Well-Defined Function-Space Variational Inference in Bayesian Neural Networks via Regularized KL-Divergence

Tristan Cinquin[1] **Robert Bamler**[1]

[1]University of Tübingen, Germany

Abstract

Bayesian neural networks (BNN) promise to combine the predictive performance of neural networks with principled uncertainty modeling crucial for safety-critical systems and decision making. However, posterior uncertainties depend on the choice of prior, and finding informative priors in weight-space has proven difficult. This has motivated variational inference (VI) methods that pose priors directly on the function represented by the BNN rather than on weights. In this paper, we address a fundamental issue with such function-space VI approaches pointed out by Wild et al. [2022a], who showed that the objective function (ELBO) is negative infinite for most priors of interest. Our solution builds on *generalized* VI with the regularized KL divergence and is, to the best of our knowledge, the first well-defined variational objective for inference in BNNs with Gaussian process (GP) priors. Experiments show that our method successfully incorporates the properties specified by the GP prior, and that it provides competitive uncertainty estimates for regression, classification and out-of-distribution detection compared to BNN baselines with both function and weight-space priors.

1 INTRODUCTION

Neural networks have shown impressive results in many fields but fail to provide well-calibrated uncertainty estimates, which are essential in applications associated with risk, such as healthcare [Abdullah et al., 2022] or finance [Bew et al., 2019]. Bayesian neural networks (BNNs) offer to combine the scalability and predictive performance of neural networks with principled uncertainty modeling by explicitly capturing epistemic uncertainty, which results from finite training data. While the choice of prior strongly affects posterior uncertainties, specifying informative priors on BNN weights has proven difficult and is hypothesized to have limited their practical applicability [Knoblauch et al., 2022, Tran et al., 2022]. For instance, the default isotropic Gaussian prior, which is often chosen for tractability rather than for the beliefs it carries [Knoblauch et al., 2022], is known to have pathological behavior in some cases [Cinquin et al., 2021, Tran et al., 2022]. A promising approach to solve this issue is to place priors directly on the function represented by the BNN instead of the weights. Function-space priors allow incorporating interpretable knowledge, for instance using the Gaussian Process (GP) literature to improve prior design and selection [Williams and Rasmussen, 2006].

A recent line of work has focused on using function-space priors in BNNs with variational inference (VI) [Sun et al., 2019]. VI is appealing because of its successful application to BNNs, its flexibility in terms of approximate posterior parameterization, and its scalability to large datasets and models [Hoffman et al., 2013, Blundell et al., 2015]. Unfortunately, for BNNs with function-space priors, the Kullbach-Leibler (KL) divergence term in the VI objective (ELBO) involves two intractabilities: (i) a supremum over infinitely many subsets and (ii) access to the density of the distribution of the BNN's function, which has no closed-form expression. Sun et al. [2019] propose to address problem (i) by approximating the supremum in the KL divergence by an expectation, and problem (ii) by using implicit score function estimators (which make this method difficult to use in practice [Ma and Hernández-Lobato, 2021]). However, the problem is actually more severe. Not only is the KL divergence intractable, it is infinite in most cases of interest [Wild et al., 2022a], such as when the prior is a non-degenerate GP or a BNN with a different architecture. Thus, in these (and many more) situations, the KL divergence cannot even be approximated. As a consequence, more recent work abandons using BNNs and instead uses deterministic neural networks to parameterize basis functions [Ma and Hernández-Lobato, 2021] or a GP mean [Wild et al., 2022b]. The only prior work [Rudner et al., 2022b] that overcomes

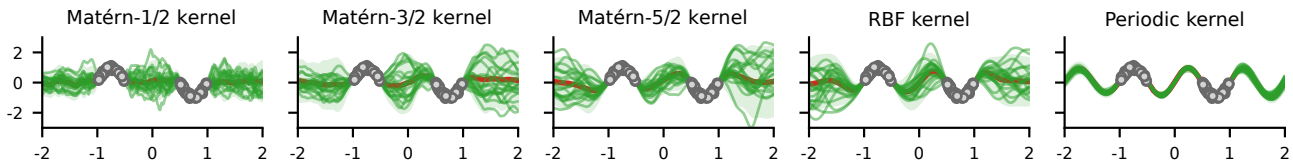


Figure 1: Inference with our GFSVI on synthetic data (gray circles) with Gaussian process priors encoding different properties such as smoothness (increasing from Matérn-1/2 to RBF) and periodicity (last panel).

the issue pointed out by Wild et al. [2022a] does so by deliberately limiting itself to cases where the KL divergence is known to be finite (by defining the prior as the pushforward of a weight-space distribution). Therefore, the method by Rudner et al. [2022b] suffers from the same issues regarding prior specification as other weight-space inference method.

In this paper, we address the argument by Wild et al. [2022a] that VI does not provide a valid objective for inference in BNNs with genuine function-space priors, and we propose to apply the framework of generalized VI [Knoblauch et al., 2022]. We present a simple method for function-space inference with GP priors that builds on the regularized KL divergence [Quang, 2019], which generalizes the conventional KL divergence and is finite for any pair of Gaussian measures. We obtain a Gaussian measure for the variational posterior by considering the linearized BNN from Rudner et al. [2022b], and we are free to choose a function-space prior from a large set of GPs which have an associated Gaussian measure on the considered function space. While the regularized KL divergence is still intractable, it can be consistently estimated from samples with a known error bound. We find that our method effectively incorporates the beliefs specified by GP priors (see Figure 1, discussed further in Section 4) and that it yields competitive performance compared to BNN baselines. To the best of our knowledge, our method is the first to provide a well-defined objective for function-space inference in BNNs with informative GP priors. Our contributions are summarized below:

1. We use generalized VI with the *regularized* KL divergence to mitigate the issue of an infinite KL divergence when using VI in BNNs with function-space priors.
2. We present a new and well-defined objective for function-space inference in the linearized BNN with GP priors, resulting in a simple algorithm.
3. We show that our method accurately captures structural properties specified by the GP prior and provides competitive uncertainty estimates for regression, classification, and out-of-distribution detection compared to baselines with both function- and weight-space priors.

The paper is structured as follows: Section 2 introduces function-space VI and the regularized KL divergence; Section 3 presents our method for generalized function-space VI (GFSVI) in BNNs; Section 4 reports experimental results; Section 5 discusses related work; and Section 6 concludes.

2 BACKGROUND

In this section, we provide background on function-space variational inference in BNNs and discuss the fundamental issue of an infinite KL divergence. We then introduce the regularized KL divergence, which is the basis for our solution presented in Section 3.

2.1 FUNCTION-SPACE VI IN BNNS

We consider a neural network $f(\cdot\,;\boldsymbol{w})$ with weights $\boldsymbol{w} \in \mathbb{R}^p$, and a data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ with features $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ and values $y_i \in \mathcal{Y}$. Bayesian Neural Networks are specified further by a likelihood function $p(\mathcal{D}\,|\,\boldsymbol{w}) = \prod_{i=1}^N p(y_i\,|\,f(\boldsymbol{x}_i;\boldsymbol{w}))$ and—traditionally—a prior $p(\boldsymbol{w})$ on the weights, and one seeks the posterior distribution $p(\boldsymbol{w}\,|\,\mathcal{D}) \propto p(\mathcal{D}\,|\,\boldsymbol{w})\,p(\boldsymbol{w})$. The method proposed in this paper builds on variational inference, which approximates $p(\boldsymbol{w}\,|\,\mathcal{D})$ with a variational distribution $q_\phi(\boldsymbol{w})$, whose variational parameters ϕ maximize the evidence lower bound (ELBO),

$$\mathcal{L}(\phi) := \mathbb{E}_{q_\phi(\boldsymbol{w})}[\log p(\mathcal{D}\,|\,\boldsymbol{w})] - D_{\mathrm{KL}}(q_\phi \,\|\, p) \tag{2.1}$$

where D_{KL} is the Kullback-Leibler divergence,

$$D_{\mathrm{KL}}(q_\phi \,\|\, p) := \mathbb{E}_{q_\phi(\boldsymbol{w})}\big[\log\big(q_\phi(\boldsymbol{w})/p(\boldsymbol{w})\big)\big]. \tag{2.2}$$

At test time, we approximate the predictive distribution for given features $\boldsymbol{x}^*$ as $p(y^*\,|\,\boldsymbol{x}^*) = \mathbb{E}_{p(\boldsymbol{w}\,|\,\mathcal{D})}\big[p(y^*\,|\,f(\boldsymbol{x}^*;\boldsymbol{w}))\big] \approx \mathbb{E}_{q_\phi(\boldsymbol{w})}\big[p(y^*\,|\,f(\boldsymbol{x}^*;\boldsymbol{w}))\big]$.

Function-space variational inference. Since neural network weights are not interpretable, we replace the weight-space prior $p(\boldsymbol{w})$ with a prior $\mathbb{P}$ directly on the function $f(\cdot\,;\boldsymbol{w})$, which we denote simply as f when there is no ambiguity. Here, the symbol $\mathbb{P}$ denotes a probability measure that does not admit a density since the function space is infinite-dimensional. We compute the expected log-likelihood as in the first term of Eq. 2.1. For the KL-term (Eq. 2.2), a naive VI-method would use the pushforward of $q_\phi(\boldsymbol{w})$ along $\boldsymbol{w} \mapsto f(\cdot\,;\boldsymbol{w})$, which defines the variational measure $\mathbb{Q}_\phi$, resulting in the ELBO in function space,

$$\mathcal{L}(\phi) := \mathbb{E}_{q_\phi(\boldsymbol{w})}[\log p(\mathcal{D}\,|\,\boldsymbol{w})] - D_{\mathrm{KL}}(\mathbb{Q}_\phi \,\|\, \mathbb{P}) \tag{2.3}$$

with D_{KL} the KL divergence between measures

$$D_{\mathrm{KL}}(\mathbb{Q}_\phi \,\|\, \mathbb{P}) := \int \log\left(\frac{d\mathbb{Q}_\phi}{d\mathbb{P}}(f)\right) d\mathbb{Q}_\phi. \tag{2.4}$$

Here, the Raydon-Nikodym derivative $d\mathbb{Q}_\phi/d\mathbb{P}$ generalizes the density ratio from Eq. 2.2. Like Eq. 2.1, the ELBO in Eq. 2.3 is a lower bound on the evidence [Wild et al., 2022a]. In fact, if $\mathbb{P}$ is the push-forward of $p(\boldsymbol{w})$ then Eq. 2.3 is a tighter bound than Eq. 2.1 by the data processing inequality, $D_{\mathrm{KL}}(\mathbb{Q}_\phi \,\|\, \mathbb{P}) \leq D_{\mathrm{KL}}(q_\phi \,\|\, p)$. However, we motivated function-space VI to avoid weight-space priors, and in this case the bound in Eq. 2.3 can be looser. We will indeed see below that the bound becomes infinitely loose in practice, and we thus propose a different objective in Section 3.

