature [e.g. 16]

The Rényi  $\alpha$ -divergence can be recovered from the  $\alpha$ -divergence by applying the transformation

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

**Definition 5** (The  $\beta$ -divergence ( $D_B^{(\beta)}$ ) [6, 55]). The  $\beta$ -divergence [6, 55] is defined as

$$\begin{aligned} D_B^{(\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) \\ = \frac{1}{\beta(\beta-1)} \int q(\boldsymbol{\theta})^\beta d\boldsymbol{\theta} + \frac{1}{\beta} \int \pi(\boldsymbol{\theta})^\beta d\boldsymbol{\theta} - \frac{1}{\beta-1} \int q(\boldsymbol{\theta})\pi(\boldsymbol{\theta})^{\beta-1} d\boldsymbol{\theta}, \end{aligned} \quad (12)$$

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.  $D_B^{(\beta)}$  has often been referred to as the Density-Power Divergence (DPD) in the statistics literature [6].

**Definition 6** (The  $\gamma$ -divergence ( $D_G^{(\gamma)}$ ) [21]). The  $\gamma$ -divergence [21] is defined as

$$D_G^{(\gamma)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\gamma(\gamma-1)} \log \frac{(\int q(\boldsymbol{\theta})^\gamma d\boldsymbol{\theta}) (\int \pi(\boldsymbol{\theta})^\gamma d\boldsymbol{\theta})^{\gamma-1}}{(\int q(\boldsymbol{\theta})\pi(\boldsymbol{\theta})^\gamma d\boldsymbol{\theta})^\gamma}, \quad (13)$$

where  $\gamma \in \mathbb{R} \setminus \{0, 1\}$ .  $D_G^{(\gamma)}$  is recovered from  $D_G^{(\alpha, \beta, r)}$  in the limit as  $r \rightarrow 0$ ,  $\alpha = 1$  and  $\beta = \gamma$ .

The  $D_G^{(\gamma)}$  can be shown to be generated from the  $D_B^{(\beta)}$  applying the following transformation

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

to all three of the  $D_B^{(\beta)}$  terms. The  $D_{AR}^{(\alpha)}$  can be shown to be generated by the  $D_A^{(\alpha)}$  under the same transformation of its two terms.

**Remark 1** (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  $\alpha = \beta = \gamma \rightarrow 1$ . This can be shown using the *replica trick*:

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

## 1.2 Loss functions

As well as being useful as a prior regulariser, statistical divergences are also attractive to measure how well a model  $p(\cdot|\boldsymbol{\theta})$  approximates a data generating process  $g$ . In fact, it is well known that Bayes' rule concentrates on the model parameter minimising  $\text{KLD}(g||p(\cdot|\boldsymbol{\theta}))$ .

Of course, one never has direct access to  $g$ . One does however have access to samples  $x_1, \dots, x_n \sim g$ . As a result, while one cannot use a statistical divergence between  $g$  and  $p(\cdot|\boldsymbol{\theta})$  directly, one can measure closeness between  $p(\cdot|\boldsymbol{\theta})$  and the empirical measure  $\hat{g}_n$  of a sample. This is particularly appealing because some divergences (including the KLD,  $D_B^{(\beta)}$  and a modification to the  $D_G^{(\gamma)}$  used in [22, 39] explained below) provide an interpretation of statistical divergences in terms of loss functions (see e.g. [11, 18]). For these divergences, there exists a function  $S : \mathcal{X} \times \mathcal{P}(\boldsymbol{\Theta}) \rightarrow \mathbb{R}$  mapping from the samples space  $\mathcal{X}$  and the space of probability measures  $\mathcal{P}(\boldsymbol{\Theta})$  on  $\boldsymbol{\Theta}$  into the real numbers such that one can write

$$D(g(\mathbf{x})||p(\mathbf{x}|\boldsymbol{\theta})) = \mathbb{E}_{g(\mathbf{x})} [S(\mathbf{x}, p(\cdot|\boldsymbol{\theta}))] - \mathbb{E}_{g(\mathbf{x})} [S(\mathbf{x}, g)]. \quad (14)$$

The loss function interpretation is then as follows: A divergence taking the form of eq. (14) is the excess expected penalty incurred for believing  $x$  was distributed according to  $p(\cdot|\boldsymbol{\theta})$  when it was actually distributed according to  $g$ . For example, the KLD can be equivalently written using the logarithmic score  $S(\mathbf{x}, p) = \ell(\boldsymbol{\theta}, \mathbf{x}) = -\log(p(\mathbf{x}|\boldsymbol{\theta}))$

$$\begin{aligned} \text{KLD}(g||p(\cdot|\boldsymbol{\theta})) &= \int g(\mathbf{x}) \log \frac{g(\mathbf{x})}{p(\mathbf{x}|\boldsymbol{\theta})} d\mathbf{x} \\ &= \mathbb{E}_g(\mathbf{x}) [-\log p(\mathbf{x}|\boldsymbol{\theta})] - \mathbb{E}_g(\mathbf{x}) [-\log g(\mathbf{x})]. \end{aligned} \quad (15)$$

$\mathbb{E}_g(\mathbf{x}) [-\log g(\mathbf{x})]$  is the entropy of the data generating process  $g$  and is unaffected by  $p$ . Hence, finding  $p$  minimising  $\text{KLD}(g||p(\cdot|\boldsymbol{\theta}))$  is equivalent to finding  $p$  minimising  $\mathbb{E}_g(\mathbf{x}) [-\log p(\mathbf{x}|\boldsymbol{\theta})]$ . Lastly the expectation under the data generating distribution  $g$  can be approximated by the empirical

distribution of the sample  $\mathbb{E}_g(\mathbf{x}) [-\log p(\mathbf{x}|\boldsymbol{\theta})] \approx \mathbb{E}_{\hat{g}_n}(\mathbf{x}) [-\log p(\mathbf{x}|\boldsymbol{\theta})] = \frac{1}{n} \sum_{i=1}^n -\log p(x_i|\boldsymbol{\theta})$ . This score function then feeds into the general Bayesian loss updating rule [see 12] as

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto \pi(\boldsymbol{\theta}) \exp \left( - \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right). \quad (16)$$

Substituting  $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$  recovers Bayes rule, demonstrating that Bayes rule is minimising the KLD.

In the Section 4.1 of the main paper we suggest minimising either the  $D_B^{(\beta)}(g||p(\cdot|\boldsymbol{\theta}))$  or  $D_G^{(\gamma)}(g||p(\cdot|\boldsymbol{\theta}))$  to provide model agnostic robust parametric inference because their corresponding loss function have convenient forms.

$$\begin{aligned} D_B^{(\beta)}(g||p(\cdot|\boldsymbol{\theta})) &= \frac{1}{\beta} \int p(\mathbf{x}|\boldsymbol{\theta})^\beta d\mathbf{x} - \frac{1}{(\beta-1)} \int p(\mathbf{x}|\boldsymbol{\theta})^{\beta-1} g(\mathbf{x}) d\mathbf{x} + \frac{1}{\beta(\beta-1)} \int g(\mathbf{x})^\beta d\mathbf{x} \\ &= \mathbb{E}_{g(\mathbf{x})} \left[ \frac{1}{\beta} \int p(\mathbf{z}|\boldsymbol{\theta})^\beta d\mathbf{z} - \frac{1}{(\beta-1)} p(\mathbf{x}|\boldsymbol{\theta})^{\beta-1} \right] \\ &\quad - \mathbb{E}_{g(\mathbf{x})} \left[ \frac{1}{\beta} \int g(\mathbf{z})^\beta d\mathbf{z} - \frac{1}{(\beta-1)} g(\mathbf{x})^{\beta-1} \right] \end{aligned} \quad (17)$$

so  $\ell(\boldsymbol{\theta}, \mathbf{x}) = \frac{1}{\beta} \int p(\mathbf{z}|\boldsymbol{\theta})^\beta d\mathbf{z} - \frac{1}{(\beta-1)} p(\mathbf{x}|\boldsymbol{\theta})^{\beta-1}$  which is available in closed form for many exponential families and only depends on the form of  $p(\mathbf{x}|\boldsymbol{\theta})$ .

Unfortunately the  $D_G^{(\gamma)}$  doesn't allow for the interpretation provided by Eq. (14)

$$\begin{aligned} D_G^{(\gamma)}(g||p(\cdot|\boldsymbol{\theta})) &= \frac{1}{\gamma} \log \int p(\mathbf{x}|\boldsymbol{\theta})^\gamma d\mathbf{x} - \frac{1}{(\gamma-1)} \log \int p(\mathbf{x}|\boldsymbol{\theta})^{\gamma-1} g(\mathbf{x}) d\mathbf{x} \\ &\quad + \frac{1}{\gamma(\gamma-1)} \log \int g(\mathbf{x})^\gamma d\mathbf{x} \\ &= \frac{1}{\gamma} \log \int p(\mathbf{x}|\boldsymbol{\theta})^\gamma d\mathbf{x} - \frac{1}{(\gamma-1)} \log \mathbb{E}_{g(\mathbf{x})} [p(\mathbf{x}|\boldsymbol{\theta})^{\gamma-1}] \\ &\quad + \frac{1}{\gamma(\gamma-1)} \log \int g(\mathbf{x})^\gamma d\mathbf{x} \end{aligned} \quad (18)$$

However minimising the  $D_G^{(\gamma)}$  for  $\boldsymbol{\theta}$  allows us to ignore the entropy term and is therefore equivalent to minimising

$$\frac{1}{\gamma} \log \int p(\mathbf{z}|\boldsymbol{\theta})^\gamma d\mathbf{z} - \frac{1}{(\gamma-1)} \log \mathbb{E}_{g(\mathbf{x})} [p(\mathbf{x}|\boldsymbol{\theta})^{\gamma-1}] = \log \frac{(\int p(\mathbf{z}|\boldsymbol{\theta})^\gamma d\mathbf{z})^{\frac{1}{\gamma}}}{(\mathbb{E}_{g(\mathbf{x})} [p(\mathbf{x}|\boldsymbol{\theta})^{\gamma-1}]^{\frac{1}{(\gamma-1)}})} \quad (19)$$

Since the purpose of a loss function is only to target a specific value of  $\boldsymbol{\theta}$ , it is valid to apply any monotonic transform to a Eq. (19) since it does not change the loss function's minimiser. To this end, note that the function  $\gamma \exp(x)$  is monotonically increasing on  $\mathbb{R}$  for  $\gamma > 0$ . Similarly, the function  $h(x) = x^{1-\gamma}$  is monotonic on  $\mathbb{R}^+$  and decreasing (increasing) for  $\gamma > 1$  ( $\gamma < 1$ ). As a result, (for  $\gamma > 0$ ) minimising Eq. (19) is equivalent to minimising,

$$-\frac{\gamma}{\gamma-1} \frac{\mathbb{E}_{g(\mathbf{x})} [p(\mathbf{x}|\boldsymbol{\theta})^{\gamma-1}]}{(\int p(\mathbf{z}|\boldsymbol{\theta})^\gamma d\mathbf{z})^{\frac{\gamma-1}{\gamma}}} \approx -\frac{\gamma}{\gamma-1} \frac{1}{n} \sum_{i=1}^n \frac{p(x_i|\boldsymbol{\theta})^{\gamma-1}}{(\int p(\mathbf{z}|\boldsymbol{\theta})^\gamma d\mathbf{z})^{\frac{\gamma-1}{\gamma}}} \text{ if } \gamma > 1 \quad (20)$$

which provides  $\gamma$  times the loss function of [22, 39] and the same form the loss takes in Eq. (4) of this main paper. Multiplying by  $\frac{1}{\gamma-1}$  cancels with the negation when  $\gamma < 1$  and  $h(x) = x^{1-\gamma}$  is actually increasing in  $x$ .

Substituting Eq. (20) into Eq. (14) results in a scalar multiple of the alternative definition of the  $\gamma$ -divergence used in [22, 39].

$$\frac{\gamma}{(\gamma-1)} \left\{ I_{g,\gamma}(\boldsymbol{\theta})^{\frac{1}{\gamma}} - \int p(\mathbf{x}|\boldsymbol{\theta})^{\gamma-1} I_{p,\gamma}(\boldsymbol{\theta})^{-\frac{\gamma-1}{\gamma}} d\mathbf{x} \right\}, \quad (21)$$

where  $I_{p,c}(\boldsymbol{\theta}) = \int p(\mathbf{z}|\boldsymbol{\theta})^c d\mathbf{z}$ . This divergence appears to be different from  $D_G^{(\gamma)}$  as defined by [16], but in fact both versions will be minimised for the same value of  $\boldsymbol{\theta}$ .



## 2 Theorems & proofs of the main paper

### 2.1 Axiomatic derivation of $P(\ell_n, D, \Pi)$

#### 2.1.1 Preliminaries & Axioms

First, recall that we argue for a representation of Bayesian inference via the optimization problem  $P(\ell_n, D, \Pi)$  over the space of probability distributions. In particular, we argue that Bayesian inference should take the form  $P(\ell_n, D, \Pi)$  given by

$$q^*(\theta) = \arg \min_{q \in \Pi} \{L(q|\mathbf{x}, \ell_n, D)\}; \quad L(q|\mathbf{x}, \ell_n, D) = \mathbb{E}_{q(\theta)} [\ell_n(\theta, \mathbf{x})] + D(q|\pi), \quad (22)$$

where the constituent parts of the form  $P(\ell_n, D, \Pi)$  are given by

- a **loss**  $\ell_n$  linking a parameter of interest  $\theta$  to the observations  $\mathbf{x} = x_{1:n}$ . Throughout, we will assume additivity, i.e.  $\ell_n(\theta, \mathbf{x}) = \sum_{i=1}^n \ell(\theta, x_i)$  for some  $\ell$ .
- a divergence  $D : \mathcal{P}(\Theta) \times \mathcal{P}(\Theta) \rightarrow \mathbb{R}_+$  regularizing the posterior with respect to the prior  $\pi$ . As  $D$  determines how uncertainty in  $q^*(\theta)$  is quantified, we call it **uncertainty quantifier**.
- a set of **admissible posteriors**  $\Pi \subseteq \mathcal{P}(\Theta)$  the regularized expected loss is minimized over.

For readability of the following Theorems, we also restate the Axioms used for deriving the form  $P(\ell_n, D, \Pi)$  and motivating GVI.

**Axiom 1** (Representation). Bayesian inference infers posteriors  $q$  on  $\Theta$  by (i) measuring how  $q$  fits a sample  $\mathbf{x}$  via the expectation of a loss  $\ell_n(\theta, \mathbf{x})$ , (ii) quantifying uncertainty about  $\theta^*$  via a divergence  $D$  between prior  $\pi$  and  $q$ , (iii) optimizing  $q$  over a space of probability distributions  $\Pi$  on  $\Theta$ .

**Axiom 2** (Information difference).  $P(\ell_n, D, \Pi)$  produces different posteriors for  $\mathbf{x} = x_{1:n}$  and  $\mathbf{x}' = x_{1:n+m}$  if there is an information difference of  $\mathbf{x}'$  relative to  $\mathbf{x}$ , i.e. if  $\ell_n(\theta, \mathbf{x}) \neq \ell_{n+m}(\theta, \mathbf{x}')$ .

**Axiom 3** (Prior regularization).  $q$  is regularized against  $\pi$  by penalizing the divergence  $D(q|\pi)$ .

**Axiom 4** (Translation Invariance). For constant  $C$  and  $\ell'_n = \ell_n + C$ ,  $P(\ell'_n, D, \Pi) = P(\ell_n, D, \Pi)$ .

#### 2.1.2 Proofs

**Theorem 1.** If Axiom 2 holds,  $\ell_n$  is additive and  $C \in \mathbb{N}$ ,  $P(\ell_n, D, \Pi) \neq P(C \cdot \ell_n, D, \Pi)$ .

*Proof.* If  $\ell_n$  is additive,  $\ell_n(\theta, \mathbf{x}) = \sum_{i=1}^n \ell(\theta, x_i)$  for some loss function  $\ell$  and  $\mathbf{x} = x_{1:n}$ . For any  $C \in \mathbb{N}$ , write  $\mathbf{x}(C) = x(C)_{1:nC}$  with  $x(C)_i = x_{(i \bmod n)+1}$ , where  $(a \bmod b)$  denotes the (integer) remainder of the division  $a/b$ . In words,  $\mathbf{x}(C)$  copies the entries of  $\mathbf{x}$  exactly  $C$  times. Now, simply note that  $C \cdot \ell_n(\theta, \mathbf{x}) = \ell_n(\theta, \mathbf{x}(C))$  to see that  $P(D, \ell_n, \Pi) = P(D, C \cdot \ell_n, \Pi)$  would violate Axiom 2 for any choice of  $D$  and  $\Pi$ .  $\square$

**Theorem 2** (Form 1). If Axiom 1 holds,  $P(\ell_n, D, \Pi)$  can be written as  $\arg \min_{q \in \Pi} \{L(q|\mathbf{x}, \ell_n, D)\}$  for  $L(q|\mathbf{x}, \ell_n, D) = f(\mathbb{E}_{q(\theta)}[\ell_n(\theta, \mathbf{x})], D(q|\pi))$ , where  $f$  is some function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ .

*Proof.* This follows directly from Axiom 1: By (iii), Bayesian inference is an optimization over  $\Pi$  producing a posterior  $q$ . Moreover, by (i) this optimization depends on the expectation of the loss  $\ell_n$  via  $q$ 's expectation, i.e. via  $\mathbb{E}_{q(\theta)}[\ell_n(\theta, \mathbf{x})]$ . Further, by (ii) it also depends on the divergence  $D$  between prior and  $q$ , i.e. on  $D(q|\pi)$ . Hence, Bayesian inference is representable as  $\arg \min_{q \in \Pi} \{L(q|\mathbf{x}, \ell_n, D)\}$  for  $L(q|\mathbf{x}, \ell_n, D) = f(\mathbb{E}_{q(\theta)}[\ell_n(\theta, \mathbf{x})], D(q|\pi))$  for a function  $f$ .  $\square$

**Theorem 3** (Form 2). For  $P(\ell_n, D, \Pi)$  being  $\arg \min_{q \in \Pi} \{L(q|\mathbf{x}, \ell_n, D)\}$  and  $\circ$  an elementary operation on  $\mathbb{R}$ ,  $L(q|\mathbf{x}, \ell_n, D) = \mathbb{E}_{q(\theta)}[\ell_n(\theta, \mathbf{x})] \circ D(q|\pi)$  satisfies Axioms 3 and 4 only if  $\circ = +$ .

*Proof.* First rewrite  $\mathbb{E}_{q(\theta)}[\ell_n(\theta, \mathbf{x})] \circ D(q|\pi)$ . The elementary operations are addition, subtraction, multiplication and division. Consider the losses  $\ell_n$  and  $\ell'_n = \ell_n + C$  for  $C \in \mathbb{R}$  a constant. It is

straightforward to see that Axiom 4 will not hold in general if  $\circ$  is multiplication, as

$$\begin{aligned}
& \arg \min_q \{L(q|\mathbf{x}, \ell'_n, D)\} \\
&= \arg \min_q \{\mathbb{E}_{q(\boldsymbol{\theta})} [\ell_n(\boldsymbol{\theta}, \mathbf{x}) + C] \cdot D(q||\pi)\} \\
&= \arg \min_q \{\mathbb{E}_{q(\boldsymbol{\theta})} [\ell_n(\boldsymbol{\theta}, \mathbf{x})] D(q||\pi) + C \cdot D(q||\pi)\} \\
&\neq \arg \min_q \{\mathbb{E}_{q(\boldsymbol{\theta})} [\ell_n(\boldsymbol{\theta}, \mathbf{x})] D(q||\pi)\} \\
&= \arg \min_q \{L(q|\mathbf{x}, \ell_n, D)\}, \tag{23}
\end{aligned}$$

and similarly if  $\circ$  is division. As  $C$  is a constant, it is however easy to show that if  $\circ$  is addition or subtraction,

$$\begin{aligned}
& \arg \min_q \{L(q|\mathbf{x}, \ell'_n, D)\} \\
&= \arg \min_q \{L(q|\mathbf{x}, \ell_n, D)\}. \tag{24}
\end{aligned}$$

Since subtracting the prior regularizer is a direct and obvious violation of Axiom 3, it follows that addition is the only elementary operation on  $\mathbb{R}$  satisfying both Axioms and the result follows.  $\square$

## 2.2 Optimality of standard Variational Inference

### 2.2.1 Preliminaries

Bissiri et al. [12] show that for arbitrary losses  $\ell_n$ ,  $P(\ell_n, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$  produces coherent belief distributions. In particular, they show that  $P(\ell_n, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$  is solved by

$$q^*(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta}) \exp \{\sum_{i=1}^n -\ell(\boldsymbol{\theta}, x_i)\}}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}) \exp \{\sum_{i=1}^n -\ell(\boldsymbol{\theta}, x_i)\} d\boldsymbol{\theta}}. \tag{25}$$

### 2.2.2 Proof

**Theorem 4** (VI: Uniquely optimal approximation). For exact and coherent Bayesian posteriors solving  $P(\ell_n, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$  and a fixed variational family  $\mathcal{Q}$ , standard VI produces the uniquely optimal  $\mathcal{Q}$ -constrained approximation to  $P(\ell_n, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$ : Having decided on approximating the Bayesian posterior with some  $q \in \mathcal{Q}$ , VI provides the uniquely optimal solution.

*Proof. Optimality:* First, note that the ELBO for VI with variational family  $\mathcal{Q}$  and with respect to a generalized Bayes Theorem as in eq. (25) takes the form  $P(\ell_n, \text{KLD}, \mathcal{Q})$ . For a detailed derivation of this fact, we refer to section 3.6 and eq. (46). As shown by Bissiri et al. [12], the prior regularizer  $D$  of a *coherent* Bayesian posterior belief has to be the KLD. It follows that any exact ( $\Pi = \mathcal{P}(\boldsymbol{\Theta})$ ) and coherent ( $D = \text{KLD}$ ) Bayesian posterior can be represented as the solution of  $P(\ell_n, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$  for some additive loss function  $\ell_n$ . We seek to find the optimal approximation  $q \in \mathcal{Q}$  to the problem  $P(\ell_n, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$  defined over  $\mathcal{P}(\boldsymbol{\Theta})$ . By definition, this means we want to solve  $P(\ell_n, \text{KLD}, \mathcal{Q})$ , which exactly corresponds to KLD-based VI for approximating  $P(\ell_n, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$ , see section 3.6 and eq. (46).

*Uniqueness:*<sup>3</sup> Moreover, VI is the unique procedure to get the optimal approximation to  $P(\ell_n, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$ : Suppose  $q^*$  solves  $P(\ell_n, \text{KLD}, \mathcal{Q})$ . Now suppose that there also exists  $q' \in \mathcal{Q}$  which gives a better approximation to the solution of  $P(\ell_n, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$  than  $q^*$ . In other words,  $q'$  is such that  $L(q'|\mathbf{x}, \ell_n, \text{KLD}) > L(q^*|\mathbf{x}, \ell_n, \text{KLD})$  in eq. (2). Since the variational posterior  $q^*$  solves  $P(\ell_n, \text{KLD}, \mathcal{Q})$  as per eq. (2), this yields a contradiction and uniqueness follows immediately.  $\square$

## 2.3 GVI modularity

### 2.3.1 Preliminaries

Here, we give some theoretical arguments for the division of responsibility in GVI's representation. In particular, GVI (i) permits robust inference by adapting the target parameter  $\boldsymbol{\theta}^*$  through  $\ell_n$  and (ii)

<sup>3</sup>This proof does not show that VI produces unique posteriors  $q^*$ . Instead, this proof makes statements about the (possibly infinite-dimensional) collection of solutions  $\{q^*\}$  of a given VI problem

permits alternative uncertainty quantification by appropriately adapting  $D$ . As robustness to model misspecification is used to describe a variety of different phenomena, we define its use in the current paper formally to avoid confusion. Our understanding of robustness to model misspecification is aligned with Hampel et al. [29] and Tukey [78]. To put it with the words of the latter, *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.* Robust statistics is the attempt at resolving this undesirable risk balance. Formalizing this, we arrive at the following definition for robustness to model misspecification for our representation of Bayesian inference.

**Definition 7 (Robustness).** Let  $M_j = P(D_j, \ell_{n,j}, \Pi)$  with  $\ell_{n,j}(\theta, \mathbf{x}) = \sum_{i=1}^n \ell_j(\theta, x_i)$  and  $\theta_j^* = \arg \min_{\theta} \{\mathbb{E}_{\mathbf{X}} [\ell_j(\theta, \mathbf{X})]\}$  for  $j = 1, 2$ . Then,  $M_1$  is more robust for  $\theta$  than  $M_2$  relative to the set of (implicit) assumptions  $A$  on  $\mathbf{X}$  if (i)  $\theta_1^*$  is a better result than  $\theta_2^*$  if  $A$  is untrue and (ii)  $\theta_1^* = \theta_2^*$  if  $A$  is true.

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  $\theta^*$ , i.e. that  $\theta_1^* \neq \theta_2^*$  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.

### 2.3.2 Proofs

**Theorem 5 (GVI modularity).** For Bayesian inference via some GVI method  $P(\ell_n, D, \mathcal{Q})$ , making it robust to model misspecification amounts to changing  $\ell_n$ . Conversely, adapting its uncertainty quantification (for fixed  $\mathcal{Q}, \pi, \theta^*, \hat{\theta}_n$ ) amounts to changing  $D$ .

*Proof. Robustness:* By Def. 7, robustness for  $\ell_n(\theta, \mathbf{x}) = \sum_{i=1}^n \ell(\theta, x_i)$  implies a change in  $\theta^* = \arg \min_{\theta} \{\mathbb{E}_{\mathbf{X}} [\ell(\theta, \mathbf{X})]\}$  if distributional assumptions about  $\mathbf{X}$  are incorrect. Notice that  $\theta^*$  is not affected by  $D$  or  $\Pi$ , but is affected by  $\ell_n$ .

*Uncertainty Quantification:*  $\Pi$  and  $\pi$  are not allowed to change by assumption and so cannot affect uncertainty quantification. Next, while  $\ell_n$  is allowed to change,  $\hat{\theta}_n$  and  $\theta^*$  are not. Notice that changing  $\ell_n$  to  $\ell'_n$  will affect  $\hat{\theta}_n = \arg \min_{\theta} \{\frac{1}{n} \ell_n(\theta, \mathbf{x})\}$  and  $\theta^* = \mathbb{E}_{\mathbf{X}} [\ell_n(\theta, \mathbf{x})]$  unless  $\ell' = C + w \cdot \ell$  for some constants  $C$  and  $w > 0$ . Since  $P(\ell_n, D, \Pi) = P(\ell_n + C, D, \mathcal{Q})$  for any  $C$  by Theorem 3, we can disregard  $C$  and turn to  $w$ . Indeed, the uncertainty quantification of  $P(\ell_n, D, \mathcal{Q})$  will be different from that of  $P(w \cdot \ell_n, D, \mathcal{Q})$  for any constant  $w \neq 1$ . However, dividing by  $w$  in eq. (2) yields that  $P(w \cdot \ell_n, D, \mathcal{Q}) = P(\ell_n, \frac{1}{w} D, \mathcal{Q})$ . (Where we define  $w^{-1} = \infty$  for  $w = 0$ .) Hence, any change in the loss that does not affect  $\hat{\theta}_n$  and  $\theta^*$  can be rewritten as a change in  $D$ . It follows that changing the uncertainty quantification amounts to changing  $D$ .  $\square$

## 2.4 Divergence recombination

**Theorem 6 (Divergence recombination).** Let  $D_l$  be divergences and  $c_l \geq 0$  scalars for  $l = 1, 2, \dots, L$ . The following are divergences between probability densities  $q, \pi$ : (i)  $\sum_{l=1}^L c_l D_l(q||\pi)$ ; (ii)  $\sum_{l=1}^L c_l D_l(q_l||\pi_l)$  if  $q = \prod_{l=1}^L q_l(\theta_l)$  and  $\pi = \prod_{l=1}^L \pi_l(\theta_l)$ ; (iii)  $D^{\theta_{-(1:L)}}(q||\pi) = \sum_{l=1}^L c_l D_l(q_l||\pi_l)$  if  $q = \prod_{l=1}^L q_l(\theta_l|\theta_{-l})$  and  $\pi = \prod_{l=1}^L \pi_l(\theta_l|\theta_{-l})$  and if  $D^{\theta_{-(1:L)}}(q||\pi) = D^{\theta'_{-(1:L)}}(q||\pi)$  for all  $\theta_{-(1:L)}, \theta'_{-(1:L)}$  for  $\theta_{-(1:L)}, \theta'_{-(1:L)}$  the conditioning sets and if a Hammersley-Clifford holds; (vi)  $f(D_l(q||\pi))$ , if  $f(x) \geq 0$  for any  $x \in \mathbb{R}$  with  $f(x) = 0$  if and only if  $x = 0$ .

