3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–17

Understanding Variational Inference in Function-Space

David R. Burt DRB62@CAM.AC.UK
Sebastian W. Ober SWO25@CAM.AC.UK
Adrià Garriga-Alonso AG919@CAM.AC.UK
Department of Engineering, University of Cambridge, UK

Mark van der Wilk M.VDWILK@IMPERIAL.AC.UK
Department of Computing, Imperial College London, UK

Abstract

Recent work has attempted to directly approximate the 'function-space' or predictive posterior distribution of Bayesian models, without approximating the posterior distribution over the parameters. This is appealing in e.g. Bayesian neural networks, where we only need the former, and the latter is hard to represent. In this work, we highlight some advantages and limitations of employing the Kullback-Leibler divergence in this setting. For example, we show that minimizing the KL divergence between a wide class of parametric distributions and the posterior induced by a (non-degenerate) Gaussian process prior leads to an ill-defined objective function. Then, we propose (featurized) Bayesian linear regression as a benchmark for 'function-space' inference methods that directly measures approximation quality. We apply this methodology to assess aspects of the objective function and inference scheme considered in Sun et al. (2018), emphasizing the quality of approximation to Bayesian inference as opposed to predictive performance.

1. Introduction

While neural networks offer a successful parametric representation of functions, performing Bayesian inference over the parameters is challenging. Since only the predictions matter, a recent line of work has focused on directly approximating the posterior predictive distribution in 'function-space' (Sun et al., 2018; Ma et al., 2019; Wang et al., 2019), with the intent of reducing the influence of the specific parameterization on the quality of inference. Similar to approximate Gaussian processes, a variational approach can be taken, where a discrepancy (e.g. a KL divergence) is minimized between approximate and posterior predictive distributions (Matthews et al., 2016). However, we are dealing with measures on functions, so care is needed to ensure that the divergence is well-defined and a useful objective function.

In this work, we investigate situations where this may not form a useful objective. Sun et al. (2018) hint that: "the function space KL divergence may be infinite, for instance if the prior assigns measure zero to the set of functions representable by a neural network". We give examples of ill-defined objective functions that arise in the context of variational inference between approximate and exact predictive posteriors. We give a proof that variational inference using the KL divergence (or any f-divergence) does not lead to a sensible objective function when the prior is a non-degenerate Gaussian process and the variational family contains only 'nice' parametric models (or vice versa). We show similar results for single hidden layer (1HL) Bayesian neural networks (BNNs) in the case when the prior has a different width than the approximate posterior and ReLU activation functions are used. Our proofs are contained in the appendix.

Having established that issues with the objective function can arise; we consider how to assess the quality of the approximation to the posterior achieved by methods motivated by performing variational inference in function-space. We propose Bayesian linear regression (BLR) as a benchmark for measuring the effectiveness of functional variational inference schemes. Since the exact posterior, as well as the KL divergence in function-space, can be computed in this case, we can directly assess the quality of *approximate inference* separately from the quality of predictive performance, which may be good even in cases where inference is not accurate. We show how this benchmark can be used to decouple the impacts of further approximations to the objective function made by Sun et al. (2018). This approach gives a principled starting point for assessing future improvements in functional variational inference.

2. Background

We begin by introducing some measure theory notation, which is required to handle the distributions on functions that we perform inference with. We review the data processing inequality, which is needed for the proofs, and discuss the work of Sun et al. (2018) which we analyze further in this work.

Notation A probability measure, P, is a function from subsets ('events') to $[0, 1]$, such that the probability of the event E is given by $P(E)$. In the case when the subsets are contained in $\mathbb{R}^k$, this can sometimes be related to a (Lebesgue) density $p : \mathbb{R}^k \to [0, \infty)$ by $P(E) = \int_{z \in E} p(z) dz$. A random variable, Z, is said to have distribution P (and we write $Z \sim P$) if for all events E, $\text{Prob}(Z \in E) = P(E)$. For measures defined on the same event space, we write $Q \ll P$ if for all E such that $P(E) = 0$, we have $Q(E) = 0$. The *Kullback-Leibler (KL) divergence* between two probability measures Q and P is given by,

$$\mathrm{D}_{KL}(Q, P) = \begin{cases} \int \log \frac{dQ}{dP} dQ & Q \ll P, \\ \infty & \text{otherwise}, \end{cases} \tag{1}$$

where $\frac{dQ}{dP}$ denotes the Radon-Nikodym derivative, which is simply the ratio of the densities of these measures when the densities exist.[1] While we focus on the KL divergence in the main text due to its wide-spread use in variational inference, our results can be naturally extended to other f-divergences as described in appendix B.

The data processing inequality Given a random variable $Z \sim P$, we can transform it by a function g to find a new random variable $g(Z)$. We refer to its distribution as the *pushforward measure* of P, which we denote $g_* P$. The *data processing inequality* states that if two random variables are transformed in this way, they cannot become easier to tell apart.

Proposition 1 (Data processing inequality Polyanskiy and Wu, 2014, Thm 6.2) *Let g be a measurable function, then* $\mathrm{D}_{KL}\left(g_* P, g_* Q\right) \leq \mathrm{D}_{KL}\left(P, Q\right)$,

[1] As we are often interested in the case when Q and P are defined on spaces without a Lebesgue measure, we use this more general formulation of KL divergence.

Background on Functional Variational Inference We consider the application of variational inference to regression. In particular, we assume data $D = \{(x_n, y_n)\}_{n=1}^N$ has been observed, with $x_n \in \mathcal{X} = \mathbb{R}^d$ and $y_n \in \mathbb{R}$. We assume an additive noise model, i.e $\hat{y}(x_*) = \hat{f}(x_*) + \epsilon_*$, where $\hat{f}$ is a stochastic process indexed by $\mathcal{X}$ and each ϵ_* is an independent mean-zero random variable. We assume a priori that $\hat{f} \sim P$. The goal is to approximate the *posterior* distribution of $\hat{f}$ given the data D, P_D. Define $\ell_D : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ to be the likelihood function given the observed data D.[2] Given an approximate distribution Q, variational inference (Blei et al., 2017) gives us the evidence lower bound (ELBO)

$$\log \int \ell_D dP \geq \int \log \ell_D dQ - \mathrm{D}_{KL}(Q, P). \tag{2}$$

Maximizing the RHS of eq. (2) over $Q \in \mathcal{Q}$ is equivalent to minimizing $\mathrm{D}_{KL}(Q, P_D)$. Moreover, the RHS of eq. (2) can often be estimated: $\int \log \ell_D dQ$ can be estimated with Monte Carlo methods so long as we can evaluate ℓ_D and sample from Q. In the context of variational inference in parameter space, Q is generally constrained to be from some family such that $\mathrm{D}_{KL}(Q, P)$ is tractable (e.g. if P is an isotropic Gaussian distribution, then Q is often chosen to be Gaussian so that this KL divergence can be evaluated in closed form).