Two intractabilities prevent directly maximizing the ELBO in function space (Eq 2.3). First, it is not obvious how to evaluate or estimate the KL divergence between two measures in Eq 2.4. Sun et al. [2019] showed that it can be expressed as a supremum of KL divergences between finite-dimensional distributions,

$$D_{\mathrm{KL}}(\mathbb{Q}_\phi \,\|\, \mathbb{P}) = \sup_{\boldsymbol{x}\in\mathcal{X}^M, M\in\mathbb{N}} D_{\mathrm{KL}}(q_\phi(f(\boldsymbol{x})) \,\|\, p(f(\boldsymbol{x}))). \tag{2.5}$$

Here, $\boldsymbol{x} = \{\boldsymbol{x}^{(i)}\}_{i=1}^M \in \mathcal{X}^M$ is a set of M points in feature space $\mathcal{X}$, and $q_\phi(f(\boldsymbol{x}))$ and $p(f(\boldsymbol{x}))$ are densities of the marginals of $\mathbb{Q}_\phi$ and $\mathbb{P}$ on $\{f(\boldsymbol{x}^{(i)})\}_{i=1}^M$ respectively. Sun et al. [2019] approximates the supremum over infinitely many sets by an expectation, and Rudner et al. [2022b] estimates it from samples.

Second, we cannot express the pushforward measure $\mathbb{Q}_\phi$ in closed form because the neural network is nonlinear. Previous work has proposed to mitigate this issue using implicit score function estimators [Sun et al., 2019] or a linearized BNN f_L to obtain a closed-form Gaussian variational measure [Rudner et al., 2022a,b]. Our proposal in Section 3 follows the linearized BNN approach as it only minimally modifies the BNN, preserving most of its inductive bias [Maddox et al., 2021] while considerably simplifying the problem by turning the pushforward of $q_\phi(\boldsymbol{w})$ into a GP. More specifically, we consider a Gaussian variational distribution $q_\phi(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{m}, \boldsymbol{S})$ with parameters $\phi = \{\boldsymbol{m}, \boldsymbol{S}\}$, and we define a linearized BNN f_L by linearizing f as a function of the weights around $\boldsymbol{w} = \boldsymbol{m}$,

$$f_L(\boldsymbol{x}; \boldsymbol{w}) := f(\boldsymbol{x}; \boldsymbol{m}) + J(\boldsymbol{x}; \boldsymbol{m})(\boldsymbol{w} - \boldsymbol{m}) \tag{2.6}$$

with $J(\boldsymbol{x}; \boldsymbol{m}) = \nabla_{\boldsymbol{w}} f(\boldsymbol{x}; \boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{m}}$. Thus, $\boldsymbol{w} \sim q_\phi(\boldsymbol{w})$ implies $f_L(\boldsymbol{x}; \boldsymbol{w}) \sim \mathcal{N}\big(f(\boldsymbol{x}; \boldsymbol{m}), J(\boldsymbol{x}; \boldsymbol{m})\boldsymbol{S}J(\boldsymbol{x}; \boldsymbol{m})^\top\big)$ for all $\boldsymbol{x} \in \mathcal{X}$, and so the function $f_L(\,\cdot\,; \boldsymbol{w})$ is a degenerate GP (as $\mathrm{rank}(J(\,\cdot\,; \boldsymbol{m})\boldsymbol{S}J(\,\cdot\,; \boldsymbol{m})^\top) \leq p < \infty$),

$$f_L \sim \mathcal{GP}\big(f(\,\cdot\,; \boldsymbol{m}), J(\,\cdot\,; \boldsymbol{m})\boldsymbol{S}J(\,\cdot\,; \boldsymbol{m})^\top\big). \tag{2.7}$$

$D_{\mathrm{KL}}(\mathbb{Q}_\phi \,\|\, \mathbb{P})$ is infinite in most relevant cases. Wild et al. [2022a] point out an even more severe issue of function-space VI in BNNs: $D_{\mathrm{KL}}(\mathbb{Q}_\phi \,\|\, \mathbb{P})$ (Eq. 2.4) is in fact infinite in most relevant cases, in particular for non-degenerate GP-priors. Thus, approximating $D_{\mathrm{KL}}(\mathbb{Q}_\phi \,\|\, \mathbb{P})$ in these settings is futile. Their proof is somewhat involved, but the fundamental reason for $D_{\mathrm{KL}}(\mathbb{Q}_\phi \,\|\, \mathbb{P}) = \infty$ is that $\mathbb{Q}_\phi$ has support on a finite-dimensional submanifold of the infinite-dimensional function space, while the measure $\mathbb{P}$ induced by a (non-degenerate) GP prior has support on the entire function space. That such a dimensionality mismatch can lead to infinite KL divergence can already be seen in a finite-dimensional example: consider the KL-divergence between two Gaussians in $\mathbb{R}^n$ for $n \geq 2$, one of which has support on the entire $\mathbb{R}^n$ (i.e., its covariance matrix $\boldsymbol{\Sigma}_1$ has full rank) while the other one has support only on a proper subspace of $\mathbb{R}^n$ (i.e., its covariance matrix $\boldsymbol{\Sigma}_2$ is singular). The KL divergence between multivariate Gaussians has a closed form expression (Eq. 2.9 with $\gamma = 0$) that contains $\log\det\big(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\big)$, which is infinite for singular $\boldsymbol{\Sigma}_2$.

We find that the fact that $D_{\mathrm{KL}}(\mathbb{Q}_\phi \,\|\, \mathbb{P}) = \infty$ has severe practical consequences even when the KL divergence is only estimated from finite samples. It naturally explains the stability issues discussed in Appendix D.1 of Sun et al. [2019] (we compare the authors' solution to this stability issue to our method in Section 3.2). Surprisingly, similar complications arise even in the setup by Rudner et al. [2022b], which performs VI in function space with the pushforward of a weight-space prior. While this makes the KL divergence technically finite because prior and variational posterior have the same support, numerical errors lead to mismatching supports and thus to stability issues even there.

In summary, the ELBO for VI in BNNs is not well-defined for most interesting function-space priors. In Section 3, we propose a solution by using the so-called regularized KL divergence, which we introduce next.

2.2 REGULARIZED KL DIVERGENCE

Our solution to the negative infinite function-space ELBO builds on a regularized KL divergence, which is expressed in terms of Gaussian measures for the variational posterior and prior. We obtain these Gaussian measures from GPs. We first discuss under which conditions a GP induces a Gaussian measure, and then present the regularized KL divergence.

Gaussian measures and Gaussian processes. The regularized KL divergence is defined in terms of Gaussian measures, and thus we need to verify that the GP variational posterior induced by the linearized BNN (Eq. 2.7) has an associated Gaussian measure. We consider the Hilbert space $L^2(\mathcal{X}, \rho)$ of square-integrable functions with respect to a probability measure ρ on a compact set $\mathcal{X} \subset \mathbb{R}^d$, with inner product $\langle f, g\rangle = \int_{\mathcal{X}} f(x)g(x)d\rho(x)$. This assumption is not restrictive since we can typically bound the region in feature space that contains the data and any points where we might want to evaluate the BNN.

Definition 2.1 (Gaussian measure, Kerrigan et al. [2023], Definition 1)**.** Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space. A measurable function $F : \Omega \mapsto L^2(\mathcal{X}, \rho)$ is called a Gaussian random element (GRE) if for any $g \in L^2(\mathcal{X}, \rho)$ the random

variable $\langle g, F\rangle$ has a Gaussian distribution on $\mathbb{R}$. For every GRE F, there exists a unique mean element $m \in L^2(\mathcal{X}, \rho)$ and a finite trace linear covariance operator $C : L^2(\mathcal{X}, \rho) \mapsto L^2(\mathcal{X}, \rho)$ such that $\langle g, F\rangle \sim \mathcal{N}(\langle g, m\rangle, \langle Cg, g\rangle)$ for all $g \in L^2(\mathcal{X}, \rho)$. The pushforward of $\mathbb{P}$ along F, denoted $\mathbb{P}^F := F_{\#}\mathbb{P}$, is a Gaussian measure on $L^2(\mathcal{X}, \rho)$.

Gaussian measures generalize Gaussian distributions to infinite-dimensional function spaces where measures do not have associated densities since there is no Lebesgue measure. Following Wild et al. [2022b], we notate the Gaussian measure obtained from the GRE F with mean element m and covariance operator C as $\mathbb{P}^F := \mathcal{N}(m, C)$. GPs provide a practical tool to specify Gaussian measures via mean and covariance functions [Kerrigan et al., 2023]. A GP $f \sim \mathcal{GP}(\mu, K)$ has an associated Gaussian measures in $L^2(\mathcal{X}, \rho)$ if its mean function satisfies $\mu \in L^2(\mathcal{X}, \rho)$ and its covariance function K is trace-class, i.e., if $\int_{\mathcal{X}} K(x, x) d\rho(x) < \infty$ [Wild et al., 2022b, Theorem 1]. The GP variational posterior induced by the linearized BNN satisfies both properties as neural networks are well-behaved functions on the compact $\mathcal{X} \subset \mathbb{R}^d$. It thus induces a Gaussian measure $\mathbb{Q}_\phi^F \sim \mathcal{N}(m_Q, C_Q)$ with mean element $m_Q = f(\,\cdot\,; \boldsymbol{m})$ and covariance operator $C_Q g(\cdot) = \int_{\mathcal{X}} J(\,\cdot\,; \boldsymbol{m}) \boldsymbol{S} J(\boldsymbol{x}', \boldsymbol{m})^\top g(\boldsymbol{x}') d\rho(\boldsymbol{x}')$. The infinite KL divergence discussed in Section 2.1 is easier to prove for the special case of Gaussian measures, and we provide the proof in Appendix A.1.