*Proof.* All claims except (iii) follow trivially from the definition of a divergence:  $D$  is a divergence if and only if (i)  $D(q||\pi) \geq 0$  for any tuple  $q, \pi$  and (i)  $D(q||\pi) = 0$  only when  $q = \pi$ .

Part (iii) requires a little more work. First, observe by definition of a divergence,  $D_l(q_l(\theta_l|\theta'_{-l})||\pi_l(\theta_l|\theta'_{-l})) = 0$  for all  $l$  and all potential conditioning sets will hold if and only if  $q_l(\theta_l|\theta'_{-l}) = \pi_l(\theta_l|\theta'_{-l})$ . Next, note that we have assumed that  $D^{\theta'_{-(1:L)}}(q||\pi) = D^{\theta_{-(1:L)}}(q||\pi)$  for all conditioning sets  $\theta'_{-(1:L)}, \theta_{-(1:L)}$ . In other words, if  $D^{\theta'_{-(1:L)}}(q||\pi) = 0$  for some  $\theta'_{-(1:L)}$ ,

then it will be 0 for *any* conditioning set. This immediately entails that for arbitrary  $\theta'_{-(1:L)}$ ,  $D^{\theta'_{-(1:L)}}(q||\pi) = 0$  if and only if  $q_l(\theta_l|\theta'_{-l}) = \pi_l(\theta_l|\theta'_{-l})$  for *all*  $l$  and for *any* choice of  $\theta'_{-l}$ . In other words, the conditionals are the same. Since the positivity condition holds, we can then apply the Hammersley-Clifford Theorem to conclude that since the conditionals fully specify the joint,  $D^{\theta'_{-(1:L)}}(q||\pi) = 0$  if and only if  $q = \pi$ .  $\square$

## 2.5 Convexity of GVI (in $q$ )

While convexity of a given problem  $P(\ell_n, D, \mathcal{Q})$  will depend on  $\mathcal{Q}$ , and  $\ell_n$ , at the most abstract level one can make statements about GVI's convexity in (unconstrained functionals)  $q$ . Clearly, the function  $\mathbb{E}_q[\cdot]$  is linear in  $q$ , and so everything hinges on the properties of the divergence  $D$ . It is well-known that the KLD is (strongly) convex (in the total variation norm). Here, we collect convexity properties of other divergences.

**Lemma 1** (Convexity properties of divergences). The divergence  $D(q||\pi)$  is *convex* in  $q$  if

$$(I) \ D = D_B^{(\beta)} \text{ and } \beta \in \mathbb{R} \setminus \{0, 1\};$$

$$(II) \ D = D_{AR}^{(\alpha)} \text{ and } \alpha \in (0, 1);$$

and *quasiconvex* in  $q$  if

$$(III) \ D = D_{AR}^{(\alpha)} \text{ and } \alpha > 1;$$

*Proof.* (II) and (III) are well-known, see e.g. Van Erven and Harremos [80]<sup>4</sup>. (I) holds trivially because the  $D_B^{(\beta)}$  is a Bregman divergence based on the strictly convex generating function  $\psi(x) = \frac{1}{\beta(\beta-1)}x^\beta$ .  $\square$

**Corollary 1** (GVI Convexity). The objective corresponding to  $P(\ell, D, \Pi)$  is *convex* in  $q$  for

$$(I) \ D = D_B^{(\beta)} \text{ and } \beta \in \mathbb{R} \setminus \{0, 1\};$$

$$(II) \ D = D_{AR}^{(\alpha)} \text{ and } \alpha \in (0, 1);$$

and *quasiconvex* in  $q$  if

$$(III) \ D = D_{AR}^{(\alpha)} \text{ and } \alpha > 1;$$

*Proof.* Since the expectation over  $q$  is a linear operator in  $q$ , it is convex in  $q$ . Thus, the Theorem holds by virtue of Lemma 1.  $\square$

## 2.6 GVI for mixtures

The convexity derived in Lemma 1 and Corollary 1 is directly applicable to tractability in closed form mixtures.

**Corollary 2** (Closed form mixtures). For a mixture model  $q(\theta|\kappa) = \sum_{i=1}^k c_i q_i(\theta|\kappa_i)$  as variational family (i.e,  $\sum_{i=1}^k c_i = 1$  and  $q_i$  the mixture components), it holds that,

$$D_{AR}^{(\alpha)} \left( \sum_{i=1}^k c_i q_i(\theta|\kappa_i) \middle\| \pi(\theta) \right) \leq \sum_{i=1}^k c_i D_{AR}^{(\alpha)} (q_i(\theta|\kappa_i) || \pi(\theta)) \text{ for } \alpha \in (0, 1) \quad (26)$$

$$D_B^{(\beta)} \left( \sum_{i=1}^k c_i q_i(\theta|\kappa_i) \middle\| \pi(\theta) \right) \leq \sum_{i=1}^k c_i D_B^{(\beta)} (q_i(\theta|\kappa_i) || \pi(\theta)) \text{ for } \beta \in \mathbb{R} \setminus \{0, 1\} \quad (27)$$

*Proof.* Apply Lemma 1  $k$  times.  $\square$

<sup>4</sup>Note that while their proof relies on an alternative parameterization of the  $D_{AR}^{(\alpha)}$ , this is not an issue here: Simply multiply both sides by  $\frac{1}{\alpha}$ .

## 2.7 Closed form $\alpha\beta\gamma$ -divergence ( $D_G^{(\alpha,\beta,r)}$ ) for exponential families

### 2.7.1 Preliminaries

The Appendix 1 states the general divergence,  $D_G^{(\alpha,\beta,r)}$ , introduced in [16]. The  $D_G^{(\alpha,\beta,r)}$  family contains many well known families as special cases. The theorems and corollaries in the following section determine the conditions under which the  $D_G^{(\alpha,\beta,r)}$  – and by implication the well known families of divergences contained within the  $D_G^{(\alpha,\beta,r)}$  family – have a closed form for two members of the same exponential family. Note that the special case of these results for the  $D_{AR}^{(\alpha)}$  has been derived before [see 26, 25, 51]. Unlike previous work, our results apply to a range of other divergences, too.

Throughout this section we will assume that the prior and variational family are exponential families that can be written down in closed form. In particular, this implies that the log-normalising constant is a closed form function of the natural parameters. To summarize some of the most important findings of this section in plain English, we find that if both  $q$  and  $\pi$  are in the same exponential family,

- $D_{AR}^{(\alpha)}(q||\pi)$  and  $D_A^{(\alpha)}(q||\pi)$  are always available in closed form if  $\alpha \in (0, 1)$  (see Corollary 3)
- $D_{AR}^{(\alpha)}(q||\pi)$  and  $D_A^{(\alpha)}(q||\pi)$  are available in closed form if  $\alpha > 1$  for most exponential families (see again Corollary 3)
- $D_B^{(\beta)}(q||\pi)$  and  $D_G^{(\gamma)}(q||\pi)$  are available in closed form for  $\beta > 1$  and  $\gamma > 1$  for most exponential families (See Corollary 5).

We note that these findings are important as closed forms for the divergence regularizer can drastically reduce the variance of black box GVI. The remainder of this section is devoted to tedious but rigorous derivations of these findings. To showcase the implications of the derived results, we use the Multivariate Gaussian (MVN) to provide examples along the way.

**Definition 8** (The MVN exponential family). The density of the MVN exponential family for vector  $\theta$  of dimension  $d$  is  $p(\theta|\eta(\kappa)) = h(\theta) \exp \{ \eta(\kappa)^T T(\theta) - A(\eta(\kappa)) \}$  where

$$\eta(\kappa) = \begin{pmatrix} \mathbf{V}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\mathbf{V}^{-1} \end{pmatrix} \quad T(\theta) = \begin{pmatrix} \theta \\ \theta\theta^T \end{pmatrix}$$

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

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

### 2.7.2 Proofs, results & examples

**Theorem 7** (Closed form  $D_G^{(\alpha,\beta,r)}$  between exponential families). The  $D_G^{(\alpha,\beta,r)}$  between a variational posterior  $q(\theta|\kappa_n)$  and prior  $\pi(\theta|\kappa_0)$  in the same exponential family,  $p(\theta|\eta(\kappa)) = h(\theta) \exp \{ \eta(\kappa)^T T(\theta) - A(\eta(\kappa)) \}$  with natural parameter space  $\mathcal{N} = \{ \eta(\kappa) : A(\eta(\kappa)) < \infty \}$ , is available in closed form under the following conditions

- $\eta(\kappa_1), \eta(\kappa_2) \in \mathcal{N} \Rightarrow (\alpha\eta(\kappa_1) + (\beta - 1)\eta(\kappa_2)) \in \mathcal{N}$ ;
- $\mathbb{E}_{p(\theta|\eta(\kappa))} [h(\theta)^{\alpha+\beta-2}]$  is a closed form function of  $\eta(\kappa) \in \mathcal{N}$ .

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

$$\begin{aligned} & \tilde{D}_G^{(\alpha,\beta)}(q(\theta|\kappa_n)||\pi(\theta|\kappa_0)) \\ &= \int (\alpha q(\theta|\kappa_n)^{\alpha+\beta-1} + (\beta - 1)\pi(\theta|\kappa_0)^{\alpha+\beta-1} - (\alpha + \beta - 1)q(\theta|\kappa_n)^\alpha \pi(\theta|\kappa_0)^{\beta-1}) d\theta \\ &= \alpha \frac{\exp \{ A((\alpha + \beta - 1)\eta(\kappa_n)) \}}{\exp \{ A(\eta(\kappa_n)) \}^{(\alpha+\beta-1)}} \mathbb{E}_{p(\theta|(\alpha+\beta-1)\eta(\kappa_n))} [h(\theta)^{\alpha+\beta-2}] \\ &+ (\beta - 1) \frac{\exp \{ A((\alpha + \beta - 1)\eta(\kappa_0)) \}}{\exp \{ A(\eta(\kappa_0)) \}^{(\alpha+\beta-1)}} \mathbb{E}_{p(\theta|(\alpha+\beta-1)\eta(\kappa_0))} [h(\theta)^{\alpha+\beta-2}] \\ &- (\alpha + \beta - 1) \frac{\exp \{ A(\alpha\eta(\kappa_n) + (\beta - 1)\eta(\kappa_0)) \}}{\exp \{ A(\eta(\kappa_n)) \}^\alpha \exp \{ A(\eta(\kappa_0)) \}^{(\beta-1)}} \mathbb{E}_{p(\theta|(\alpha\eta(\kappa_n) + (\beta-1)\eta(\kappa_0)))} [h(\theta)^{\alpha+\beta-2}] \end{aligned} \quad (28)$$

*Proof.* The  $D_G^{(\alpha, \beta, r)}$  in equation (8) is a closed form function of  $\tilde{D}_G^{(\alpha, \beta)}$  given by (9). As a result if the  $\tilde{D}_G^{(\alpha, \beta)}$  is available in closed form then so is  $D_G^{(\alpha, \beta)}$ . 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 (29)-(31) are available in closed form for our exponential families.

$$G_1 := \int q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)^{\alpha+\beta-1} d\boldsymbol{\theta} \quad (29)$$

$$G_2 := \int \pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)^{\alpha+\beta-1} d\boldsymbol{\theta} \quad (30)$$

$$G_3 := \int q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)^\alpha \pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)^{\beta-1} d\boldsymbol{\theta} \quad (31)$$

First we tackle  $G_1$ : Eq. (29) simplifies to

$$\begin{aligned} & \int q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)^{\alpha+\beta-1} d\boldsymbol{\theta} \\ &= \int h(\boldsymbol{\theta})^{\alpha+\beta-1} \exp\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)^T T(\boldsymbol{\theta}) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_n))\} d\boldsymbol{\theta} \\ &= \exp\{-(\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_n))\} \int h(\boldsymbol{\theta})^{\alpha+\beta-1} \exp\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)^T T(\boldsymbol{\theta})\} d\boldsymbol{\theta} \\ &= \exp\{A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_n))\} \\ & \quad \cdot \int h(\boldsymbol{\theta})^{\alpha+\beta-1} \exp\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)^T T(\boldsymbol{\theta}) - A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n))\} d\boldsymbol{\theta} \\ &= \exp\{A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_n))\} \mathbb{E}_{p(\boldsymbol{\theta} | (\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n))} [h(\boldsymbol{\theta})^{\alpha+\beta-2}] \end{aligned} \quad (32)$$

where Condition (i) taking  $\eta(\boldsymbol{\kappa}_1) = \eta(\boldsymbol{\kappa}_2) = \eta(\boldsymbol{\kappa}_n)$  ensures that

$$A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)) = \int h(\boldsymbol{\theta}) \exp\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)^T T(\boldsymbol{\theta})\} d\boldsymbol{\theta} \quad (33)$$

which in turn ensures that  $\mathbb{E}_{p(\boldsymbol{\theta} | (\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n))} [h(\boldsymbol{\theta})^{\alpha+\beta-2}]$  is a valid expectation and Condition (ii) guarantees this is a closed form function of  $\eta(\boldsymbol{\kappa}_n)$ . Similarly for  $G_2$ : Eq. (30) simplifies to

$$\begin{aligned} & \int \pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)^{\alpha+\beta-1} d\boldsymbol{\theta} \\ &= \int h(\boldsymbol{\theta})^{\alpha+\beta-1} \exp\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0)^T T(\boldsymbol{\theta}) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_0))\} d\boldsymbol{\theta} \\ &= \exp\{-(\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_0))\} \int h(\boldsymbol{\theta})^{\alpha+\beta-1} \exp\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0)^T T(\boldsymbol{\theta})\} d\boldsymbol{\theta} \\ &= \exp\{A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0)) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_0))\} \\ & \quad \cdot \int h(\boldsymbol{\theta})^{\alpha+\beta-1} \exp\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0)^T T(\boldsymbol{\theta}) - A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0))\} d\boldsymbol{\theta} \\ &= \exp\{A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0)) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_0))\} \mathbb{E}_{p(\boldsymbol{\theta} | (\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0))} [h(\boldsymbol{\theta})^{\alpha+\beta-2}] \end{aligned} \quad (34)$$

where in analogy to  $G_1$ , conditions (i) and (ii) with  $\eta(\boldsymbol{\kappa}_1) = \eta(\boldsymbol{\kappa}_2) = \eta(\boldsymbol{\kappa}_0)$  ensure this has a closed form. Lastly for  $G_3$ : Eq. (31) becomes



$$\begin{aligned}
& \int q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)^\alpha \pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)^{\beta-1} d\boldsymbol{\theta} \\
&= \int h(\boldsymbol{\theta})^\alpha \exp \{ \alpha \eta(\boldsymbol{\kappa}_n)^T T(\boldsymbol{\theta}) - \alpha A(\eta(\boldsymbol{\kappa}_n)) \} \\
&\quad \cdot h(\boldsymbol{\theta})^{\beta-1} \exp \{ (\beta-1) \eta(\boldsymbol{\kappa}_0)^T T(\boldsymbol{\theta}) - (\beta-1) A(\eta(\boldsymbol{\kappa}_0)) \} d\boldsymbol{\theta} \\
&= \exp \{ -\alpha A(\eta(\boldsymbol{\kappa}_n)) - (\beta-1) A(\eta(\boldsymbol{\kappa}_0)) \} \\
&\quad \cdot \int h(\boldsymbol{\theta})^{\alpha+\beta-1} \exp \{ (\alpha \eta(\boldsymbol{\kappa}_n) + (\beta-1) \eta(\boldsymbol{\kappa}_0))^T T(\boldsymbol{\theta}) \} d\boldsymbol{\theta} \\
&= \exp \{ A(\alpha \eta(\boldsymbol{\kappa}_n) + (\beta-1) \eta(\boldsymbol{\kappa}_0)) - \alpha A(\eta(\boldsymbol{\kappa}_n)) - (\beta-1) A(\eta(\boldsymbol{\kappa}_0)) \} \\
&\quad \cdot \mathbb{E}_{p(\boldsymbol{\theta} | (\alpha \eta(\boldsymbol{\kappa}_n) + (\beta-1) \eta(\boldsymbol{\kappa}_0)))} [h(\boldsymbol{\theta})^{\alpha+\beta-2}] \tag{35}
\end{aligned}$$

where Condition (i) with  $\eta(\boldsymbol{\kappa}_1) = \eta(\boldsymbol{\kappa}_n)$  and  $\eta(\boldsymbol{\kappa}_2) = \eta(\boldsymbol{\kappa}_0)$  ensures that

$$A(\alpha \eta(\boldsymbol{\kappa}_n) + (\beta-1) \eta(\boldsymbol{\kappa}_0)) = \int h(\boldsymbol{\theta}) \exp \{ (\alpha \eta(\boldsymbol{\kappa}_n) + (\beta-1) \eta(\boldsymbol{\kappa}_0)) \eta(\boldsymbol{\kappa}_n)^T T(\boldsymbol{\theta}) \} d\boldsymbol{\theta} \tag{36}$$

which in turn ensures that  $\mathbb{E}_{p(\boldsymbol{\theta} | (\alpha \eta(\boldsymbol{\kappa}_n) + (\beta-1) \eta(\boldsymbol{\kappa}_0)))} [h(\boldsymbol{\theta})^{\alpha+\beta-2}]$  is a valid expectation and Condition (ii) guarantees this is a closed form function of  $\eta(\boldsymbol{\kappa}_n)$  and  $\eta(\boldsymbol{\kappa}_0)$ .

Therefore, provided Condition (i)-(ii) hold then integrals  $G_1, G_2$  and  $G_3$  are available in closed and thus so is  $D_G^{(\alpha, \beta, r)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n) || \pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0))$ .  $\square$

**Remark 2** (Conditions of Theorem 7 for the MVN exponential family). Substituting the definition of the MVN exponential family into the conditions of Theorem 7 provides:

i)

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

$$\text{where } \boldsymbol{\mu}^* := \left\{ \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 - \left( (\frac{1}{\alpha} \mathbf{V}_1)^{-1} + (\frac{1}{\beta-1} \mathbf{V}_2)^{-1} \right)^{-1} \left( (\frac{1}{\alpha} \mathbf{V}_1)^{-1} \boldsymbol{\mu}_2 + (\frac{1}{\beta-1} \mathbf{V}_2)^{-1} \boldsymbol{\mu}_1 \right) \right\}$$

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.

**Corollary 3** (Closed form  $D_A^{(\alpha)}$  and  $D_{AR}^{(\alpha)}$  for exponential families). The  $D_A^{(\alpha)}$  or  $D_{AR}^{(\alpha)}$  between a variational posterior  $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)$  and prior  $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)$  in the same exponential family,  $p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa})) = h(\boldsymbol{\theta}) \exp \{ \eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa})) \}$  with natural parameter space  $\mathcal{N} = \{ \eta(\boldsymbol{\kappa}) : A(\eta(\boldsymbol{\kappa})) < \infty \}$ , 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  $D_A^{(\alpha)}$  and  $D_{AR}^{(\alpha)}$  can be written as

$$\begin{aligned}
D_A^{(\alpha)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n) || \pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)) &= \frac{1}{\alpha(1-\alpha)} \left[ 1 - \frac{\exp \{ A(K(\alpha, \boldsymbol{\kappa}_n, \boldsymbol{\kappa}_0)) \}}{\exp \{ A(\eta(\boldsymbol{\kappa}_n)) \}^\alpha \exp \{ A(\eta(\boldsymbol{\kappa}_0)) \}^{(1-\alpha)}} \right] \\
D_{AR}^{(\alpha)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n) || \pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)) &= \frac{1}{\alpha(\alpha-1)} [A(K(\alpha, \boldsymbol{\kappa}_n, \boldsymbol{\kappa}_0)) - \alpha A(\eta(\boldsymbol{\kappa}_n)) - (1-\alpha) A(\eta(\boldsymbol{\kappa}_0))]
\end{aligned}$$

where  $K(\alpha, \boldsymbol{\kappa}_n, \boldsymbol{\kappa}_0) = (\alpha \eta(\boldsymbol{\kappa}_n) + (1-\alpha) \eta(\boldsymbol{\kappa}_0))$

*Proof.*  $D_A^{(\alpha)}$ : Following [16] 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 7 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  $\eta(\kappa_n) \in \mathcal{N}$  and  $\eta(\kappa_0) \in \mathcal{N}$  then  $\alpha\eta(\kappa_n) + (1 - \alpha)\eta(\kappa_0) \in \mathcal{N}$ . If  $\alpha < 0$  or  $\alpha > 1$ , then this can no longer be guaranteed.

$D_{AR}^{(\alpha)}$ : Following [16], the  $D_{AR}^{(\alpha)}$  can be found when taking the limit as  $r \rightarrow 0$  of the  $D_G^{(\alpha, \beta, r)}$ . Alternatively the  $D_{AR}^{(\alpha)}$  can be written as a function of the  $D_A^{(\alpha)}$

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

Since  $D_A^{(\alpha)}(q(\theta) || \pi(\theta))$  is available in closed form and  $D_{AR}^{(\alpha)}(q(\theta) || \pi(\theta))$  is simply a closed form transformation of  $D_A^{(\alpha)}(q(\theta) || \pi(\theta))$ ,  $D_{AR}^{(\alpha)}(q(\theta) || \pi(\theta))$  must be available in closed form. (Note however that technically, the  $D_{AR}^{(\alpha)}$  is no longer necessarily finite for  $\alpha \in (0, 1)$ .)  $\square$

**Remark 3** (Conditions for Corollary 3 for the MVN exponential family). The condition that  $\alpha\eta(\kappa_n) + (1 - \alpha)\eta(\kappa_0) \in \mathcal{N}$  can only be guaranteed for  $\alpha \in (0, 1)$ . However we can see from Remark 2 that provided  $\mathbf{V}^* = \left( \left( \frac{1}{\alpha} \mathbf{V}_1 \right)^{-1} + \left( \frac{1}{\beta - 1} \mathbf{V}_2 \right)^{-1} \right)^{-1}$  is a symmetric semi-positive definite (SPD) matrix for  $\beta = 2 - \alpha$  then this condition will be satisfied.  $\mathbf{V}^*$  being an SPD matrix is enough to guarantee  $\mu^*$  is a real valued vector.  $\alpha > 1$  or  $\alpha < 0$  mean we cannot guarantee that  $\mathbf{V}^*$  is SPD. However, we implement the  $D_{AR}^{(\alpha)}$  to quantify uncertainty for  $\alpha > 1$  in the main paper. This will still generally provide a closed form divergence provided the variational posterior has a sufficiently smaller variance than the prior, which can always be guaranteed to hold in practice.

**Corollary 4** (Closed form  $D_B^{(\beta)}$  and  $D_G^{(\gamma)}$  for exponential families). The  $D_B^{(\beta)}$  and  $D_G^{(\gamma)}$  between a variational posterior  $q(\theta | \kappa_n)$  and prior  $\pi(\theta | \kappa_0)$  in the same exponential family,  $p(\theta | \eta(\kappa)) = h(\theta) \exp \{ \eta(\kappa)^T T(\theta) - A(\eta(\kappa)) \}$  with natural parameter space  $\mathcal{N} = \{ \eta(\kappa) : A(\eta(\kappa)) < \infty \}$ , is available in closed form under the following conditions with  $\gamma = \beta$

- i)  $\eta(\kappa_1), \eta(\kappa_2) \in \mathcal{N} \Rightarrow ((\beta - 1)\eta(\kappa_1) + \eta(\kappa_2)) \in \mathcal{N}$
- ii)  $\mathbb{E}_{p(\theta | \eta(\kappa))} [h(\theta)^{\beta - 1}]$  is a closed form function of  $\eta(\kappa) \in \mathcal{N}$ .

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

$$D_B^{(\beta)}(q(\theta | \kappa_n) || \pi(\theta | \kappa_0)) = \frac{1}{\beta(\beta - 1)} \frac{\exp \{ A(\beta\eta(\kappa_n)) \}}{\exp \{ A(\eta(\kappa_n)) \}^\beta} E(\beta, \kappa_n) + \frac{1}{\beta} \frac{\exp \{ A(\beta\eta(\kappa_0)) \}}{\exp \{ A(\eta(\kappa_0)) \}^\beta} E(\beta, \kappa_0) - \frac{1}{(\beta - 1)} \frac{\exp \{ A(\eta(\kappa_n) + (\beta - 1)\eta(\kappa_0)) \}}{\exp \{ A(\eta(\kappa_n)) \} \exp \{ A(\eta(\kappa_0)) \}^{(\beta - 1)}} E(\beta, (\eta(\kappa_n) + (\beta - 1)\eta(\kappa_0))) \quad (37)$$

$$D_G^{(\gamma)}(q(\theta | \kappa_n) || \pi(\theta | \kappa_0)) = \frac{1}{\gamma(\gamma - 1)} (A(\gamma\eta(\kappa_n)) + \log E(\gamma, \kappa_n)) + \frac{1}{\gamma} (A(\gamma\eta(\kappa_0)) + \log E(\gamma, \kappa_0)) - \frac{1}{(\gamma - 1)} (A(\eta(\kappa_n) + (\gamma - 1)\eta(\kappa_0)) + \log E(\gamma, (\eta(\kappa_n) + (\gamma - 1)\eta(\kappa_0)))) \quad (38)$$

where  $E(\beta, \kappa) = \mathbb{E}_{p(\theta | \beta\eta(\kappa))} [h(\theta)^{\beta - 1}]$

*Proof.*  $D_B^{(\beta)}$  Following [16], 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 7 become (i)-(ii) above.

$D_G^{(\gamma)}$  Similarly the  $D_{AR}^{(\alpha)}$  the  $D_G^{(\gamma)}$  can be obtained from the  $D_G^{(\alpha, \beta, r)}$  by taking the limit as  $r \rightarrow 0$  with  $\alpha = 1$  and  $\beta = \gamma$ . More simply [16] illustrate that the  $D_G^{(\gamma)}$  can be recovered by taking the following closed form transformation of each of the three terms in the  $D_B^{(\beta)}$  and combining them additively

$$c_0 \int p^{c_1}(x) q^{c_2}(x) dx \rightarrow \log \left( \int p^{c_1}(x) q^{c_2}(x) dx \right)^{c_0}$$

So providing that  $D_B^{(\beta)}$  with  $\beta = \gamma$  is available in closed form, so is the  $D_G^{(\gamma)}$ .  $\square$

**Remark 4** (Conditions for Corollary 4 under the MVN exponential family). Following Remark 2:

- i) of Corollary 4 is satisfied providing  $\mathbf{V}^* = \left( (\mathbf{V}_n)^{-1} + \left( \frac{1}{\beta-1} \mathbf{V}_0 \right)^{-1} \right)^{-1}$  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 3, when  $\beta < 1$  closed forms will require that the variational posterior has a sufficiently smaller variance than the prior.
- ii) is trivially satisfied as for the MVN family  $h(\boldsymbol{\theta})$  doesn't depend on  $\boldsymbol{\theta}$  or  $\eta(\boldsymbol{\kappa})$ .

In fact Remark 4 can be extended to many other exponential families if we constrain  $\beta = \gamma > 1$ , this is formalised in Corollary 5.

**Corollary 5** (Closed form  $D_B^{(\beta)}$  and  $D_G^{(\gamma)}$  for exponential families when  $\beta = \gamma > 1$ ). When  $\beta = \gamma > 1$ , the conditions for Corollary 4 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 5 follows straight from that of Corollary 4.  $\square$

### 3 Examples of existing methods encompassed by $P(\ell_n, D, \Pi)$

This section presents a selection of Bayesian inference approaches and how they relate to the axiomatically derived form  $P(\ell_n, D, \Pi)$  defined in the main paper. Clearly, this list is non-exhaustive and many more Bayesian inference approaches admit our representation.

#### 3.1 Standard Bayesian inference

Standard Bayesian inference solves  $P(-\sum_{i=1}^n \log(p(\boldsymbol{\theta}|x_i)), \text{KLD}, \mathcal{P}(\boldsymbol{\theta}))$ . We note that using the log score  $\ell_n(\boldsymbol{\theta}, \mathbf{x}) = -\sum_{i=1}^n \log(p(x_i|\boldsymbol{\theta}))$  is equivalent to learning the parameter  $\boldsymbol{\theta}^*$  whose corresponding model minimizes the KLD to the true DGP as mentioned in the main paper (see also Jewson et al. [40]). To see this, simply note that for a *fixed* parameter value  $\boldsymbol{\theta}$  and a sample  $\mathbf{x} \stackrel{iid}{\sim} g$ ,

$$\text{KLD}(g(\mathbf{x})||p(\mathbf{x}|\boldsymbol{\theta})) = \mathbb{E}_g[-\log(p(\mathbf{x}|\boldsymbol{\theta}))] - \mathbb{E}_g[-\log(g(\mathbf{x}))]. \quad (39)$$

The second term is called entropy and does not depend on  $\boldsymbol{\theta}$ , so it can be ignored for the inference task. Yet, as one does not know  $g$ , the expectation in the first term is replaced by sample averages. This then yields the loss  $\ell_n$  of standard Bayesian inference. Using this together with the KLD inside the form  $P(\ell_n, \text{KLD}, \mathcal{P}(\boldsymbol{\theta}))$  then yields Bayes' Theorem [see 12, for a proof].