Sun et al. (2018) proposed using eq. (2) with Q and P the approximate predictive and prior predictive distributions, in which case $D_{KL}(Q, P)$ is a divergence between measures on the infinite product space $\mathbb{R}^{\mathcal{X}}$. The starting point for their work is:

Proposition 2 (Sun et al. 2018, Theorem 1) *For measures Q, P on (the product σ-algebra of) $\mathbb{R}^{\mathcal{X}}$,*

$$\mathrm{D}_{KL}(Q, P) = \sup_{\mathrm{X} \subset \mathcal{X},\, |X| < \infty} \mathrm{D}_{KL}(Q_X, P_X), \tag{3}$$

where Q_X, P_X are the marginals of the measures Q and P on the set X.[3]

In other words, the KL divergence between the stochastic processes is equal to the supremum of the KL divergence between the measures restricted to finite marginals. Substituting eq. (3) into eq. (2) introduces two sources of intractability. First, the supremum is over uncountably many subsets, and will be generally intractable. Second, the distributions Q_X, P_X are often defined implicitly through a tractable sampling procedure, which does not provide closed form densities (with the notable exception when either Q_X or P_X is a Gaussian measure).

Sun et al. (2018) propose replacing the supremum with an expectation, and using this in eq. (2) to address the first intractability. This involves defining a distribution over finite subsets of $\mathcal{X}$, e.g. sampling points from the data as well as uniformly from a subset of $\mathcal{X}$, and using the KL divergence between the approximate posterior and the prior restricted to this index set. The second intractability can be addressed using any form of implicit inference (Huszár, 2017). Sun et al. (2018) use the spectral Stein gradient estimator (SSGE; Shi et al. (2018)) to obtain estimates of the gradient of the KL.

[2] Commonly this is written $\prod_{n=1}^N p(y_n | x_n, \hat{f})$.

[3] Precisely, $Q_X = \pi_{X*}Q, P_X = \pi_{X*}P$, where π_X denotes the canonical projection (definition 1) onto $x \in X$.

3. Properties of the KL divergence in function-space

In this section we discuss properties of KL divergences in function-space, noting that our results generalize with minor modifications to all f-divergences (appendix B). We focus on finding conditions under which eq. (2) is a well-defined objective. We discuss the case when parametric models are used to describe both the prior and the approximate posterior, and then move to the case when a Gaussian process is used as either the approximate posterior or the prior.

Parametric distributions We call a distribution parametric if it is described by a probability distribution over a parameter space, which we assume is $\mathbb{R}^k$, as well as a mapping from parameters Θ to functions, i.e. $\hat{f}(x) = h(x, \Theta)$ for all $x \in \mathcal{X} = \mathbb{R}^d$. Note that $\hat{f}$ is a random function (stochastic process) indexed by $\mathcal{X}$. Moreover, from its definition, we see that given $\Theta \sim P_\Theta$, upon defining $g(\Theta)(x) = h(x, \Theta)$, we have $\hat{f} = g(\Theta) \sim g_* P_\Theta$. We call such a distribution parameterized by the pair (P_Θ, g). The function $g : \mathbb{R}^k \to \mathbb{R}^{\mathcal{X}}$ is the mapping from parameters to functions and is assumed to be measurable. From the data processing inequality (proposition 1), we make the following observation:

Proposition 3 *Suppose that the approximate posterior is parameterized by (Q_θ, g) and that the prior is parameterized by (P_θ, g). Define $Q = g_* Q_\theta$ and $P = g_* P_\theta$ to be the approximate posterior predictive and the prior predictive respectively. Then,*

$$\mathrm{D}_{KL}(Q, P) \leq \mathrm{D}_{KL}(Q_\theta, P_\theta). \tag{4}$$

If g is injective (each set of parameters leads to a unique predictive function) equality holds.

Remark 1 *Proposition 3 implies that in cases where we can perform variational inference in parameter space, variational inference in function-space is also well-defined. While the ELBO using KL-divergence in parameter-space depends on the choice of parameterization (unless g is injective), the ELBO in function-space does not.*

Ma et al. (2019) observed the inequality in eq. (4) in the context of performing inference with stochastic processes; it is an immediate consequence of the data processing inequality. The equality can be derived by noting that if g is injective, it has a left inverse and the data-processing inequality can be applied in the opposite direction. We note that it is possible that for specific Q, P equality can hold even if g is not injective: in other words, the converse is not generally true.

When applied to variational inference in function-space, this inequality tells us that the KL divergence in function-space is no larger than the KL divergence in parameter space, implying that the evidence lower bound (ELBO) obtained in function-space must be closer to the log marginal likelihood than the ELBO in parameter space.

Bayesian neural networks The above discussion applies immediately to variational inference with Bayesian neural networks (after establishing measurability of g), given the approximate posterior and the prior have the same architecture. This can be seen as motivation for using functional variational inference, particularly since BNNs are highly non-identifiable (e.g. permuting neurons leads to identical predictions), and equality between parameter-space and function-space KLs does not generally hold in these models. A natural question is whether variational inference can be applied when the prior and candidate approximate posteriors have different architectures. We show in appendix C that in the case of single hidden ReLU networks, if the prior and approximate posterior have different widths, and both Q_Θ and P_Θ have densities (with respect to Lebesgue measure), then the KL divergence between the approximate posterior and the prior is infinite. We conjecture that this result is true much more broadly for Bayesian neural networks of different architectures, excluding cases where architectures are trivially the same (e.g. if a neuron is added but the outgoing weight is 0 with probability 1 so that the additional neuron is always pruned).

Remark 2 *Proposition 9 (appendix C) shows that variational inference in function space is not always well-defined when both the prior and approximate posteriors are defined using neural networks, if the architectures are not the same.*

Parametric distributions and Gaussian processes A Gaussian process (GP) is a random function such that when the function is indexed at any finite collection of points, the distribution of the function values is multivariate Gaussian. We call a Gaussian process non-degenerate if there exist arbitrarily large collections of points where, when we evaluate the function at these points, the resulting multivariate Gaussian has a full-rank covariance matrix (equivalently has a density with respect to the appropriate Lebesgue measure).[4]

Gaussian processes have been proposed for use in functional inference schemes both as priors (Sun et al., 2018) and as approximate posteriors (Ma et al., 2019). However, under quite general conditions, we show that the KL divergence between Gaussian processes and parametric models is not a useful objective.

Proposition 4 *Let (Q_θ, g) parameterize the approximate posterior of a parametric model and let P be a non-degenerate Gaussian process. Assume that $g(\cdot)(x)$ is locally Lipschitz for all x. Let $Q = g_* Q_\theta$. Then, $\mathrm{D}_{KL}(Q, P) = \infty$ and $\mathrm{D}_{KL}(P, Q) = \infty$.*

The assumption that g is locally Lipschitz in each output is very weak; it holds for most commonly used Bayesian machine learning models, including (deep) BNNs with ReLU, tanh or sigmoid non-linearities.

Remark 3 *Proposition 4 tells us that using KL divergences as an objective function to approximate Gaussian processes with parametric models does not lead to a useful objective. However, this does not mean that parametric models cannot approximate Gaussian processes well. This is evidently false from the success of methods such as Random Fourier features (Rahimi and Recht, 2007) and Subset of Regressors (Wahba, 1990, §7).*

While propositions 4 and 9 highlight limitations of variational objective defined in function space, we believe that the overall idea of approximating the predictive posterior as opposed to the parameter-space posterior is well-motivated. Propositions 4 and 9 suggest the need for other objective functions for this task, as well as carefully assessing whether the predictive posterior obtained by variational inference in function space resembles the exact predictive posterior.

[4] To be non-degenerate, the GP needs infinite basis functions (Rasmussen and Williams, 2006, §4.3).