Definition 2.2 (Regularized KL divergence, Quang [2022] Definition 5)**.** Let $\nu_1 = \mathcal{N}(m_1, C_1)$ and $\nu_2 = \mathcal{N}(m_2, C_2)$ be two Gaussian measures with $m_1, m_2 \in L^2(\mathcal{X}, \rho)$ and C_1, C_2 bounded, self-adjoint, positive and trace-class linear operators on $L^2(\mathcal{X}, \rho)$. Let $\gamma \in \mathbb{R}_{>0}$ be fixed. The regularized KL divergence is defined as follows,

$$\begin{aligned} D_{\mathrm{KL}}^{\gamma}(\nu_1 \,\|\, \nu_2) := & \frac{1}{2}\langle m_1 - m_2, (C_2 + \gamma\mathbb{I})^{-1}(m_1 - m_2)\rangle \\ & + \frac{1}{2}\operatorname{Tr}_X\left[(C_2 + \gamma\mathbb{I})^{-1}(C_1 + \gamma\mathbb{I}) - \mathbb{I}\right] \\ & - \frac{1}{2}\log\det{}_X\left[(C_2 + \gamma\mathbb{I})^{-1}(C_1 + \gamma\mathbb{I})\right]. \end{aligned} \tag{2.8}$$

Here Tr_X and $\det_X$ are the extended trace and extended Fredholm determinant [Quang, 2022]. For any $\gamma > 0$, the regularized KL divergence is well-defined and finite (following Quang [2017, Proposition 1]), even if the Gaussian measures are singular [Quang, 2019]. It converges to the conventional KL divergence (if it is well-defined) for $\gamma \to 0$ (Quang, 2022, Theorem 6). Furthermore, if the Gaussian measures ν_1 and ν_2 are induced by GPs $\mathcal{GP}(\mu_i, K_i)$ for $i = 1, 2$, respectively, then $D_{\mathrm{KL}}^{\gamma}(\nu_1 \,\|\, \nu_2)$ is consistently estimated [Quang, 2022] by

$$\begin{aligned} \hat{D}_{\mathrm{KL}}^{\gamma}(\nu_1 \,\|\, \nu_2) := & \frac{1}{2}(\boldsymbol{m}_1 - \boldsymbol{m}_2)^\top (\boldsymbol{\Sigma}_2^{(\gamma)})^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \\ & + \frac{1}{2}\operatorname{Tr}\left[(\boldsymbol{\Sigma}_2^{(\gamma)})^{-1}\boldsymbol{\Sigma}_1^{(\gamma)} - \mathbb{I}_M\right] \\ & - \frac{1}{2}\log\det\left[(\boldsymbol{\Sigma}_2^{(\gamma)})^{-1}\boldsymbol{\Sigma}_1^{(\gamma)}\right] \end{aligned} \tag{2.9}$$

with $\boldsymbol{m}_i := \mu_i(\boldsymbol{x})$ and $\boldsymbol{\Sigma}_i^{(\gamma)} := K_i(\boldsymbol{x}, \boldsymbol{x}) + \gamma M \mathbb{I}_M$ where $\mu_i(\boldsymbol{x})$ and $K_i(\boldsymbol{x}, \boldsymbol{x})$ are the mean vector and the covariance matrix obtained by evaluating μ_i and K_i respectively, at measurement points $\boldsymbol{x} = \{\boldsymbol{x}^{(i)}\}_{i=1}^M \overset{\text{i.i.d}}{\sim} \rho(\boldsymbol{x})$. The right-hand side of Eq. 2.9 is the expression for the KL-divergence between Gaussian distributions $\mathcal{N}(\boldsymbol{m}_1, \boldsymbol{\Sigma}_1^{(\gamma)})$ and $\mathcal{N}(\boldsymbol{m}_2, \boldsymbol{\Sigma}_2^{(\gamma)})$. Quang [2022] shows that the absolute error of the estimator is bounded by $\mathcal{O}(\sqrt{1/M})$ with high probability with constants depending on γ and properties of the GP mean and covariance functions (see Appendix A.2 for the exact bound).

3 GENERALIZED FUNCTION-SPACE VI WITH THE REGULARIZED KL DIVERGENCE

This section presents our proposed generalized function-space variational inference (GFSVI) method, which addresses the problem of the infinite KL divergence discussed in Section 2.1, which we take for an indication that VI is too restrictive if one wants to use genuine function-space priors. We instead consider generalized variational inference [Knoblauch et al., 2022], which reinterprets the ELBO in Eq. 2.1 as a regularized expected log-likelihood and explores alternative divergences for the regularizer. Specifically, we propose to use the regularized KL divergence. This section builds heavily on tools introduced in Section 2, which turn out to fit together perfectly: the pushforward of a Gaussian variational distribution in weight-space through the linearized neural network (Eq. 2.6) induces a GP variational posterior (Eq. 2.7) that admits a Gaussian measure on $L^2(\mathcal{X}, \rho)$. Further, selecting a GP prior which has an associated Gaussian measure on $L^2(\mathcal{X}, \rho)$ allows us to use the regularized KL divergence (Eq. 2.8). We present GFSVI in Section 3.1 and compare it to prior work in Section 3.2.

3.1 GENERALIZED FUNCTION-SPACE VI

We present a well-defined objective for function-space inference, and a simple algorithm for its optimization.

Objective function. We start from the ELBO in Eq. 2.3, where we use the Gaussian variational measure $\mathbb{Q}_\phi^F$ induced by the pushforward of a Gaussian variational distribution $q_\phi(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \,|\, \boldsymbol{m}, \boldsymbol{S})$ along the linearized network f_L (Eq. 2.6). The function-space prior may be any GP that has an associated Gaussian measure $\mathbb{P}^F$ on $L^2(\mathcal{X}, \rho)$. We now replace the KL divergence in the ELBO with the regularized KL divergence D_{KL}^{γ} (Eq. 2.8), which is well-defined and finite for any pair of Gaussian measures. For a likelihood

function $p(\mathcal{D} \mid \boldsymbol{w}) = \prod_{i=1}^{N} p(y_i \mid f_L(\boldsymbol{x}_i; \boldsymbol{w}))$, we obtain

$$\mathcal{L}(\phi) := \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\boldsymbol{w})}[\log p(y_i | f_L(\boldsymbol{x}_i; \boldsymbol{w}))] - D_{\mathrm{KL}}^{\gamma}\big(\mathbb{Q}_\phi^F \,\|\, \mathbb{P}^F\big). \tag{3.1}$$

Estimation and optimization. The expected log-likelihood (first term in Eq. 3.1) can be estimated by sampling from $q_\phi(\boldsymbol{w})$. For a Gaussian likelihood, it can also be computed in closed form as (unlike Rudner et al. [2022b]) we use the linearized network f_L, which made training more stable in our experiments. We estimate the regularized KL divergence (second term in Eq. 3.1) using its consistent estimator (see Eq. 2.9), with $\boldsymbol{m}_1 = f(\boldsymbol{x}; \boldsymbol{m})$, $\boldsymbol{\Sigma}_1^{(\gamma)} = J(\boldsymbol{x}; \boldsymbol{m}) \boldsymbol{S} J(\boldsymbol{x}; \boldsymbol{m})^\top + \gamma M \mathbb{I}_M$, $\boldsymbol{m}_2 = \mu(\boldsymbol{x})$, and $\boldsymbol{\Sigma}_2^{(\gamma)} = K(\boldsymbol{x}, \boldsymbol{x}) + \gamma M \mathbb{I}_M$, where μ and K are the mean and covariance functions of the GP prior, and $\boldsymbol{x} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{M} \overset{\text{i.i.d}}{\sim} \rho(\boldsymbol{x})$ are measurement points. We maximize the estimated objective over the mean $\boldsymbol{m}$ and covariance $\boldsymbol{S}$ of the Gaussian variational distribution $q_\phi(\boldsymbol{w})$, and over any likelihood parameter (e.g., the variance of a Gaussian likelihood), see Algorithm 1. Appendix B provides expressions for the estimator with Gaussian and Categorical likelihoods as well as an analysis of their computational complexity.

Technical details (γ and ρ). It turns out that increasing γ reduces the influence of the prior on inference (see Figure 20). At the same time, γ acts as jitter that prevents numerical errors (see Section 3.2). We recommend setting γ large enough to avoid numerical errors but sufficiently small to strongly regularize the objective in Eq. 3.1 (see Figure 18 in appendix) and setting M to the largest value allowed by the computational budget. We found that the estimator $\hat{D}_{\mathrm{KL}}^{\gamma}\big(\mathbb{Q}_\phi^F \,\|\, \mathbb{P}^F\big)$ converges quickly to a finite value (especially for smooth kernels, see Figure 20 in appendix), and that GFSVI is robust to a wide range of values (we fixed $\gamma = 10^{-10}$). The probability measure ρ for $L^2(\mathcal{X}, \rho)$ has to assign non-zero probability to any open set of $\mathcal{X}$ to regularize the BNN on all of its support. Following Rudner et al. [2022b], we draw measurement points from a uniform distribution over $\mathcal{X}$ when using tabular data and explore different configurations (samples from other data sets) for high-dimensional image data (see Appendix C.4).

3.2 CONNECTIONS TO PRIOR WORK

TFSVI [Rudner et al., 2022b] and FVI [Sun et al., 2019] solve stability issues by introducing jitter/white noise, which has a similar effect as the regularization in Eq. 2.8. However, TFSVI introduces jitter only to overcome numerical issues and is fundamentally restricted to prior specification in weight space since its function-space prior is the pushforward of a weight-space prior. Conversely, FVI adds white noise to prevent the KL divergence (Eq. 2.4) to blow up as M increases. However, FVI does not linearize the BNN, and hence does not have access to an explicit variational measure in function space. This severely complicates the estimation of (gradients of) the KL divergence in FVI, and the authors resort to implicit score function estimators, which make their method difficult to use in practice [Ma and Hernández-Lobato, 2021]. Our proposed GFSVI does not suffer from these difficulties as the variational posterior is an explicit Gaussian measure. This allows us to estimate the regularized KL divergence without sampling any noise or using implicit score function estimators.

Algorithm 1 Generalized function-space variational inference (GFSVI)

Require: Linearized BNN f_L with measure $\mathbb{Q}_\phi^F$, GP prior $\mathcal{GP}(\mu, K)$ with measure $\mathbb{P}^F$, measurement point distribution $\rho(\boldsymbol{x})$, data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, batch size B, $\gamma > 0$.

for all minibatch $(\boldsymbol{x}_\mathcal{B}, y_\mathcal{B}) \sim \mathcal{D}$ **do**
  Calculate $\hat{\ell}_1 = \frac{N}{B} \mathbb{E}_{q_\phi(\boldsymbol{w})}[\log p(y_\mathcal{B} \mid f_L(\boldsymbol{x}_\mathcal{B}, \boldsymbol{w}))]$;
  Draw measurement points $\boldsymbol{x} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{M} \overset{\text{i.i.d.}}{\sim} \rho(\boldsymbol{x})$;
  Calculate $\hat{\ell}_2 = \hat{D}_{\mathrm{KL}}^{\gamma}\big(\mathbb{Q}_\phi^F \,\|\, \mathbb{P}^F\big)$ using $\boldsymbol{x}$ (Eq. 2.9);
  Calculate $\hat{\mathcal{L}}(\phi) = \hat{\ell}_1 - \hat{\ell}_2$
  Update ϕ using a step in the direction $\nabla_\phi \hat{\mathcal{L}}(\phi)$

4 EXPERIMENTS

In this section, we evaluate our generalized function-space variational inference (GFSVI) method qualitatively on synthetic data and quantitatively on real-world data. GFSVI accurately captures structural properties specified by the GP prior, and that it performs competitively on regression, classification and out-of-distribution detection tasks. We also discuss the influence of the BNN's inductive biases.

Baselines. We compare GFSVI to two weight-space inference methods: mean-field VI (MFVI) [Blundell et al., 2015] and linearized Laplace [Immer et al., 2021]; and to three function-space inference methods: FVI [Sun et al., 2019], TFVSI [Rudner et al., 2022b] and VIP [Ma et al., 2019] (TFSVI performs inference in function space but with the pushforward of a weight-space prior; VIP uses a BNN prior). All BNNs have the same architecture and fully-factorized Gaussian approximate posterior. We also include results for a sparse GP with a posterior mean parameterized by a neural network (GWI) [Wild et al., 2022b], and for a Gaussian Process (GP) [Williams and Rasmussen, 2006] (when the size of the dataset allows it), and for a sparse GP [Hensman et al., 2013] for regression tasks. We consider the GP, sparse GP and GWI as gold standards as they represent the exact (or near exact) posterior for models with GP priors.