#### 3.2 Generalized Bayesian Inference (GBI) [12]

GBI encompasses standard Bayesian inference as a special case. Using the notation introduced in the main paper, GBI produces posterior beliefs over the set of problems

$$P_{\text{GBI}} = \left\{ P(\ell_n, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta})) \text{ for } \ell_n(\boldsymbol{\theta}, \mathbf{x}) = \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \text{ s.t. } \int_{\boldsymbol{\Theta}} \exp\{-\ell_n(\boldsymbol{\theta}, \mathbf{x})\} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} < \infty \right\}$$

In fact, additivity of  $\ell_n$  together with the requirement that  $D = \text{KLD}$  and  $\Pi = \mathcal{P}(\boldsymbol{\Theta})$  is all that is needed to ensure the updating mechanism is coherent. Coherence implies that for  $\mathbf{x} = (\mathbf{x}_A, \mathbf{x}_B)$ ,  $q^*$  will be the same if one updates the posterior sequentially with  $\mathbf{x}_A$  and then  $\mathbf{x}_B$  or alternatively directly computes it with  $\mathbf{x}$ . Coherence is a property that holds for all  $P \in P_{\text{GBI}}$ . In some sense, this observation is trivial: One simply notes that Bayesian inference problems  $P \in P_{\text{GBI}}$  are solved by a generalized Bayesian posterior that is exponentially additive and obeys the Generalized Bayes Theorem:

$$q^*(\boldsymbol{\theta}) = \frac{\exp\{-\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\} \pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \exp\{-\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \quad (40)$$

We note that distributions of this kind have been studied as Gibbs or pseudo posteriors before [e.g. 24, 2], but Bissiri et al. [12] show that they provide a principled Bayesian rule for belief updating in their own right. In situations where  $\ell_n$  does not involve a likelihood function, it can sometimes be important to calibrate the magnitude of the loss against the prior [see e.g. 52]. This stems from the fact that while likelihood functions have to integrate to one and hence are well-behaved, the magnitude of general losses  $\ell_n$  may dominate  $\text{KLD}(q||\pi)$  in magnitude, leading to over-concentrated posteriors. Such loss calibration considerations are natural extensions of power-likelihood inference to likelihood-free losses[52].

Lastly, we give a brief sketch of the elegant proof in Bissiri et al. [12]. In fact, once one can motivate why the form  $P(\ell_n, \text{KLD}, \mathcal{P}(\Theta))$  is the only form producing coherent posteriors, the derivation of the corresponding generalized Bayes Theorem is simply based on the realization that one may rewrite the objective for  $P(\ell_n, \text{KLD}, \mathcal{P}(\Theta))$  as

$$\begin{aligned} q^*(\theta) &= \arg \min_{q \in \mathcal{P}(\Theta)} \left\{ \int_{\Theta} \left[ \log \left( \exp \left\{ \sum_{i=1}^n \ell(\theta, x_i) \right\} \right) + \log \left( \frac{q(\theta)}{\pi(\theta)} \right) \right] q(\theta) \right\} \\ &= \arg \min_{q \in \mathcal{P}(\Theta)} \left\{ \int_{\Theta} \log \left( \frac{q(\theta)}{\pi(\theta) \exp \left\{ -\sum_{i=1}^n \ell(\theta, x_i) \right\}} \right) q(\theta) \right\}. \end{aligned} \quad (41)$$

Because we only care about the minimizer  $q^*$  (and not the objective value), it also holds that for any constant  $C > 0$ , the above is equal to

$$\begin{aligned} q^*(\theta) &= \arg \min_{q \in \mathcal{P}(\Theta)} \left\{ \int_{\Theta} \log \left( \frac{q(\theta)/C}{\pi(\theta) \exp \left\{ -\sum_{i=1}^n \ell(\theta, x_i) \right\} / C} \right) q(\theta)/C \right\} \\ &= \arg \min_{q \in \mathcal{P}(\Theta)} \left\{ \text{KLD} \left( q(\theta)/C \parallel \pi(\theta) \exp \left\{ -\sum_{i=1}^n \ell(\theta, x_i) \right\} / C \right) \right\}. \end{aligned} \quad (42)$$

Lastly, one now sets  $C = \int_{\Theta} \exp \left\{ -\sum_{i=1}^n \ell(\theta, x_i) \right\} \pi(\theta) d\theta$  and notes that by the definition of a (proper) divergence, the KLD is minimized if and only if its two arguments are the same. This recovers eq. (40).

### 3.3 Power-likelihood inference

Power-likelihood inference is a special case of GBI and concerns likelihood-based losses of form  $\ell(\theta, \mathbf{x}) = -\sum_{i=1}^n \log(p(x_i|\theta)^w) = -\sum_{i=1}^n w \log(p(x_i|\theta))$  with  $w > 0$ . This recovers an updating rule of the form

$$q^*(\theta) \propto \pi(\theta) \prod_{i=1}^n p(x_i|\theta)^w. \quad (43)$$

Typically, inference of this kind is motivated by model misspecification, which leads standard Bayesian uncertainty quantification to become inappropriate [see e.g. 34, 28, 57]. In particular, if the observed data is overdispersed relative to the probabilistic model,  $w$  should be chosen to be  $< 1$  to slow the posterior concentration and increase the posterior variances. Conversely, if the observed data is underdispersed relative to the model,  $w$  should be chosen  $> 1$ . Such corrections are especially important in the Bayesian non-parametric setting [e.g., 57]. However, if  $\theta$  is finite-dimensional (as for GVI), raising the likelihood to some power does not change the inference target  $\theta^* = \arg \min_{\theta} \mathbb{E}_{\mathbf{X}} [\ell_n(\theta, \mathbf{x})]$ . This is why in our formulation  $P(\ell_n, D, \Pi)$ , we assign the role of uncertainty quantification to the divergence  $D$ . Thus conceptually, rather than considering power-likelihood inference as an operation weighting  $\ell_n$  by  $w$ , we instead consider it as an operation weighting  $D$  by  $\frac{1}{w}$ .

### 3.4 Divergence-minimizing Bayesian inference

Divergence-minimizing Bayesian inference is yet another special case of GBI with a particular kind of (divergence-induced) loss function. In particular, these methods are inspired by the fact that standard Bayesian inference uses the negative log likelihood loss to minimize the KLD between the model and the Data Generating Process as represented by the sample  $\mathbf{x}$ . Using the same idea, these methods seek to derive losses minimizing alternative divergences  $D$ . Important examples in this

line of work include Ghosh and Basu [24] and Futami et al. [22]<sup>5</sup> who propose the Tsallis score minimizing the robust  $\beta$ -divergence, which is given by

$$\mathcal{L}_p^\beta(\boldsymbol{\theta}, x_i) = -\frac{1}{\beta-1}p(x_i|\boldsymbol{\theta})^{\beta-1} + \frac{I_{p,\beta}(\boldsymbol{\theta})}{\beta}, \quad (44)$$

where  $I_{p,\beta}(\boldsymbol{\theta}) = \int p(x|\boldsymbol{\theta})^\beta dx$ . It was originally introduced in Basu et al. [6] for Maximum Likelihood type inference. Since then, it has been successfully used for robust inference in problems ranging from blind source separation [56] over spatial filtering [75] to on-line changepoint detection [46]. More recently, an additive loss for the  $\gamma$ -divergence has been proposed [39], which can be deployed in similar fashion by using

$$\mathcal{L}_p^\gamma(\boldsymbol{\theta}, x_i) = -\frac{1}{\gamma-1}p(x_i|\boldsymbol{\theta})^{\gamma-1} \cdot \frac{\gamma}{I_{p,\gamma}(\boldsymbol{\theta})^{\frac{\gamma-1}{\gamma}}}, \quad (45)$$

For derivations of both losses, we refer to section 1.2. Similarly, the work of Hooker and Vidyashankar [35] shows how to use losses derived from the Hellinger divergence loss. This is less straightforward because it requires non-parametric estimation of the density  $g$  of  $\mathbf{x}$ , which scales poorly and effectively means that one has to estimate the loss function itself. For a survey of these methods, we refer to Jewson et al. [40].

### 3.5 Regularized Bayesian Inference (RegBayes) [23]

RegBayes formulates Variational Inference (VI) with a regularizer. Specifically, for  $\ell_n$  some negative log likelihood, its posteriors are derived as

$$q^*(\boldsymbol{\theta}) = \arg \min_{q \in \mathcal{Q}(\boldsymbol{\xi}), \boldsymbol{\xi} \in E} L(q|\mathbf{x}, \ell_n, R, \boldsymbol{\xi})$$

$$L(q|\mathbf{x}, \ell, R, \boldsymbol{\xi}) = \mathbb{E}_q[\ell_n(\boldsymbol{\theta}, \mathbf{x})] + \text{KLD}(q||\pi) + R(\boldsymbol{\xi}).$$

For a variational family  $\mathcal{Q}$ , the variables  $\boldsymbol{\xi}$  determine the subspace  $\mathcal{Q}(\boldsymbol{\xi}) \subset \mathcal{Q} \subset \mathcal{P}$  that  $q^*$  can lie in. Making the set  $\mathcal{Q}(\boldsymbol{\xi})$  more flexible incurs a cost expressed via a regularizer  $R(\boldsymbol{\xi})$ . Many regularizers of interest fit into the eqs. (2) of the main paper. All that is required for this is that duality theory admits a representation for the above problem as optimization over all of  $\mathcal{Q}$  regularized by a function  $R'(\boldsymbol{\theta}) = \mathbb{E}_q[\phi(\boldsymbol{\theta}, \mathbf{x})]$ . In this case,  $\boldsymbol{\xi}$  can be dropped from the objective altogether. One can then define  $\ell'_n(\boldsymbol{\theta}, \mathbf{x}) = \ell_n(\boldsymbol{\theta}, \mathbf{x}) + \phi(\boldsymbol{\theta}, \mathbf{x})$  to see that RegBayes solves  $P(\ell'_n, \text{KLD}, \mathcal{Q})$ . This logic can for instance be applied to the Infinite Latent SVM models in Zhu et al. [88].

### 3.6 Variational Inference (for generalized posteriors)

Variational Inference (VI) is traditionally defined as finding the optimal approximation to the exact posterior  $q^*$  by finding the member  $q$  of the variational family  $\mathcal{Q}$  that minimizes the KLD from it to the standard Bayesian posterior  $q^*$ . This does not hinge on  $q^*$  being a function of likelihoods and thus straightforwardly extends to any  $q^*$  solving a generalized Bayesian problem  $P(\ell_n, \text{KLD}, \mathcal{P}(\boldsymbol{\theta}))$ , see eq. (40). The derivation of the Evidence Lower Bound (ELBO) now proceeds by observing that

$$\begin{aligned} \text{KLD}(q||q^*) &= \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \left( \frac{q(\boldsymbol{\theta})}{q^*(\boldsymbol{\theta})} \right) \right] \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \left( \frac{q(\boldsymbol{\theta})}{\frac{\exp \{ -\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \} \pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \exp \{ -\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}} \right) \right] \\ &= \underbrace{\mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \left( \frac{q(\boldsymbol{\theta})}{\exp \{ -\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \} \pi(\boldsymbol{\theta})} \right) \right]}_{\text{Generalized ELBO}} + \underbrace{\log \left( \int_{\boldsymbol{\theta}} \exp \left\{ -\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right\} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \right)}_{\text{Generalized log evidence}} \quad (46) \end{aligned}$$

<sup>5</sup>Despite calling their method a variational method, Futami et al. [22] actually take exact inference with the  $\beta$ -divergence induced loss as their starting point

Note that for  $\ell(\boldsymbol{\theta}, x_i) = -\log(p(\boldsymbol{\theta}|x_i))$  for some likelihood function  $p$ , these two terms recover the familiar ELBO and log evidence terms of VI. Rewriting the Generalized ELBO term further using log-additivity, we find that

$$\mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \left( \frac{q(\boldsymbol{\theta})}{\exp \left\{ -\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right\} \pi(\boldsymbol{\theta})} \right) \right] = \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right] + \text{KLD}(q||\pi). \quad (47)$$

Minimizing this last term will be equivalent to solving the problem  $P(\ell_n, \text{KLD}, \mathcal{Q})$ .

### 3.7 ( $\beta$ -)Variational Autoencoders

Variational Autoencoders (VAEs) [44] and  $\beta$ -VAEs [32] (which recover standard VAEs for  $\beta = 1$ ) are special cases of GVI, too. Here,  $\boldsymbol{\theta}$  is a *latent variable* (typically denoted by  $\mathbf{z}$  in the literature) rather than a parameter in the traditional sense. Specifically, denoting the encoder  $q(\boldsymbol{\theta}|\boldsymbol{\kappa})$  and the parameters of the decoder as  $\boldsymbol{\theta}$ , the loss function of a  $\beta$ -VAE is given by

$$\ell_n^\zeta(\boldsymbol{\theta}, \mathbf{x}) = -\log(p_\zeta(\mathbf{x}|\boldsymbol{\theta})) \quad (48)$$

while its  $D$ -argument is given for a prior  $\pi$  by

$$D_{\beta\text{-VAE}}(q||\pi) = \beta \cdot \text{KLD}(q||\pi). \quad (49)$$

With a given variational family  $\mathcal{Q}$ , the resulting  $\beta$ -VAE objective can then be written in GVI-objective form as  $P(\ell_n^\zeta, D_{\beta\text{-VAE}}, \mathcal{Q})$ , i.e.

$$\arg \min_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})} [\ell_n^\zeta(\boldsymbol{\theta}, \mathbf{x})] + D_{\beta\text{-VAE}}(q||\pi) \right\}. \quad (50)$$

Notice that the parameters  $\zeta$  are also inferred, though via point estimates further minimizing the above objective. The terminology traditionally used for this procedure in the literature on VI is *optimizing the lower bound*. Because another interpretation is more generally applicable in the context of GVI and for additional reasons we outline in section 7.1.3, we prefer to think of this procedure as *choosing the optimal loss* (inside a class indexed by  $\zeta$ ) for a given GVI objective. Lastly, note that the  $\beta$  serves the exact same function as the  $\frac{1}{w}$  in the power likelihood case. Thus, the same duality between loss and uncertainty quantifier arises. Specifically, instead of redefining the uncertainty quantifier, one could alternatively see  $\beta$ -VAE methods as using the standard KLD uncertainty quantifier with the redefined loss  $-\log(p_\zeta(\mathbf{x}|\boldsymbol{\theta})^{1/\beta})$ .

## 4 A comparison of different divergences for uncertainty quantification

### 4.1 Notation

For simplicity, in what follows we abbreviate the GVI objective function associated with  $P(D, \ell_n, \mathcal{Q})$  as  $L_D(q) = L(q|\mathbf{x}, D, \ell_n)$ . Moreover, the results of this section are agnostic to the specific loss function  $\ell_n$  up to multiplication with a constant  $w$ . We leverage this to simplify notation and write the GBI [12] posterior corresponding to  $P(\text{KLD}, w \cdot \ell_n, \mathcal{P}(\boldsymbol{\Theta}))$  as

$$q_w^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp \left\{ -\sum_{i=1}^n w \cdot \ell(\boldsymbol{\theta}, x_i) \right\}. \quad (51)$$

For obvious reasons, we also suppress the subscript if  $w = 1$  and write  $q_1^*(\boldsymbol{\theta}) = q^*(\boldsymbol{\theta})$ . Notice that the above recovers the power likelihood of eq. (43) if  $\ell$  is some negative log likelihood function. GVI does not restrict attention to likelihood-based losses. However, if the loss  $w \cdot \ell_n$  is not a likelihood, a necessary condition for the results presented below to hold is that the normalizing constant for the posterior of eq. (51) exists.

### 4.2 High-level overview

While we motivate GVI as a constrained optimisation rather than a posterior approximation, in this section, we show that GVI objectives for  $D \neq \text{KLD}$  can be shown to have meaningful interpretations – even if viewed as posterior approximations. Inspired by the interpretation of the ELBO for VI, we



derive corresponding bounds on (generalized) marginal loss-likelihoods for GVI. Conceptually, it is instructive to think of these bounds as inequalities of the form

$$L_D(q) \geq C_D + S_D(q). \quad (52)$$

Here, the three quantities have the following interpretations:

- $L_D(q) = L(q|\mathbf{x}, D, \ell_n)$  is the objective function associated with principled Bayesian problem  $P(D, \ell_n, Q)$ .
- $C_D$  is the (generalized) log-marginal likelihood associated with the exact posterior and does not depend on the (variational) posterior  $q$ .
- $S_D(q)$  is a quantity that we call the approximate target of GVI.

Note that we can decompose the approximate target  $S_D(q)$  into a sum of two terms: The KLD between approximate and exact posterior corresponding to  $P(\text{KLD}, w\ell_n, \mathcal{P})$  plus an adjustment term  $T_D(q)$ . I.e., we can write the approximate target as

$$S_D(q) = \text{KLD}(q||q_w^*) + T_D(q). \quad (53)$$

Here the adjustment term  $T_D(q)$  as well as the weight  $w$  depend on the prior regularising divergence  $D$ . More generically, the specific form of the approximate target  $S_D(q)$  will depend on  $D$  and enforce different behaviours for the constructed posteriors. In this sense, this term provides GVI with its additional flexibility with regard to uncertainty quantification. Specifically, it is this term that controls (i) how tightly marginal variances will be fitted as well as (ii) how robust GVI is to badly specified priors.

The inequalities/upper bounds of the form in eq. (52) will allow us to interpret GVI analogously to the traditional interpretation of standard VI. To see this, first recall that standard VI is typically interpreted as minimizing  $\text{KLD}(q||q^*)$ , i.e. the Kullback-Leibler divergence between the exact Bayesian posterior  $q^*$  and the variational posterior  $q \in \mathcal{Q}$ . The bounds derived in the remainder of this section for  $P(\ell_n, D, \mathcal{Q})$  with  $D \in \{D_{AR}^{(\alpha)}, D_B^{(\beta)}, D_G^{(\gamma)}\}$  allow for two related interpretations:

- The first interpretation is that GVI of this form minimizes the *same* objective as VI, except for the addition of the adjustment term encouraging the type of behaviour one wants GVI posteriors to have – namely prior robustness and more/less conservative marginals than VI.
- The second interpretation is that GVI minimizes the KLD-discrepancy to a *generalized* Bayesian posterior  $q_w^*$ . Here,  $w$  takes the value of the relevant of  $D \in \{D_{AR}^{(\alpha)}, D_B^{(\beta)}, D_G^{(\gamma)}\}$ .

From this, we conclude that in addition to being theoretically appealing, the form  $P(\ell_n, D, \Pi)$  also is interpretable as an approximation to (generalized) exact Bayesian posteriors.

The remainder of this section gives a high-level overview of the different upper bounds for  $D \in \{D_{AR}^{(\alpha)}, D_B^{(\beta)}, D_G^{(\gamma)}\}$ . Further, each subsection provides a short interpretation that explains why and how certain upper bounds encourage desirable behaviours for the variational posterior. The bounds are proven in Section 4.3

We note at this stage that VI is often associated with maximizing a lower bound on the log-marginalised likelihood (aka. the ELBO), while Eq. (52) interprets GVI as minimizing an upper bound on this. We believe it is more conceptually appealing to consider the minimization approach of GVI. That being said, a simple negation of the GVI objective function and a reversion of the bounds stated below will enable the wide-spread interpretation of maximizing an upper bound on the (generalized) log-marginalised likelihood.

#### 4.2.1 Kullback-Leibler Divergence/standard VI ( $D = \text{KLD}$ )

To explicitly link to standard VI, we will first phrase well-known facts about standard VI in terms of the more general representation used in eq. (52). Specifically, to recover eq. (52) from eq. (46), one

simply rearranges terms and sets

$$\begin{aligned} L_{\text{KLD}}(q) &= \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right] + \text{KLD}(q||\pi) \\ C_{\text{KLD}} &= -\log \left( \int_{\boldsymbol{\theta}} \exp \left\{ -\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right\} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \right) \\ S_{\text{KLD}}(q) &= \text{KLD}(q||q^*). \end{aligned}$$

Unlike most GVI methods, VI is a special case where the inequality of eq. (52) is in fact an equality and there is no adjustment term, i.e.  $T_{\text{KLD}}(q) = 0$ :

$$L_{\text{KLD}}(q) = C_{\text{KLD}} + S_{\text{KLD}}(q).$$

*Interpretation:* The above equation shows that minimizing  $L_{\text{KLD}}(q)$  is equivalent to minimizing  $\text{KLD}(q||q^*) = S_{\text{KLD}}(q)$ . This corresponds to the most common interpretation of VI.

As we shall see next, conceptually similar versions of this insight with different interpretations hold for GVI with  $D \neq \text{KLD}$ , too. For these alternative divergences, we first provide intuition and then prove the bounds in the theorems below. Unsurprisingly, GVI yields different bounds for different values of the parameter either side of the KLD parametrization. For example, GVI with  $D = D_{\text{AR}}^{(\alpha)}$  gives *more* conservative uncertainty quantification than  $D = \text{KLD}$  if  $\alpha \in (0, 1)$ , but *less* conservative uncertainty quantification than  $D = \text{KLD}$  for  $\alpha > 1$ . Accordingly, one obtains different bounds for  $\alpha > 1$  and  $\alpha \in (0, 1)$ . Similar results hold for the  $D_B^{(\beta)}$  and  $D_G^{(\gamma)}$ . To make these differences obvious in the interpretations of the bounds, we separate these cases throughout the remainder of this section.

#### 4.2.2 Rényi's $\alpha$ -divergence ( $D = D_{\text{AR}}^{(\alpha)}$ ) with $\alpha \in (0, 1)$

If  $\alpha \in (0, 1)$ , one finds

$$\begin{aligned} L_{\text{AR}}^{(0,1)}(q) &= \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right] + D_{\text{AR}}^{(\alpha)}(q||\pi) \\ C_{\text{AR}}^{(0,1)} &= -\log \int \pi(\boldsymbol{\theta}) \exp(-\ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta} \\ S_{\text{AR}}^{(0,1)}(q) &= \text{KLD}(q||q^*) + T_{\text{AR}}^{(0,1)}(q) \end{aligned} \tag{54}$$

$$T_{\text{AR}}^{(0,1)}(q) = D_{\text{AR}}^{(\alpha)}(q||\pi) - \text{KLD}(q||\pi). \tag{55}$$

This is the second special case (besides standard VI) where the relation holds with equality:

$$L_{\text{AR}}^{(0,1)}(q) = C_{\text{AR}}^{(0,1)} + S_{\text{AR}}^{(0,1)}(q).$$

*Interpretation:* Eq. (54) shows that minimising  $L_{\text{AR}}^{(0,1)}(q)$  encodes a trade-off between minimizing  $\text{KLD}(q||q^*)$  and minimizing the adjustment term  $T_{\text{AR}}^{(0,1)}$ .  $\text{KLD}(q||q^*)$  is the same target as in traditional VI. However the adjustment term will induce a different uncertainty quantification: In particular, it will typically encourage the posterior to be regularised *more strongly* against the prior, which is to say that  $D_{\text{AR}}^{(\alpha)}(q||\pi) \geq \text{KLD}(q||\pi)$ . The attentive reader may wonder how one would hope to attain the desired prior-robustness if the  $D_{\text{AR}}^{(\alpha)}$  for  $\alpha \in (0, 1)$  regularizes  $q$  against  $\pi$  more strongly than the KLD. To answer this question, one needs to note that the notion of closeness to the prior is determined by the geometry induced by the divergence. It is clear that  $D_{\text{AR}}^{(\alpha)}(q||\pi)$  and  $\text{KLD}(q||\pi)$  produce very different geometries. Specifically, the demonstrations in Section 4.4 show that closeness in terms of  $D_{\text{AR}}^{(\alpha)}$  favours an increase in variance of the posterior relative to the KLD, but focuses less on the location of the posterior. It is precisely this behaviour that allows the  $D_{\text{AR}}^{(\alpha)}$  to achieve prior robustness.

#### 4.2.3 Rényi's $\alpha$ -divergence ( $D = D_{AR}^{(\alpha)}$ ) with $\alpha > 1$

If  $\alpha > 1$ , one finds

$$\begin{aligned} L_{AR}^{(1,\infty)}(q) &= \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right] + D_{AR}^{(\alpha)}(q||\pi) \\ C_{AR}^{(1,\infty)} &= -\frac{1}{\alpha} \log \int \pi(\boldsymbol{\theta}) \exp(-\alpha \ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta} \\ S_{AR}^{(1,\infty)}(q) &= \frac{1}{\alpha} \text{KLD}(q||q_\alpha^*). \end{aligned} \quad (56)$$

*Interpretation:* Eq. (56) shows that minimising  $L_{AR}^{(1,\infty)}(q)$  makes  $\text{KLD}(q||q_\alpha^*)$  small. Since  $\alpha > 1$ , this implies that  $P(D_{AR}^{(\alpha)}, \ell_n, Q)$  produces posteriors that are *more* concentrated than  $P(\text{KLD}, \ell_n, Q)$ . This is so because  $q_\alpha^*$  for  $\alpha > 1$  up-weights the loss relative to  $q^*$ , making it more concentrated around the in-sample minimizer  $\hat{\boldsymbol{\theta}}_n$  of  $\ell_n$  than  $q^*$ .

#### 4.2.4 $\beta$ -divergence ( $D = D_B^{(\beta)}$ ) with $\beta \in (0, 1)$

If  $\beta \in (0, 1)$ , one finds

$$\begin{aligned} L_B^{(0,1)}(q) &= \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right] + \frac{1}{\beta(\beta-1)} \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\beta-1}] - \frac{1}{(\beta-1)} \mathbb{E}_{q(\boldsymbol{\theta})} [\pi(\boldsymbol{\theta})^{\beta-1}] \\ C_B^{(0,1)} &= -\log \int \pi(\boldsymbol{\theta}) \exp(-\ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta} \\ S_B^{(0,1)}(q) &= \text{KLD}(q||q^*) + T_B^{(0,1)}(q) \end{aligned} \quad (57)$$

$$T_B^{(0,1)}(q) = \frac{1}{\beta(\beta-1)} \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\beta-1}] - \mathbb{E}_{q(\boldsymbol{\theta})} [\log q(\boldsymbol{\theta})] - \frac{1}{\beta-1}. \quad (58)$$

*Interpretation:* Eq. (58) shows that minimising  $L_B^{(0,1)}(q)$  encodes a trade-off between minimizing  $\text{KLD}(q||q^*)$  and minimizing  $T_B^{(0,1)}(q)$ . These two terms being traded off against one another has a clear interpretation: In particular,  $\text{KLD}(q||q^*)$  is simply the same target as in traditional VI. Further, it is also straightforward to show that the adjustment term encourages the solution to  $P(D_B^{(\beta)}, \ell_n, Q)$  with  $0 < \beta < 1$  to have greater variance than that of  $P(\text{KLD}, \ell_n, Q)$  (i.e., VI). We can see this by rewriting

$$T_B^{(0,1)}(q) = -\frac{1}{\beta} h_T^{(\beta)}(q(\boldsymbol{\theta})) + h_{\text{KLD}}(q(\boldsymbol{\theta})) + \frac{1-\beta}{\beta},$$

where  $h_{\text{KLD}}(q(\boldsymbol{\theta}))$  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  $\beta$ . Now Lemma 2 can again be applied to show that for  $0 < \beta < 1$ ,  $h_T^{(\beta)}(q(\boldsymbol{\theta})) > h_{\text{KLD}}(q(\boldsymbol{\theta}))$ . It follows that minimising  $-\frac{1}{\beta} h_T^{(\beta)}(q(\boldsymbol{\theta})) + h_{\text{KLD}}(q(\boldsymbol{\theta}))$  for  $0 < \beta < 1$  will lead to making  $h_T^{(\beta)}(q(\boldsymbol{\theta}))$  large – an effect that is achieved by increasing the variance of  $q(\boldsymbol{\theta})$ .