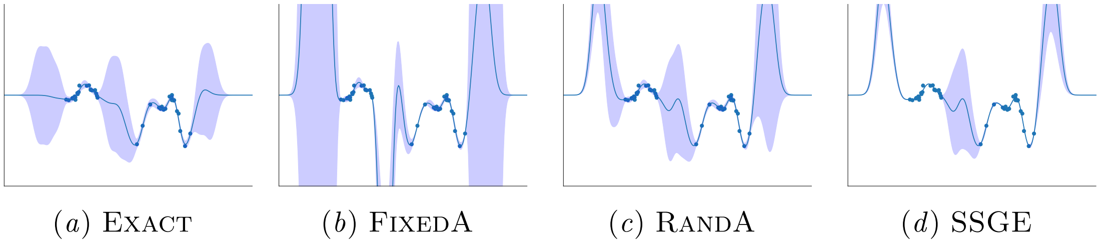
(*a*) Exact (*b*) FixedA (*c*) RandA (*d*) SSGE

Figure 1: Predictive posteriors for each method using all (full-covariance) Gaussian distributions as $\mathcal{Q}$, in a toy 1d regression.

4. Benchmarking Functional Approaches to Variational Inference

Bayesian linear regression with Gaussian priors can be used as a tool for assessing the quality of variational inference in function-space. We consider the model,

$$\hat{y}(x_i) = \Theta^\intercal \phi(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \Theta \sim \mathcal{N}(0, I), \tag{5}$$

where $\phi : \mathcal{X} \to \mathbb{R}^k$ is a feature mapping, and Θ is the (random) vector of weights. Recall the notation introduced in Section 3, $g(\Theta)(x) = \Theta^T \phi(x)$. In this case, we can verify that g is injective for a given set of features by finding a set of k inputs $A = \{a_1, \dots, a_k\}$ such that the vectors $\{\phi(a_1), \dots, \phi(a_k)\}$ are linearly independent. Therefore, proposition 3 implies that the KL divergence in parameter and function-space are *exactly the same*. We can therefore expect that successful inference methods with identical variational families should obtain the same approximate posteriors, regardless of whether they are represented in the parameter space or function-space.

Since the exact solution for BLR is Gaussian and can be computed analytically, we can compare different function-space inference methods by seeing which method finds the solution with the smallest KL divergence to the exact posterior. This allows us to assess the quality of *inference* while avoiding potential issues of model mis-specification, whereby it is possible to achieve good test performance with a poor model by using poor inference.

Following Sun et al. (2018), we consider the modified variational objective,

$$\max_{Q \in \mathcal{Q}} \left(\mathbb{E}_{\Theta \sim Q}[\log p(y|x, \Theta)] - \mathbb{E}_{A \sim \mu}[\mathrm{D}_{KL}(Q_A, P_A)] \right) \tag{6}$$

where P_A, Q_A denote marginals of P, Q at points indexed by A and μ is a measure on subsets of $\mathcal{X}$. We note that eq. (6) may be finite, even in cases where eq. (2) is not. We compare four algorithms using both full-covariance Gaussians (Full) and fully-factorised Gaussians (FFG) as the approximating families (noting that Full will contain the true posterior). As a baseline, we consider using the Exact KL divergence, i.e. $\mathrm{D}_{KL}(Q, P)$, which we can obtain since the KLs are the same in weight and function-space. In FixedA, we randomly select a set of input points A and keep it fixed throughout training in eq. (6). For RandA, we use the approach proposed in Sun et al. (2018) and sample a different set A at each iteration, so that we Monte Carlo evaluate $\mathbb{E}_{A \sim \mu}[\mathrm{D}_{KL}(Q_A, P_A)]$ in eq. (6). However,

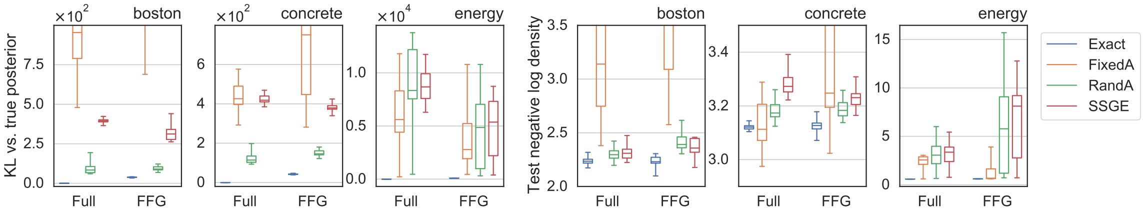

Figure 2: NLPD and KL for the first UCI data sets (alphabetic order). Lower is better. In BOSTON and CONCRETE, the KL divergence to the posterior of SSGE is much larger than the one from RandA.

as we can evaluate the KL exactly in this case, we do not use SSGE, instead leaving it for the final algorithm, SSGE, which uses the random sampling scheme as well. Therefore, SSGE is similar to the implementation in Sun et al. (2018), although we do not use their heuristic for re-scaling the KL term to reduce over-fitting (as they note that the modified objective will underestimate the KL term since it cannot achieve the supremum over all finite inputs). We consider two experiments, for which we provide additional experimental details in appendix F.

Toy experiment We generate a synthetic 1D dataset by sampling a 20-dimensional weight vector from the prior and applying it to 20 radial basis function features. We then use this model to sample 40 noisy (x, y) pairs with a noise standard deviation of 0.1. We perform full-covariance inference with each method and plot the predictive posteriors in Figure 1. The posteriors found using the approximate KL divergence are prone to over-fitting; in the case of FixedA it can be shown the optimal mean behaves like a combination of maximum likelihood estimation (MLE) and maximum a posteriori (MAP) inference (appendix E).

UCI regression task We fit a sparse variational GP (Titsias, 2009) with a squared-exponential kernel, with a separate length-scale for each input dimension. The learned kernel parameters and inducing points form a good representation of the training data (Wahba, 1990, §7). We use these features for linear regression. We show the negative log predictive densities (NLPDs) and KLs to the true posterior for the BOSTON, CONCRETE, and ENERGY datasets in Figure 2, providing additional results in appendix F.

Conclusions In general, we observe from both experiments that, of the approximate methods RandA, performs the best in terms of matching the true posterior. However, we note that all the approximate methods exhibit some amount of overfitting, which we would expect since the true functional KL divergence is obtained by taking the supremum over *all* finite marginals. Finally, we note that SSGE tends to perform worse than RandA, which we would expect as it introduces an additional approximation. It is our hope that using this benchmark will help researchers find ways of improving these methods; for example, using other methods of implicit inference to reduce the discrepancy between RandA and a version of inference that does not rely on Gaussianity.

Acknowledgments

The authors would like to thank Andrew Y.K. Foong and Wessel Bruinsma for providing feedback on a draft of this work, as well as Christos Louizos for useful conversations related to this research. DRB would like to acknowledge funding from the Qualcomm innovation fellowship and the Herchel Smith fellowship. SWO acknowledges the support of the Gates Cambridge Trust in funding his doctoral studies.

References

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Keith Conrad. The fundamental theorem of algebra via proper maps. URL https://kconrad.math.uconn.edu/blurbs/fundthmalg/propermaps.pdf.

Ferenc Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Chao Ma, Yingzhen Li, and José Miguel Hernández-Lobato. Variational implicit processes. In *International Conference on Machine Learning*, 2019.