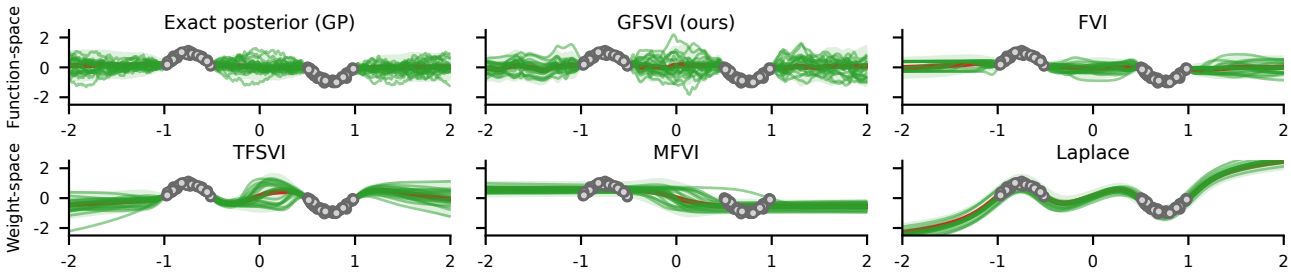


Figure 2: Inference on synthetic data (gray circles) using a Matérn-1/2 prior for function-space methods GFSVI and FVI. The proposed GFSVI provides the best approximation of the exact GP posterior.

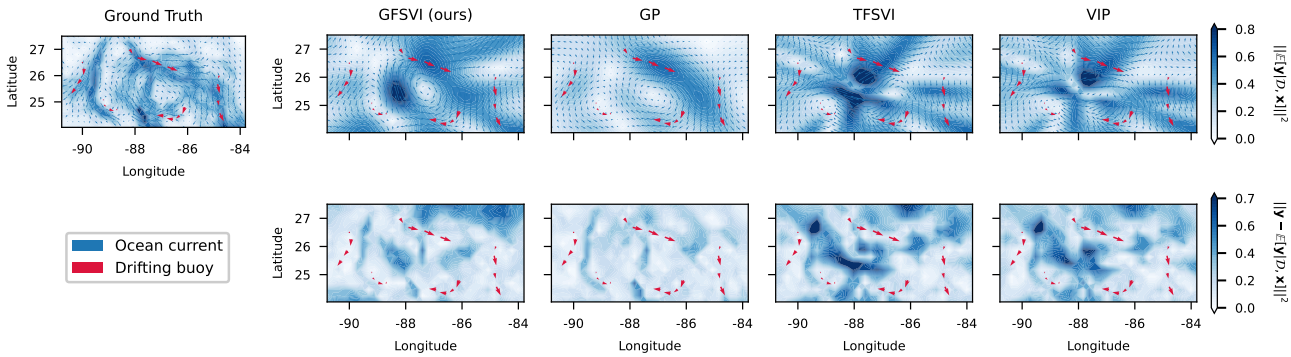


Figure 3: Results for the ocean current modeling experiment. We report the norm of the mean velocity vectors and the squared errors. Unlike TFSVI, we find that GFSVI accurately captures ocean current dynamics.

Qualitative results on synthetic data. We consider a 1-dimensional regression task where the values y_i are sampled around $\sin(2\pi x_i)$ (circles in Figures 1-8 and 11) and the two moons 2-dimensional binary classification task [Pedregosa et al., 2011] (see Figures 9 and 10). For regression, the green lines show functions sampled from the (approximate) posteriors, and the red lines are the inferred mean functions. For classification, the first and second row show the inferred mean probability of class 1 (blue dots) and its 2-standard deviations with respect to posterior samples. More details in Appendix C.1. We find that GFSVI captures the beliefs of the RBF and Matérn-1/2 GP priors better than BNN-baselines in the regression setting (see Figures 2 and 5) as well as in classification (see Figures 9 and 10), and shows greater uncertainty outside of the support of the data. Figures 1 and 6 show that GFSVI notably adapts to varying prior assumptions (varying smoothness and length scale, respectively). In addition, Figures 4 and 8 in the Appendix show that GFSVI provides strong regularization when the data generative process is noisy, and that it can be trained with fewer measurement points M than FVI without significant degradation.

Inductive biases. Figure 11 in the Appendix compares GFSVI to the exact GP-posterior across two different priors and three model architectures (details in Appendix C.1). We find that, with ReLU activations, small models are prone to underfitting for smooth priors (RBF), and to collapsing uncertainty for rough priors (Matérn-1/2). By contrast, with smooth activations (Tanh), smaller models suffice, and they are compatible with most standard GP priors (the results shown in Figure 11 extend to RBF, Matérn, and Rational Quadratic in our experiments). We also analyzed how the number M of measurement points affects performance. Figures 7 and 17 in the appendix show that capturing the beliefs of rough GP priors and estimating D_{KL}^{γ} with these priors requires larger M.

4.1 QUANTITATIVE RESULTS ON REAL-WORLD DATA

We evaluate GFSVI on regression, classification, and out-of-distribution detection. In all tables, we bold the highest score and any score whose error bar (standard error) overlaps with the highest score's error bar.

Ocean current modeling. We measure how well GFSVI can incorporate knowledge specified via a GP prior on real-world data by considering the problem of modeling ocean currents in the Gulf of Mexico. We follow the setup by Shalashilin [2024] and use the GulfDrifters dataset [Lilly and Pérez-Brunius, 2021] to estimate ocean currents from 20 2-dimensional velocity vectors collected from drifter buoys. We embed physical properties of fluid motions into the GP prior and to the neural networks by applying the Helmholtz decomposition [Berlinghieri et al., 2023, Cinquin et al.,

Table 1: Test expected log-likelihood (higher is better) of evaluated methods on regression datasets. GFSVI performs competitively compared to all BNN baselines and obtains the best mean rank.

Dataset	Function-space priors		Weight-space priors				Gaussian Processes (Gold Standards)		
	GFSVI (ours)	FVI	TFSVI	MFVI	VIP	Laplace	GWI	Sparse GP	GP
Boston	**-0.733 ± 0.144**	**-0.571 ± 0.113**	-1.416 ± 0.046	-1.308 ± 0.052	**-0.722 ± 0.196**	**-0.812 ± 0.205**	-0.940 ± 0.145	-0.884 ± 0.182	-1.594 ± 0.556
Concrete	**-0.457 ± 0.041**	**-0.390 ± 0.017**	-0.983 ± 0.012	-1.353 ± 0.018	**-0.427 ± 0.050**	-0.715 ± 0.025	-0.744 ± 0.079	-0.966 ± 0.025	-2.099 ± 0.421
Energy	**1.319 ± 0.052**	**1.377 ± 0.042**	0.797 ± 0.098	-0.926 ± 0.197	**1.046 ± 0.378**	**1.304 ± 0.043**	0.461 ± 0.093	-0.206 ± 0.027	-0.205 ± 0.022
Kin8nm	-0.136 ± 0.013	-0.141 ± 0.023	-0.182 ± 0.011	-0.641 ± 0.225	**-0.102 ± 0.013**	-0.285 ± 0.014	-0.708 ± 0.054	-0.443 ± 0.014	*(infeasible)*
Naval	**3.637 ± 0.132**	2.165 ± 0.194	2.758 ± 0.044	1.034 ± 0.160	1.502 ± 0.061	3.404 ± 0.084	-0.301 ± 0.254	4.951 ± 0.014	*(infeasible)*
Power	**0.044 ± 0.011**	**0.031 ± 0.021**	0.007 ± 0.013	-0.003 ± 0.015	**0.036 ± 0.018**	-0.002 ± 0.019	0.043 ± 0.009	-0.100 ± 0.010	*(infeasible)*
Protein	-1.036 ± 0.005	-1.045 ± 0.005	-1.010 ± 0.004	-1.112 ± 0.007	-0.994 ± 0.007	-1.037 ± 0.006	-1.050 ± 0.009	-1.035 ± 0.002	*(infeasible)*
Wine	-1.289 ± 0.040	**-1.215 ± 0.007**	-2.138 ± 0.221	-1.248 ± 0.018	-1.262 ± 0.025	-1.249 ± 0.025	-1.232 ± 0.038	-1.240 ± 0.037	-1.219 ± 0.035
Yacht	**1.058 ± 0.080**	**0.545 ± 0.735**	-1.187 ± 0.064	-1.638 ± 0.030	-0.062 ± 1.378	0.680 ± 0.171	0.441 ± 0.138	-0.976 ± 0.092	-0.914 ± 0.045
Wave	5.521 ± 0.036	6.612 ± 0.008	5.148 ± 0.117	**6.883 ± 0.008**	4.043 ± 0.093	4.658 ± 0.027	1.566 ± 0.123	4.909 ± 0.001	*(infeasible)*
Denmark	**-0.487 ± 0.013**	-0.801 ± 0.005	-0.513 ± 0.013	-0.675 ± 0.007	-0.583 ± 0.021	-0.600 ± 0.008	-0.841 ± 0.026	-0.768 ± 0.001	*(infeasible)*
Mean rank	1.545	2.000	2.727	3.455	2.091	2.455	-	-	-

Table 2: Test expected log-likelihood, accuracy, expected calibration error and OOD detection accuracy on MNIST and Fashion MNIST.

	Metric	Function-space priors				Weight-space priors					GP-based
		GFSVI (rnd)	GFSVI (kmnist)	FVI (rnd)	FVI (kmnist)	TFSVI (rnd)	TFSVI (kmnist)	MFVI	VIP	Laplace	GWI
MNIST	Log-like. (↑)	**-0.033 ± 0.000**	-0.041 ± 0.000	-0.145 ± 0.005	-0.238 ± 0.006	-0.047 ± 0.003	-0.041 ± 0.001	-0.078 ± 0.001	**-0.033 ± 0.001**	-0.108 ± 0.002	-0.090 ± 0.003
	Acc. (↑)	**0.992 ± 0.000**	0.991 ± 0.000	0.976 ± 0.001	0.943 ± 0.001	0.989 ± 0.000	0.989 ± 0.000	0.990 ± 0.000	0.989 ± 0.000	0.984 ± 0.000	0.971 ± 0.001
	ECE (↓)	**0.002 ± 0.000**	0.006 ± 0.000	0.064 ± 0.001	0.073 ± 0.003	0.007 ± 0.000	0.006 ± 0.000	0.021 ± 0.000	**0.002 ± 0.001**	0.048 ± 0.001	0.003 ± 0.000
	OOD acc. (↑)	0.921 ± 0.008	**0.980 ± 0.004**	0.894 ± 0.010	0.891 ± 0.006	0.887 ± 0.011	0.893 ± 0.005	0.928 ± 0.002	0.871 ± 0.012	0.903 ± 0.007	0.829 ± 0.007
FMNIST	Log-like. (↑)	-0.260 ± 0.003	-0.294 ± 0.006	-0.300 ± 0.002	-0.311 ± 0.005	-0.261 ± 0.001	-0.261 ± 0.002	-0.290 ± 0.002	**-0.252 ± 0.001**	-0.426 ± 0.009	-0.260 ± 0.001
	Acc. (↑)	0.910 ± 0.001	0.909 ± 0.001	0.910 ± 0.002	0.906 ± 0.002	0.909 ± 0.001	0.907 ± 0.001	**0.913 ± 0.001**	0.911 ± 0.001	0.886 ± 0.001	0.906 ± 0.000
	ECE (↓)	0.020 ± 0.003	0.042 ± 0.002	0.027 ± 0.005	0.024 ± 0.002	0.022 ± 0.002	0.021 ± 0.002	**0.010 ± 0.001**	0.024 ± 0.001	0.060 ± 0.004	0.016 ± 0.001
	OOD acc. (↑)	0.853 ± 0.005	**0.997 ± 0.001**	0.925 ± 0.005	0.975 ± 0.002	0.802 ± 0.006	0.779 ± 0.010	0.805 ± 0.010	0.790 ± 0.010	0.826 ± 0.006	0.792 ± 0.005

Table 3: Results for the ocean current modeling task.