#### 4.2.5 $\beta$ -divergence ( $D = D_B^{(\beta)}$ ) with $\beta > 1$

If  $\beta > 1$ , one finds

$$\begin{aligned} L_B^{(1,\infty)}(q) &= \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right] + \frac{1}{\beta(\beta-1)} \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\beta-1}] - \frac{1}{(\beta-1)} \mathbb{E}_{q(\boldsymbol{\theta})} [\pi(\boldsymbol{\theta})^{\beta-1}] \\ C_B^{(1,\infty)} &= -\frac{1}{\beta} \log \int \pi(\boldsymbol{\theta}) \exp(-\beta \ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta} \\ S_B^{(1,\infty)}(q) &= \frac{1}{\beta} \text{KLD}(q||q_\beta^*) + T_B^{(1,\infty)}(q) \end{aligned} \quad (59)$$

$$T_B^{(1,\infty)}(q) = \frac{1}{\beta} \mathbb{E}_{q(\boldsymbol{\theta})} [\log \pi(\boldsymbol{\theta})] - \frac{1}{(\beta-1)} \mathbb{E}_{q(\boldsymbol{\theta})} [\pi(\boldsymbol{\theta})^{\beta-1}] + \frac{1}{\beta(\beta-1)}. \quad (60)$$

*Interpretation:* Eq. (59) shows that minimising  $L_B^{(1,\infty)}(q)$  encodes a trade-off between minimizing  $\text{KLD}(q||q^*)$  with minimizing  $T_B^{(1,\infty)}(q)$ . Minimising  $\text{KLD}(q||q^*)$  for  $\beta > 1$  will encourage the solution of  $P(D_B^{(\beta)}, \ell_n, Q)$  to be more concentrated around the minimiser  $\hat{\theta}_n$  of  $\ell_n$  than  $P(\text{KLD}, \ell_n, Q)$  (i.e., VI). Additionally, one can show that minimising the adjustment term also favours shrinking the variance of  $q(\theta)$ . To see this, rewrite

$$T_B^{(1,\infty)}(q) = \frac{1}{\beta} \mathbb{E}_{q(\theta)} [\log(\pi(\theta))] - \frac{1}{\beta-1} \mathbb{E}_{q(\theta)} [\pi(\theta)^{\beta-1} - 1] - \frac{1}{\beta}. \quad (61)$$

Applying Lemma 2 then shows that for  $\beta > 1$ ,

$$\frac{1}{\beta-1} \mathbb{E}_{q(\theta)} [\pi(\theta)^{\beta-1} - 1] \geq \mathbb{E}_{q(\theta)} [\log(\pi(\theta))] \geq \frac{1}{\beta} \mathbb{E}_{q(\theta)} [\log(\pi(\theta))]. \quad (62)$$

From this, it follows that minimising Eq. (61) will make  $\frac{1}{\beta-1} \mathbb{E}_{q(\theta)} [\pi(\theta)^{\beta-1}]$  large. Fixing  $\pi(\theta)$ , maximising  $\frac{1}{\beta-1} \mathbb{E}_{q(\theta)} [\pi(\theta)^{\beta-1}]$  plus  $\frac{1}{\beta}$  times the Tsallis entropy of  $q(\theta)$  is equivalent to minimising  $D_B^{(\beta)}(q(\theta)||\pi(\theta))$ . Because  $D_B^{(\beta)}$  is a divergence, this maximization would seek to choose  $q(\theta)$  close to  $\pi(\theta)$ . The Tsallis entropy term in this formulation would have acted to increase the variance of  $q(\theta)$ . Conversely, since we maximize only  $\frac{1}{\beta-1} \mathbb{E}_{q(\theta)} [\pi(\theta)^{\beta-1}]$  (i.e. without adding the Tsallis entropy of  $q(\theta)$ ), choices of  $\beta > 1$  will lead to shrinking the variance of  $q(\theta)$  relative to standard VI.

#### 4.2.6 $\gamma$ -divergence ( $D = D_G^{(\gamma)}$ ) with $\gamma \in (0, 1)$

If  $\gamma \in (0, 1)$ , one finds

$$L_G^{(0,1)}(q) = \mathbb{E}_{q(\theta)} \left[ \sum_{i=1}^n \ell(\theta, x_i) \right] + \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\theta)} [q(\theta)^{\gamma-1}] - \frac{1}{(\gamma-1)} \log \mathbb{E}_{q(\theta)} [\pi(\theta)^{\gamma-1}]$$

$$C_G^{(0,1)} = -\log \int \pi(\theta) \exp(-\ell_n(\theta, \mathbf{x})) d\theta$$

$$S_G^{(0,1)}(q) = \text{KLD}(q||q^*) + T_G^{(0,1)}(q) \quad (63)$$

$$T_G^{(0,1)}(q) = \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\theta)} [q(\theta)^{\gamma-1}] - \mathbb{E}_{q(\theta)} [\log q(\theta)]. \quad (64)$$

*Interpretation:* Eq. (64) shows that minimising  $L_G^{(0,1)}(q)$  trades off minimizing  $\text{KLD}(q||q^*)$  with minimizing  $T_G^{(0,1)}(q)$ . While  $\text{KLD}(q||q^*)$  is the same target as in traditional VI, it is straightforward to show that the adjustment term encourages the solution to  $P(D_G^{(\gamma)}, \ell_n, Q)$  with  $0 < \gamma < 1$  to have greater variance than that of  $P(\text{KLD}, \ell_n, Q)$  (i.e., VI). Specifically, one can rewrite

$$T_G^{(0,1)}(q) = -\frac{1}{\gamma} h_R^{(\gamma)}(q(\theta)) + h_{\text{KLD}}(q(\theta)), \quad (65)$$

where  $h_{\text{KLD}}(q(\theta))$  is the Shannon entropy of  $q(\theta)$  and  $h_R^{(\gamma)}(q(\theta))$  is the Rényi entropy of  $q(\theta)$  with parameter  $\gamma$ . Now Theorem 3 in Van Erven and Harremoës [80] can be extended to show that  $h_R^{(\gamma)}(q(\theta))$  is decreasing in  $\gamma$ . Since it is also well-known that  $\lim_{\gamma \rightarrow 1} h_R^{(\gamma)}(q(\theta)) = h_{\text{KLD}}(q(\theta))$ , it follows that minimising  $-\frac{1}{\gamma} h_R^{(\gamma)}(q(\theta)) + h_{\text{KLD}}(q(\theta))$  for  $0 < \gamma < 1$  will make  $h_R^{(\gamma)}(q(\theta))$  large – an effect that is achieved by increasing the variance of  $q(\theta)$ .

#### 4.2.7 $\gamma$ -divergence ( $D = D_G^{(\gamma)}$ ) with $\gamma > 1$

If  $\gamma > 1$ , one finds

$$L_G^{(1,\infty)}(q) = \mathbb{E}_{q(\theta)} \left[ \sum_{i=1}^n \ell(\theta, x_i) \right] + \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\theta)} [q(\theta)^{\gamma-1}] - \frac{1}{(\gamma-1)} \log \mathbb{E}_{q(\theta)} [\pi(\theta)^{\gamma-1}]$$

$$C_G^{(1,\infty)} = -\frac{1}{\gamma} \log \int \pi(\theta) \exp(-\gamma \ell_n(\theta, \mathbf{x})) d\theta$$

$$S_G^{(1,\infty)}(q) = \frac{1}{\gamma} \text{KLD}(q||q^*) + T_G^{(1,\infty)}(q) \quad (66)$$

$$T_G^{(1,\infty)}(q) = \frac{1}{\gamma} \mathbb{E}_{q(\theta)} [\log \pi(\theta)] - \frac{1}{(\gamma-1)} \log \mathbb{E}_{q(\theta)} [\pi(\theta)^{\gamma-1}]. \quad (67)$$

*Interpretation:* Eq. (66) shows that minimising  $L_G^{(1,\infty)}(q)$  is trading off making  $\text{KLD}(q||q_\gamma^*)$  small with also making  $T_G^{(1,\infty)}(q)$  small. Minimising  $\text{KLD}(q||q_\gamma^*)$  for  $\gamma > 1$  will encourage the solution of  $P(D_G^{(\gamma)}, \ell_n, Q)$  to be more concentrated than minimising  $\text{KLD}(q||q^*)$ . We can also show that the adjustment terms  $T_G^{(1,\infty)}(q)$  encourage shrinkage of the variance of  $q(\theta)$ .

Jensen's inequality shows that for  $\gamma > 1$

$$\frac{1}{\gamma-1} \log \mathbb{E}_{q(\theta)} [\pi(\theta)^{\gamma-1}] \geq \mathbb{E}_{q(\theta)} [\log(\pi(\theta))] \geq \frac{1}{\gamma} \mathbb{E}_{q(\theta)} [\log(\pi(\theta))]. \quad (68)$$

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

### 4.3 Theorems and Proofs

This section is devoted to the rigorous (albeit slightly tedious) derivation of the bounds presented above. Some of these proofs rely on a technical Lemma that we state and prove first. Its function in the proofs will be to related the polynomial  $\frac{Z^x}{x}$  to  $\log(Z)$

**Lemma 2** (A Taylor series bound for the natural logarithm). The natural logarithm of a positive real number  $Z$  can be bounded as follows

- If  $x > 0$

$$\log(Z) \leq \frac{Z^x - 1}{x} \quad (69)$$

- If  $x < 0$

$$\log(Z) \geq \frac{Z^x - 1}{x}. \quad (70)$$

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

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

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  $\log(Z)$ .  $\square$

With this, we are ready to prove the lower bounds for three divergences: Rényi's  $\alpha$ -divergence ( $D_{AR}^{(\alpha)}$ ), the  $\beta$ -divergence ( $D_B^{(\beta)}$ ) and the  $\gamma$ -divergence ( $D_G^{(\gamma)}$ ).

#### 4.3.1 Bounds for Rényi's $\alpha$ -divergence ( $D_{AR}^{(\alpha)}$ )

**Theorem 8** (Lower bounding the marginal loss-likelihood using the  $D_{AR}^{(\alpha)}$  uncertainty quantifier). The objective function,  $L(q|x, D_{AR}^{(\alpha)}, \ell_n)$ , associated with Bayesian problem,  $P(D_{AR}^{(\alpha)}, \ell_n, Q)$ , can be used to lower bound the marginal loss-likelihood (normalising constant) of the GBIPosterior

- If  $\alpha > 1$

$$\frac{1}{\alpha} \log \int \pi(\theta) \exp(-\alpha \ell_n(\theta, x)) d\theta \geq \frac{1}{\alpha} \text{KLD}(q(\theta)||\pi^{\alpha\ell}(\theta|x)) - L(q|D_{AR}^{(\alpha)}, \ell_n) \quad (71)$$

$$\text{where } \pi^{\alpha\ell}(\theta|x) = \frac{\pi(\theta) \exp(-\alpha \ell_n(\theta, x))}{\int \pi(\theta) \exp(-\alpha \ell_n(\theta, x)) d\theta}.$$

- If  $0 < \alpha < 1$

$$\log \int \pi(\boldsymbol{\theta}) \exp(-\ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta} = \text{KLD}(q(\boldsymbol{\theta}) || \pi^\ell(\boldsymbol{\theta} | \mathbf{x})) - L(q | D_{AR}^{(\alpha)}, \ell_n) \quad (72)$$

$$+ D_{AR}^{(\alpha)}(q(\boldsymbol{\theta}) || \pi(\boldsymbol{\theta})) - \text{KLD}(q(\boldsymbol{\theta}) || \pi(\boldsymbol{\theta}))$$

$$\text{where } \pi^\ell(\boldsymbol{\theta} | \mathbf{x}) = \frac{\pi(\boldsymbol{\theta}) \exp(-\ell_n(\boldsymbol{\theta}, \mathbf{x}))}{\int \pi(\boldsymbol{\theta}) \exp(-\ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta}}$$

*Proof. Case 1)  $\alpha > 1$*

Jensen's inequality and the concavity of the natural logarithm give us that

$$\log \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \left( \frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right)^{\alpha-1} \right] \geq \mathbb{E}_{q(\boldsymbol{\theta})} \left[ (\alpha-1) \log \left( \frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right) \right] \quad (73)$$

As a result we can write the objective function associated with the Bayesian decision problem as

$$\begin{aligned} & L(q | \mathbf{x}, D_{AR}^{(\alpha)}, \ell_n) \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} [\ell_n(\boldsymbol{\theta}, \mathbf{x})] + \frac{1}{\alpha(\alpha-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \left( \frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right)^{\alpha-1} \right] \\ &\geq \mathbb{E}_{q(\boldsymbol{\theta})} [\ell_n(\boldsymbol{\theta}, \mathbf{x})] + \frac{1}{\alpha(\alpha-1)} \mathbb{E}_{q(\boldsymbol{\theta})} \left[ (\alpha-1) \log \left( \frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right) \right] \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} [\ell_n(\boldsymbol{\theta}, \mathbf{x})] + \frac{1}{\alpha} \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \left( \frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right) \right] \\ &= \frac{1}{\alpha} \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \left( \frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}) \exp(-\alpha \ell_n(\boldsymbol{\theta}, \mathbf{x}))} \right) \right] \\ &= \frac{1}{\alpha} \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \left( \frac{q(\boldsymbol{\theta})}{\frac{\pi(\boldsymbol{\theta}) \exp(-\alpha \ell_n(\boldsymbol{\theta}, \mathbf{x}))}{\int \pi(\boldsymbol{\theta}) \exp(-\alpha \ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta}}} \right) \right] - \frac{1}{\alpha} \log \int \pi(\boldsymbol{\theta}) \exp(-\alpha \ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta} \\ &= \frac{1}{\alpha} \text{KLD}(q(\boldsymbol{\theta}) || \pi^{\alpha \ell}(\boldsymbol{\theta} | \mathbf{x})) - \frac{1}{\alpha} \log \int \pi(\boldsymbol{\theta}) \exp(-\alpha \ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta} \end{aligned}$$

Which can be rearranged to give Eq. (71).

**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} & L(q | \mathbf{x}, D_{AR}^{(\alpha)}, \ell_n) \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} [\ell(\boldsymbol{\theta}, x)] + D_{AR}^{(\alpha)}(q(\boldsymbol{\theta}) || \pi(\boldsymbol{\theta})) \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} [\ell(\boldsymbol{\theta}, x)] + \mathbb{E}_{q(\boldsymbol{\theta})} [\log(\pi(\boldsymbol{\theta}))] - \mathbb{E}_{q(\boldsymbol{\theta})} [\log(\pi(\boldsymbol{\theta}))] + D_{AR}^{(\alpha)}(q(\boldsymbol{\theta}) || \pi(\boldsymbol{\theta})) \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} [\log(\pi(\boldsymbol{\theta}))] - \mathbb{E}_{q(\boldsymbol{\theta})} [\log(\pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x)))] + D_{AR}^{(\alpha)}(q(\boldsymbol{\theta}) || \pi(\boldsymbol{\theta})) \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} [\log(\pi(\boldsymbol{\theta}))] - \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \left( \frac{\pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x))}{\int \pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta}} \right) \right] \\ &\quad - \log \int \pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta} + D_{AR}^{(\alpha)}(q(\boldsymbol{\theta}) || \pi(\boldsymbol{\theta})) \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} [\log(\pi(\boldsymbol{\theta}))] - \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \left( \frac{\pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x))}{\int \pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta}} \right) \right] - \mathbb{E}_{q(\boldsymbol{\theta})} [\log(q(\boldsymbol{\theta}))] \\ &\quad + \mathbb{E}_{q(\boldsymbol{\theta})} [\log(q(\boldsymbol{\theta}))] - \log \int \pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta} + D_{AR}^{(\alpha)}(q(\boldsymbol{\theta}) || \pi(\boldsymbol{\theta})) \\ &= -\text{KLD}(q(\boldsymbol{\theta}) || \pi(\boldsymbol{\theta})) + \text{KLD}(q(\boldsymbol{\theta}) || \pi^\ell(\boldsymbol{\theta} | \mathbf{x})) - \log \int \pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta} + D_{AR}^{(\alpha)}(q(\boldsymbol{\theta}) || \pi(\boldsymbol{\theta})) \end{aligned}$$



Which can be rearranged to give Eq. (73).

□

#### 4.3.2 Bounds for the $\beta$ -divergence ( $D_B^{(\beta)}$ )

**Theorem 9** (Lower bounding the marginal loss-likelihood using the  $D_B^{(\beta)}$  uncertainty quantifier). The objective function,  $L(q|\mathbf{x}, D_B^{(\beta)}, \ell_n)$ , associated with Bayesian problem,  $P(D_B^{(\beta)}, \ell_n, Q)$ , can be used to lower bound the marginal loss-likelihood (normalising constant) of the GB1posterior

- If  $0 < \beta < 1$

$$\begin{aligned} \log \int \pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta} &\geq \text{KLD}(q(\boldsymbol{\theta}) || \pi^\ell(\boldsymbol{\theta}|\mathbf{x})) - L(q|D_B^{(\beta)}, \ell_n) \\ &\quad + \frac{1}{\beta(\beta-1)} \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\beta-1}] - \mathbb{E}_{q(\boldsymbol{\theta})} [\log q(\boldsymbol{\theta})] - \frac{1}{\beta-1} \end{aligned} \quad (74)$$

$$\text{where } \pi^\ell(\boldsymbol{\theta}|\mathbf{x}) = \frac{\pi(\boldsymbol{\theta}) \exp(-\ell_n(\boldsymbol{\theta}, \mathbf{x}))}{\int \pi(\boldsymbol{\theta}) \exp(-\ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta}}.$$

- If  $\beta > 1$

$$\begin{aligned} \frac{1}{\beta} \log \int \pi(\boldsymbol{\theta}) \exp(-\beta \ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta} &\geq \frac{1}{\beta} \text{KLD}(q(\boldsymbol{\theta}) || \pi^{\beta\ell}(\boldsymbol{\theta}|\mathbf{x})) - L(q|D_B^{(\beta)}, \ell_n) \\ &\quad + \frac{1}{\beta} \mathbb{E}_{q(\boldsymbol{\theta})} [\log \pi(\boldsymbol{\theta})] - \frac{1}{(\beta-1)} \mathbb{E}_{q(\boldsymbol{\theta})} [\pi(\boldsymbol{\theta})^{\beta-1}] + \frac{1}{\beta(\beta-1)} \end{aligned} \quad (75)$$

$$\text{where } \pi^{\beta\ell}(\boldsymbol{\theta}|\mathbf{x}) = \frac{\pi(\boldsymbol{\theta}) \exp(-\beta \ell_n(\boldsymbol{\theta}, \mathbf{x}))}{\int \pi(\boldsymbol{\theta}) \exp(-\beta \ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta}}.$$

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

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

We have to consider two cases for  $\beta$  as the positivity and negativity of  $\beta - 1$  affect which part of Lemma 2 we use.

**Case 1)**  $0 < \beta < 1$

Lemma 2 gives us that for  $\beta - 1 < 0$ ,  $\frac{Z^{\beta-1}}{\beta-1} \leq \log(Z) + \frac{1}{\beta-1}$  therefore

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

Which can be rearranged to give Eq. (75).

**Case 2)**  $\beta > 1$

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

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

Which can be rearranged to give Eq. (76).

□

#### 4.3.3 Bounds for the $\gamma$ -divergence ( $D_G^{(\gamma)}$ )

**Theorem 10** (Lower bounding the marginal loss-likelihood using the  $D_G^{(\gamma)}$  uncertainty quantifier). The objective function,  $L(q|\mathbf{x}, D_G^{(\gamma)}, \ell_n)$ , associated with Bayesian problem,  $P(D_G^{(\gamma)}, \ell_n, Q)$ , can be used to lower bound the marginal loss-likelihood (normalising constant) of the GBIPosterior in the following ways:

- If  $0 < \gamma < 1$

$$\log \int \pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta} \geq \text{KLD}(q(\boldsymbol{\theta}) || \pi^\ell(\boldsymbol{\theta} | \mathbf{x})) - L(q | D_G^{(\gamma)}, \ell_n) \quad (76)$$

$$+ \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\gamma-1}] - \mathbb{E}_{q(\boldsymbol{\theta})} [\log q(\boldsymbol{\theta})]$$

$$\text{where } \pi^\ell(\boldsymbol{\theta} | \mathbf{x}) = \frac{\pi(\boldsymbol{\theta}) \exp(-\ell_n(\boldsymbol{\theta}, \mathbf{x}))}{\int \pi(\boldsymbol{\theta}) \exp(-\ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta}}.$$

- If  $\gamma > 1$

$$\frac{1}{\gamma} \log \int \pi(\boldsymbol{\theta}) \exp(-\gamma \ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta} \geq \frac{1}{\gamma} \text{KLD}(q(\boldsymbol{\theta}) || \pi^{\gamma \ell}(\boldsymbol{\theta} | \mathbf{x})) - L(q | D_G^{(\gamma)}, \ell_n) \quad (77)$$

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

$$\text{where } \pi^{\gamma \ell}(\boldsymbol{\theta} | \mathbf{x}) = \frac{\pi(\boldsymbol{\theta}) \exp(-\gamma \ell_n(\boldsymbol{\theta}, \mathbf{x}))}{\int \pi(\boldsymbol{\theta}) \exp(-\gamma \ell_n(\boldsymbol{\theta}, \mathbf{x})) d\boldsymbol{\theta}}.$$

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

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

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

$$= \arg \min_{q \in \mathcal{Q}} \{ L(q | \mathbf{x}, \ell_n, D_G^{(\gamma)}) \}$$

We have to consider two cases for  $\gamma$  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  $\mathbb{E}_{q(\boldsymbol{\theta})} [\pi(\boldsymbol{\theta})^{\gamma-1}]$  provides

$$L(q | \mathbf{x}, D_G^{(\gamma)}, \ell_n)$$

$$= \mathbb{E}_{q(\boldsymbol{\theta})} [\ell(\boldsymbol{\theta}, x)] + \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\gamma-1}] - \frac{1}{(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [\pi(\boldsymbol{\theta})^{\gamma-1}]$$

$$= \mathbb{E}_{q(\boldsymbol{\theta})} [\ell(\boldsymbol{\theta}, x)] + \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\gamma-1}] + \frac{1}{(1-\gamma)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [\pi(\boldsymbol{\theta})^{\gamma-1}]$$

$$\geq \mathbb{E}_{q(\boldsymbol{\theta})} [\ell(\boldsymbol{\theta}, x)] + \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\gamma-1}] + \frac{1}{(1-\gamma)} \mathbb{E}_{q(\boldsymbol{\theta})} [(\gamma-1) \log \pi(\boldsymbol{\theta})]$$

$$= \mathbb{E}_{q(\boldsymbol{\theta})} [\ell(\boldsymbol{\theta}, x)] + \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\gamma-1}] - \mathbb{E}_{q(\boldsymbol{\theta})} [\log \pi(\boldsymbol{\theta})]$$

$$= \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\gamma-1}] - \mathbb{E}_{q(\boldsymbol{\theta})} [\log (\pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x)))]$$

$$= \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\gamma-1}] - \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \frac{\pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x))}{\int \pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta}} \right] - \log \int \pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta}$$

$$= \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\gamma-1}] + \text{KLD}(q(\boldsymbol{\theta}) || \pi^\ell(\boldsymbol{\theta} | \mathbf{x})) - \log \int \pi(\boldsymbol{\theta}) \exp(-\ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta} - \mathbb{E}_{q(\boldsymbol{\theta})} [\log q(\boldsymbol{\theta})].$$

Which when rearranged gives Eq. (77). Unfortunately we cannot perform the same trick when  $\gamma > 1$  as  $\frac{1}{1-\gamma}$  is no longer positive and the inequality would be reversed.

**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}
& L(q|\mathbf{x}, D_G^{(\gamma)}, \ell_n) \\
&= \mathbb{E}_{q(\boldsymbol{\theta})} [\ell(\boldsymbol{\theta}, x)] + \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [q(\boldsymbol{\theta})^{\gamma-1}] - \frac{1}{(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [\pi(\boldsymbol{\theta})^{\gamma-1}] \\
&= \mathbb{E}_{q(\boldsymbol{\theta})} [\ell(\boldsymbol{\theta}, x)] + \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} \left[ q(\boldsymbol{\theta})^{\gamma-1} \frac{\pi(\boldsymbol{\theta})^{\gamma-1}}{\pi(\boldsymbol{\theta})^{\gamma-1}} \right] - \frac{1}{(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [\pi(\boldsymbol{\theta})^{\gamma-1}] \\
&\geq \mathbb{E}_{q(\boldsymbol{\theta})} [\ell(\boldsymbol{\theta}, x)] + \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})} [\pi(\boldsymbol{\theta})^{\gamma-1}] \\
&= \frac{1}{\gamma} \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}) \exp(-\gamma \ell(\boldsymbol{\theta}, x))} \right] + \frac{1}{\gamma} \mathbb{E}_{q(\boldsymbol{\theta})} [\log \pi(\boldsymbol{\theta})] - \frac{1}{(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [\pi(\boldsymbol{\theta})^{\gamma-1}] \\
&= \frac{1}{\gamma} \mathbb{E}_{q(\boldsymbol{\theta})} \left[ \log \frac{q(\boldsymbol{\theta})}{\frac{\pi(\boldsymbol{\theta}) \exp(-\gamma \ell(\boldsymbol{\theta}, x))}{\int \pi(\boldsymbol{\theta}) \exp(-\gamma \ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta}}} \right] - \frac{1}{\gamma} \int \pi(\boldsymbol{\theta}) \exp(-\gamma \ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta} \\
&\quad + \frac{1}{\gamma} \mathbb{E}_{q(\boldsymbol{\theta})} [\log \pi(\boldsymbol{\theta})] - \frac{1}{(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [\pi(\boldsymbol{\theta})^{\gamma-1}] \\
&= \frac{1}{\gamma} \text{KLD}(q(\boldsymbol{\theta}) || \pi^{\gamma \ell}(\boldsymbol{\theta}|x)) - \frac{1}{\gamma} \int \pi(\boldsymbol{\theta}) \exp(-\gamma \ell(\boldsymbol{\theta}, x)) d\boldsymbol{\theta} \\
&\quad + \frac{1}{\gamma} \mathbb{E}_{q(\boldsymbol{\theta})} [\log \pi(\boldsymbol{\theta})] - \frac{1}{(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})} [\pi(\boldsymbol{\theta})^{\gamma-1}] \tag{78}
\end{aligned}$$

Which when rearranged gives Eq. (78). □

#### 4.4 Demonstrations

In order to understand the impact the choice of divergence used for regularization and its hyperparameter have on the inference, we consider the following simple Bayesian linear regression example with two predictors and no intercept

$$\begin{aligned}
\sigma^2 &\sim \mathcal{IG}(a_0, b_0) \\
\boldsymbol{\theta}|\sigma^2 &\sim \mathcal{N}_2(\boldsymbol{\mu}_0, \sigma^2 V_0) \\
y_i|\boldsymbol{\theta}, \sigma^2 &\sim \mathcal{N}(X_i \boldsymbol{\theta}, \sigma^2). \tag{79}
\end{aligned}$$

We choose this example because it provides a closed form exact Bayesian posterior as well as a closed form VI and GVI objective. In other words, both the uncertainty quantifier term as well as the expected loss term are available in closed form for  $\ell(\boldsymbol{\theta}, x) = -\log(p(x|\boldsymbol{\theta}))$ . Consequently, no sampling is required, neither for calculating the exact posterior nor for the optimization of the GVI and VI posteriors. Studying the exact posterior for  $\beta$  reveals that if the two variables  $x_1$  and  $x_2$  are correlated, the corresponding exact Bayesian posterior will be strongly correlated, too. As we wish to investigate the underestimation of marginal variances for standard VI as well as the way in which GVI can address this, we simulate the highly correlated variables

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

and compare the performance of the different GVI methods with VI with the variational family  $\mathcal{Q} = \{q(\theta_1, \theta_2, \sigma^2) | q(\theta_1, \theta_2, \sigma^2) = q(\theta_1|\sigma^2, \boldsymbol{\kappa}_n)q(\theta_2|\sigma^2, \boldsymbol{\kappa}_n)q(\sigma^2|\boldsymbol{\kappa}_n), \boldsymbol{\kappa}_n \in \mathbf{K}\}$ , where  $\boldsymbol{\kappa}_n = (a_n, b_n, \mu_{1,n}, \mu_{2,n}, v_{1,n}, v_{2,n})^T$  and

$$\begin{aligned}
q(\sigma^2|\boldsymbol{\kappa}_n) &= \mathcal{IG}(\sigma^2|a_n, b_n) \\
q(\theta_1|\sigma^2, \boldsymbol{\kappa}_n) &= \mathcal{N}(\theta_1|\mu_{1,n}, \sigma^2 v_{1,n}) \\
q(\theta_2|\sigma^2, \boldsymbol{\kappa}_n) &= \mathcal{N}(\theta_2|\mu_{2,n}, \sigma^2 v_{2,n}). \tag{80}
\end{aligned}$$

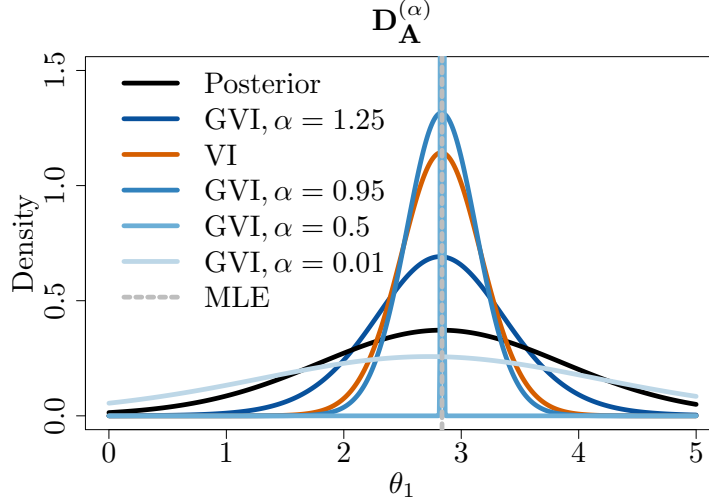


Figure 6: Marginal **VI** and **GVI** posterior for the  $\theta_1$  coefficient of a Bayesian linear model under the  $D_A^{(\alpha)}$  prior regulariser for different values of the divergence hyperparameters. The boundedness of the  $D_A^{(\alpha)}$  causes **GVI** to severely over-concentrate if  $\alpha$  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  $\theta_2 | \sigma^2 \sim \mathcal{N}(0, 25\sigma^2)$ .