Alexander G de G Matthews, James Hensman, Richard E. Turner, and Zoubin Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 231–239, 2016.

Alexander G. de G. Matthews, Mark van der Wilk, Tom Nickson, Keisuke. Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 2017.

James R. Munkres. *Topology*. Prentice Hall, Incorporated, 2000.

Jorge Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of computation*, 1980.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC)*, 2014.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Walter Rudin. *Real and complex analysis*. McGraw-Hill, 1966.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. A spectral approach to gradient estimation for implicit distributions. In *International Conference on Machine Learning*, 2018.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational Bayesian neural networks. In *International Conference on Learning Representations*, 2018.

Terence Tao. *An introduction to measure theory.* American Mathematical Society, 2011.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, 2009.

Grace Wahba. *Spline models for observational data.* SIAM, 1990.

Ziyu Wang, Tongzheng Ren, Jun Zhu, and Bo Zhang. Function space particle optimization for Bayesian neural networks. In *International Conference on Learning Representations*, 2019.

Appendix A. Measure-theoretic Definitions and Lemmas

In this section, we recall several definitions and lemmas that will be useful in formalizing the results in the main text.

We say two measures P, Q on a common measurable space are *equivalent* and write $P \sim Q$ if $P \ll Q$ and $Q \ll P$. We say P and Q are *mutually singular* and write $P \perp Q$ if there exists a (measurable) event E such that $P(E) = 0$ and $Q(E^c) = 0$ (where E^c denotes the complement of E). In the case of probability measures, this is the same as $P(E) = 0$ and $Q(E) = 1$.

Lemma 1 *Let Q, Q', and P be three measures on the measurable space (A, Σ_A). Then $Q \sim Q'$ and $Q' \perp P$ implies that $Q \perp P$.*

Proof By the assumption $Q' \perp P$, there exists an event $E \in \Sigma_A$ such that $Q'(E) = 0$ and $P(E^c) = 0$. Since $Q \sim Q'$, $Q'(E) = 0$ implies $Q(E) = 0$. ■

Definition 1 (Canonical projection) *For any $A \subset \mathcal{X}$, let $\pi_A : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^A$ denote the canonical projection onto A, i.e. $\pi_A(f) = (f(a))_{a \in A}$.*

A.1. Product σ-algebra

For any finite $A \subset \mathcal{X}$, we let λ_A denote Lebesgue measure on $\mathbb{R}^A$, restricted to the Borel σ-algebra. As in Sun et al. (2018), we consider the product σ-algebra on $\mathbb{R}^{\mathcal{X}}$ i.e. the coarsest σ-algebra on $\mathbb{R}^{\mathcal{X}}$ such that $\pi_{\{x\}} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ is measurable as a map to $\mathbb{R}$ equipped with the Borel σ-algebra for all $x \in \mathcal{X}$. For arbitrary $A \subset \mathcal{X}$ the map π_A is measurable when $\mathbb{R}^A$ and $\mathbb{R}^{\mathcal{X}}$ are both equipped with their respective product σ-algebras (see Tao (2011, Exercise 2.4.1.2)).

Lemma 2 *Let Q, P denote measures on Σ. Suppose there exists an $A \subset \mathcal{X}$ such that $\pi_{A*}Q \perp \pi_{A*}P$. Then $Q \perp P$.*

Proof Let $E \in \Sigma_A$ denote a witness to the orthogonality of $\pi_{A*}P \perp \pi_{A*}Q$. Then $\pi_A^{-1}(E) \in \Sigma$ by the measurability of projections. It follows from the definition of a pushforward measure that $\pi_A^{-1}(E)$ is a witness to the orthogonality of P and Q. ■

A.2. Topological Lemmas

As we work in the product σ-algebra generated by the Borel σ-algebra, we use several lemmas from topology in order to prove sets are measurable. We note that in general the product σ-algebra on $\mathbb{R}^{\mathcal{X}}$ generated by the Borel σ-algebra on $\mathbb{R}$ is not the same as the Borel σ-algebra generated by the product topology on $\mathbb{R}^{\mathcal{X}}$ when $\mathcal{X}$ is uncountable (Tao, 2011, Exercise 2.4.1.5-6).

Lemma 3 (Closed Mapping Lemma (Conrad, Theorem 2.6)) *Suppose X is compact and Y is Hausdorff. Let $\phi : X \to Y$ continuous, then ϕ is a closed map.*

Lemma 4 *Suppose X can be written as a countable union of compact subspaces and Y is Hausdorff. Let $\phi : X \to Y$ continuous. Let $A \subset X$ be a closed set, the $\phi(A)$ is a countable union of closed sets.*

Proof As X is countable union of compact spaces, we can write $X = \bigcup_{i\in\mathbb{N}} X_i$, with X_i compact. Define $A_i = A \cap X_i$, and note that A_i is closed in the subspace topology of X_i. Define $\phi_i : X_i \to Y$ to be the restriction of ϕ to X_i; it follows from the definition of the subspace topology ϕ_i is continuous. Then,

$$\phi(A) = \bigcup_{i\in\mathbb{N}} \phi_i(A_i).$$

By the closed mapping lemma $\phi_i(A_i)$ is closed for all i. ■

Appendix B. f-divergences

In this appendix, we briefly recall the definition of an f-divergence, as well as the necessary results to generalize our claims from the Kullback-Leibler divergence to other f-divergences. We use the definition from Polyanskiy and Wu (2014),

Definition 2 *Given a measurable space (Ω, Σ) and a convex function $f : [0, \infty) \to \mathbb{R}$ satisfying $f(1) = 0$ which is strictly convex at 1. For any two probability measures P, Q on Σ,*

$$\mathrm{D}_f(Q, P) := \int_{\{z:p(z)>0\}} f\left(\frac{q(z)}{p(z)}\right) p(z)d\mu(z) + f'(\infty)Q(\{z : p(z) = 0\})$$

with $p(z) = \frac{dP}{d\mu}(z)$ and $q(z) = \frac{dQ}{d\mu}(z)$, μ is an arbitrary dominating measure (e.g. $(P+Q)/2$), $f'(\infty) = \lim_{z\to 0^+} zf(1/z)$ and the understanding that if $Q(\{z : p(z) = 0\})$ the second term is 0 (even if $f'(\infty) = \infty$).

Examples of f-divergences include KL divergence, total variation distance, squared Hellinger distance and α-divergence. The data processing inequality holds for general f-divergences:

Proposition 5 (Data processing inequality Polyanskiy and Wu (2014, Thm 6.2)) *Let (A, Σ_A) and (B, Σ_B) measurable spaces and $g : A \to B$ a (Σ_A, Σ_B)-measurable function, then for any f-divergence D_f,*

$$\mathrm{D}_f\left(g_* P, g_* Q\right) \leq \mathrm{D}_f\left(P, Q\right),$$

where $g_* P$ indicates the pushforward measure of P by g.

From this the analogue of proposition 3 holds for general f-divergences.

Proposition 6 *Suppose the approximate posterior is parameterized by (Q_θ, g) and the prior is parameterized by (P_θ, g). Define $Q = g_* Q_\theta$ and $P = g_* P_\theta$ to be the approximate posterior predictive and the prior predictive respectively. Then for any f-divergence D_f,*

$$\mathrm{D}_f(Q, P) \leq \mathrm{D}_f(Q_\theta, P_\theta). \tag{7}$$

Moreover, if g is injective (each set of parameters corresponds to a unique predictive function) then equality holds in eq. (4).