Metric	GFSVI (ours)	TFSVI	VIP	GP
Log-like.	-6.627 ± 0.753	-22.651 ± 2.947	-11.631 ± 3.171	-0.507 ± 0.000
MSE	0.021 ± 0.002	0.034 ± 0.003	0.026 ± 0.001	0.013 ± 0.000

2024]. We compare our GFSVI to a GP, to TFSVI and to VIP. More details can be found in Appendix C.2. We find that incorporating knowledge via an informative GP prior in GFSVI improves performance over weight-space priors in TFSVI and VIP (see Table 3 and Figure 3). However, the GP outperforms both BNNs, which suggests that the physically motivated kernel describes the fluid dynamics well enough that the additional inductive bias introduced by a neural network hurts performance rather than helping it. In the following, we consider experiments with larger datasets (making exact GP inference computationally infeasible in many cases), and where structural prior knowledge in function space exists but is not derived from laws of nature.

Regression. We assess the predictive performance of GFSVI on data sets from the UCI repository [Dua and Graff, 2017]. Table 1, and Table 6 in the appendix, show expected log-likelihood and mean squared error, respectively. We perform 5-fold cross validation and report means and standard errors across the test folds. We also rank the methods for each dataset and report the mean rank of each method across all datasets. See Appendix C.3 for more details. We find that GFSVI performs competitively compared to baselines and obtains the best mean rank for both metrics, matching the top performing methods on nearly all datasets. In particular, we find that using GP priors in the linearized BNN with GFSVI yields improvements over the weight-space priors used in TFSVI, and that GFSVI performs slightly better than FVI despite being simpler. Further, we find that GFSVI approximates the exact GP-posterior more accurately that FVI (see Table 7 and Appendix D.3), and that it converges in slightly more steps than TFSVI (Figure 15).

Classification. We further evaluate classification performance of our method on the MNIST [LeCun et al., 2010] and FashionMNIST [Xiao et al., 2017] image data sets. We fit the models on a random subset of 90% of the training set, use the remaining 10% as validation data, and evaluate on the provided test split. We repeat with 5 different random seeds and report the mean and standard error of the expected log-likelihood, accuracy, and expected calibration error (ECE) in Table 2. For GFSVI, FVI, and TFSVI, we tested measurement points from both a uniform random (rnd) distribution $\rho(\boldsymbol{x})$ and from kmnist. Details in Appendix C.4. We find that GFSVI performs competitively on MNIST, exceeding the expected log-likelihood and accuracy of top-scoring baselines and similarly to best baselines on FashionMNIST. GFSVI also yields well-calibrated models with low ECE.

Out-of-distribution detection. We next evaluate our method by testing if its epistemic uncertainty is predictive of out-of-distribution (OOD) data. We consider two settings: (i) with tabular data and a Gaussian likelihood [Malinin et al., 2021], and (ii) with image data and a categorical likelihood

Table 4: Out-of-distribution accuracy (higher is better) of evaluated methods on regression datasets. GFSVI (ours) performs competitively on OOD detection and obtains the highest mean rank.

Dataset	Function-space priors		Weight-space priors				Gaussian Processes (Gold Standards)		
	GFSVI (ours)	FVI	TFSVI	MFVI	VIP	Laplace	GWI	Sparse GP	GP
Boston	$\mathbf{0.893 \pm 0.011}$	0.594 ± 0.024	0.705 ± 0.107	0.563 ± 0.013	0.628 ± 0.010	0.557 ± 0.009	0.817 ± 0.017	0.947 ± 0.011	0.952 ± 0.003
Concrete	$\mathbf{0.656 \pm 0.016}$	0.583 ± 0.022	0.511 ± 0.003	0.605 ± 0.012	0.601 ± 0.024	0.578 ± 0.015	0.730 ± 0.020	0.776 ± 0.006	0.933 ± 0.004
Energy	$\mathbf{0.997 \pm 0.002}$	0.696 ± 0.017	$\mathbf{0.997 \pm 0.001}$	0.678 ± 0.014	0.682 ± 0.037	0.782 ± 0.020	0.998 ± 0.001	0.998 ± 0.001	0.998 ± 0.001
Kin8nm	0.588 ± 0.007	$\mathbf{0.604 \pm 0.023}$	0.576 ± 0.008	0.570 ± 0.009	0.563 ± 0.015	$\mathbf{0.606 \pm 0.009}$	0.602 ± 0.011	0.608 ± 0.014	*(infeasible)*
Naval	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.919 ± 0.017	0.621 ± 0.059	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	1.000 ± 0.000	*(infeasible)*
Power	$\mathbf{0.698 \pm 0.006}$	0.663 ± 0.021	0.676 ± 0.008	0.636 ± 0.019	0.514 ± 0.004	0.654 ± 0.013	0.754 ± 0.004	0.717 ± 0.004	*(infeasible)*
Protein	$\mathbf{0.860 \pm 0.011}$	0.810 ± 0.022	$\mathbf{0.841 \pm 0.018}$	0.693 ± 0.020	$\mathbf{0.549 \pm 0.020}$	0.629 ± 0.013	0.942 ± 0.002	0.967 ± 0.001	*(infeasible)*
Wine	0.665 ± 0.013	0.517 ± 0.004	0.549 ± 0.015	0.542 ± 0.009	$\mathbf{0.706 \pm 0.028}$	0.531 ± 0.007	0.810 ± 0.008	0.781 ± 0.014	0.787 ± 0.007
Yacht	0.616 ± 0.030	0.604 ± 0.025	$\mathbf{0.659 \pm 0.043}$	$\mathbf{0.642 \pm 0.035}$	$\mathbf{0.688 \pm 0.040}$	0.612 ± 0.024	0.563 ± 0.014	0.762 ± 0.018	0.787 ± 0.011
Wave	$\mathbf{0.975 \pm 0.005}$	0.642 ± 0.004	0.835 ± 0.034	0.658 ± 0.026	0.500 ± 0.000	0.529 ± 0.005	0.903 ± 0.001	0.513 ± 0.001	*(infeasible)*
Denmark	0.521 ± 0.006	$\mathbf{0.612 \pm 0.008}$	0.519 ± 0.006	0.513 ± 0.003	0.500 ± 0.000	0.529 ± 0.008	0.688 ± 0.003	0.626 ± 0.002	*(infeasible)*
Mean rank	1.455	2.364	1.909	2.909	3.364	2.909	-	-	-

[Osawa et al., 2019]. We report the accuracy of classifying OOD vs. in-distribution (ID) data using a (learned) threshold on the predictive uncertainty. More details in Appendix C.5. In setting (i), GFSVI performs competitively and obtains the highest mean rank (Table 4). Likewise in setting (ii), GFSVI strongly outperforms all baselines when using the KMNIST measurement point distribution $\rho(\boldsymbol{x})$ (Figure 12, Tables 2 and 8). We find that with high-dimensional image data, the choice of measurement point distribution highly influences OOD detection accuracy (see Appendix D.5 for a discussion). In both settings, using GP priors with GFSVI rather than weight-space priors with TFSVI is beneficial, and GFSVI also improves over FVI. GFSVI's uncertainty is also well-calibrated under distribution shift of the input features (see Appendix D.6).

5 RELATED WORK

In this section, we review related work on function-space VI with neural networks, and on approximating functions-space measures with weight-space priors.

Function-space inference with neural networks. Prior work on function-space VI in BNNs has addressed issues (i) intractable variational posterior in function space and (ii) intractable KL divergence discussed in Section 2.1. Sun et al. [2019] address (i) by using implicit score function estimators, and (ii) by replacing the supremum with an expectation. Rudner et al. [2022b] address (i) by using a linearized BNN [Khan et al., 2019, Immer et al., 2021, Maddox et al., 2021], and (ii) by replacing the supremum with a maximum over a finite set. Other work abandons approximating the neural network's posterior and instead uses a BNN to specify a prior [Ma et al., 2019], or deterministic neural networks as features for Bayesian linear regression [Ma and Hernández-Lobato, 2021] or the mean of a generalized sparse GP [Wild et al., 2022b]. Unlike our more expressive GP posterior covariance, Wild et al. [2022b] uses a simple stationary sparse GP posterior covariance (Table 2) which has higher sampling cost and can lead to model misspecification (Figure 14). Our work combines linearized BNNs with generalized VI, but we use the regularized KL divergence [Quang, 2019], which naturally generalizes the KL divergence and allows for informative GP priors.

Approximating function-space measures with weight-space priors. Flam-Shepherd et al. [2017], Tran et al. [2022] minimize a divergence between the BNN's prior predictive and a GP before performing inference on weights, while Wu et al. [2023] directly incorporate the bridging divergence inside the inference objective. Alternatively, Pearce et al. [2020] derive BNN architectures mirroring GPs, and Matsubara et al. [2021] use the Ridgelet transform to design weight-spaces priors approximating a GP in function space. Similarly, Rudner et al. [2023] and Sam et al. [2024] use empirical weight-space priors to regularize in function space and encode domain knowledge specified via a loss function, respectively. Yang et al. [2020] instead imposes functional constraints directly via the prior.

6 DISCUSSION

We proposed a simple inference method with a well-defined variational objective function for BNNs with GP priors in function-space. As standard VI with functions-space priors suffers from an infinite KL divergence problem, we propose to follow the generalized VI framework. Specifically, we substitute the conventional KL divergence in the ELBO by the regularized KL divergence, which is always finite, and which can be estimated consistently within the linearized BNN approximation. We demonstrated that our method allows to incorporate interpretable structural properties via a GP prior, accurately approximates the true GP posterior on synthetic and small real-world data sets, and provides competitive uncertainty estimates for regression, classification and out-of-distribution detection compared to BNNs with both function-space and weight-space priors. Future work should investigate the use of more expressive variational distributions, such as Gaussian with low-rank plus diagonal covariance proposed by Tomczak et al. [2020].

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy – EXC number 2064/1 – Project number 390727645, and by the German Research Foundation (DFG) under project 448588364 of the Emmy Noether Programme. Additional support was provided by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors extend their gratitude to the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Tristan Cinquin. Finally, Tristan Cinquin thanks Marvin Pförtner for very useful discussions on the theory and experiments, and Vincent Fortuin for feedback.