For the experiments,  $n = 25$  observations are simulated from eq. (80) with  $\theta = (2, 3)$  and  $\sigma^2 = 4$ . We use the negative log-likelihood corresponding to eq. (80) as loss function and investigate GVI's behaviour across different ranges of  $D$ , specifically  $D \in \{D_A^{(\alpha)}, D_B^{(\beta)}, D_{AR}^{(\alpha)}, D_G^{(\gamma)}\}$ . The results are depicted in Figs. 6 and 8-12. We summarize the most interesting results from these plots in the following three subsections.

#### 4.4.1 The boundedness of the $\alpha$ -divergence ( $D_A^{(\alpha)}$ )

Of the alternative divergences to the KLD contained within the  $D_G^{(\alpha, \beta, \gamma)}$  family [16],  $D_A^{(\alpha)}$  is arguably the most well known. Here we demonstrate that it is not necessarily suitable to quantify uncertainty in a Bayesian problem specified via  $P(\ell_n, D_A^{(\alpha)}, \mathcal{Q})$ . In particular, Fig. 6 shows that the solutions to  $P(\ell_n, D_A^{(\alpha)}, \mathcal{Q})$  can produce degenerate posteriors. For example, when  $\alpha = 0.5$ ,  $P(\ell_n, D_A^{(\alpha)}, \mathcal{Q})$  essentially collapses to the Maximum Likelihood Estimate. This is a consequence of the boundedness of  $D_A^{(\alpha)}$  for  $\alpha \in (0, 1)$ : One can show that  $D_A^{(\alpha)} \leq \frac{1}{\alpha(1-\alpha)}$ . As  $\alpha$  decreases from 1, this upper-bound initially decreases and as a result decreases the maximal penalty for uncertainty quantification far from the prior – this allows the optimisation to focus solely on minimising the in-sample loss. This phenomenon is also depicted in Fig. 7. However, Fig. 7 also shows that the magnitude increases again as  $\alpha$  approaches 0 and for  $\alpha > 1$ , where the divergence is no longer bounded. For these values of the hyperparameter, it is possible to achieve more conservative uncertainty quantification. In Fig. 6 for example,  $\alpha = 1.25$  and  $\alpha = 0.01$  are able to achieve marginal variances that more closely correspond to the exact posterior. In spite of this, the  $D_A^{(\alpha)}$  stands as a cautionary tale: Without understanding the properties of the uncertainty quantifier  $\bar{D}$  sufficiently well, GVI may well yield unsatisfactory posteriors.

#### 4.4.2 Increasing the magnitude of the divergence results in posteriors with large variances

In this section, we summarize the impact that a selection of robust divergences can have on the marginal variances of the solution to  $P(\ell_n, D, \mathcal{Q})$ .

Fig. 7 provides some idea of how the magnitude of the uncertainty quantifier changes with the hyperparameter. Fig. 8 illustrates the impact this has on the marginal variances of the resulting posteriors. The latter plot shows that  $D_B^{(\beta)}$ ,  $D_{AR}^{(\alpha)}$  and  $D_G^{(\gamma)}$  are able to produce more conservative posterior variance for  $\beta, \alpha, \gamma < 1$  and less conservative posterior variance for  $\beta, \alpha, \gamma > 1$ . This is a manifestation of the posterior being penalized more heavily ( $\beta, \alpha, \gamma < 1$ ) or less heavily ( $\beta, \alpha, \gamma > 1$ ) for deviating from the prior than under the traditional VI. This is also illustrated by Fig. 7 which

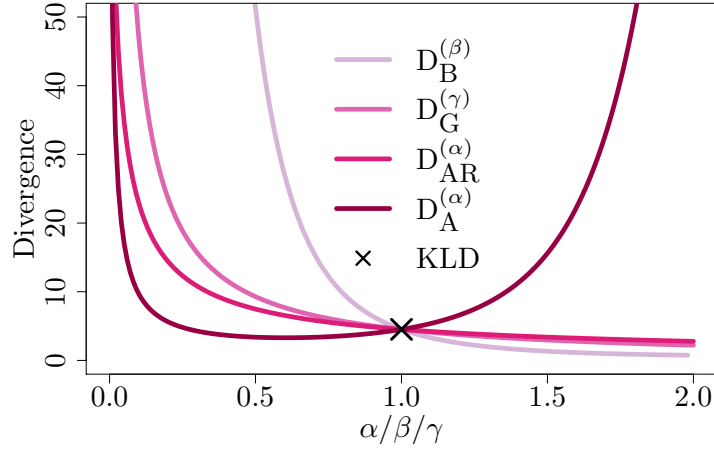


Figure 7: A comparison of the sizes of the  $D_A^{(\alpha)}$ ,  $D_B^{(\beta)}$ ,  $D_{AR}^{(\alpha)}$  and  $D_G^{(\gamma)}$  between two bivariate NIG families with  $a_n = 512$ ,  $b_n = 543$ ,  $\mu_n = (2.5, 2.5)$ ,  $\mathbf{V}_n = \text{diag}(0.3, 2)$  and  $a_0 = 500$ ,  $b_0 = 500$ ,  $\mu_0 = (0, 0)$ ,  $V_0 = \text{diag}(25, 2)$  for various values of the hyperparameter with the KLD.

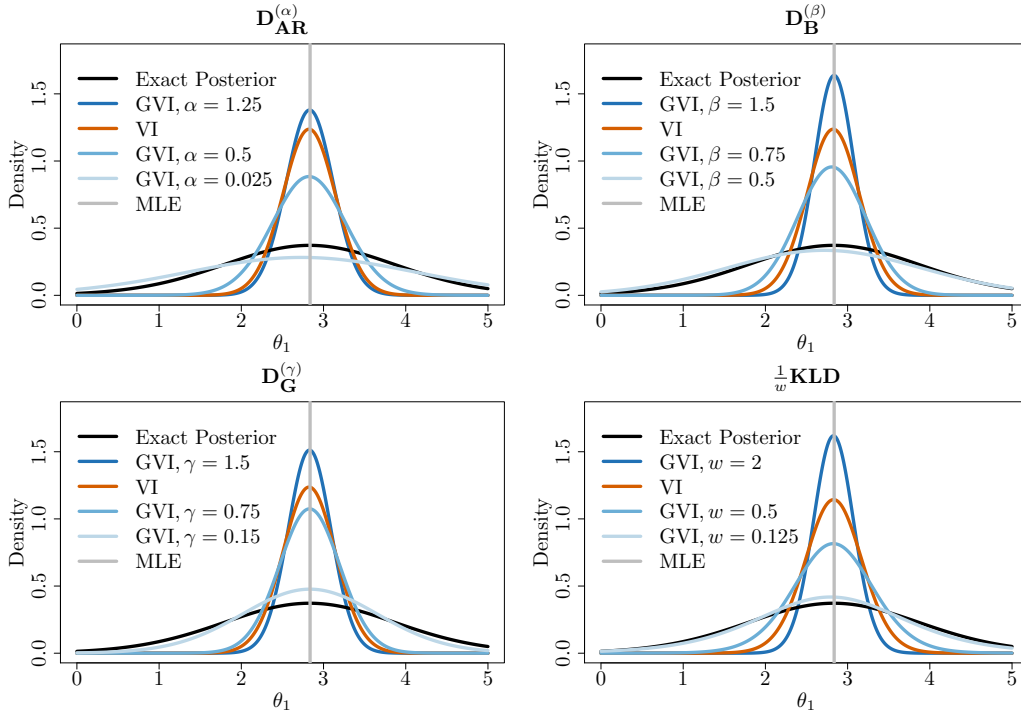


Figure 8: Marginal **VI** and **GVI** posterior for the  $\theta_1$  coefficient of a Bayesian linear model under the  $D_{AR}^{(\alpha)}$ ,  $D_B^{(\beta)}$ ,  $D_G^{(\gamma)}$  and  $\frac{1}{w}$  KLD prior regularisers for different values of the divergence hyperparameters. Correlated covariates cause dependency in the exact posterior of the coefficients  $\theta$ , and as a result **VI** underestimates marginal variances. **GVI** has the flexibility to more accurately capture the exact marginal variances. Prior Specification:  $\sigma^2 \sim \text{IG}(20, 50)$ ,  $\theta_1 \sim \mathcal{N}(0, 5^2)$  and  $\theta_2 \sim \mathcal{N}(0, 5^2)$ .

shows that the magnitude of these divergences increases as the hyperparameters decrease below 1 and decreases as the hyperparameters increase above 1.

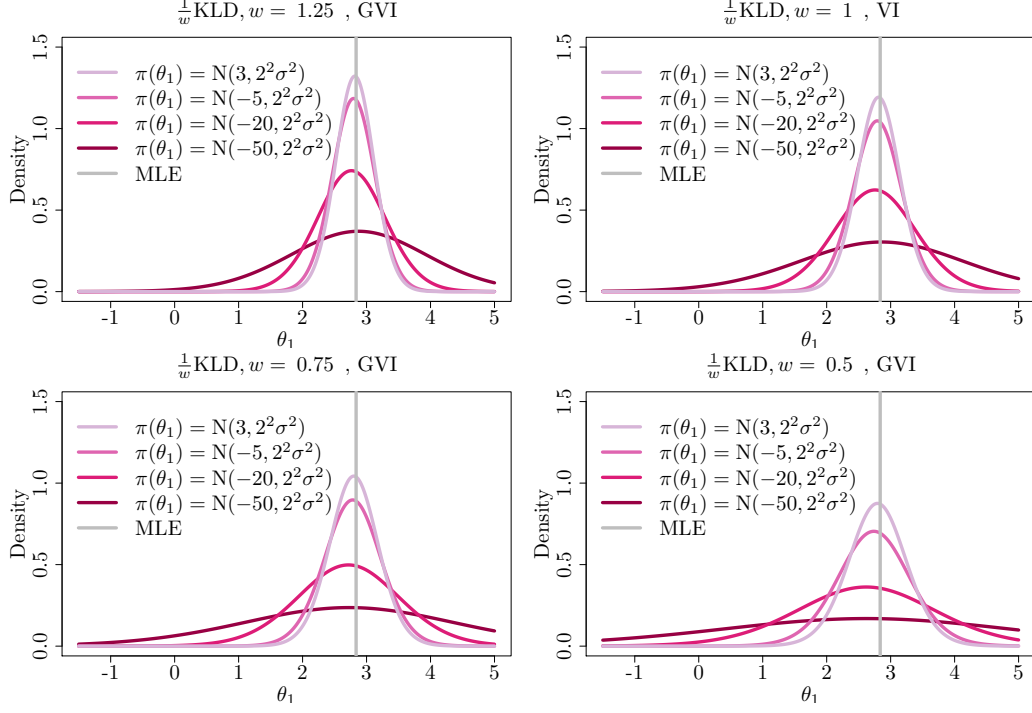


Figure 9: Marginal VI and GVI posterior for the  $\theta_1$  coefficient of a Bayesian linear model under different prior specifications and the using the  $\frac{1}{w}$ KLD as the uncertainty quantifying divergence for several values of  $w$ . Prior specification:  $\sigma^2 \sim \mathcal{IG}(3, 5)$ .

As a result, by choosing the divergence and its hyperparameter appropriately, greater control can be exerted over the resulting posterior than is possible with standard VI. Specifically, it allows desirable properties for the posteriors (such as conservative uncertainty quantification or prior robustness) to be directly and transparently incorporated into the form  $P(\ell_n, D, \mathcal{Q})$  via  $D$ .

#### 4.4.3 Robustness to the prior

In this section we compare the impact of changing the uncertainty quantifier 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  $\alpha = \gamma = w$ . This follows the results and intuitions from Theorem 8 and ensures a fair comparison. The  $D_B^{(\beta)}$  uncertainty quantifier required the selection of different values to ensure its availability in a closed form.

**$\frac{1}{w}$ KLD:** Firstly, Fig. 9 examines how weighting the KLD impacts the solution to  $P(\ell_n, \frac{1}{w}\text{KLD}, \mathcal{Q})$ . Choosing  $w < 1$  leads to posteriors that encourage larger variances, making them amenable to conservative uncertainty quantification. However, this comes at the price of making them *more* sensitive to the prior. Finally, we note that  $w > 1$  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 the  $\frac{1}{w}$ KLD uncertainty quantifier unattractive: In essence, one has to choose between wider variances (at the expense of being robust to the prior) and prior robustness (at the expense of 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 provide a way to have your cake and eat it, too.

**$D_{AR}^{(\alpha)}$ :** Fig. 10 demonstrates the sensitivity of  $P(\ell_n, D_{AR}^{(\alpha)}, \mathcal{Q})$  to prior specification. For  $0 < \alpha < 1$ , the  $D_{AR}^{(\alpha)}$  is able to provide more conservative marginal variances than standard VI while being more robust to badly specified priors. That being said, for  $\alpha > 1$  the  $D_{AR}^{(\alpha)}$  is more sensitive to the prior

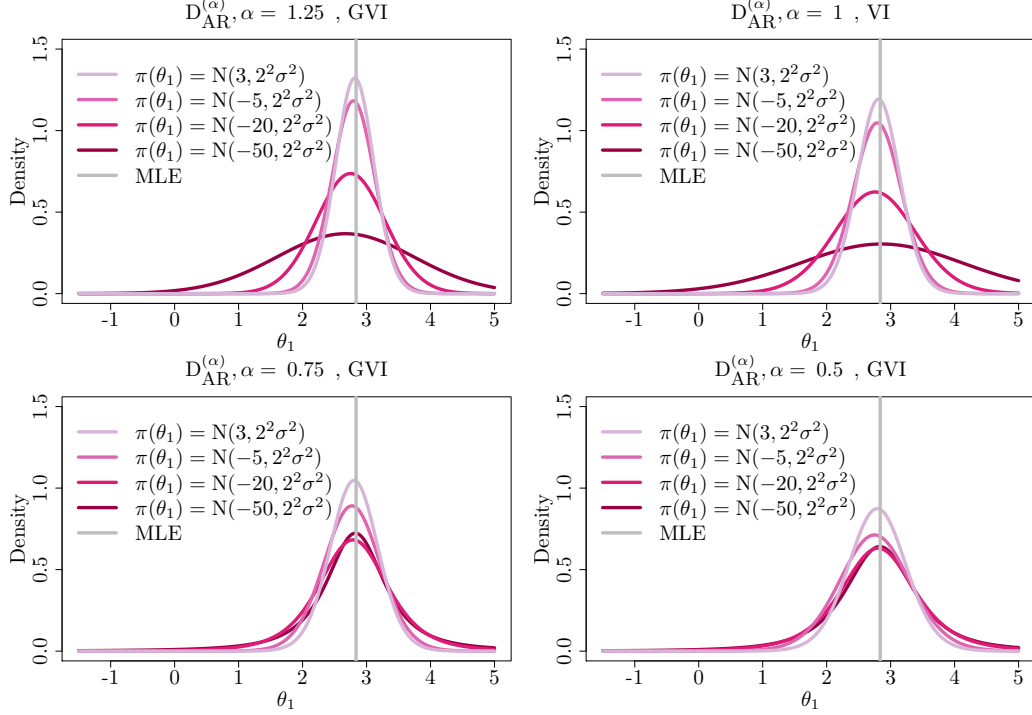


Figure 10: Marginal VI and GVI posterior for the  $\theta_1$  coefficient of a Bayesian linear model under different prior specifications and the using the  $D_{AR}^{(\alpha)}$  as the uncertainty quantifying divergence for several values of  $\alpha$ . Prior specification:  $\sigma^2 \sim \mathcal{IG}(3, 5)$ .

than for  $\alpha \in (0, 1)$ . This can be seen by examining the form of the  $D_{AR}^{(\alpha)}$ :

$$D_{AR}^{(\alpha)}(q(\theta) || \pi(\theta)) = \frac{1}{\alpha(\alpha-1)} \log \int q(\theta)^\alpha \pi(\theta)^{1-\alpha} d\theta \quad (81)$$

$$= \frac{1}{\alpha(\alpha-1)} \log \int \frac{q(\theta)^\alpha}{\pi(\theta)^{\alpha-1}} d\theta \quad (82)$$

where we rearrange to ensure all of the powers are positive. For  $\alpha > 1$  there is now a ratio of densities in the  $D_{AR}^{(\alpha)}$ . This means that if  $q(\theta)$  is large in an area where  $\pi(\theta)$  is not, then a severe penalty is incurred. This limits how far the  $q(\theta)$  can move from the prior and thus results in lack of prior robustness. In fact, this is also implicitly stated in Theorem 8.

$D_B^{(\beta)}$ : Fig. 11 demonstrates the sensitivity of  $P(\ell_n, D_B^{(\beta)}, \mathcal{Q})$  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  $q(\theta)$  and one containing an interaction between them.

$$D_B^{(\beta)}(q(\theta) || \pi(\theta)) = \frac{1}{\beta} \int \pi(\theta)^\beta d\theta - \frac{1}{\beta-1} \int \pi(\theta)^{\beta-1} q(\theta) d\theta + \frac{1}{\beta(\beta-1)} \int q(\theta)^\beta d\theta \quad (83)$$

The integral depending only on the prior is constant in  $q$  and so we can ignore it. Now if  $\beta$  increases substantially above 1, the interaction term between  $\pi(\theta)$  and  $q(\theta)$  will have a smaller weight in the optimisation than the term only involving  $q(\theta)$ . As a result, the optimisation will focus on decreasing  $\int q^\beta(\theta) d\theta$  rather than increasing  $\int \pi^{\beta-1}(\theta) q(\theta) d\theta$ . This is closely linked to the *ignorance to the data* phenomenon discussed in [40] when  $\beta$  gets too big. The uncertainty quantification part of the Bayesian decision problem is therefore largely controlled by a term only involving  $q(\theta)$ . This integral is very large if the variance of  $q(\theta)$  gets very small, which prevents it from converging to a point mass at the MLE as the  $D_A^{(\alpha)}$  did in Fig. 6. Therefore, the  $D_B^{(\beta)}$  is able to provide almost prior-invariant uncertainty quantification.



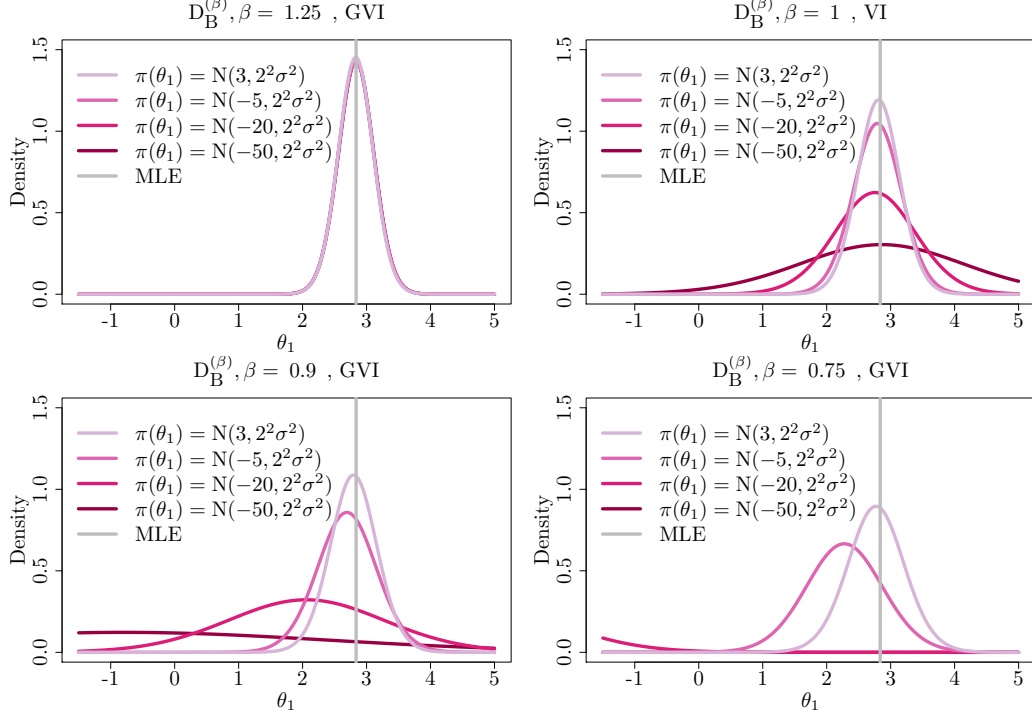


Figure 11: Marginal VI and GVI posterior for the  $\theta_1$  coefficient of a Bayesian linear model under different prior specifications and using the  $D_B^{(\beta)}$  as the uncertainty quantifying divergence for several values of  $\beta$ . Prior specification:  $\sigma^2 \sim \mathcal{IG}(3, 5)$ .

For  $\beta < 1$ , the opposite effect is observed: Here, the integral term based only on  $q(\theta)$  ( $\int q^\beta(\theta) d\theta$ ) has smaller weight relative to the interaction between  $\pi(\theta)$  and  $q(\theta)$  ( $\int \pi^{\beta-1}(\theta) q(\theta) d\theta$ ). As a result, the corresponding posterior will be very close to the prior. (In fact, notice that that two of the four posteriors for  $\beta = 0.75$  favour the prior so much that there is virtually *no* mass at the Maximum Likelihood Estimate.)

$D_G^{(\gamma)}$ : Lastly, Fig. 12 demonstrates the sensitivity of  $P(\ell_n, D_G^{(\gamma)}, \mathcal{Q})$  to prior specification. The  $D_G^{(\gamma)}$  with  $\gamma > 1$  produces greater robustness to the prior than the  $\frac{1}{w}$  KLD uncertainty quantifier with  $w > 1$ . However, this robustness is not as extreme as was seen for the  $D_B^{(\beta)}$ . The reason for this is that although the  $D_G^{(\gamma)}$  consists of the exact same three terms as the  $D_B^{(\beta)}$ , these terms are now logarithms. 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  $\gamma$  much smoother than across  $\beta$ : Unlike for the  $D_B^{(\beta)}$ , minimising the  $D_G^{(\gamma)}$  can no longer disregard any one term in order to minimise the others. For  $\gamma < 1$  it appears as though the  $D_G^{(\gamma)}$  reacts similarly to the  $\frac{1}{w}$  KLD for  $w < 1$ .

#### 4.5 Comparing GVI and F-VI for a multimodal posterior

One major focus of F-VI methods is on the *zero-forcing* or *zero-avoiding* of the divergence F. [58] and [31] demonstrate these aspects of  $D_A^{(\alpha)}$  by approximating a bimodal distribution with a unimodal one. We seek a similar analysis to compare F-VI with GVI. First, note that F-VI is only motivated as a  $\mathcal{Q}$ -constrained approximation to the (exact) Bayesian posterior. In contrast, GVI defines its own Bayesian inference problem via  $P(\ell_n, D, \mathcal{Q})$ . In fact, by extending the arguments of Theorem 4, GVI gives the  $\mathcal{Q}$ -optimal posterior for a given choice of  $\ell_n$  and  $D$ . The takeaway here is that F-VI will seek to mimic the exact Bayesian posterior with some (typically insufficiently flexible)  $q \in \mathcal{Q}$ , while GVI will seek to solve a very different but interpretable Bayesian inference problem  $P(\ell_n, D, \mathcal{Q})$ .

To study the implication of this difference between F-VI and GVI, we consider an exact Bayesian posterior that exhibits bimodality. A prime example of how bimodal posteriors are induced is the

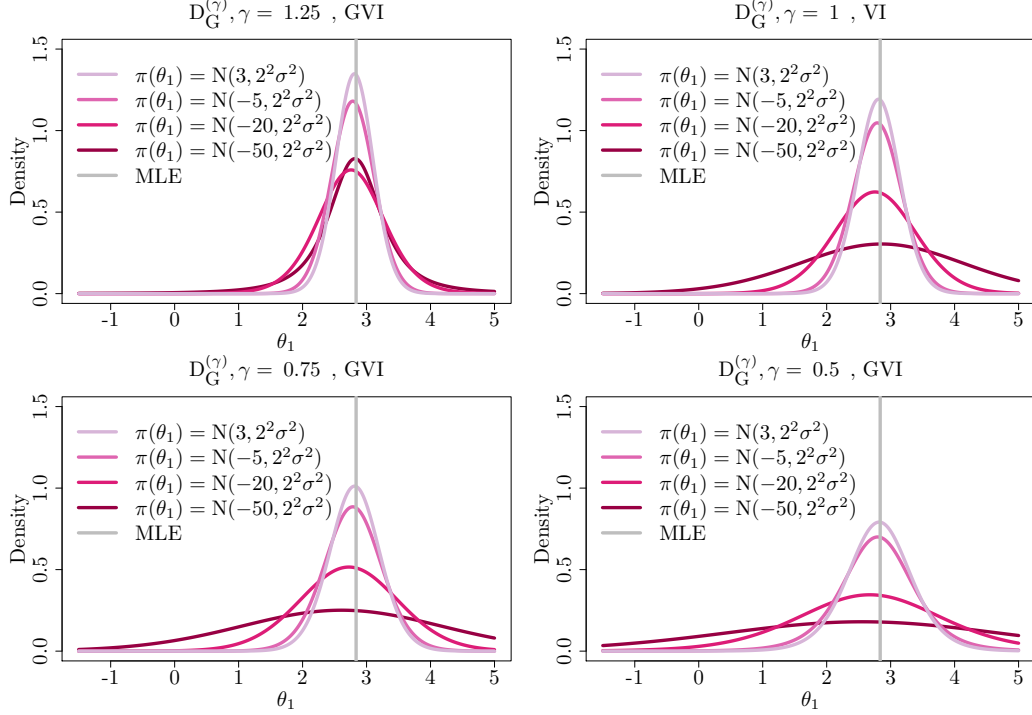


Figure 12: Marginal VI and GVI posterior for the  $\theta_1$  coefficient of a Bayesian linear model under different prior specifications and the using the  $D_G^{(\gamma)}$  as the uncertainty quantifying divergence for several values of  $\gamma$ . Prior specification:  $\sigma^2 \sim \mathcal{IG}(3, 5)$ .

label switching phenomenon. This phenomenon occurs if the likelihood function is invariant to switching parameter labels. One straightforward example of this which is of great practical importance are Bayesian mixture models. Consequently, we use a Bayesian mixture model to investigate the differences between F-VI and GVI when faced with multi-modal posteriors.

In particular, we conduct inference for the coefficients  $\theta = (\mu_1, \mu_2)$  in the model

$$p(x|\theta) = 0.5\mathcal{N}(x|\mu_1, 0.65^2) + 0.5\mathcal{N}(x|\mu_2, 0.65^2). \quad (84)$$

The bimodality in the posteriors for  $\mu$  is a consequence of the fact that  $\mu = (a, b)$  has exactly the same likelihood as  $\mu = (b, a)$ . In order to compare the performance of VI, F-VI and GVI we generate  $n = 100$  observations from the model with  $\mu = (0, 1)$ . We plot the exact and approximate posteriors in Fig. 13. All approximate posteriors come from a mean field Gaussian variational family, and all inference was based on the priors  $\pi(\mu_1) = \pi(\mu_2) = \mathcal{N}(0, 2^2)$ .