Proof For the inequality, using proposition 5,

$$\mathrm{D}_f(Q, P) = \mathrm{D}_f(g_* Q_\Theta, g_* P_\Theta) \leq \mathrm{D}_f(Q_\Theta, P_\Theta).$$

Suppose g is injective, then there exists a $g' : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^k$ such that $g' \circ g(\theta) = \theta$ for all $\theta \in \mathbb{R}^k$ (g has a left inverse). Then,

$$\begin{aligned} \mathrm{D}_f(Q, P) &= \mathrm{D}_f(g_* Q_\Theta, g_* P_\Theta) \\ &\geq \mathrm{D}_f(g'_*(g_* Q_\Theta), g'_*(g_* P_\Theta)) \\ &= \mathrm{D}_f((g' \circ g)_* Q_\Theta), (g' \circ g)_* P_\Theta)) \\ &= \mathrm{D}_f(Q_\Theta, P_\Theta)). \end{aligned}$$

■

If $P \perp Q$, then $Q(p = 0) = 1$, so that $\mathrm{D}_f(Q, P) = f'(\infty)$. Note that this value is the same for all $P \perp Q$ (and by the convexity of f is the maximum value that can be obtained by the f-divergence), so that if all $Q \in \mathcal{Q}$ are mutually singular to P, then any f-divergence is entirely independent of which $Q \in \mathcal{Q}$ is selected.

This leads to the generalizations of propositions 4 and 9 for other f-divergences:

Proposition 7 *Suppose the approximate posterior is parameterized by (Q_{Θ_1}, g_1) and the prior is parameterized by (P_{Θ_2}, g_2), where both Q_θ and P_θ have densities (with respect to Lebesgue measure). Further suppose that $g_1(\Theta_1)$ is the mapping defined by a 1HL BNN with ReLU activation functions, k neurons and parameters Θ_1 and that $g_2(\Theta_2)$ is defined similarly, but with $j \neq k$ neurons. Let $Q = g_{1*} Q_{\Theta_1}$ and $P = g_{2*} P_{\Theta_2}$ denote the approximate posterior predictive and the prior predictive respectively, then*

$$\mathrm{D}_f(Q, P) = f'(\infty).$$

Proposition 8 *Let (Q_Θ, g) parameterize the approximate posterio and P be a (non-degenerate) Gaussian process. Assume that $g(\cdot)(x)$ is locally Lipschitz for all x. Let $Q = g_* Q_\theta$. Then,*

$$\mathrm{D}_f(Q,P) = f'(\infty) \quad \textit{and} \quad \mathrm{D}_f(P,Q) = f'(\infty). \tag{8}$$

By choosing $f(z) = z\log(z)$, we obtain the KL divergence used above. In this case $f'(\infty) = \lim_{z\to 0^+} \log(1/z) = \infty$.

Appendix C. ReLU BNNs are mutually singular

Proposition 9 *Suppose the approximate posterior is parameterized by (Q_{Θ_1}, g_1) and the prior is parameterized by (P_{Θ_2}, g_2), where both Q_{Θ_1} and P_{Θ_2} have densities (with respect to the appropriate Lebesgue measures). Further suppose that $g_1(\Theta_1)$ is the mapping defined by a 1HL BNN with ReLU activation functions, k neurons and parameters Θ_1 and $g_2(\Theta_2)$ is defined similarly, but with $j \neq k$ neurons. Let $Q = g_{1*}Q_{\Theta_1}$ and $P = g_{2*}P_{\Theta_2}$ denote the approximate posterior predictive and the prior predictive respectively, then $\mathrm{D}_{KL}(Q,P) = \infty$.*

C.1. Preliminary Definitions and Lemmas

We will construct a measurable event which one neural network assigns probability 1 to and the other probability 0. This event will roughly be functions $f \in \mathbb{R}^{\mathcal{X}}$ such that $f(x, 0, \ldots, 0)$ is continuous with $k+1$ linear pieces. There are two steps: constructing an event that is measurable and captures this behavior; and showing that a ReLU network with a distribution over parameters with Lebesgue density and k neurons assigns probability 1 to this event.

Definition 3 *Let $f \in \mathbb{R}^{\mathbb{R}}$, we say f* is continuous with k-linear pieces *if there exists an $x_1 \leq x_2 \leq \ldots \leq x_{k-1}$ and $b, a_1, \ldots, a_k$ such that*

$$f(x) = a_{i+1}x + b_i \quad \text{for } x \in [x_i, x_{i+1}) \tag{9}$$

with the understanding that $x_0 = -\infty, x_k = \infty$, $b_0 = b$, and where b_i is selected so that the resulting function is continuous for $0 < i \leq k$.

We say a function $\tilde{f} \in \mathbb{R}^{\mathbb{Q}}$ is *continuous with k-linear pieces* if it can be extended to a function $f \in \mathbb{R}^{\mathbb{R}}$ that is continuous with k-linear pieces.

Proposition 10 *Define $E_k := \{f \in \mathbb{R}^{\mathbb{R}} : \pi_{\mathbb{Q}}(f)$ is continuous with k-linear pieces$\}$. Then E_k is measurable in the product σ-algebra on $\mathbb{R}^{\mathbb{R}}$ induced by the Borel σ-algebra.*

Proof By the measurability of $\pi_{\mathbb{Q}}$, it suffices to show that

$$\tilde{E}_k := \{\tilde{f} \in \mathbb{R}^{\mathbb{Q}} : \tilde{f} \text{ is continuous with } k\text{-linear pieces}\}$$

is measurable in the product σ-algebra on $\mathbb{R}^{\mathbb{Q}}$.

We will use the previous lemma to show that $\tilde{E}_k$ is a countable union of closed sets and is hence measurable in the Borel σ-algebra induced by the product topology on $\mathbb{R}^{\mathbb{Q}}$. As this coincides with the product σ-algebra for countable products of $\mathbb{R}$ (Tao, 2011, Exercise 2.4.1.5), this suffices.

We will do this by showing that $\tilde{E}_k$ is the image of $\mathbb{R}^{2k+1}$ under a continuous function, and applying lemma 4. As $\mathbb{R}^{2k+1}$ is closed, $\mathbb{R}^{\mathbb{Q}}$ is a product of Hausdorff spaces, hence Hausdorff and $\mathbb{R}^{2k+1} = \bigcup_{i\in\mathbb{N}}[-i,i]^{2k+1}$ (i.e it is a countable union of compact set), all that remains is to construct a continuous $\phi : \mathbb{R}^{2k+1} \to \mathbb{R}^{\mathbb{Q}}$ such that $\phi(\mathbb{R}^{2k+1}) = \tilde{E}_k$.