References

Abdullah Abdullah, Masoud Hassan, and Yaseen Mustafa. A review on Bayesian deep learning in healthcare: Applications and challenges. *IEEE Access*, 10:1–1, 01 2022. doi: 10.1109/ACCESS.2022.3163384.

Mikhail Belkin. Approximation beats concentration? an approximation view on inference with smooth radial kernels. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1348–1361. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/belkin18a.html.

Renato Berlinghieri, Brian L. Trippe, David R. Burt, Ryan James Giordano, Kaushik Srinivasan, Tamay Özgökmen, Junfei Xia, and Tamara Broderick. Gaussian processes at the helm(holtz): A more fluid model for ocean currents. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2113–2163. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/berlinghieri23a.html.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

David Bew, Campbell R. Harvey, Anthony Ledford, Sam Radnor, and Andrew Sinclair. Modeling analysts' recommendations via Bayesian machine learning. *The Journal of Financial Data Science*, 1(1):75–98, 2019. ISSN 2640-3943. doi: 10.3905/jfds.2019.1.1.075. URL https://jfds.pm-research.com/content/1/1/75.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/blundell15.html.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Hao Chen, Lili Zheng, Raed Al Kontar, and Garvesh Raskutti. Gaussian process parameter estimation using mini-batch stochastic gradient descent: convergence guarantees and empirical benefits. *J. Mach. Learn. Res.*, 23 (1), January 2022. ISSN 1532-4435.

Tristan Cinquin, Alexander Immer, Max Horn, and Vincent Fortuin. Pathologies in priors and inference for bayesian transformers. In *I (Still) Can't Believe It's Not Better! NeurIPS 2021 Workshop*, 2021. URL https://openreview.net/forum?id=Gv41ucDhhlY.

Tristan Cinquin, Marvin Pförtner, Vincent Fortuin, Philipp Hennig, and Robert Bamler. FSP-laplace: Function-space priors for the laplace approximation in bayesian deep learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=83vxe8alV4.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Yamamoto Kazuaki, and David Ha. Deep learning for classical japanese literature, 12 2018.

Dheeru Dua and Casey Graff. Uci machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Daniel Flam-Shepherd, James Requeima, and David Duvenaud. Mapping gaussian process priors to bayesian neural networks, 2017.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and

Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Tom Hennigan, Trevor Cai, Tamara Norman, Lena Martens, and Igor Babuschkin. Haiku: Sonnet for JAX, 2020. URL http://github.com/deepmind/dm-haiku.

James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 282290, Arlington, Virginia, USA, 2013. AUAI Press.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(40):1303–1347, 2013. URL http://jmlr.org/papers/v14/hoffman13a.html.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 703–711. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/immer21a.html.

Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion generative models in infinite dimensions. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 9538–9563. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/kerrigan23a.html.

Mohammad Emtiyaz Khan, Alexander Immer, Ehsan Abedi, and Maciej Korzepa. Approximate inference turns deep networks into gaussian processes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/b3bbccd6c008e727785cb81b1aa08ac5-Paper.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022. URL http://jmlr.org/papers/v23/19-1047.html.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Jonathan M. Lilly and Paula Pérez-Brunius. GulfDrifters: A consolidated surface drifter dataset for the Gulf of Mexico, January 2021. URL https://doi.org/10.5281/zenodo.4421585.

Chao Ma and José Miguel Hernández-Lobato. Functional variational inference based on stochastic process generators. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21795–21807. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/b613e70fd9f59310cf0a8d33de3f2800-Paper.pdf.

Chao Ma, Yingzhen Li, and Jose Miguel Hernandez-Lobato. Variational implicit processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4222–4233. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ma19b.html.

Wesley Maddox, Shuai Tang, Pablo Moreno, Andrew Gordon Wilson, and Andreas Damianou. Fast adaptation with linearized neural networks. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2737–2745. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/maddox21a.html.

Andrey Malinin, Liudmila Prokhorenkova, and Aleksei Ustimenko. Uncertainty in gradient boosting via ensembles. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=1Jv6b0Zq3qi.

Takuo Matsubara, Chris J. Oates, and FranÃğois-Xavier Briol. The ridgelet prior: A covariance function approach to prior specification for bayesian neural networks. *Journal of Machine Learning Research*, 22(157):1–57, 2021. URL http://jmlr.org/papers/v22/20-1300.html.

Dimitrios Milios, Raffaello Camoriano, Pietro Michiardi, Lorenzo Rosasco, and Maurizio Filippone. Dirichlet-based gaussian processes for large-scale calibrated

classification. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/b6617980ce90f637e68c3ebe8b9be745-Paper.pdf.

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical deep learning with bayesian principles. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/b53477c2821c1bf0da5d40e57b870d35-Paper.pdf.

Tim Pearce, Russell Tsuchida, Mohamed Zaki, Alexandra Brintrup, and Andy Neely. Expressive priors in bayesian neural networks: Kernel combinations and periodic functions. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 134–144. PMLR, 22–25 Jul 2020. URL https://proceedings.mlr.press/v115/pearce20a.html.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Thomas Pinder and Daniel Dodd. Gpjax: A gaussian process framework in jax. *Journal of Open Source Software*, 7(75):4455, 2022. doi: 10.21105/joss.04455. URL https://doi.org/10.21105/joss.04455.

Minh Ha Quang. Infinite-dimensional log-determinant divergences ii: Alpha-beta divergences, 2017. URL https://arxiv.org/abs/1610.08087.

Minh Ha Quang. Regularized divergences between covariance operators and gaussian measures on hilbert spaces. *Journal of Theoretical Probability*, 2019. URL https://api.semanticscholar.org/CorpusID:118638845.

Minh Ha Quang. Kullback-leibler and renyi divergences in reproducing kernel hilbert space and gaussian process settings, 2022.

Tim G. J. Rudner, Freddie Bickford Smith, Qixuan Feng, Yee Whye Teh, and Yarin Gal. Continual learning via sequential function-space variational inference. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18871–18887. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/rudner22a.html.

Tim G. J. Rudner, Zonghao Chen, Yee Whye Teh, and Yarin Gal. Tractabe Function-Space Variational Inference in Bayesian Neural Networks. In *Advances in Neural Information Processing Systems*, 2022b.

Tim G. J. Rudner, Sanyam Kapoor, Shikai Qiu, and Andrew Gordon Wilson. Function-space regularization in neural networks: A probabilistic perspective. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29275–29290. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/rudner23a.html.

Dylan Sam, Rattana Pukdee, Daniel P. Jeong, Yewon Byun, and J. Zico Kolter. Bayesian neural networks with domain knowledge priors, 2024. URL https://arxiv.org/abs/2402.13410.

G. Santin and R. Schaback. Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics*, 42(4):973–993, jan 2016. doi: 10.1007/s10444-015-9449-5. URL https://doi.org/10.1007%2Fs10444-015-9449-5.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a981f2b708044d6fb4a71a1463242520-Paper.pdf.

Ivan Shalashilin. Gaussian processes for vector fields and ocean current modelling, March 2024. URL https://docs.jaxgaussianprocesses.com/examples/oceanmodelling/.

Dan Simpson. Priors for the parameters in a Gaussian process, 2022. URL https://dansblog.netlify.app/posts/2022-09-07-priors5/priors5.html.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. FUNCTIONAL VARIATIONAL BAYESIAN

NEURAL NETWORKS. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rkxacs0qY7.

Marcin Tomczak, Siddharth Swaroop, and Richard Turner. Efficient low rank gaussian variational inference for neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4610–4622. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/310cc7ca5a76a446f85c1a0d641ba96d-Paper.pdf.

Ba-Hien Tran, Simone Rossi, Dimitrios Milios, and Maurizio Filippone. All you need is a good functional prior for bayesian deep learning. *Journal of Machine Learning Research*, 23(74):1–56, 2022. URL http://jmlr.org/papers/v23/20-1340.html.

Veit D. Wild, Robert Hu, and Dino Sejdinovic. Generalized variational inference in function spaces: Gaussian measures meet bayesian deep learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022a. Curran Associates Inc. ISBN 9781713871088.

Veit David Wild, Robert Hu, and Dino Sejdinovic. Generalized variational inference in function spaces: Gaussian measures meet bayesian deep learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=mMT8bhVBoUa.

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

Andrew Gordon Wilson, Pavel Izmailov, Matthew D Hoffman, Yarin Gal, Yingzhen Li, Melanie F Pradier, Sharad Vikram, Andrew Foong, Sanae Lotfi, and Sebastian Farquhar. Evaluating approximate inference in bayesian deep learning. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 113–124. PMLR, 06–14 Dec 2022. URL https://proceedings.mlr.press/v176/wilson22a.html.

Mengjing Wu, Junyu Xuan, and Jie Lu. Indirect functional bayesian neural networks. In *Fifth Symposium on Advances in Approximate Bayesian Inference*, 2023. URL https://openreview.net/forum?id=-xfwSaifkU.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Wanqian Yang, Lars Lorch, Moritz Graule, Himabindu Lakkaraju, and Finale Doshi-Velez. Incorporating interpretable output constraints in bayesian neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12721–12731. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/95c7dfc5538e1ce71301cf92a9a96bd0-Paper.pdf.

Supplementary Material

Tristan Cinquin[1] **Robert Bamler**[1]

[1]University of Tübingen, Germany

A DIVERGENCES BETWEEN GAUSSIAN MEASURES

A.1 THE KL DIVERGENCE IS INFINITE

In this section, we show that the Kullbach-Liebler (KL) divergence between the Gaussian measures $\mathbb{Q}_\phi^F \sim \mathcal{N}(m_Q, C_Q)$ and $\mathbb{P}^F \sim \mathcal{N}(m_P, C_P)$, respectively induced by the linearized BNN in Eq 2.7 and by a non-degenerate Gaussian process satisfying conditions given in Section 2.2, is infinite. While this has already been shown by Wild et al. [2022a], the proof is easier for Gaussian measures. We first need the Feldman-Hàjek theorem which tells us when the KL divergence between two Gaussian measures is well-defined.

Theorem A.1 (Feldman-Hàjek, Quang [2022] Theorem 2, Simpson [2022] Theorem 7)**.** *Consider two Gaussian measures $\nu_1 = \mathcal{N}(m_1, C_1)$ and $\nu_2 = \mathcal{N}(m_2, C_2)$ on $L^2(\mathcal{X}, \rho)$. Then ν_1 and ν_2 are called equivalent if and only if the following holds:*

1. $m_1 - m_2 \in Im(C_2^{1/2})$
2. *The operator T such that $C_1 = C_2^{1/2}(I - T)C_2^{1/2}$ is Hilbert-Schmidt, that is T has a countable set of eigenvalues λ_i that satisfy $\lambda_i < 1$ and $\sum_{i=1}^\infty \lambda_i^2 < \infty$.*

otherwise ν_1 and ν_2 are singular. If ν_1 and ν_2 are equivalent, then the Radon-Nikodym derivative exists and $D_{\mathrm{KL}}(\nu_1 \,\|\, \nu_2)$ admits an explicit formula. Otherwise, $D_{\mathrm{KL}}(\nu_1 \,\|\, \nu_2) = \infty$.