Fig. 13 shows the danger of not carefully constructing the inference problem under a multimodal posterior. The invariance to label switching in the likelihood means  $-\log(p(x_i|\theta))$  is equally minimised at either mode. The modular formulation of VI and GVI ensures that if  $q$  is constrained to be unimodal then the optimal  $q$  still focuses on one of two equally good combinations of the parameter values minimising the loss. Predictively, the resulting parameter inference will therefore still perform well. Here, GVI uses the  $D_{AR}^{(\alpha)}$  with  $\alpha = 0.5$  and thus fits a larger posterior variance than VI. However F-VI is formulated as a posterior approximation rather than a principled inference problem. As a result, the optimal unimodal approximation focuses on approximating the exact posterior rather than minimising a regularised loss function. This causes F-VI to miss either local optimum and smooth between the two. In fact, it does the worst possible thing and concentrates on the values of  $\mu_1$  and  $\mu_2$  which locally maximize the loss/minimize the likelihood.

## 5 Variance Reduction in Black box GVI

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

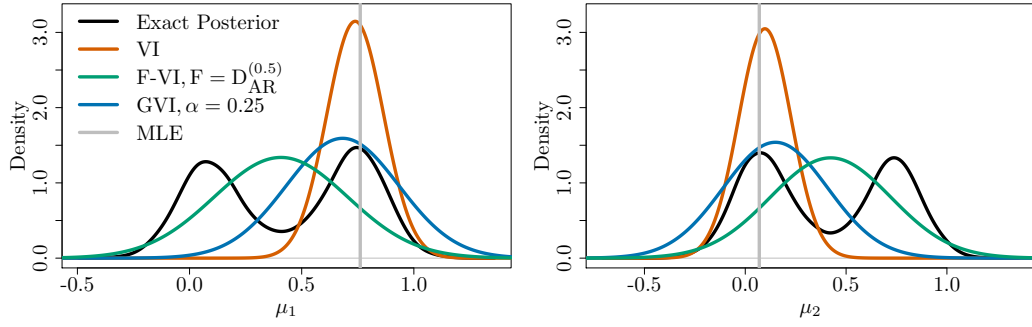


Figure 13: Marginal **Exact**, **VI**,  $D_{AR}^{(\alpha)}$ -**VI** and **GVI** using the  $-\log(p(x_i|\theta))$  and  $D_{AR}^{(\alpha)}$  prior regulariser posteriors for the coefficients  $\theta = (\mu_1, \mu_2)$  of a 2-component mixture model. The exact posterior is bimodal as a result of label-switching. **VI** and **GVI** are able to concentrate on one of the parameter minimising the loss. In contrast,  $D_{AR}^{(\alpha)}$ -**VI** is smoothing between the two modes. This demonstrates an implicit change to the loss function. Note in particular that the highest posterior mass is placed at a locally (least) likely combination of  $\mu_1$  and  $\mu_2$ .

corresponding methods, their special cases and the relevant black box variance reduction techniques are then derived and elaborated upon.

## 5.1 Preliminaries and assumptions

The variance reduction techniques in Ranganath et al. [65] crucially rely on three implicit assumptions that are reasonable for many applications of standard VI, but that we cannot always make for GVI:

- (A1) (Structured) mean-field variational inference is used, i.e.  $\mathcal{Q} = \{q(\theta|\kappa) = \prod_{j=1}^k q_j(\theta_j|\kappa_j) : \kappa_j \in K_j \text{ for all } j\}$ ;
- (A2) For all factors  $\theta_j$ , we have that  $\ell_n(\theta, x) = \ell_n^j(\theta, x) + \ell_n^{-j}(\theta, x)$ , where  $\ell_n^j$  are the components of the loss that include terms of the  $j$ -th factor and  $\ell_n^{-j}$  are the components of the loss not including terms of the  $j$ -th factor. Note that such additivity typically holds for standard VI with  $-\sum_{i=1}^n \log(p(x_i|\theta))$ . This is so because for most likelihood functions  $p$  used in practice, the components  $\theta_j$  are conditionally independent. Additionally, the priors for the factors  $\theta_j$  factorize similarly due to (conditional) independence;
- (A3)  $D = \frac{1}{w} \cdot \text{KLD}$  (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 conditional independence of the  $\theta_j$ . For GVI, (A2) and (A3) do not necessarily hold.

Yet if they do, (A2) and (A3) can greatly simplify the computations. To leverage this, we derive multiple black box algorithms for different problem classes. To ensure that the algorithm remains black box, we introduce a common wrapper function with a Boolean logic. This implies that the practitioner will not only specify  $D$ ,  $\ell_n$  and  $\mathcal{Q}$ , but also two boolean values  $a_2$  and  $a_3$  to indicate whether (A2) and (A3) are satisfied. Depending on  $a_2$  and  $a_3$ , the algorithm can then internally execute different sub-variants of black box inference depending on the simplifications available for each case. The Black Box Wrapper pseudo code below illustrates this. Next, we explain the four black box inference methods in the wrapper function.

## 5.2 Families of variance-reduced black box GVI methods

### 5.2.1 VRBBVI

**VRBBVI** stands for variance-reduced black box VI. This is the algorithm of Ranganath et al. [65] if (A1), (A2) and (A3) hold. In this case, both their Rao-Blackwellization and control variate

---

**Black Box Wrapper**


---

**Input:**  $\mathbf{x}, \pi, D, \ell_n, \mathcal{Q}, a_2, a_3$   
**if**  $a_2 == \text{True}$  and  $a_3 == \text{True}$  **then**  
     **return** VRBBVI( $\mathbf{x}, \pi, D, \ell_n, \mathcal{Q}$ )  
**if**  $a_2 == \text{False}$  and  $a_3 == \text{True}$  **then**  
     **return** BBVI( $\mathbf{x}, \pi, D, \ell_n, \mathcal{Q}$ )  
**if**  $a_2 == \text{True}$  and  $a_3 == \text{False}$  **then**  
     **return** VRBBGVI( $\mathbf{x}, \pi, D, \ell_n, \mathcal{Q}$ )  
**if**  $a_2 == \text{False}$  and  $a_3 == \text{False}$  **then**  
     **return** BBGVI( $\mathbf{x}, \pi, D, \ell_n, \mathcal{Q}$ )

---

techniques for variance reduction are applicable. Without any further modification, their algorithm can accommodate non-likelihood based losses  $\ell_n$  that are decomposable as  $\ell_n = \ell_n^j + \ell_n^{-j}$  as per (A2). Similarly, the case  $D = \frac{1}{w}\text{KLD}$  can be accomodated for any  $w \neq 0$  by simple rescaling of the objective. In particular, it yields a modification of eq. (5) in Ranganath et al. [65]

$$\mathbb{E}_{q(j)} \left[ \nabla_{\kappa_j} \log(q(\theta_j | \kappa_j)) \left( - \sum_{i=1}^n \log(p_j(x_i | \theta_{(j)})) - w\pi(\theta_{(j)}) - w \log(q(\theta_j | \kappa_j)) \right) \right] \quad (85)$$

which we derive below and give in eq. (93).

### 5.2.2 BBVI

**BBVI** is the black box VI method if  $D$  is a multiple of KLD, say  $D = \frac{1}{w}\text{KLD}$ , but  $\ell_n$  is not decomposable along the factorization. As before, this recovers the standard method up to the same scaling of eq. (85). The difference is that the scaling now enters on the global level rather than the factor level. Thus, the new objective is a simple modification of eq. (3) in Ranganath et al. [65]:

$$\mathbb{E}_q \left[ \nabla_{\kappa} \log(q(\theta | \kappa)) \left( - \sum_{i=1}^n \log(p(x_i | \theta)) - w\pi(\theta) - w \log(q(\theta | \kappa)) \right) \right]. \quad (86)$$

### 5.2.3 BBGVI

**BBGVI** stands for black box GVI. It is a modification of BBVI for the case where  $D \neq \frac{1}{w}\text{KLD}$ . As remarked before, the ELBO is just a special case of  $L(q|\mathbf{x}, \ell_n, D)$ . Thus, for  $D \neq \frac{1}{w}\text{KLD}$ , we simply compute the gradient  $\nabla_{\kappa} L(q|\mathbf{x}, \ell_n, D)$  more generally. Depending on the uncertainty quantifier  $D$ , this can take two different forms. To discern these two forms clearly, we introduce the following two assumptions on the Uncertainty Quantifier:

(UQ1) For the prior  $\pi$  and  $q \in \mathcal{Q}$ ,  $D(q|\pi)$  has a closed form and is differentiable with respect to  $\kappa$ .

(UQ2) One can write  $D(q|\pi) = \mathbb{E}_{q(\theta|\kappa)} [\ell_{\kappa, \pi}^D(\theta)]$  for some function  $\ell_{\kappa, \pi}^D : \mathbb{R} \rightarrow \mathbb{R}$ .

The derivation for the weaker condition (UQ2) is

$$\begin{aligned}
 \nabla_{\kappa} L(q|\mathbf{x}, \ell_n, D) &= \nabla_{\kappa} \left[ \int_{\theta} [\ell_n(\theta, \mathbf{x}) + \ell_{\kappa, \pi}^D(\theta)] q(\theta | \kappa) d\theta \right] \\
 &= \int_{\theta} [\ell_n(\theta, \mathbf{x}) + \ell_{\kappa, \pi}^D(\theta)] \nabla_{\kappa} q(\theta | \kappa) d\theta + \int_{\theta} [\nabla_{\kappa} \ell_{\kappa, \pi}^D(\theta)] q(\theta | \kappa) d\theta \\
 &= \mathbb{E}_{q(\theta|\kappa)} [(\ell_n(\theta, \mathbf{x}) + \ell_{\kappa, \pi}^D(\theta)) \nabla_{\kappa} \log(q(\theta | \kappa))] + \mathbb{E}_{q(\theta|\kappa)} [\nabla_{\kappa} \ell_{\kappa, \pi}^D(\theta)]. \quad (87)
 \end{aligned}$$

This derivation is a more general case of the one given in Ranganath et al. [65], but further simplifies to the one therein if  $D = \text{KLD}$ . The gradient is estimated without bias by sampling  $\theta^{(1:S)}$  from  $q(\theta | \kappa)$  and computing for each  $\theta^{(s)}$  the quantity

$$L(\theta^{(s)}) = [\ell_n(\theta^{(s)}, \mathbf{x}) + \ell_{\kappa, \pi}^D(\theta^{(s)})] \nabla_{\kappa} \log(q(\theta^{(s)} | \kappa)) + \nabla_{\kappa} \ell_{\kappa, \pi}^D(\theta^{(s)}). \quad (88)$$

Under the stronger condition (UQ1), the closed form for  $D$  provides not only a simpler objective, but an automatic reduction in variance. In this case, one may simply use

$$\begin{aligned}\nabla_{\kappa} L(q|\mathbf{x}, \ell_n, D) &= \nabla_{\kappa} \left[ \int_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}, \mathbf{x}) q(\boldsymbol{\theta}|\kappa) d\boldsymbol{\theta} + D(q||\pi) \right] \\ &= \int_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}, \mathbf{x}) \nabla_{\kappa} q(\boldsymbol{\theta}|\kappa) d\boldsymbol{\theta} + \nabla_{\kappa} D(q||\pi) \\ &= \mathbb{E}_{q(\boldsymbol{\theta}|\kappa)} [\ell_n(\boldsymbol{\theta}, \mathbf{x}) \nabla_{\kappa} \log(q(\boldsymbol{\theta}|\kappa))] + \nabla_{\kappa} D(q||\pi).\end{aligned}\quad (89)$$

Correspondingly, the gradient can then be estimated without bias by instead computing for each  $\boldsymbol{\theta}^{(s)}$  the quantity

$$L(\boldsymbol{\theta}^{(s)}) = \ell_n(\boldsymbol{\theta}^{(s)}, \mathbf{x}) \nabla_{\kappa} \log(q(\boldsymbol{\theta}^{(s)})) + \nabla_{\kappa} D(q||\pi) \quad (90)$$

Under both (UQ1) or (UQ2), one could estimate  $\nabla_{\kappa} L(q|\mathbf{x}, \ell_n, D)$  directly via  $\frac{1}{S} \sum_{i=1}^S L(\boldsymbol{\theta}^{(s)})$ . To decrease the variance of the gradient estimate, one can instead also apply the control variate strategy of Ranganath et al. [65]. Using the (black box) control variate  $\hat{a}^* \cdot h(\boldsymbol{\theta}) = \nabla_{\kappa} \log(q(\boldsymbol{\theta}|\kappa))$  with the estimated optimal scaling

$$\hat{a}^* = \frac{\sum_{s=1}^S \widehat{\text{Cov}}(L(\boldsymbol{\theta}^{(s)}), h(\boldsymbol{\theta}^{(s)}))}{\sum_{s=1}^S \widehat{\text{Var}}(h(\boldsymbol{\theta}^{(s)}))}, \quad (91)$$

we can reduce the variance even if (A2) and (A3) are not satisfied. Putting this all together, we get the following black box GVI algorithm with a single layer of variance reduction.

---

#### Black box GVI (BBGVI)

---

**Input:**  $\mathbf{x}, \pi, D, \ell_n, \mathcal{Q}$   
Randomly initialize  $\kappa$   
set  $t \leftarrow 1$   
**while**  $\|\Delta\kappa\| > \varepsilon$  **do**  
  **for**  $s = 1, 2, \dots, S$  **do**  
     $\boldsymbol{\theta}^{(s)} \sim q$   
     $h(\boldsymbol{\theta}^{(s)}) \leftarrow \log(q(\boldsymbol{\theta}^{(s)}|\kappa))$   
    Compute  $L(\boldsymbol{\theta}^{(s)})$   
  Compute  $\hat{a}^*$  as per eq. (91)  
   $\rho_t \leftarrow \text{LearningRate}(t)$   
   $\Delta\kappa \leftarrow \rho_t \cdot \frac{1}{S} \sum_{s=1}^S [L(\boldsymbol{\theta}^{(s)}) - \hat{a}^* h(\boldsymbol{\theta}^{(s)})]$   
   $\kappa \leftarrow \kappa + \Delta\kappa$   
   $t \leftarrow t + 1$

---

#### 5.2.4 VRBBGVI

**VRBBGVI** stands for the (doubly) variance-reduced version of BBGVI. In addition to the control variate strategy from before, one now adds the Rao-Blackwellization variance reduction. This is done by rewriting for  $q_{-j}(\boldsymbol{\theta}_{-j}|\kappa_{-j}) = \prod_{l=1, l \neq j}^k q_l(\boldsymbol{\theta}_l|\kappa_l)$  the partial derivatives as

$$\nabla_{\kappa_j} L(q|\mathbf{x}, \ell_n, D) = \nabla_{\kappa_j} \mathbb{E}_{q_j} [\mathbb{E}_{q_{-j}} [L(q|\mathbf{x}, \ell_n, D)|\boldsymbol{\theta}_j]].$$

The hope is then to get around computing as many expectations over  $q_{-j}$  as possible. Investigating this expression assuming that both (A1) and (A2) hold, and writing  $q_j = q_j(\boldsymbol{\theta}_j|\kappa_j)$  as well as  $\ell_n = \ell_n(\boldsymbol{\theta}, \mathbf{x})$  and  $\ell^D = \ell_{\kappa, \pi}^D(\boldsymbol{\theta})$ , one can write it as

$$\mathbb{E}_{q_j} [\nabla_{\kappa_j} \log(q_j) (\mathbb{E}_{q_{-j}} [\ell_n^j] + \mathbb{E}_{q_{-j}} [\ell_n^{-j}] + \mathbb{E}_{q_{-j}} [\ell^D])] + \mathbb{E}_{q_{-j}} [\nabla_{\kappa_j} \ell^D]. \quad (92)$$

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

$$\begin{aligned}&\mathbb{E}_{q_j} [\nabla_{\kappa_j} \log(q_j) \mathbb{E}_{q_{-j}} [\ell_n^j] + \mathbb{E}_{q_{-j}} [\ell^D + \nabla_{\kappa_j} \ell^D]] \\ &= \mathbb{E}_{q_{(j)}} [\nabla_{\kappa_j} \log(q_j) \ell_n^{(j)}(\boldsymbol{\theta}_{(j)}, \mathbf{x}) + \mathbb{E}_{q_{-(j)}} [\ell^D + \nabla_{\kappa_j} \ell^D]],\end{aligned}\quad (93)$$

where we have denoted  $\theta_{(j)}$  as the group of parameters constituting the Markov blanket of  $\theta_j$ . I.e.,  $\theta_{(j)}$  is the group of parameters (including  $\theta_j$ ) for which  $\ell_n^{(j)} = \ell_n^A + \ell_n^B$  so that  $\ell_n^A$  is a function of  $\theta_j$  and  $\ell_n^B$  is only a function of a subset of  $\theta_{(j)} \setminus \theta_j$  such that  $\ell_n^{(j)}$  is *not* further linearly decomposable. Similarly,  $q_{(j)}$  is the (variational) distribution over  $\theta_{(j)}$ . In the context of conditional independence (i.e., when  $\ell_n$  consists of likelihood terms),  $\theta_{(j)}$  has the interpretation of the collection of variables appearing in the full conditional of  $\theta_j$ . While the derivations in the supplement of Ranganath et al. [65] are restricted to likelihood losses, the more general version presented here holds for arbitrary decomposable losses as per assumption (A2).

The expectation in eq. (93) then directly and naturally leads to a Rao-Blackwellized (and thus variance-reduced) estimator. We can estimate its gradient with respect to  $\kappa_j$  as in the standard black box VI method. For each factor  $j$ , one then samples  $\theta_{(j)}^{(1:S)}$  from  $q_{(j)}$  and replaces the relevant parts of  $\theta^{(1:S)}$  with that new sample. With this, if (UQ1) holds one can compute the Rao-Blackwellized estimator for  $\kappa_j$  via

$$L_j(\theta^{(s)}) = \nabla_{\kappa_j} \log(q_j(\theta_j^{(s)} | \kappa_j)) \cdot \ell_n^{(j)}(\theta_{(j)}^{(s)}, \mathbf{x}) + \nabla_{\kappa} D(q || \pi) \quad (94)$$

If (UQ1) does not hold but (UQ2) does, one additionally needs to sample  $\theta^{(1:S)}$  from the global posterior  $q$  for approximating the uncertainty quantifier.

$$L_j(\theta^{(s)}) = \nabla_{\kappa_j} \log(q_j(\theta_j^{(s)} | \kappa_j)) \cdot \ell_n^{(j)}(\theta_{(j)}^{(s)}, \mathbf{x}) + \ell_{\kappa, \pi}^D(\theta^{(s)}) + \nabla_{\kappa_j} \ell_{\kappa, \pi}^D(\theta^{(s)}). \quad (95)$$

The control variate is implemented similarly to before using  $h_j(\theta^{(s)}) = \nabla_{\kappa_j} \log(q(\theta_j | \kappa_j))$  and

$$\hat{a}_j^* = \frac{\sum_{s=1}^S \widehat{\text{Cov}}(L_j(\theta^{(s)}), h_j(\theta^{(s)}))}{\sum_{s=1}^S \widehat{\text{Var}}(h_j(\theta^{(s)}))}, \quad (96)$$

**Example with decomposable loss:** While the interpretation of conditional independence is straightforward, it is perhaps less clear what kind of non-likelihood based losses the required decomposability of (A2) applies to. To illustrate this for a very simple non-likelihood based scenario, suppose each  $x_i = (x_{i,1}, x_{i,2}, x_{i,3})'$  consists of 3 measurements that we wish to relate to some other observable  $y_i$  with  $l = 1, 2, 3$  through

$$\begin{aligned} x_{i,1} &= \theta_1 + y_i \theta_2 + \xi_1 \\ x_{i,2} &= \theta_2 + y_i \theta_3 + \xi_2 \\ x_{i,3} &= \theta_4 + \xi_3 \end{aligned}$$

where  $\xi_j$  are slack variables (or errors) we wish to minimize by some prediction loss

$$\ell_n(\theta, \mathbf{x}) = \sum_{i=1}^n \sum_{l=1}^3 \tilde{\ell}(\widehat{x_{i,l}}(\theta) - x_{i,l}). \quad (97)$$

Regardless of the choice of  $\tilde{\ell}$  (possible choices could be the absolute or squared loss), this provides an illustrative example for the decomposability implications of (A2). Using the notation from before, it is clear that  $\ell_n^{(1)}(\theta, \mathbf{x}) = \ell_n^{(1)}(\theta_1, \mathbf{x}) = \sum_{i=1}^n \tilde{\ell}(\widehat{x_{i,1}}(\theta) - x_{i,1})$ , since  $\theta_1$  only appears in the first of the three equations. Similarly for  $\theta_3$  and  $\theta_4$ . The second factor is the only one in this example where  $\theta_{(j)} \neq \theta_j$ . In particular,  $\ell_n^{(2)}(\theta, \mathbf{x}) = \ell_n^{(2)}(\theta_{1:3}, \mathbf{x}) = \sum_{i=1}^n \sum_{l=1}^2 \tilde{\ell}(\widehat{x_{i,l}}(\theta) - x_{i,l})$ .

In summary, as VRBBVI, its GVI counterpart VRBBGVI relies on assumption (A2). Unlike VRBBVI however, it is applicable to the case where  $D \neq \frac{1}{w}$  KLD. In contrast to the objective of VRBBVI in eq. (85), extending variance reduction to non-standard uncertainty quantifiers requires computing expectations over  $\prod_{l=1, l \neq j}^k q_l(\theta_l | \kappa_l)$ . This is not desirable, but cannot be avoided unless  $\ell_{\kappa, \pi}^D$  has a similar additive decomposition as  $\ell_n$  for each factor  $j$ . As this is the case for the log function,  $D = \frac{1}{w}$  KLD is the unique case not requiring the computation of expectations over  $\prod_{l=1, l \neq j}^k q_l(\theta_l | \kappa_l)$ . For  $D \neq \frac{1}{w}$  KLD, if the expectation estimation is implemented naively, it will require sampling from all of  $\theta$  rather than only the factor  $\theta_j$  at each iteration  $S$  for all factors  $k$ . We will avoid this and leverage (A1) so that one only needs to sample  $\theta$  from  $q$  exactly once. In particular, since the draws from  $q_j$  are independent of the draws from  $q_l$  for  $l \neq j$ , it suffices to replace samples factor by factor as one updates them.

---

**Variance-reduced black box GVI (VRBBGVI)**


---

**Input:**  $x, \pi, D, \ell_n, \mathcal{Q}$   
 Randomly initialize  $\kappa$   
 set  $t \leftarrow 1$   
 // Draw initial sample from current  $q$   
 $\theta^{(1:S)} \sim q(\theta|\kappa)$   
**while**  $\|\Delta\kappa\| > \varepsilon$  **do**  
     **for**  $s = 1, 2, \dots, S$  **do**  
         // Draw new sample from  $q_j$  to replace  $\theta_j^{(s)}$   
          $\theta_{j,\text{new}}^{(s)} \sim q_j$   
          $\theta_j^{(s)} \leftarrow \theta_{j,\text{new}}^{(s)}$   
          $h_j(\theta^{(s)}) \leftarrow \log(q_j(\theta_j^{(s)}|\kappa))$   
         Compute  $L_j(\theta^{(s)})$   
     Compute  $\hat{a}_j^*$  as per eq. (96) for all  $j$   
      $\rho_t \leftarrow \text{LearningRate}(t)$   
      $\tilde{L}(\theta^{(s)}) \leftarrow (L_1(\theta^{(s)}), L_2(\theta^{(s)}), \dots, L_k(\theta^{(s)}))'$   
      $\tilde{a}^* \leftarrow (\hat{a}_1^*, \hat{a}_2^*, \dots, \hat{a}_k^*)'$   
      $\tilde{h}(\theta^{(s)}) \leftarrow (h_1(\theta^{(s)}), h_2(\theta^{(s)}), \dots, h_k(\theta^{(s)}))'$   
      $\Delta\kappa \leftarrow \rho_t \cdot \frac{1}{S} \sum_{s=1}^S [\tilde{L}(\theta^{(s)}) - \tilde{a}^* \tilde{h}(\theta^{(s)})]$   
      $\kappa \leftarrow \kappa + \Delta\kappa$   
      $t \leftarrow t + 1$

---

## 6 Derivations for experiments

### 6.1 Robust Deep Gaussian Processes with alternative uncertainty quantifiers

For a technical report summarizing the modifications for the DGP without going into detailed calculations, we refer the interested reader to Knoblauch [45]. As we use the inference of Salimbeni and Deisenroth [74], we adapt their notation in the remainder of the following derivation without re-explaining all the terms and variables. Further, to keep notation simple, we will derive the univariate result in eq. (10) of Salimbeni and Deisenroth [74] only. The multivariate extension in eq. (17) in [74] then follows trivially by the same steps as in their paper.

#### 6.1.1 Changing the uncertainty quantifier

We need to derive under which conditions the uncertainty quantifier  $D$  only depends on the prior  $p(\mathbf{u})$  and the variational posterior  $q(\mathbf{u})$  of the inducing points  $\mathbf{u}$  (rather than on  $p(\mathbf{f}, \mathbf{u})$  and  $q(\mathbf{f}, \mathbf{u})$ ). This follows for a range of divergences  $D$ , including f-divergences and functions of f-divergences that are still divergences. To see this, one simply needs to re-examine eq. (10) – (12) in Bonilla et al. [13]. In particular, note that for any divergence  $D'(q||\pi)$  that can be written as  $D'(q||\pi) = g(D(q||\pi))$  for some function  $g(x)$  such that  $g(x) \geq 0$  and  $g(x) = 0$  if and only if  $x = 0$  and for some f-divergence  $D(q||\pi) = \int_{\theta} q(\theta) f\left(\frac{q(\theta)}{\pi(\theta)}\right) d\theta$ , it holds that

$$\begin{aligned}
 D'(q(\mathbf{f}, \mathbf{u})||p(\mathbf{f}, \mathbf{u})) &= g\left(\mathbb{E}_{q(\mathbf{f}, \mathbf{u})}\left[f\left(\frac{q(\mathbf{f}, \mathbf{u})}{p(\mathbf{f}, \mathbf{u})}\right)\right]\right) = g\left(\mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})}\left[f\left(\frac{q(\mathbf{u})}{p(\mathbf{u})}\right)\right]\right) \\
 &= g\left(\mathbb{E}_{q(\mathbf{u})}\left[f\left(\frac{q(\mathbf{u})}{p(\mathbf{u})}\right)\right]\right) = D'(q(\mathbf{u})||p(\mathbf{u})),
 \end{aligned}$$

where the first equality holds by definition, the second because  $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u})$  and  $p(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})p(\mathbf{u})$ , the third by integrating out  $\mathbf{f}$  and the last again by definition.

As additivity between  $D$  and the loss is implied by definition of GVI, it immediately follows that by modifying the objective of eq. (17) in [74] as

$$\mathcal{L}_{\text{DGP}}^{D'} = - \sum_{i=1}^N \mathbb{E}_{q(\mathbf{f}_i^L)} [-\log p(\mathbf{y}_i | \mathbf{f}_i^L)] + \sum_{l=1}^L D'(q(\mathbf{U}^l) || p(\mathbf{U}^l)), \quad (98)$$

one can perform GVI inference on Deep Gaussian Processes with a global uncertainty quantifier  $D'$  that is a (functional of) an f-divergence. In GVI notation, this problem then would solve  $P(\sum_{i=1}^N -\log p(\mathbf{y}_i | \mathbf{f}_i^L), D', \mathcal{Q})$ , where  $\mathcal{Q}$  is the same (sparse) variational family described by Salimbeni and Deisenroth [74].

Moreover, following Theorem 6, one can choose different local divergences  $D^l$  for each layer as long as all of them are (functions of) f-divergences. To see this, first recall that

$$\begin{aligned} q(\{\mathbf{U}^l\}_{l=1}^L, \{\mathbf{F}^l\}_{l=1}^L) &= \prod_{l=1}^L p(\mathbf{F}^l | \mathbf{U}^l, \mathbf{F}^{l-1}) q(\mathbf{U}^l) \\ p(\{\mathbf{U}^l\}_{l=1}^L, \{\mathbf{F}^l\}_{l=1}^L) &= \prod_{l=1}^L p(\mathbf{F}^l | \mathbf{U}^l, \mathbf{F}^{l-1}) p(\mathbf{U}^l) \end{aligned}$$

and write for a *fixed* conditioning set  $\{\mathbf{F}_*^l\}_{l=1}^L$  the new customized (and conditioning-set specific) divergence

$$\begin{aligned} D^{\{\mathbf{F}_*^l\}_{l=1}^L} (q(\{\mathbf{U}_l\}_{l=1}^L, \{\mathbf{F}_l\}_{l=1}^L) || p(\{\mathbf{U}_l\}_{l=1}^L, \{\mathbf{F}_l\}_{l=1}^L)) \\ = \sum_{l=1}^L D^l (p(\mathbf{F}^l | \mathbf{U}^l, \mathbf{F}_*^{l-1}) q(\mathbf{U}^l) || p(\mathbf{F}^l | \mathbf{U}^l, \mathbf{F}_*^{l-1}) p(\mathbf{U}^l)) \end{aligned} \quad (99)$$

$$= \sum_{l=1}^L D^l (q(\mathbf{U}^l) || p(\mathbf{U}^l)) \quad (100)$$

The first equality follows by definition and the second one by the property of (functions of) f-divergences we exploited above. Note that *independently* of the conditioning set  $\{\mathbf{F}_*^l\}_{l=1}^L$ , the same (conditioning-set invariant) form emerges as long as all  $D^l$  are (functions of) f-divergences. Moreover, the conditionals satisfy the positivity condition required for the Hammersley-Clifford Theorem to hold, which means that eq. (100) is a well-defined divergence according to Theorem 6 (iii).