Let

$$\phi(\tilde{s}_0, \tilde{s}_1, \dots, \tilde{s}_{2k}) = \tilde{f}^{(\tilde{s}_0, \tilde{s}_1, \dots, \tilde{s}_{2k})} \tag{10}$$

with $\tilde{f}$ defined as in definition 3 (restricted to $\mathbb{Q}$) with $b = \tilde{s}_0, a_i = \tilde{s}_i$ and $x_1 = \tilde{s}_{k+1}$ and $x_{i+1} = x_i + |\tilde{s}_{i+k+1}|$. From this definition, it is clear that $\phi(\mathbb{R}^{2k+1}) \subset \tilde{E}_k$. The reverse inclusion follows from the noting that any function of the form in definition 3 can be written in this form. It remains to show ϕ is continuous. By the universal property of the product topology (Munkres, 2000, Theorem 19.6), we need only show $\phi_{\{q\}}$ is continuous for all $q \in \mathbb{Q}$, which can be shown from the metric space definition of continuity (with some care for cases when q is on the boundary of two linear regions). ■

C.2. Proof in the case when input space is one-dimensional

We first prove proposition 9 under the assumption that $\mathcal{X} = \mathbb{R}$; the generalization to multidimensional inputs is straightforward.

Proof All that remains to show that the implied measures for two 1HL ReLU BNNs mapping from $\mathbb{R} \to \mathbb{R}$ are orthogonal is showing that if a 1HL ReLU BNN has k neurons, then it produces a function in $\tilde{E}_{k+1} \setminus \tilde{E}_k$ with probability 1.

We first show that with probability 1, the implied function is in E_{k+1} (in fact, this holds surely). Let $w^1, w^2, b^1 \in \mathbb{R}^k$ and $b^2 \in \mathbb{R}$ be an arbitrary realization of weights and biases, then $f(x) = b^{(2)} + \langle w^{(2)}, \max(0, w^{(1)} \circ x + b^{(1)})\rangle = b^{(2)}$, where $\circ$ denotes an element-wise vector product. We can rewrite this as

$$f(x) = b^{(2)} + \sum_{\substack{i=1 \\ w_i^{(1)}x + b_i^{(1)} > 0}}^{k} w_i^{(1)} w_i^{(2)} x + w_2 b_1.$$

Note that this is piecewise linear, with boundaries at $\tilde{x}_i = -\frac{b_i}{w_i}$ for $i \leq k$, and is continuous in x as it can be written as a composition of continuous functions. As this holds for arbitrary realizations of parameters, f is surely in E_{k+1}.

On the other hand, for f to be in E_k it must be the case that either:

- $\tilde{x}_i = \tilde{x}_j$ for some $i \neq j$ (i.e. boundaries coincide).
- There exists an i such that $w_i^{(1)} w_i^{(2)} = 0$ (adjacent regions have the same slope).

The above conditions define a Lebesgue null set; hence under the assumption that the distribution over parameters has density with respect to Lebesgue measure, this is a probability 0 event. ■

C.3. Extension to multidimensional inputs

The extension to multidimensional inputs is almost immediate up considering the set $E'_k := \{f \in \mathbb{R}^{\mathbb{R}^d} : f(\cdot, 0, 0, \ldots, 0) \text{ is continuous with k linear pieces}\}$. Defining $\tilde{E}'_k := \{f \in \mathbb{R}^{\mathbb{Q}} \times \mathbb{R}^{\mathbb{R}^{d-1}} : f(\cdot, 0, 0, \ldots, 0) \text{ is continuous with } k \text{ linear pieces on rational } x\}$, we see that $\tilde{E}'_k$ is measurable for the same reason $\tilde{E}_k$ is measurable. Moreover, viewed along this slice of input space, the neural network is identical to a 1HL network mapping from $\mathbb{R} \to \mathbb{R}$, so the proof in the previous subsection holds without modification.

Appendix D. Parametric models and non-degenerate Gaussian measures are mutually singular

D.1. Preliminaries

The main ingredient in the result is the following lemma:

Lemma 5 (Rudin (1966, Lemma 7.25)) *Let $E \subset \mathbb{R}^k$ a Lebesgue null set. Suppose $f : \mathbb{R}^k \to \mathbb{R}^k$ satisfies, for all $x \in E$ there exists a $\delta > 0$ and $M > 0$ such that*

$$\frac{\|f(x) - f(y)\|}{\|x - y\|} \leq M$$

for all $y \in E \cap B(x; \delta)$ where $B(x; \delta)$ denotes the ball of radius δ centered at x. Then $\lambda_k(f(E)) = 0$.

Note that the condition on f holds if f is locally Lipschitz. Moreover, if a function from $\mathbb{R}^k \to \mathbb{R}^k$ is locally Lipschitz for every output index, it is locally Lipschitz as a function from $\mathbb{R}^k \to \mathbb{R}^k$.

D.2. Statement and Proof

Proposition 4 *Let (Q_θ, g) parameterize the approximate posterior of a parametric model and let P be a non-degenerate Gaussian process. Assume that $g(\cdot)(x)$ is locally Lipschitz for all x. Let $Q = g_* Q_\theta$. Then, $\mathrm{D}_{KL}(Q, P) = \infty$ and $\mathrm{D}_{KL}(P, Q) = \infty$.*

Proof Fix a set $A \subset \mathcal{X}, |A| = k+1$ so that $\pi_{A*} P \sim \lambda_A$. Such an A exists by the assumption that P is non-degenerate and $\mathcal{X}$ is infinite. By lemma 1 and lemma 2 it then suffices to show $\pi_A Q \perp \lambda_A$.

Define the event $E = (\pi_A \circ g)(\mathbb{R}^k)$. As $\pi_A \circ g$ is continuous, by the closed mapping lemma, E is a countable union of closed sets (lemma 4), hence Borel measurable. Also,

$$\pi_{A_*} Q(E) = Q_\theta(g^{-1}(\pi_A^{-1}(\pi_A(g(\mathbb{R}^k))))) \geq Q_\theta(g^{-1}(g(\mathbb{R}^k))) \geq Q_\theta(\mathbb{R}^k) = 1.$$

The inequalities follow from the pre-image of the image of a set containing the original set.

All that is left to show is that E is a Lebesgue null set. Let $\tilde{g}_A : \mathbb{R}^{k+1} \to \mathbb{R}^A$ be defined by $\tilde{g}_A(x_1, \ldots, x_{k+1}) = (\pi_A \circ g)(x_1, \ldots, x_k)$. Then $\tilde{g}_A(\mathbb{R}^k \times \{0\}) = E$. As $\pi_A \circ g$ is locally Lipschitz, so is $\tilde{g}_A$. The proof is then completed by lemma 5, noting that $\mathbb{R}^k \times \{0\}$ is a Lebesgue null set in $\mathbb{R}^{k+1}$. ■

Appendix E. Bayesian Linear Regression and Variational Inference

Consider the Bayesian linear regression model eq. (5) and the modified variational objective,

$$\max_{Q\in\mathcal{Q}} \mathbb{E}_{\Theta\sim Q}[\log p(y|x,\Theta)] - \mathrm{D}_{KL}(Q_A, P_A),$$

which is a special case of eq. (6) when μ is a point mass on A (i.e. FixedA). Let ϕ_X be the $n \times k$ feature matrix with $[\phi_X]_{ij} = \phi_j(x_i)$. Let $A = \{a_1, \cdots, a_m\}$ and ϕ_A be the $m \times k$ feature matrix with $[\phi_A]_{ij} = \phi_j(a_i)$. Without loss of generality, we assume that ϕ_A has linearly independent rows, so that Q_A, P_A are non-degenerate. Then,

$$\log p(y|x,\Theta) = -\frac{n}{2}\log(2\pi\sigma^2)^{-n/2} - \frac{1}{2\sigma^2}(y - \phi_X\Theta)^\intercal(y - \phi_X W)$$