Let us now show that the KL divergence between $\mathbb{Q}_\phi^F$ and $\mathbb{P}^F$ is indeed infinite.

Proposition 1. The Gaussian measures $\mathbb{Q}_\phi^F$ and $\mathbb{P}^F$ are mutually singular and $D_{\mathrm{KL}}(\mathbb{Q}_\phi^F || \mathbb{P}^F) = \infty$.

Proof. The proof follows from the Feldman-Hàjek theorem (Theorem A.1). In our case, C_Q has at most p non-zero eigenvalues as the covariance function of the GP induced by the BNN is degenerate, while C_P has a set of (countably) infinite non-zeros eigenvalues (prior is non-degenerate as per assumption). Hence, for the equality in condition (2) to hold, T must have eigenvalue 1 which violates the requirement that T is Hilbert-Schmidt i.e. that its eigenvalues $\{\lambda_i\}_{i=1}^\infty$ satisfy $\lambda_i < 1$ and $\sum_{i=1}^\infty \lambda_i^2 < \infty$. Therefore, $\mathbb{Q}_\phi^F$ and $\mathbb{P}^F$ are mutually singular and $D_{\mathrm{KL}}(\mathbb{Q}_\phi^F || \mathbb{P}^F) = \infty$. □

A.2 THE REGULARIZED KL DIVERGENCE

We provide the bound describing the asymptotic convergence of the regularized KL divergence estimator in Equation (A.1). The error results from the fact that taking a finite number M of context points effectively cuts off the spectra of the covariance operators and the estimator $\hat{D}_{\mathrm{KL}}^\gamma$ converges to D_{KL}^γ as $M \to \infty$ with high probability.

Theorem A.2 (Convergence of estimator, Quang [2022] Theorem 45)**.** *Assume the following:*

1. *Let T be a $\sigma -$ compact metric space, that is $T = \cup_{i=1}^{\infty} T_i$, where $T_1 \subset T_2 \subset \cdots$ with each T_i being compact.*
2. *ρ is a non-degenerate Borel probability measure on T, that is $\rho(B) > 0$ for each open set $B \subset T$.*
3. *$K_1, K_2 : T \times T \to \mathbb{R}$ are continuous, symmetric, positive definite kernels and there exists $\kappa_1 > 0, \kappa_2 > 0$ such that $\int_T K_i(x,x) d\rho(x) \leq \kappa_i^2$ for $i = 1, 2$.*
4. *$\sup_{x \in T} K_i(x,x) \leq \kappa_i^2$ for $i = 1, 2$.*
5. *$f_i \sim GP(\mu_i, K_i)$, where $\mu_i \in L^2(T, \rho)$ for $i = 1, 2$.*
6. *$\exists B_i > 0$ such that $\|\mu_i\|_\infty \leq B_i$ for $i = 1, 2$.*

Let $\boldsymbol{x} = \{\boldsymbol{x}^{(i)}\}_{i=1}^M$, $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)} \overset{\text{i.i.d}}{\sim} \rho(\boldsymbol{x})$. If Gaussian measures $\mathcal{N}(m_i, C_i)$ are induced by GPs $f_i \sim \mathcal{GP}(\mu_i, K_i)$ for $i = 1, 2$, then for any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$
\begin{aligned}
&|D_{\mathrm{KL}}(\mathcal{N}(\mu_1(\mathbf{x}), K_1(\mathbf{x},\mathbf{x}) + M\gamma\mathbb{I}_M) \,\|\, \mathcal{N}(\mu_2(\mathbf{x}), K_2(\mathbf{x},\mathbf{x}) + M\gamma\mathbb{I}_M)) - D_{\mathrm{KL}}^{\gamma}(\mathcal{N}(m_1, C_1) \,\|\, \mathcal{N}(m_2, C_2))| \\
&\qquad \leq \frac{1}{2\gamma}(B_1 + B_2)^2[1 + \kappa_2^2/\gamma]^2 \left(\frac{2\log\frac{48}{\delta}}{M} + \sqrt{\frac{2\log\frac{48}{\delta}}{M}} \right) \\
&\qquad\qquad + \frac{1}{2\gamma^2}[\kappa_1^4 + \kappa_2^4 + \kappa_1^2\kappa_2^2(2 + \kappa_2^2/\gamma)] \left(\frac{2\log\frac{12}{\delta}}{M} + \sqrt{\frac{2\log\frac{12}{\delta}}{M}} \right)
\end{aligned}
\tag{A.1}
$$

Note that Equation (A.1) provides a very general bound on the error that does not make assumptions on the spectral decay, and it may therefore dramatically overestimate the error. Indeed, we analyze convergence empirically in Figure 20 and observe that the estimator converges quickly except for very rough priors (e.g., Matérn-1/2) with very small γ.

B ADDITIONAL DETAILS ON THE GFSVI OBJECTIVE ESTIMATOR

In this section, we present details on the estimation of the generalized function-space variational inference (GFSVI) objective. Let $f_L(\,\cdot\,; \boldsymbol{w})$ be the linearized BNN (Eq 2.6) with weights $\boldsymbol{w} \in \mathbb{R}^p$, and $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ a data set with features $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ and associated values $y_i \in \mathcal{Y}$. Assuming a likelihood $p(\mathcal{D} \,|\, \boldsymbol{w}) = \prod_{i=1}^N p(y_i \,|\, f(\boldsymbol{x}_i; \boldsymbol{w}))$ and a Gaussian variational distribution on model weights $q_\phi(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \,|\, \boldsymbol{m}, \boldsymbol{S})$, the GFSVI objective function is

$$
\mathcal{L}(\phi) = \sum_{i=1}^N \mathbb{E}_{q_\phi(\boldsymbol{w})}[\log p(y_i \,|\, f_L(\boldsymbol{x}_i; \boldsymbol{w}))] - D_{\mathrm{KL}}^{\gamma}\big(\mathbb{Q}_\phi^F \,\|\, \mathbb{P}^F\big) \tag{B.1}
$$

where $\mathbb{Q}_\phi^F$ and $\mathbb{P}^F$ are the Gaussian measures induced by the linearized BNN and a Gaussian process prior respectively.

Expected log-likelihood When considering a Gaussian likelihood, we use the closed form expression available due to the Gaussian variational measure over functions induced by the linearized BNN

$$
\mathbb{E}_{q_\phi(\boldsymbol{w})}\big[\log \mathcal{N}\big(y_i \,|\, f_L(\boldsymbol{x}_i; \boldsymbol{w}), \sigma_y^2\big)\big] = -\frac{1}{2}\log\big(2\pi\sigma_y^2\big) - \frac{(y_i - f(\boldsymbol{x}_i; \boldsymbol{m}))^2 + J(\boldsymbol{x}_i; \boldsymbol{m})\boldsymbol{S}J(\boldsymbol{x}_i; \boldsymbol{m})^\top}{2\sigma_y^2}. \tag{B.2}
$$

When considering a Categorical likelihood with C different classes, we estimate the expected log-likelihood term using Monte-Carlo integration as

$$
\mathbb{E}_{q_\phi(\boldsymbol{w})}[\log \mathrm{Cat}(y_i \,|\, \sigma(f_L(\boldsymbol{x}_i; \boldsymbol{w})))] = \frac{1}{K}\sum_{k=1}^K \sum_{c=1}^C \mathbb{I}[y_i = c] \left[f_L^c(\boldsymbol{x}_i; \boldsymbol{w}^{(k)}) - \log\left[\sum_{c'=1}^C \exp\left(f_L^{c'}(\boldsymbol{x}_i; \boldsymbol{w}^{(k)}) \right) \right] \right] \tag{B.3}
$$

where $\boldsymbol{w}^{(k)} \sim q_\phi(\boldsymbol{w})$ for $k = 1, \ldots, K$, $\mathbb{I}[\cdot]$ is the indicator function, $\sigma(\cdot)$ is the softmax function and $f_L^c(\,\cdot\,; \boldsymbol{w})$ is the logit for class c obtained from f_L.

Regularized KL divergence We estimate the regularized KL divergence using its consistent estimator (Eq. 2.9)

$$
\hat{D}_{\mathrm{KL}}^{\gamma}\big(\mathbb{Q}_\phi^F \,\|\, \mathbb{P}^F\big) = \frac{1}{2}(f(\boldsymbol{x}; \boldsymbol{m}) - \mu(\boldsymbol{x}))^\top (K(\boldsymbol{x}, \boldsymbol{x}) + \gamma M \mathbb{I}_M)^{-1} (f(\boldsymbol{x}; \boldsymbol{m}) - \mu(\boldsymbol{x}))
$$

$$+ \frac{1}{2} \operatorname{Tr}\left[(K(\boldsymbol{x},\boldsymbol{x}) + \gamma M \mathbb{I}_M)^{-1}(J(\boldsymbol{x};\boldsymbol{m}) S J(\boldsymbol{x};\boldsymbol{m})^\top + \gamma M \mathbb{I}_M) - \mathbb{I}_M\right]$$
$$- \frac{1}{2} \log\det\left[(K(\boldsymbol{x},\boldsymbol{x}) + \gamma M \mathbb{I}_M)^{-1}(J(\boldsymbol{x};\boldsymbol{m}) S J(\boldsymbol{x};\boldsymbol{m})^\top + \gamma M \mathbb{I}_M)\right] \quad \text{(B.4)}$$

with measurement points $\boldsymbol{x} = \{\boldsymbol{x}^{(i)}\}_{i=1}^M$, $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)} \overset{\text{i.i.d}}{\sim} \rho(\boldsymbol{x})$ sampled from a probability measure on $\mathcal{X}$.

Computational complexity Evaluating the objective in Eq. B.1 has complexity $O(BKC+M^3)$ for Categorical likelihoods and $O(B+M^3)$ for Gaussian likelihoods, where B is the batch size, K the number of variational posterior samples, C the number of classes, and M the number of context points. The first term corresponds to the expected log-likelihood in our objective and the second term to the regularized KL divergence estimator. We note that evaluating the linearized neural network can be efficiently done in about 3x the cost of one forward pass using the Jacobian-vector product computational primitive.