In our implementation, this implies for example that we can use Rényi's  $\alpha$ -divergence in all or some of the Deep Gaussian Process layers. In each layer  $l$ ,  $q(\mathbf{U}^l)$  and  $p(\mathbf{U}^l)$  are (multivariate) normal distributions. Following Corollary 3, it is clear what form the regularizer must take. In particular, it depends on the natural parameters and the normalizing constants of  $q(\mathbf{U}^l)$  and  $p(\mathbf{U}^l)$ . Notice that for a  $d$ -dimensional multivariate normal  $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ , the natural parameter space is given by  $\{\boldsymbol{\eta} = (\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}, -\frac{1}{2} \boldsymbol{\Sigma}^{-1})' : \boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{M}^d\}$  for  $\mathbb{M}^d$  the set of symmetric, semi-positive definite  $d \times d$  matrices. For  $q(\mathbf{U}^l) = \mathcal{N}(\mathbf{U}^l | \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$  and  $p(\mathbf{U}^l) = \mathcal{N}(\mathbf{U}^l | \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ , denoting  $\boldsymbol{\eta}_q = (\boldsymbol{\Sigma}_q^{-1} \boldsymbol{\mu}_q, -\frac{1}{2} \boldsymbol{\Sigma}_q^{-1})'$  and  $\boldsymbol{\eta}_p = (\boldsymbol{\Sigma}_p^{-1} \boldsymbol{\mu}_p, -\frac{1}{2} \boldsymbol{\Sigma}_p^{-1})'$ , with  $A(\boldsymbol{\eta}) = \frac{1}{2} [\boldsymbol{\mu}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \ln |\boldsymbol{\Sigma}|]$  the log normalizing constant, Corollary 3 implies that for  $\alpha \in (0, 1)$ ,

$$\begin{aligned} D_{AR}^{(\alpha)} &= \frac{1}{\alpha(1-\alpha)} [A(\alpha \boldsymbol{\eta}_q + (1-\alpha) \boldsymbol{\eta}_p) - \alpha A(\boldsymbol{\eta}_q) - (1-\alpha) A(\boldsymbol{\eta}_p)] \quad (101) \\ A(\boldsymbol{\eta}_q) &= \frac{1}{2} [\boldsymbol{\mu}_q' \boldsymbol{\Sigma}_q^{-1} \boldsymbol{\mu}_q + \ln |\boldsymbol{\Sigma}_q|] \\ A(\boldsymbol{\eta}_p) &= \frac{1}{2} [\boldsymbol{\mu}_p' \boldsymbol{\Sigma}_p^{-1} \boldsymbol{\mu}_p + \ln |\boldsymbol{\Sigma}_p|] \\ A(\alpha \boldsymbol{\eta}_q + (1-\alpha) \boldsymbol{\eta}_p) &= \frac{1}{2} [\boldsymbol{\mu}^{*'} (\boldsymbol{\Sigma}^*)^{-1} \boldsymbol{\mu}^* + \ln |\boldsymbol{\Sigma}^*|] \\ (\boldsymbol{\Sigma}^*)^{-1} &= \alpha \boldsymbol{\Sigma}_q^{-1} + (1-\alpha) \boldsymbol{\Sigma}_p^{-1} \\ \boldsymbol{\mu}^* &= \boldsymbol{\Sigma}^* (\alpha \boldsymbol{\Sigma}_q^{-1} \boldsymbol{\mu}_q + (1-\alpha) \boldsymbol{\Sigma}_p^{-1} \boldsymbol{\mu}_p) \end{aligned}$$

Notice that computing this is of the same order as computing the KLD uncertainty quantifier, since one performs a cholesky decomposition of  $\boldsymbol{\Sigma}_q$  and  $\boldsymbol{\Sigma}_p$  for  $D = \text{KLD}$ , too.



### 6.1.2 Changing the loss

Next, we show that the robust losses  $\mathcal{L}_p^\beta, \mathcal{L}_p^\gamma$  introduced in the main paper can be efficiently deployed for Deep Gaussian Processes. In fact, one simply replaces  $-\log p(\mathbf{y}_n|\mathbf{f}_n^L)$  with its robustified counterpart. With these losses, one can show that as with the log likelihood loss, the expectations over  $q(\mathbf{f}_i^L)$  are still available in closed form if one has drawn the sequential samples  $\mathbf{x}_i = \mathbf{f}_i^0, \mathbf{f}_i^1, \dots, \mathbf{f}_i^{L-1}$  as described in [74]. Showing this boils down to a simplified version of the derivation in the Appendix accompanying Knoblauch et al. [46], applied to Gaussian Processes. First, we redefine

$$\begin{aligned}\mathcal{L}_p^\beta(\mathbf{f}_i^L, \mathbf{y}_i) &= -\left(\frac{1}{\beta-1}p(\mathbf{y}_i|\mathbf{f}_i^L)^{\beta-1} - \frac{I_{p,\beta}(\mathbf{f}_i^L)}{\beta}\right), \\ \mathcal{L}_p^\gamma(\mathbf{f}_i^L, \mathbf{y}_i) &= -\left(\frac{1}{\gamma-1}p(\mathbf{y}_i|\mathbf{f}_i^L)^{\gamma-1} \cdot \frac{\gamma}{I_{p,\gamma}(\mathbf{f}_i^L)^{\frac{\gamma-1}{\gamma}}}\right)\end{aligned}$$

for Gaussian Processes, where as before  $I_{p,c}(\mathbf{f}) = \int p(\mathbf{y}|\mathbf{f})^c d\mathbf{y}$ . The likelihood is Gaussian with a fixed variance parameter  $\sigma^2$ , i.e. for  $\mathbf{y}_i \in \mathbb{R}^d$  with  $i = 1, 2, \dots, N$

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

With this, note that integrating out the normal density yields

$$I_{p,c}(\mathbf{f}_i^L) = (2\pi\sigma^2)^{-0.5dc} c^{-0.5d}. \quad (102)$$

Note in particular that this is a constant and does not depend on  $\mathbf{f}$ , which makes computing the expectation over  $q(\mathbf{f}_i^L)$  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 [46]. To simplify notation, we write  $\mathbf{f} = \mathbf{f}_i^L$ . Note also that the variational parameters  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  are (stochastic) functions of the draws of  $\mathbf{f}_i^{1:L-1}$  from the previous layers, but we suppress this dependency, again for readability.

$$\begin{aligned}& \mathbb{E}_{q(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} \left[ \frac{1}{c} p(\mathbf{y}_i|\mathbf{f})^c \right] \\ &= \frac{1}{c} (2\pi\sigma^2)^{-0.5dc} \cdot \mathbb{E}_{q(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} \left[ \exp\left\{-\frac{c}{2\sigma^2}(\mathbf{y}_i^T \mathbf{y}_i + \mathbf{f}^T \mathbf{f} - 2\mathbf{f}^T \mathbf{y}_i)\right\} \right] \\ &= \frac{1}{c} (2\pi\sigma^2)^{-0.5dc} \exp\left\{-\frac{c}{2\sigma^2} \mathbf{y}_i^T \mathbf{y}_i\right\} \cdot \mathbb{E}_{q(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} \left[ \exp\left\{-\frac{c}{2\sigma^2}(\mathbf{f}^T \mathbf{f} - 2\mathbf{f}^T \mathbf{y}_i)\right\} \right] \\ &= \frac{1}{c} (2\pi\sigma^2)^{-0.5dc} (2\pi\sigma^2)^{-0.5d} |\boldsymbol{\Sigma}|^{-0.5} \exp\left\{-\frac{c}{2\sigma^2} \mathbf{y}_i^T \mathbf{y}_i\right\} \times \\ & \quad \int \exp\left\{-\frac{1}{2} \left( \frac{c}{\sigma^2} \mathbf{f}^T \mathbf{f} - \frac{2c}{\sigma^2} \mathbf{f}^T \mathbf{y}_i + (\mathbf{f} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{f} - \boldsymbol{\mu}) \right)\right\} d\mathbf{f} \\ &= \frac{1}{c} (2\pi\sigma^2)^{-0.5dc} (2\pi)^{-0.5d} |\boldsymbol{\Sigma}|^{-0.5} \exp\left\{-\frac{1}{2} \left( \frac{c}{\sigma^2} \mathbf{y}_i^T \mathbf{y}_i + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right)\right\} \times \\ & \quad \int \exp\left\{-\frac{1}{2} \left( \frac{c}{\sigma^2} \mathbf{f}^T \mathbf{f} - \frac{2c}{\sigma^2} \mathbf{f}^T \mathbf{y}_i + \mathbf{f}^T \boldsymbol{\Sigma}^{-1} \mathbf{f} - 2\mathbf{f}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right)\right\} d\mathbf{f} \quad (103)\end{aligned}$$

The integral suggests one can obtain a closed form through the Gaussian integral by completing the squares. Defining  $\tilde{\boldsymbol{\Sigma}}^{-1} = (\frac{c}{\sigma^2} \mathbf{I}_d + \boldsymbol{\Sigma}^{-1})$ ,  $\tilde{\boldsymbol{\mu}} = (\frac{c}{\sigma^2} \mathbf{y}_i + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu})$  and  $\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\mu}}$ , one can do this via

$$\begin{aligned}& \frac{c}{\sigma^2} \mathbf{f}^T \mathbf{f} - \frac{2c}{\sigma^2} \mathbf{f}^T \mathbf{y}_i + \mathbf{f}^T \boldsymbol{\Sigma}^{-1} \mathbf{f} - 2\mathbf{f}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ &= \mathbf{f}^T \left( \mathbf{I}_d \frac{c}{\sigma^2} + \boldsymbol{\Sigma}^{-1} \right) \mathbf{f} - 2\mathbf{f}^T \left( \frac{c}{\sigma^2} \mathbf{y}_i + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right) \\ &= (\mathbf{f} - \hat{\boldsymbol{\mu}})^T \tilde{\boldsymbol{\Sigma}}^{-1} (\mathbf{f} - \hat{\boldsymbol{\mu}}) - \tilde{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\mu}}, \quad (104)\end{aligned}$$

which allows us to finally rewrite the integral as

$$\begin{aligned}& \int \exp\left\{-\frac{1}{2} \left( \frac{c}{\sigma^2} \mathbf{f}^T \mathbf{f} - \frac{2c}{\sigma^2} \mathbf{f}^T \mathbf{y}_i + \mathbf{f}^T \boldsymbol{\Sigma}^{-1} \mathbf{f} - 2\mathbf{f}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right)\right\} d\mathbf{f} \\ &= \exp\left\{-\frac{1}{2} \tilde{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\mu}}\right\} \int \exp\left\{-\frac{1}{2} (\mathbf{f} - \hat{\boldsymbol{\mu}})^T \tilde{\boldsymbol{\Sigma}}^{-1} (\mathbf{f} - \hat{\boldsymbol{\mu}})\right\} d\mathbf{f} \\ &= \exp\left\{\frac{1}{2} \tilde{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\mu}}\right\} (2\pi)^{0.5d} |\tilde{\boldsymbol{\Sigma}}|^{0.5}. \quad (105)\end{aligned}$$

Putting everything together and simplifying expressions, this means that

$$\mathbb{E}_{q(\mathbf{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})} \left[ \frac{1}{c} p(\mathbf{y}_i|\mathbf{f})^c \right] = \frac{1}{c} (2\pi\sigma^2)^{-0.5dc} \frac{|\tilde{\boldsymbol{\Sigma}}|^{0.5}}{|\boldsymbol{\Sigma}|^{0.5}} \exp \left\{ -\frac{1}{2} \left( \frac{c}{\sigma^2} \mathbf{y}_i^T \mathbf{y}_i + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\mu}} \right) \right\} \quad (106)$$

Depending on whether one uses the  $\beta$ - or  $\gamma$ -divergence for robustifying the loss, one thus obtains the closed form expressions

$$\mathbb{E}_{q(\mathbf{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})} \left[ -\frac{1}{\beta-1} p(\mathbf{y}_i|\mathbf{f})^{\beta-1} + \frac{I_{p,\beta}(\mathbf{f})}{\beta} \right] = \mathbb{E}_{q(\mathbf{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})} \left[ -\frac{1}{\beta-1} p(\mathbf{y}_i|\mathbf{f})^{\beta-1} \right] + \frac{I_{p,\beta}(\mathbf{f})}{\beta} \quad (107)$$

$$\mathbb{E}_{q(\mathbf{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})} \left[ -\frac{1}{\gamma-1} p(\mathbf{y}_i|\mathbf{f})^{\gamma-1} \cdot \frac{\gamma}{I_{p,\gamma}(\mathbf{f})^{\frac{\gamma-1}{\gamma}}} \right] = \mathbb{E}_{q(\mathbf{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})} \left[ -\frac{1}{\gamma-1} p(\mathbf{y}_i|\mathbf{f})^{\gamma-1} \right] \cdot \frac{\gamma}{I_{p,\gamma}(\mathbf{f})^{\frac{\gamma-1}{\gamma}}}, \quad (108)$$

with the expectation over  $q(\mathbf{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})$  as in eq. (109) and integrals  $I_{p,\beta}(\mathbf{f})$  and  $I_{p,\gamma}(\mathbf{f})$  as in eq. (102). Note that we have derived the general case for  $\mathbf{y}_i \in \mathbb{R}^d$ , where  $\boldsymbol{\Sigma}$ ,  $\mathbf{f}$  and  $\boldsymbol{\mu}$  are matrix- and vector-valued. Since the derivation of Salimbeni and Deisenroth [74] shows that one in fact only needs to integrate over the marginals  $\mathbf{f}_i^L$ , if  $d = 1$  (as in all experiments in both this paper and [74]), 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  $\mu$ ,  $\Sigma$  and defining  $\tilde{\Sigma} = \frac{1}{\frac{c}{\sigma^2} + \frac{1}{\Sigma}}$  and  $\tilde{\mu} = \left( \frac{c\mathbf{y}_i}{\sigma^2} + \frac{\mu}{\Sigma} \right)$ , eq. (109) simplifies to

$$\mathbb{E}_{q(\mathbf{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})} \left[ \frac{1}{c} p(\mathbf{y}_i|\mathbf{f})^c \right] = \frac{1}{c} s (2\pi\sigma^2)^{-0.5c} \sqrt{\frac{\tilde{\Sigma}}{\Sigma}} \cdot \exp \left\{ -\frac{1}{2} \left( \frac{c\mathbf{y}_i^2}{\sigma^2} + \frac{\mu^2}{\Sigma} - \tilde{\mu}^2 \tilde{\Sigma} \right) \right\} \quad (109)$$

## 6.2 Robust Bayesian Neural Nets with alternative uncertainty quantifiers

For robustifying the normal likelihoods used within the layers of the Bayesian Neural Nets, one again needs the integral term derived in eq. (102). As for the uncertainty quantifiers, one again uses Rényi's  $\alpha$ -divergence with the univariate version of the derivation in eq. (101).

## 7 Experiments

This section contains further details and results on the experiments. First, we provide a more detailed overview of the methods used. Next, we show additional results that complement the analysis in the main paper. For both the Deep Gaussian Process and Bayesian Neural Net examples we exclusively use data sets from the UCI repository [50]. All code will be made publicly available upon publication at <https://github.com/JeremiasKnoblauch/GVIPublic>.

### 7.1 Bayesian Neural Nets

#### 7.1.1 Details

We use the model and code of Li and Turner [49] and the same probabilistic back-propagation [30] with 100 samples per iteration. All BNNs have a single hidden layer with 50 ReLU units [as in 49, 31]. The priors are standard normal and the variational posterior over the parameters is a completely factorized normal. For inference, we use 500 epochs, a batch size of 32 and the ADAM [42] optimizer with default values. To compute the test values, we evaluate them on the test set of 50 random 90:10 training:test splits. Test metrics are computed by averaging 100 samples taken from the variational posterior. The Python implementation uses autograd [53].

#### 7.1.2 Additional results

See Fig. 14 for further GVI results varying the loss as well as the uncertainty quantifier. The results are comparable to only varying  $D$ . In particular, GVI with  $D = D_{AR}^{(\alpha)}$ ,  $\alpha > 1$  outperforms F-VI on the test RMSE metric across all data sets because – unlike F-VI – it does not implicitly change the loss.

Moreover, Fig. 15 shows results for four additional data sets. These largely conform with the findings in the main paper: GVI produces banana-shaped performance curves with  $\alpha$  and typically beats standard VI. Moreover, more concentration typically yields better test errors than less concentration.

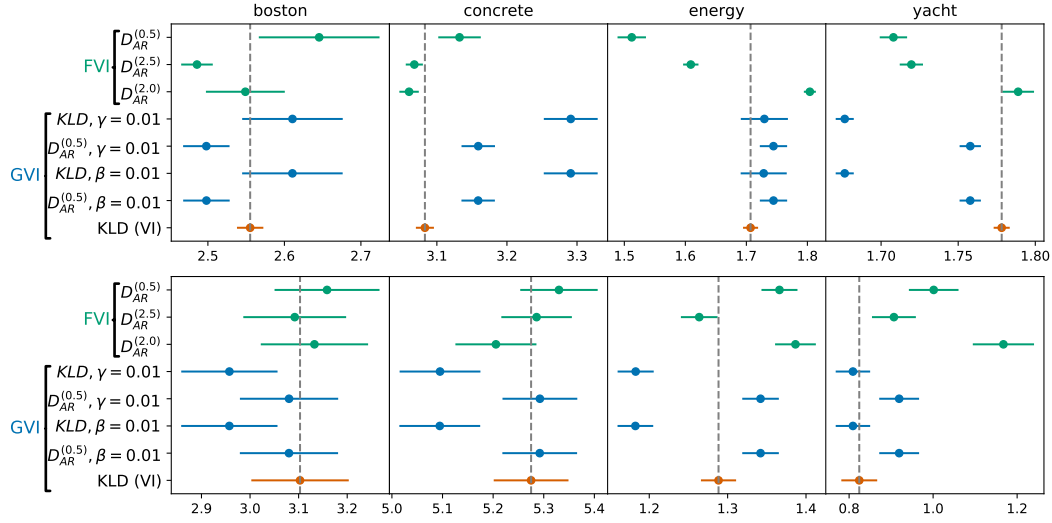


Figure 14: Comparing performance on the BNN for **GVI** with different loss functions  $\ell_n(\theta, \mathbf{x})$  and different uncertainty quantifiers  $D$  against **VI** and **F-VI**. **Top row**: Negative test log likelihoods. **Bottom row**: Test RMSE. The lower the better.

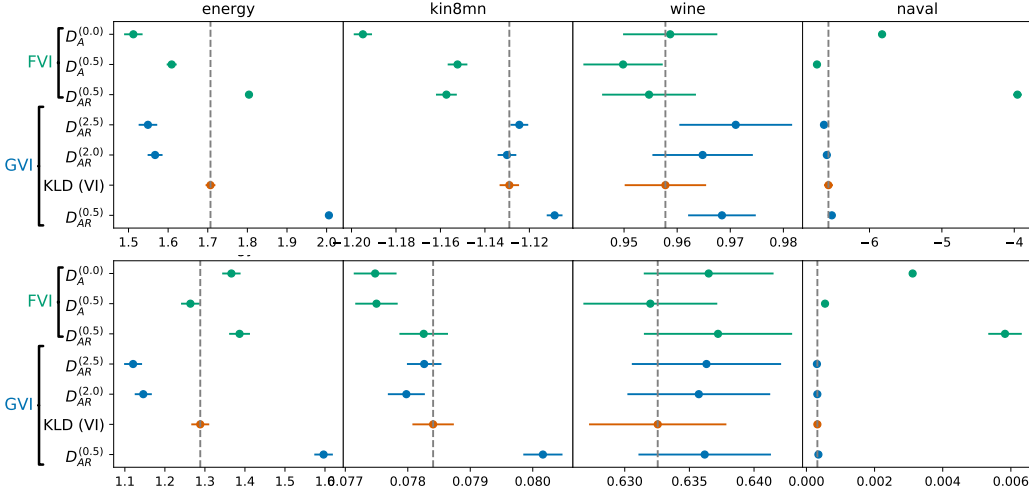


Figure 15: Comparing test set performance on the BNN between **F-VI** and **GVI** with a range of additional standard test sets with the same settings as in the main paper.

We note that on the kin8mn data, F-VI manages to outperform GVI noticeably, indicating that the conflation of loss and uncertainty quantification (and the resulting implicit loss function transformation) applied by F-VI is beneficial on this data set. On the naval data on the other hand, F-VI does very poorly. This is likely another reflection of the fact that it conflates loss and uncertainty quantification: As the test prediction errors for standard VI indicate, this data set is easy to learn and well-modelled by the network deployed. In other words, the normal log likelihood is an excellent loss on naval. Thus, GVI and VI do better than F-VI methods.

### 7.1.3 Posterior predictives: F-VI vs GVI methods

We provide the predictive distributions for some GVI and F-VI methods for three test points, see Fig. 16. Note that the variational posterior predictives take form

$$q(y|\mathbf{x}, \hat{\sigma}^2) = \int_{\theta} \mathcal{N}(y|F(\theta, \mathbf{x}), \hat{\sigma}^2) q(\theta|\kappa) d\theta \quad (110)$$

where  $F$  are the ReLU transformations of  $\mathbf{x}$  to  $y$  parameterized by  $\theta$  and  $\kappa$  are the variational parameters of the variational posterior  $q(\theta|\kappa)$ . Also note that for all methods, inference on  $\sigma^2$  is

performed using a point estimate  $\hat{\sigma}^2$  obtained by optimizing the objective function over  $\sigma^2$ . In particular, the variational lower bounds (for the F-VI methods) and the expected loss function (in the GVI methods) are minimized over  $\sigma^2$ . To disentangle the approximate Bayesian parameter inference on  $\theta$  from the point estimate  $\hat{\sigma}^2$ , we plot both the posterior predictive  $q(y'|\mathbf{x}', \hat{\sigma}^2)$  for test inputs  $\mathbf{x}'$  as well as the posterior over the mean function  $F(\theta, \cdot)$  for test inputs  $\mathbf{x}'$ . Crucially, the latter does not depend on  $\hat{\sigma}^2$ . It is given by

$$q(F(\theta, \mathbf{x}')|\mathbf{x}) = \int_{\theta} F(\theta, \mathbf{x}') q(\theta|\kappa) d\theta. \quad (111)$$

Because the posterior on  $F(\theta, \cdot)$  does not depend on the point estimate  $\hat{\sigma}^2$ , this allows us to check whether the smaller variances of the F-VI methods are mainly due to  $\sigma^2$ , or the posteriors on  $\theta$ . Plotting  $q(F(\theta, \mathbf{x}')|\mathbf{x})$  in Fig. 17 and comparing it to Fig. 16 allows for the following conclusions: (1) While the F-VI parameter posterior  $q(F(\theta, \mathbf{x}')|\mathbf{x})$  seems to become flatter for extreme enough choices of  $\alpha$ , it does not seem to do so for more moderate values of  $\alpha$ . (2) In contrast, the F-VI posterior predictives are even more concentrated than the VI posterior predictives. This is due to  $\hat{\sigma}^2$  not even neutralizing, but in fact reversing the effect of more conservative inference on  $\theta$ , see Table 2. We explain why GVI does not have this problem below. (3) GVI provides more conservative parameter (and predictive) posteriors as one would expect.

Why does  $\hat{\sigma}^2$  take smaller and smaller values for nominally more conservative choices of the F-VI hyperparameters? The reason this happens is that the F-VI methods optimize  $\sigma^2$  as follows: For a given discrepancy measure  $F$  and  $q^*$  being the exact Bayesian posterior, one wishes to solve

$$\hat{\sigma}^2, q(\theta|\hat{\sigma}^2, \kappa) = \arg \min_{\sigma^2} \left\{ \arg \min_{q' \in \mathcal{Q}} F(q'(\theta|\sigma^2, \kappa) || q^*(\theta|\sigma^2, \mathbf{x}, \mathbf{y})) \right\}. \quad (112)$$

Now if  $F$  is zero-avoiding, having areas of the variational posterior  $q(\theta|\hat{\sigma}^2, \kappa) \approx 0$  will incur a large  $F$ -discrepancy value/penalty if the true posterior  $q^*(\theta|\hat{\sigma}^2, \mathbf{x}, \mathbf{y})$  is (far) away from zero for the same values of  $\theta$ . But making  $\hat{\sigma}^2$  smaller will change the *target*  $q^*(\theta|\hat{\sigma}^2, \mathbf{x}, \mathbf{y})$  itself via the likelihood  $p(y_i|\theta, \hat{\sigma}^2, x_i) = \mathcal{N}(y_i|F(\theta, \mathbf{x}), \hat{\sigma}^2)$ . In particular, it will make  $p(y_i|\theta, \hat{\sigma}^2, x_i)$  a much more extreme loss function (consider the limiting case as  $\hat{\sigma}^2 \rightarrow 0$ ), which in turn makes the posterior (i.e. the target  $q^*$ ) a much more concentrated posterior. Reversing this logic, imagine now that all other things equal, one lowers  $\hat{\sigma}^2$  to  $\hat{\sigma}_c^2$ ,  $\hat{\sigma}^2 < \hat{\sigma}_c^2$  with the corresponding targets  $q^*, q_c^*$ . Here,  $q_c^*$  is strictly more concentrated than  $q^*$ , but their modes are identical. This means that for a given variational posterior  $q$ , the penalty of concentrating around the (same) mode of the target will be smaller for the more concentrated posterior  $q_c^*$  than it will be for  $q^*$ . In other words, picking  $\hat{\sigma}_c^2$  rather than  $\hat{\sigma}^2$  lowers the  $F$ -discrepancy penalty associated with over-concentrating. Hence, optimizing hyperparameters over the F-VI lower bounds should be treated with care, as one changes the (exact) Bayesian posterior to be approximated in ways that affect the posterior predictive in the *exact opposite* way of how one originally intended the method to behave.

	GVI, $D = D_A^{(1.25)}$	VI	GVI, $D = D_A^{(0.5)}$	GVI, $D = D_A^{(0.01)}$	F-VI, $D_{AR}^{(0.5)}$	F-VI, $D_A^{(0.5)}$	F-VI, $D_A^{(0.0)}$
$\hat{\sigma}^2$	9.225	10.797	16.811	39.533	4.016	5.856	0.911

Table 2: Comparing the value of  $\hat{\sigma}^2$  for different  $\mathcal{Q}$ -constrained posterior inference methods (GVI with Rényi’s  $\alpha$ -divergence uncertainty quantifier and the F-VI methods of [49] and [31]). For F-VI methods,  $\sigma^2$  produces a substitution effect because it directly affects the target about which uncertainty is quantified. For GVI methods, uncertainty quantification and loss are additively separated, which prevents this substitution effect.