For $Q = \mathcal{N}(\mu_Q, \Sigma_Q)$, we have

$$\begin{aligned}
\mathbb{E}_{W\sim Q}[\log p(y|W)] &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathbb{E}_{W\sim Q}[(y-\phi_X W)^\intercal(y-\phi_X W)] \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(y^\intercal y - 2y^\intercal\phi_X\mathbb{E}[W] + \mathbb{E}[W^\intercal\phi_X^\intercal\phi_X W]\right) \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(y^\intercal y - 2y^\intercal\phi_X\mu_Q + \mathbb{E}[W^\intercal\phi_X^\intercal\phi_X W]\right) \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(y^\intercal y - 2y^\intercal\phi_X\mu_Q + \mathrm{tr}(\phi_X^\intercal\phi_X\Sigma_Q) + \mu_Q^\intercal\phi_X^\intercal\phi_X\mu_Q\right) \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left((y-\phi_X\mu_Q)^\intercal(y-\phi_X\mu_Q) + \mathrm{tr}(\phi_X^\intercal\phi_X\Sigma_Q)\right).
\end{aligned}$$

The gradient of this term with respect to both μ_Q is,

$$\nabla_{\mu_Q}\mathbb{E}_{W\sim Q}[\log p(y|W)] = \frac{1}{\sigma^2}\phi_X^\intercal(y - \phi_X\mu_Q).$$

We now turn out attention to the KL divergence between Q_A and P_A. We assume $P = \mathcal{N}(0, I)$. We then have $Q_A = \mathcal{N}(\phi_A\mu_Q, \phi_A\Sigma_Q\phi_A^\intercal)$ and $P_A = \mathcal{N}(0, \phi_A\phi_A^\intercal)$. The KL divergence is then,

$$\begin{aligned}
\mathrm{D}_{KL}(Q_A, P_A) = \frac{1}{2}\Bigg(&-m + \mu_Q^\intercal\phi_A^\intercal(\phi_A\phi_A^\intercal)^{-1}\phi_A\mu_Q + \mathrm{tr}((\phi_A\phi_A^\intercal)^{-1}\phi_A\Sigma_Q\phi_A^\intercal) \\
&- \log\left|(\phi_A\phi_A^\intercal)^{-1}\phi_A\Sigma_Q\phi_A^\intercal\right|\Bigg),
\end{aligned}$$

and its gradient with respect to μ_Q is,

$$\nabla_{\mu_Q}\mathrm{D}_{KL}(Q_A, P_A) = \phi_A^\intercal(\phi_A\phi_A^\intercal)^{-1}\phi_A\mu_Q.$$

Note that $\phi_A^\intercal(\phi_A\phi_A^\intercal)^{-1}\phi_A$ is the projection of μ_Q onto the column space of ϕ_A.

We can find the optimal μ_Q solutions to our optimization problem by setting gradient equal to 0. This yields,

$$\frac{1}{\sigma^2}\phi_X^\intercal y - \frac{1}{\sigma^2}\phi_X^\intercal\phi_X\mu_Q - \phi_A^\intercal(\phi_A\phi_A^\intercal)^{-1}\phi_A\mu_Q = 0.$$

Rearranging,

$$\left(\frac{1}{\sigma^2}\phi_X^\intercal\phi_X + \phi_A^\intercal(\phi_A\phi_A^\intercal)^{-1}\phi_A\right)\mu_Q = \frac{1}{\sigma^2}\phi_X^\intercal y.$$

A solution for μ_Q (though this solution is not in general unique) is given by,

$$\mu_Q = (\phi_X^\intercal\phi_X + \sigma^2\phi_A^\intercal(\phi_A\phi_A^\intercal)^{-1}\phi_A)^\dagger\phi_X^\intercal y.$$

where we use † to denote the pseudo-inverse.

The maximum likelihood solution could be found by removing the term $\phi_A^\intercal(\phi_A\phi_A^\intercal)^{-1}\phi_A$, while map inference is recovered when $\phi_A^\intercal(\phi_A\phi_A^\intercal)^{-1}\phi_A = I$. As $\phi_A^\intercal(\phi_A\phi_A^\intercal)^{-1}\phi_A$ is a projection matrix, we see this has the intuitive interpretation of regularizing the MLE solution, but only in the directions that effect predictions on the points in A.

Appendix F. Experimental Details

In this section, we present details on the experiments we performed.

F.1. Toy experiment

We generate the data using 20 linearly-spaced radial basis function features between -2 and 2, with each feature having lengthscale 0.2. The input data are sampled by taking 20 samples from $\mathcal{N}(-1.2, 0.3^2)$ and 20 from $\mathcal{N}(1.2, 0.3^2)$. We sample the weights from a standard normal prior and generate the output data as described earlier.

For each method, we optimize the ELBO using Adam (Kingma and Ba, 2015) using 5000 gradient steps with a learning rate of 0.01. We evaluate the likelihood term using the full dataset every iteration without minibatching. For the FixedA, RandA, and SSGE methods we use 10 points for A, with 5 from selected from the dataset, and the other 5 sampled uniformly from a box with bounds determined by the bounds of the training data.

F.2. UCI experiments

We begin by training sparse Gaussian process regression (SGPR) (Titsias, 2009) for each split of the dataset we use, using 100 inducing points initialized by sampling from the dataset. We use the GPflow (Matthews et al., 2017) implementation of SGPR, initializing the ARD lengthscales to 1 and the noise variance to 1e-4. We use the built-in Scipy optimizer with a maximum of 5000 iterations. After the SGPR model is trained, we use the 100 inducing points to obtain 100 features to use for Bayesian linear regression (Wahba, 1990), with a standard normal prior over the weights. We then train each of the methods for estimating the KL term, keeping the learned noise variance from the SGPR models fixed.

For Exact, we simply use the closed-form solutions to the inference problem, which exist for both Full and FFG. For the algorithms that make use of eq. (6), we set $|A| = 80$ and choose half of its points from the training set (note this differs slightly from Sun et al. (2018), who always include the current mini-batch used to evaluate the expected log likelihood term in A), and the other half uniformly randomly over the bounding box of the training set. For FixedA, we optimize using L-BFGS (Nocedal, 1980), whereas for RandA and SSGE we optimize using Adam. In each case we optimize for (up to) 15000 steps. We note that we add jitter as necessary to ensure that the Cholesky decompositions we perform succeed. Finally, we average each run over 20 seeds.

Extended results are displayed in figures 3 and 4.