The issue discussed in the paragraph above does not occur with GVI. Optimizing  $\sigma^2$  in GVI is done similarly to the procedure in F-VI (but with the GVI objective instead of the  $F$ -discrepany between  $q'$  and  $q^*$ ):

$$\hat{\sigma}^2, q(\theta|\hat{\sigma}^2, \mathbf{x}, \mathbf{y}) = \arg \min_{\sigma^2} \left\{ \arg \min_{q' \in \mathcal{Q}} \mathbb{E}_{q'} [\ell_n(\theta, \mathbf{x}|\mathbf{y}, \sigma^2)] + D(q' || \pi) \right\}. \quad (113)$$

This does not lead to the same problems that F-VI exhibits because the uncertainty quantifier  $D$  does *not* depend on  $\sigma^2$ . So unlike for the F-VI methods, there is no direct interaction or substitution effect

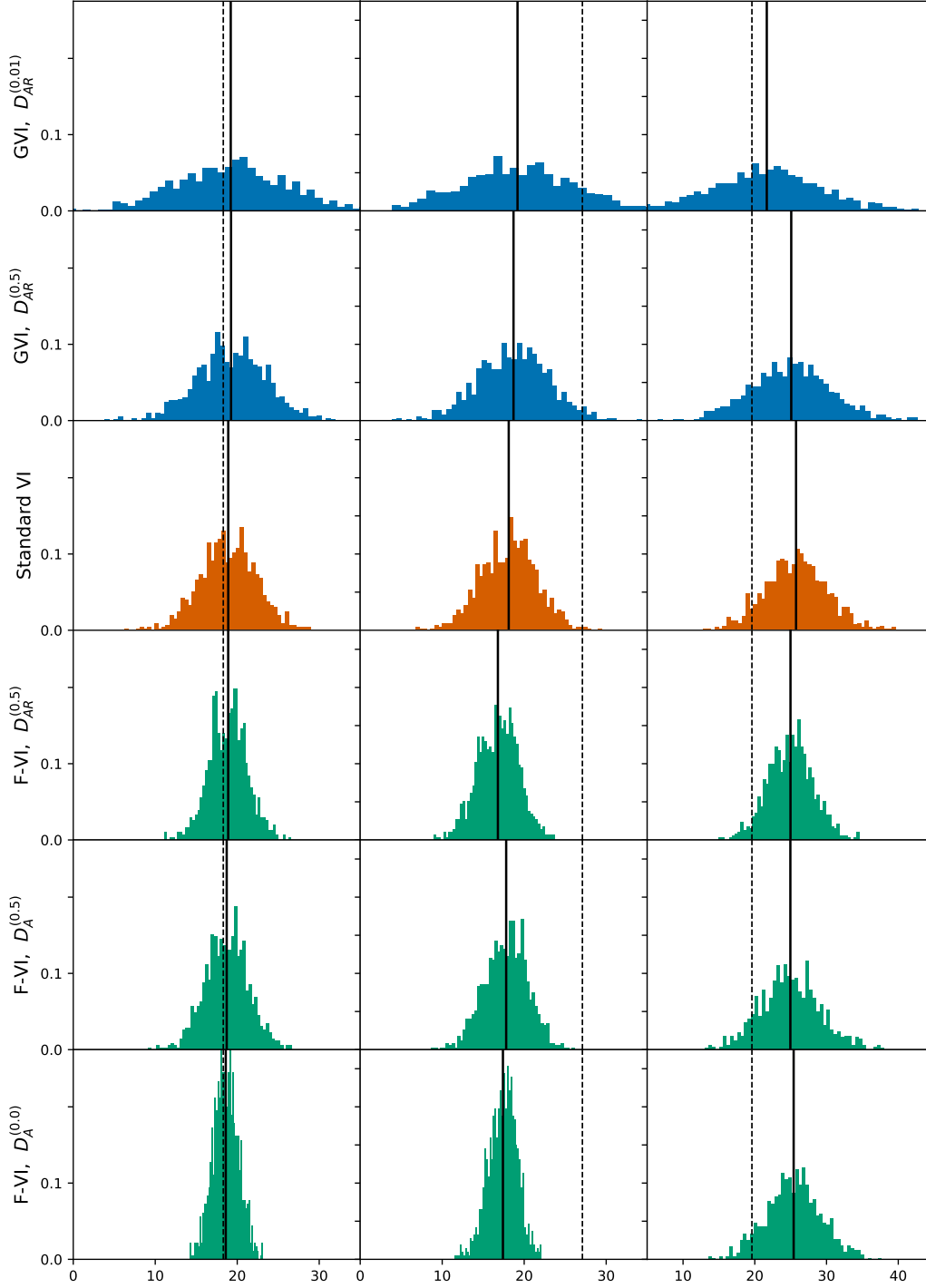


Figure 16: Posterior predictives  $q(y|\mathbf{x})$  for **VI**, the  $D_A^{(\alpha)}$ -**VI** method of Hernández-Lobato et al. [31], the  $D_{AR}^{(\alpha)}$ -**VI** method of Li and Turner [49] and **GVI** for  $D = D_{AR}^{(\alpha)}$  on three test points on the boston data sets; based on 1,000 samples each. Notice that relative to standard **VI**, all **F-VI** posterior predictives are *more* contracted. In contrast, **GVI** with a more conservative uncertainty quantifier does what one would expect zero-avoiding **F-VI** methods to do. Thus, while **F-VI** may provide flatter marginal variances in the (variational) posterior for  $\theta$ , this does not translate into the predictive.

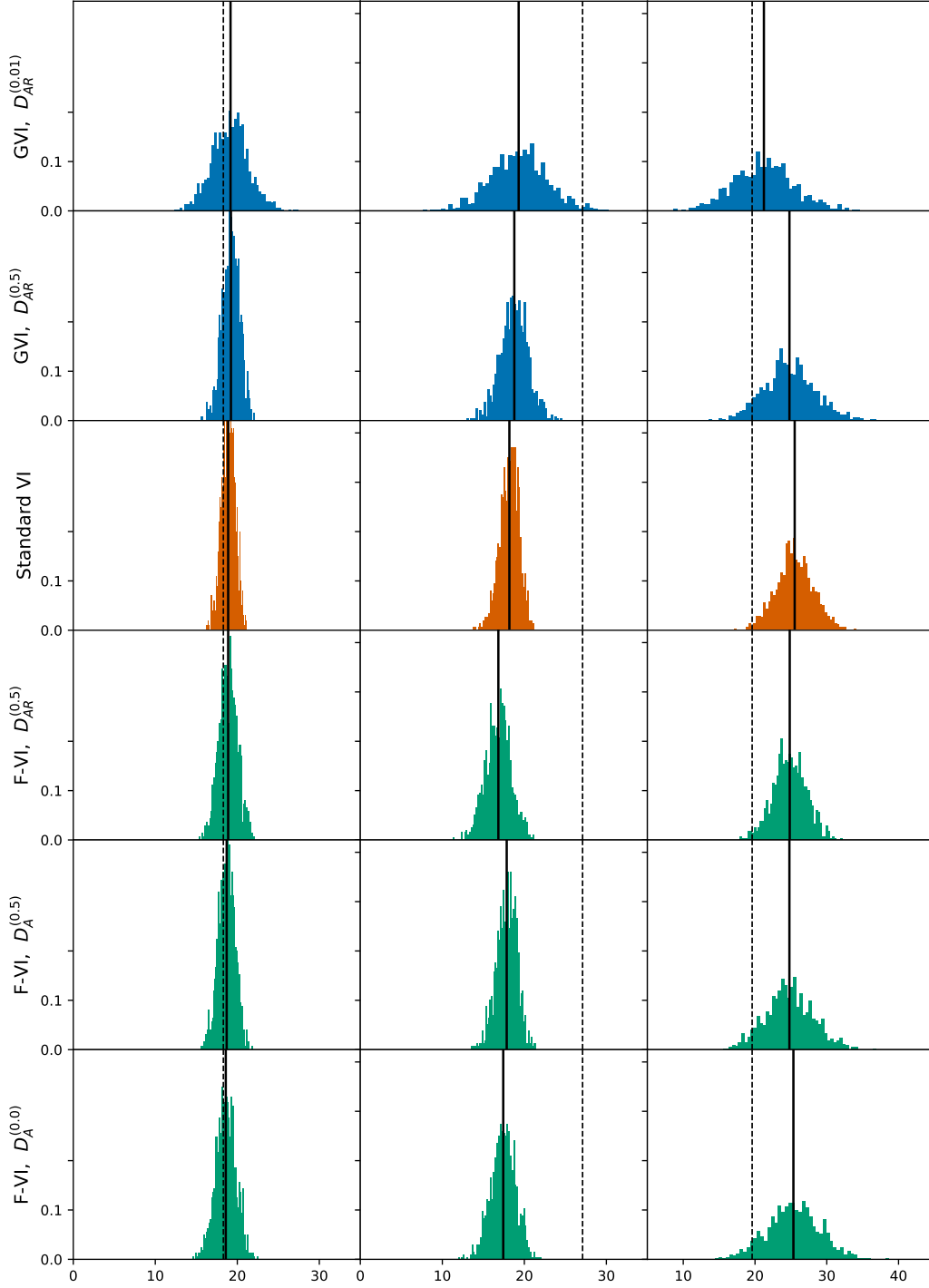


Figure 17: Variational posterior  $q(F(\theta, x)|x)$  over  $F$  for **VI**, the  $D_A^{(\alpha)}$ -**VI** method of Hernández-Lobato et al. [31], the  $D_{AR}^{(\alpha)}$ -**VI** method of Li and Turner [49] and **GVI** for  $D = D_{AR}^{(\alpha)}$  on three test points on the boston data sets; based on 1,000 samples each. One can see that the posteriors over  $\theta$  inherit the zero-avoiding properties as expected. Thus, they produce flatter variances. Note that the **GVI** methods with more conservative uncertainty quantification also provide flatter posterior variances over  $\theta$ .

between  $\sigma^2$  and the uncertainty quantification about  $\theta$ . Moreover, since GVI is specifically set up as an expected loss-minimizing objective with regularizer  $D$ , it is immediately clear what interpretation optimizing  $\sigma^2$  has. In particular, since it only enters  $\ell_n(\cdot, \cdot, \sigma^2) = \ell_n^{\sigma^2}$ , optimizing for it means that one seeks to find the one producing the *optimal loss* in  $\{\ell_n^{\sigma^2} : \sigma > 0\}$  for a given tuple  $(D, \mathcal{Q})$ . Depending on how conservative we want the uncertainty quantification to be (which we regulate via the fixed value  $\alpha$  for the  $D_{AR}^{(\alpha)}$ ), the optimal value for  $\sigma^2$  will be different. In particular, Table 2 shows that for more conservative  $D$ , more diffuse (i.e. larger) values of  $\sigma^2$  are preferred and vice versa. To conclude, we note that the optimizations of  $\sigma^2$  within our GVI examples behave according to our expectations precisely because of the separation of responsibilities inherent in GVI. This leads to the desired outcome: The more conservative our uncertainty quantification (i.e., the smaller  $\alpha$ ), the larger  $\sigma^2$ .

## 7.2 Interpreting performance improvements with over-concentration (relative to VI)

Throughout a wide range of data sets, test performance seem to benefit from more contracted posterior predictives. Conversely, less contracted posterior predictives seem to affect test performance adversely. This appears puzzling at first: Is not one of VI’s biggest issues the over-concentration of marginal variances?

While it is true that VI produces very narrow marginal variances within the mean-field family if the true posteriors are highly correlated, whether or not this is a problem will depend on two main factors: (1) how much information the correlation contains, (2) how expressive the model is. For instance, the often-cited paper of Turner and Sahani [79] investigates poor performance of VI for time series models with very few parameters. By their very nature, the parameter correlation inside time series models will encode a lot of informative and important information. Consequently, losing it will yield over-confident inference detrimental to out-of-sample performance. Moreover, this effect will be stronger if there are fewer parameters to begin with. In fact, in traditional statistical models with moderately many parameters  $\theta$ , over-concentration is a problem *even for exact inference* [12, 83, 57]. More precisely, this phenomenon occurs if there is model misspecification. Now, observe that misspecification simply means that the true Data Generating Mechanism  $g$  for  $x$  and the statistical model  $f(x|\theta)$  for  $x$  are not a good match. I.e., there exists no  $\theta$  such that  $g \approx f(x|\theta)$ . Again, this is a situation that typically holds for the time series models discussed by Turner and Sahani [79].

However, the situation is completely different from the BNNs on which we compare F-VI and GVI. If anything, BNNs are over-parameterized:  $|\theta| = (50 + 1) \cdot (D + 1)$  where  $D$  is the number of features. Consequently, the pairwise correlations between the parameters are much less informative than their (joint) mode and model misspecification issues are not typically a problem in the regression setting. In fact, our empirical finding that over-concentrated posteriors aid out-of-sample performance actually tells us that there is a direct trade-off between focusing on the (joint) mode and focusing on conservative uncertainty quantification. In particular, focusing on the latter will cause the the mode to be estimated less precisely, yielding a performance decrease on the test data.

## 7.3 Deep Gaussian Processes

### 7.3.1 Details

We use the variational family and code base of Salimbeni and Deisenroth [74]. Except for choosing 50 (instead of 20) -fold cross validation with a 10% randomly selected held out test set, all settings are the same as in Salimbeni and Deisenroth [74]: As in their paper, each experiment runs with ADAM [42] and a learning rate of 0.01 with 20,000 iterations. For the kernel, we choose the RBF kernel with a lengthscale for each dimension. The number of inducing points is 100 for all settings, and they are run after normalization with a whitened representation of the Gaussian Process. The batch size is  $\min(1000, n)$ , where  $n$  is the number of observations in the training set. Test metrics are computed by averaging 100 samples taken from the variational posterior. The Python implementation uses tensorflow [1] and gpflow [54].

### 7.3.2 Additional results

Using the same settings as for the results reported in the main paper, we also compute the test scores on five additional UCI data sets. The results are depicted in Fig. 18 and have the same implications

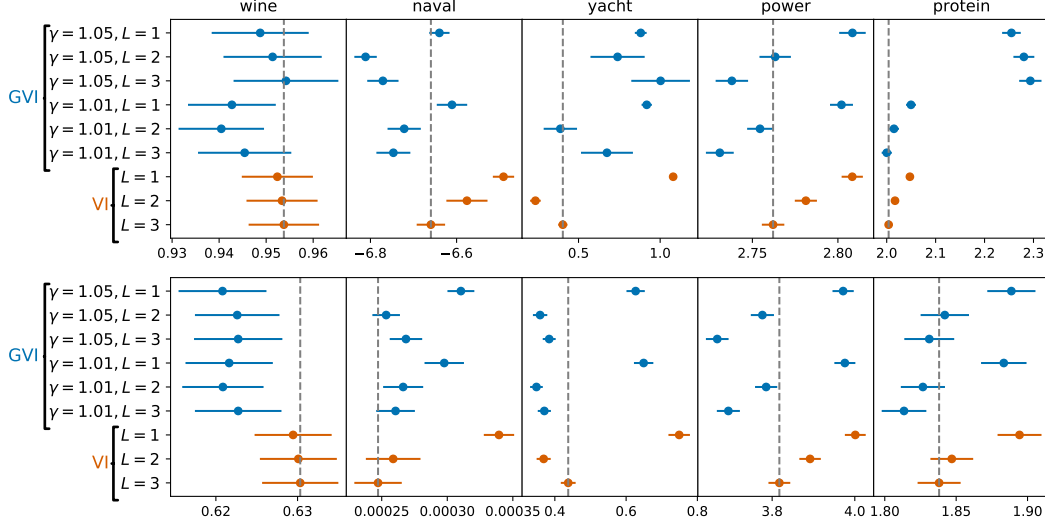


Figure 18: Comparing performance in DGPs with 3 layers for **GVI** with  $\ell_n(\theta, x) = \sum_{i=1}^n \mathcal{L}_p^\gamma(\theta, x_i)$  and **VI**. **Top row**: Negative test log likelihoods. **Bottom row**: Test RMSE. The lower the better.

as those in the main paper: Overall, one can expect a significant performance increase from mild robustification (i.e., for  $\gamma = 1.01$ ). Robustifying too much can result in a loss of efficiency and impact performance adversely (see e.g. the likelihood scores on the *protein* data set). This is especially true if the data set does not have a complicated structure and produces very good regression fits (e.g., *yacht* and *naval*).

In addition to varying the loss function, one can also consider varying the uncertainty quantification for the DGP. The results are shown in Fig. 19.

We consider the  $D_{AR}^{(\alpha)}$  for  $\alpha = 0.5$  in the last layer as well as  $\frac{1}{w}$ KLD:. For more conservative uncertainty quantification with the DGP, we focus on  $\frac{1}{w}$ KLD: for two reasons: Firstly, the prior  $\pi$  in the DGPs of Salimbeni and Deisenroth [74] is constructed in a very informative way. Thus, assigning more weight to it is not expected to lead to robustness issues. Secondly, for  $\alpha \notin (0, 1)$ , Theorem 3 does not generally apply for multivariate normal distributions. The reason for this is that the space of precision matrices is not closed under arbitrary linear combinations. Put simply, for two covariance matrices  $\Sigma_1, \Sigma_2$ , it typically does not hold that  $\alpha\Sigma_1^{-1} + (1 - \alpha)\Sigma_2^{-1}$  is again the inverse of a valid covariance matrix if  $(1 - \alpha) < 0$  or if  $\alpha < 0$ . This is not an issue for the completely factorized variational family on the BNNs: Here,  $\Sigma_1, \Sigma_2$  are diagonal. Thus, the conditions can be formulated entry-wise. In particular, it needs to hold that  $\alpha/\sigma_{1,ii}^2 + (1 - \alpha)/\sigma_{2,ii}^2 > 0$  for all  $i$  on the diagonal. Since the prior is significantly more uninformative than the posterior, this condition is satisfied for the BNN examples.

## References

- [1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283.
- [2] Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414.
- [3] Amari, S.-i. (2012). *Differential-geometrical methods in statistics*, volume 28. Springer Science & Business Media.
- [4] Ambrogioni, L., Güçlü, U., Güçlütürk, Y., Hinne, M., van Gerven, M. A. J., and Maris, E. (2018). Wasserstein variational inference. In *Advances in Neural Information Processing Systems 31*, pages 2478–2487.



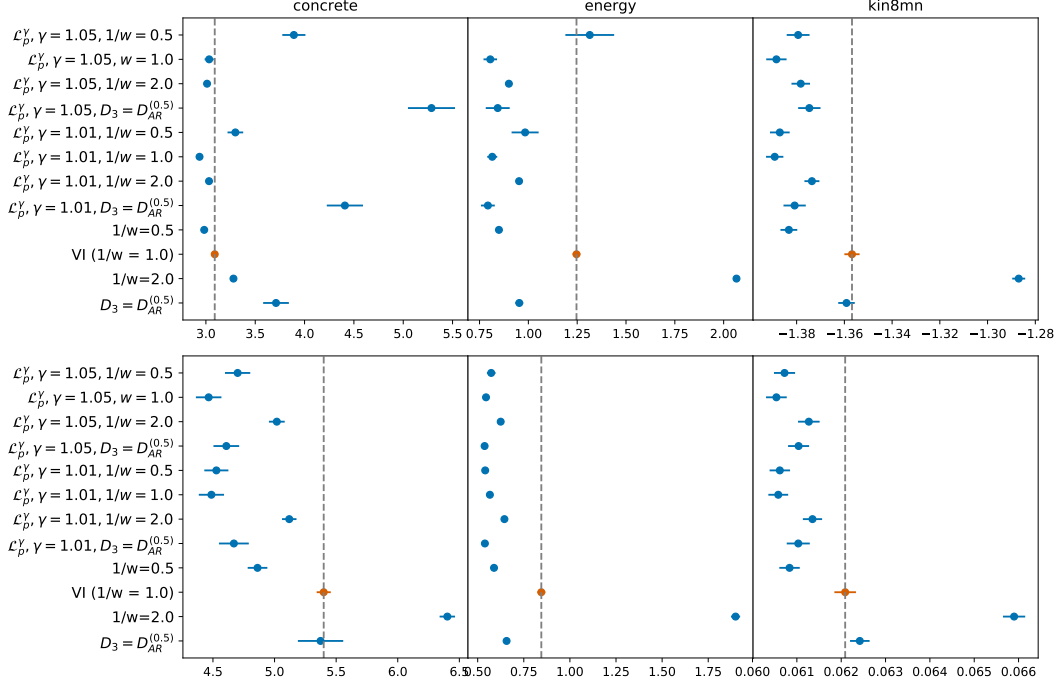


Figure 19: Comparing performance in DGPs with 3 layers for **GVI** with  $\ell_n(\theta, \mathbf{x}) = \sum_{i=1}^n \mathcal{L}_p^\gamma(\theta, x_i)$  and **VI** on some additional data sets. **Top row**: Negative test log likelihoods. **Bottom row**: Test RMSE. The lower the better.

- [5] Bamler, R., Zhang, C., Oppner, M., and Mandt, S. (2017). Perturbative black box variational inference. In *Advances in Neural Information Processing Systems*, pages 5079–5088.
- [6] Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.
- [7] Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. University College London.
- [8] Berg, R. v. d., Hasenclever, L., Tomczak, J. M., and Welling, M. (2018). Sylvester normalizing flows for variational inference. *arXiv preprint arXiv:1803.05649*.
- [9] Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.
- [10] Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., et al. (1994). An overview of robust Bayesian analysis. *Test*, 3(1):5–124.
- [11] Bernardo, J. (2000). Bayesian theory. *Wiley Series in Probability and Statistics*. 23 cm. 586 p.
- [12] Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130.
- [13] Bonilla, E. V., Krauth, K., and Dezfouli, A. (2016). Generic inference in latent Gaussian process models. *arXiv preprint arXiv:1609.00577*.
- [14] Bui, T., Hernández-Lobato, D., Hernandez-Lobato, J., Li, Y., and Turner, R. (2016). Deep Gaussian processes for regression using approximate expectation propagation. In *International Conference on Machine Learning*, pages 1472–1481.

- [15] Chernoff, H. et al. (1952). 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.
- [16] Cichocki, A. and Amari, S.-i. (2010). Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568.
- [17] Damianou, A. and Lawrence, N. (2013). Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215.
- [18] Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93.
- [19] Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. (2017). Variational inference via  $\chi$  upper bound minimization. In *Advances in Neural Information Processing Systems*, pages 2732–2741.
- [20] Eguchi, S. et al. (1985). A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima mathematical journal*, 15(2):341–391.
- [21] Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081.
- [22] Futami, F., Sato, I., and Sugiyama, M. (2017). Variational inference based on robust divergences. *arXiv preprint arXiv:1710.06595*.
- [23] Ganchev, K., Gillenwater, J., Taskar, B., et al. (2010). Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11(Jul):2001–2049.
- [24] Ghosh, A. and Basu, A. (2016). Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437.
- [25] Gil, M. (2011). *On Rényi divergence measures for continuous alphabet sources*. PhD thesis.
- [26] Gil, M., Alajaji, F., and Linder, T. (2013). Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131.
- [27] Giordano, R., Broderick, T., and Jordan, M. I. (2018). Covariances, robustness, and variational bayes. *Journal of Machine Learning Research*, 19:1–49.
- [28] Grünwald, P., Van Ommen, T., et al. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.
- [29] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons.
- [30] Hernández-Lobato, J. M. and Adams, R. (2015). Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869.
- [31] Hernández-Lobato, J. M., Li, Y., Rowland, M., Hernández-Lobato, D., Bui, T. D., and Turner, R. E. (2016). Black-box  $\alpha$ -divergence minimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1511–1520.
- [32] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3.
- [33] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.
- [34] Holmes, C. and Walker, S. (2017). Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503.
- [35] Hooker, G. and Vidyashankar, A. N. (2014). Bayesian model robustness via disparities. *Test*, 23(3):556–584.

- [36] Huang, C.-W., Tan, S., Lacoste, A., and Courville, A. C. (2018). Improving explorability in variational inference with annealed variational objectives. In *Advances in Neural Information Processing Systems 31*, pages 9724–9734.
- [37] Huber, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer.
- [38] Huber, P. J. et al. (1964). Robust estimation of a location parameter. *The annals of mathematical statistics*, 35(1):73–101.
- [39] Hung, H., Jou, Z.-Y., and Huang, S.-Y. (2018). Robust mislabel logistic regression without modeling mislabel probabilities. *Biometrics*, 74(1):145–154.
- [40] Jewson, J., Smith, J., and Holmes, C. (2018). Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442.
- [41] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.
- [42] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [43] Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751.
- [44] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.
- [45] Knoblauch, J. (2019). Robust deep gaussian processes. *arXiv preprint arXiv:1904.02303*.
- [46] Knoblauch, J., Jewson, J., and Damoulas, T. (2018). Doubly robust Bayesian inference for non-stationary streaming data using  $\beta$ -divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 64–75.
- [47] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- [48] Kurtek, S. and Bharath, K. (2015). Bayesian sensitivity analysis with the Fisher–Rao metric. *Biometrika*, 102(3):601–616.
- [49] Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081.
- [50] Lichman, M. et al. (2013). Uci machine learning repository.
- [51] Liese, F. and Vajda, I. (1987). Convex statistical distances.
- [52] Lyddon, S., Holmes, C., and Walker, S. (2017). Generalized Bayesian updating and the loss-likelihood bootstrap. *arXiv preprint arXiv:1709.07616*.
- [53] Maclaurin, D., Duvenaud, D., and Adams, R. P. (2015). Autograd: Effortless gradients in numpy. In *ICML 2015 AutoML Workshop*.
- [54] Matthews, D. G., Alexander, G., Van Der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrà, P., Ghahramani, Z., and Hensman, J. (2017). Gpflow: A Gaussian process library using tensorflow. *The Journal of Machine Learning Research*, 18(1):1299–1304.
- [55] Mihoko, M. and Eguchi, S. (2002a). Robust blind source separation by beta divergence. *Neural computation*, 14(8):1859–1886.
- [56] Mihoko, M. and Eguchi, S. (2002b). Robust blind source separation by beta divergence. *Neural computation*, 14(8):1859–1886.
- [57] Miller, J. W. and Dunson, D. B. (2018). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, (just-accepted):1–31.

- [58] Minka, T. (2004). Power ep. Technical report, Technical report, Microsoft Research, Cambridge.
- [59] Minka, T. et al. (2005). Divergence measures and message passing. Technical report, Technical report, Microsoft Research.
- [60] Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc.
- [61] Nakagawa, T. and Hashimoto, S. (2019). Robust bayesian inference via  $\gamma$ -divergence. *Communications in Statistics-Theory and Methods*, pages 1–18.
- [62] Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.
- [63] Oppor, M. and Winther, O. (2000). Gaussian processes for classification: Mean-field algorithms. *Neural computation*, 12(11):2655–2684.
- [64] Paisley, J., Blei, D. M., and Jordan, M. I. (2012). Variational Bayesian inference with stochastic search.
- [65] Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822.
- [66] Ranganath, R., Tran, D., Altosaar, J., and Blei, D. (2016). Operator variational inference. In *Advances in Neural Information Processing Systems*, pages 496–504.
- [67] Regli, J.-B. and Silva, R. (2018). Alpha-beta divergence for variational inference. *arXiv preprint arXiv:1805.01045*.
- [68] Rényi, A. et al. (1961). 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.
- [69] Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*.
- [70] Rue, H., Martino, S., and Chopin, N. (2009). 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.
- [71] Saha, A., Bharath, K., and Kurtek, S. (2017). A geometric variational approach to Bayesian inference. *arXiv preprint arXiv:1707.09714*.
- [72] Salimans, T. and Knowles, D. A. (2014). On using control variates with stochastic approximation for variational Bayes and its connection to stochastic linear regression. *arXiv preprint arXiv:1401.1022*.
- [73] Salimbeni, H., Cheng, C.-A., Boots, B., and Deisenroth, M. (2018). Orthogonally decoupled variational Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 8725–8734.
- [74] Salimbeni, H. and Deisenroth, M. (2017). Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599.
- [75] Samek, W., Blythe, D., Müller, K.-R., and Kawanabe, M. (2013). Robust spatial filtering with beta divergence. In *Advances in Neural Information Processing Systems*, pages 1007–1015.
- [76] Titsias, M. and Lázaro-Gredilla, M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*, pages 1971–1979.
- [77] Tran, D., Ranganath, R., and Blei, D. M. (2015). The variational Gaussian process. *arXiv preprint arXiv:1511.06499*.

- [78] Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485.
- [79] Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In *Bayesian time series models*. Cambridge University Press.
- [80] Van Erven, T. and Harremoës, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.
- [81] Wang, D., Liu, H., and Liu, Q. (2018). Variational inference with tail-adaptive f-divergence. In *Advances in Neural Information Processing Systems*, pages 5742–5752.
- [82] Wang, Y., Kucukelbir, A., and Blei, D. M. (2017). Reweighted data for robust probabilistic models. *International Conference on Machine Learning*.
- [83] Watson, J., Holmes, C., et al. (2016). Approximate models and robust decisions. *Statistical Science*, 31(4):465–489.
- [84] Wu, M., Goodman, N., and Ermon, S. (2018). Differentiable antithetic sampling for variance reduction in stochastic variational inference. *arXiv preprint arXiv:1810.02555*.
- [85] Yang, Y., Martin, R., and Bondell, H. (2019). Variational approximations using Fisher divergence. *arXiv preprint arXiv:1905.05284*.
- [86] Yang, Y., Pati, D., and Bhattacharya, A. (2017).  $\alpha$ -variational inference with statistical guarantees. *arXiv preprint arXiv:1710.03266*.
- [87] Zellner, A. (1988). Optimal information processing and Bayes’s theorem. *The American Statistician*, 42(4):278–280.
- [88] Zhu, J., Chen, N., and Xing, E. P. (2014). Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1):1799–1847.