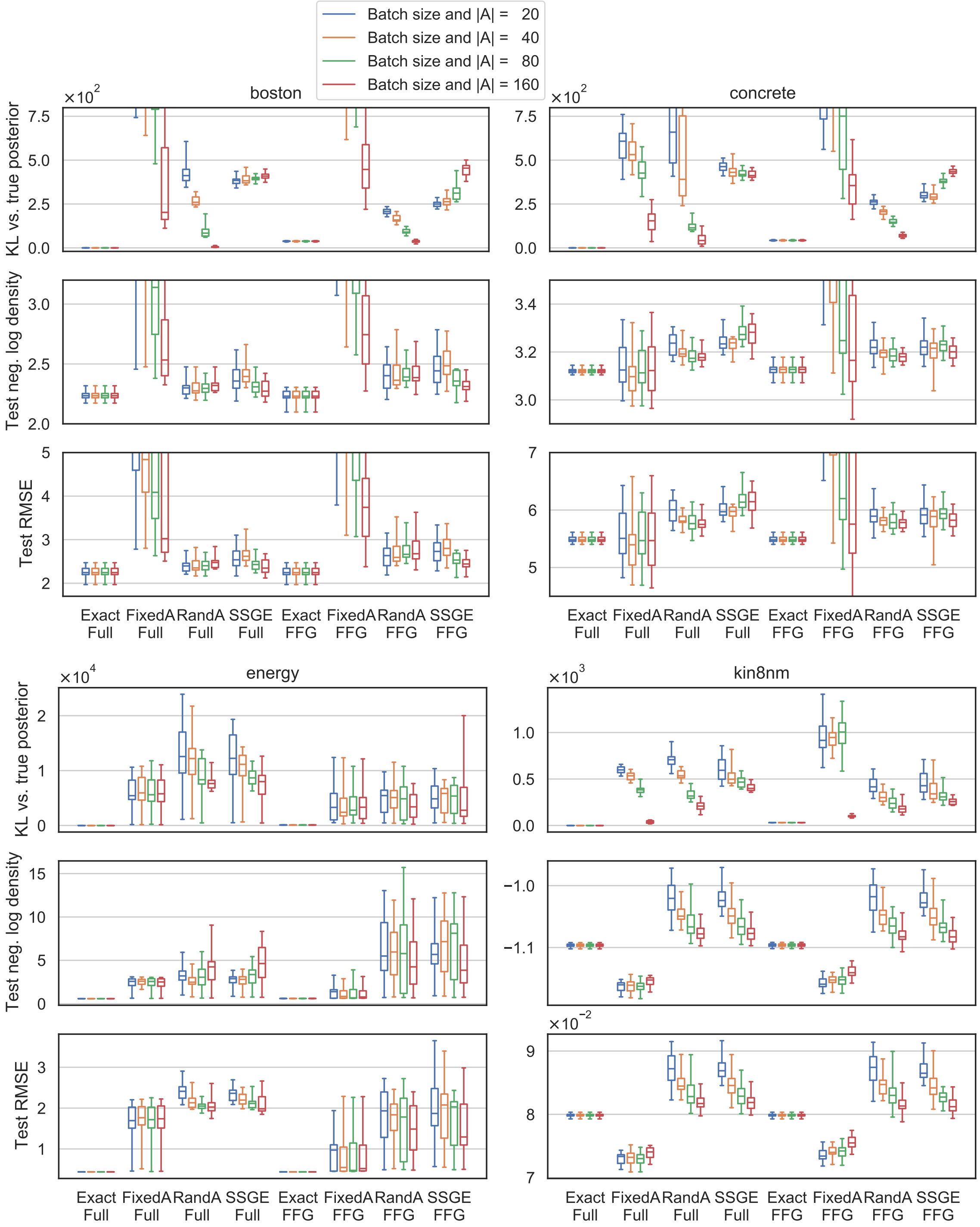

Figure 3: Results for datasets boston, concrete, energy and kin8nm. The different colors represent different sizes of $|A|$ and the minibatch. From left to right: 20, 40, 80, 160.

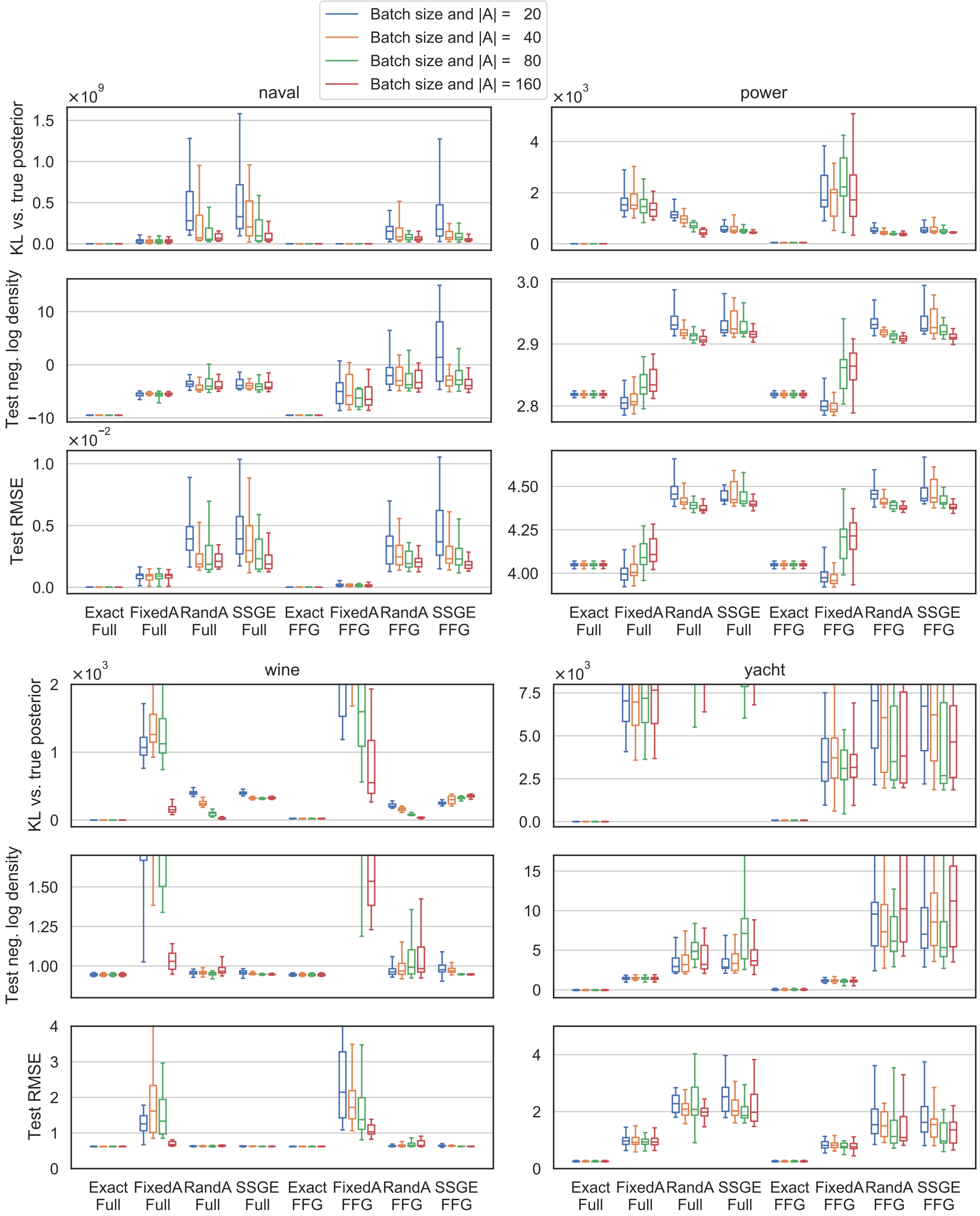

Figure 4: Results for datasets naval, power, wine and yacht. The different colors represent different sizes of $|A|$ and the minibatch. From left to right: 20, 40, 80, 160.