C ADDITIONAL DETAILS ON THE EXPERIMENTAL SETUP

C.1 EXPERIMENTS ON SYNTHETIC DATA

Regression We consider the following generative model for the toy data

$$y_i = \sin(2\pi x_i) + \epsilon \quad \text{with} \quad \epsilon \sim \mathcal{N}\big(0, \sigma_n^2\big) \quad \text{(C.1)}$$

and draw $x_i \sim \mathcal{U}([-1,-0.5] \cup [0.5,1])$. When not otherwise specified, we use $\sigma_n = 0.1$. On the plots, the data points are shown as gray circles, inferred mean functions as red lines, their 2-standard-deviations interval around the mean in light green, and functions sampled from the approximate posterior as green lines. In general, we consider two hidden-layer BNNs with 30 neurons per layer and hyperbolic tangent activation (Tanh) functions. Specifically in Figure 11, the small BNN has the same architecture as above while the large BNN has 100 neurons per layer. All the BNN baselines have the same architecture and fully-factorized Gaussian approximate posterior. The prior scale of TFSVI [Rudner et al., 2022b] is set to $\sigma_p = 0.2$ and $\sigma_p = 0.75$ for MFVI [Blundell et al., 2015] and Laplace [Immer et al., 2021]. For the Gaussian process posterior baseline, we fit the prior parameters by maximizing the log-marginal likelihood [Williams and Rasmussen, 2006]. Apart from the cases where the parameters of the GP prior used for GFSVI (our method) and FVI [Sun et al., 2019] are explicitly stated, we consider a constant zero-mean function and find the parameters of the covariance function by maximizing the log-marginal likelihood from mini-batches [Chen et al., 2022]. Except where otherwise stated, we estimate the functional KL divergences with 500 measurement points and use the regularized KL divergence with $\gamma = 10^{-10}$.

Classification We sample 100 data points perturbed by Gaussian noise with $\sigma_n = 0.1$ from the two moons data [Pedregosa et al., 2011]. On the plots, the data points are shown as red (class 0) and blue (class 1) dots. We plot the mean and 2-standard-deviations of the probability that $\boldsymbol{x}$ belongs to class 1 with respect to the posterior (i.e. $p(y = 1 \mid \boldsymbol{w}^{(k)}, \boldsymbol{x})$) which we estimate from $K = 100$ samples $\boldsymbol{w}^{(k)} \sim q_\phi(\boldsymbol{w})$ for $k = 1, \ldots, K$. We consider two hidden-layer BNNs with 100 neurons per layer and hyperbolic tangent activation (Tanh) functions. All the BNN baselines have the same architecture and fully-factorized Gaussian approximate posterior. The prior scale of MFVI [Blundell et al., 2015] is set to $\sigma_p = 0.8$ and $\sigma_p = 1.0$ for TFSVI [Rudner et al., 2022b] and Laplace [Immer et al., 2021]. For the Gaussian process posterior baseline, we approximate the intractable posterior using the Laplace approximation and find the prior parameters by maximizing the log-marginal likelihood [Williams and Rasmussen, 2006]. The GP prior for GFSVI (our method) and FVI [Sun et al., 2019] has a constant zero-mean function and we find the parameters of the covariance function by maximizing the log-marginal likelihood from mini-batches [Chen et al., 2022] using the method to transform classifications labels into regression targets from Milios et al. [2018]. We estimate the functional KL divergences with 500 measurement points and use the regularized KL divergence with $\gamma = 10^{-10}$.

C.2 OCEAN CURRENT MODELING EXPERIMENT

Following Cinquin et al. [2024], we apply the Helmholtz decomposition to the neural network f as

$$f(\cdot, \boldsymbol{w}) = \operatorname{grad} \Phi(\cdot, \boldsymbol{w}_1) + \operatorname{rot} \Psi(\cdot, \boldsymbol{w}_2) \quad \text{(C.2)}$$

where $\boldsymbol{w} = \{\boldsymbol{w}_1, \boldsymbol{w}_2\}$ and, $\Phi(\cdot, \boldsymbol{w}_1)$ and $\Psi(\cdot, \boldsymbol{w}_2)$ are 2-layer fully-connected neural networks with 50 hidden units per layer and hyperbolic tangent activation functions. GFSVI and TFSVI both use 160 fixed context points. The prior scale

Table 5: UCI regression dataset description

DATASET	BOSTON	NAVAL	POWER	PROTEIN	YACHT	CONCRETE	ENERGY	KIN8NM	WINE	WAVE	DENMARK
NUMBER SAMPLES	506	11 934	9 568	45 730	308	1 030	768	8 192	1 599	288 000	434 874
NUMBER FEATURES	13	16	4	9	6	8	8	8	11	49	2

of TFSVI is set to $\sigma_p = 0.5$. We fit the neural networks on the entire dataset and average the scores with respect to five different random seeds.

C.3 REGRESSION EXPERIMENTS WITH TABULAR DATA

Datasets and pre-processing We evaluate the predictive performance of our model on regression datasets from the UCI repository [Dua and Graff, 2017] described in Table 5. These datasets are also considered in Sun et al. [2019], Wild et al. [2022b] but we include two additional larger ones (Wave and Denmark). We perform 5-fold cross validation, leave out one fold for testing, consider 10% of the remaining 4 folds as validation data and the rest as training data. We report mean and standard-deviation of the average expected log-likelihood and average mean square error on the test fold. We also report the mean rank of the methods across all datasets by assigning rank 1 to the best scoring method as well as any method who's error bars overlap with the highest score's error bars, and recursively apply this procedure to the methods not having yet been assigned a rank. The expected log-likelihood is estimated by Monte Carlo integration when it is not available in closed form (MFVI, TFSVI and FVI) with 100 posterior samples. We preprocess the dataset by encoding categorical features as one-hot vectors and standardizing the features and labels.

Baseline specification We compare our GFSVI method to two weight-space inference methods (mean-field variational inference [Blundell et al., 2015] and linearized Laplace [Immer et al., 2021]) and two function-space inference methods (FVI [Sun et al., 2019] and TFSVI [Rudner et al., 2022b]). While FVI uses GP priors, TFSVI performs inference in function space but with the pushforward to function space of the variational distribution and prior on the weights. We compute the function-space (regularized) KL divergence using a set of 500 measurement points sampled from a uniform distribution for GFSVI and TFSVI, and 50 points drawn from a uniform distribution along with 450 samples from the training batch for FVI as specified in Sun et al. [2019]. All the BNN baselines have the same architecture and fully-factorized Gaussian approximate posterior. We also provide results with a GP [Williams and Rasmussen, 2006] when the size of the dataset allows it, and a sparse GP [Hensman et al., 2013]. As we restrict our comparison to BNNs, we do not consider the GP and sparse GP as baselines but rather as gold-standards. All models have a Gaussian homoskedastic noise model with a learned scale parameter. All the BNNs are fit using the Adam optimizer [Kingma and Ba, 2017] using a mini-batch size of 2000 samples. We also perform early stopping when the validation loss stops decreasing.

Model selection Hyper-parameter optimization is conducted using the Bayesian optimization tool provided by Wandb [Biewald, 2020]. BNN parameters are selected to maximize the average validation expected log-likelihood across the 5 cross-validation folds. We optimize over prior parameters (kernel and prior scale), learning-rate and activation function. We select priors for GFSVI, FVI, sparse GP and GP among the RBF, Matérn-1/2, Matérn-3/2, Matérn-5/2, Linear and Rational Quadratic covariance functions. The GP prior parameters used with GFSVI and FVI are selected by maximizing the log-marginal likelihood from batches as proposed by Chen et al. [2022] and done in Sun et al. [2019]. Hyper-parameters for GPs and sparse GPs (kernel parameters and learning-rate) are selected to maximize the mean log-marginal likelihood of the validation data across the 5 cross-validation folds.

C.4 CLASSIFICATION EXPERIMENTS WITH IMAGE DATA

Datasets and pre-processing We further evaluate the predictive performance of our model on classification tasks with the MNIST [LeCun et al., 2010] and Fashion MNIST [Xiao et al., 2017] image data sets. We fit the models on a random subset of 90% of the provided training split, consider the remaining 10% as validation data and evaluate on the provided test split. We repeat this procedure 5 times with different random seeds and report the mean and standard-deviation of the average expected log-likelihood, accuracy and expected calibration error (ECE) of the mean of the predictive distribution on the test set. The expected log-likelihood is estimated by Monte Carlo integration with 100 posterior samples when it is not available in closed form (MFVI, TFSVI and FVI). We estimate the mean of the predictive distribution to compute the accuracy and the ECE with 100 posterior samples. We preprocess the dataset by standardizing the images.

Baseline specification We compare our GFSVI method to the same baselines as for the regression experiments (see C.3). All the BNN baselines have the same architecture and fully-factorized Gaussian approximate posterior. More specifically, we consider a CNN with three convolutional layers (with output channels 16, 32 and 64) before two fully connected layers (with output size 128 and 10). The convolutional layers use 3×3 shaped kernels. Each pair of convolutional layers is interleaved with a max-pooling layer. We consider three different measurement point distributions ρ to estimate the (regularized) KL divergence in GFSVI, FVI and TFSVI: RANDOM, RANDOM PIXEL and KMNIST. The RANDOM measurement point distribution is sampled from by drawing 50% of the samples from the training data batch and 50% of the samples from a uniform distribution over $[p_{min}, p_{max}]^{H \times W \times C}$, where H, W and C are respectively the height, width and number of channels of the images, and $p_{min} = v_{min} - 0.5 \times \Delta$ and $p_{max} = v_{max} + 0.5 \times \Delta$ where $\Delta = v_{max} - v_{min}$ is the difference between the minimal (v_{min}) and maximal (v_{max}) pixel values of the data set. The RANDOM PIXEL measurement point distribution is taken from Rudner et al. [2022b] and is sampled from by randomly choosing each pixel value among the ones available from the training data batch at the same position in the 28×28 pixel grid. Finally, the KMNIST measurement point distribution is also taken from Rudner et al. [2022b] and is drawn from by randomly sampling data points from the Kuzushiji-MNIST (KMNIST) dataset [Clanuwat et al., 2018]. The KMNIST dataset is a collection of 70'000 gray-scale images of size 28×28 which we preprocess by standardizing the images. We sample 25 measurement points when using RANDOM, 25 measurement points when using RANDOM PIXEL and 20 when using KMNIST. All the BNNs are trained using the Adam optimizer [Kingma and Ba, 2017] using a mini-batch size of 100. We also perform early stopping when the validation loss stops decreasing.

Model selection Hyper-parameter optimization is conducted just like for the regression tasks (see C.3). The Gaussian process prior parameters used with GFSVI and FVI are selected by maximizing the log-marginal likelihood from batches [Chen et al., 2022] using the method to transform classifications labels into regression targets from Milios et al. [2018]. We optimize the same hyper-parameters as for the regression experiments with the exception of the additional α_ϵ parameter introduced by Milios et al. [2018] for the function-space VI methods with GP priors (FVI and GFSVI).

C.5 OOD DETECTION

Tabular data with a Gaussian likelihood Following the setup from Malinin et al. [2021] we take epistemic uncertainty to be the variance of the mean prediction with respect to samples from the posterior. We consider the test data to be in-distribution (ID) data and a subset of the song dataset [Bertin-Mahieux et al., 2011] of equal length and with an equal number of features as out-of-distribution (OOD) data. We use the same preprocessing as for regression as well as the same baselines with the same hyper-parameters (see Appendix C.3). We first fit a model, then evaluate the extend by which the epistemic uncertainty under the model is predictive of the ID and OOD data using a single threshold obtained by a depth-1 decision tree fit to minimize the classification loss. We report the mean and standard error of the accuracy of the threshold to classify OOD from ID data based on epistemic uncertainty across the 5 folds of cross-validation. We also provide results obtained using a GP and sparse GP as gold standard.

Image data with a Categorical likelihood Following the setup by Osawa et al. [2019], we take the epistemic uncertainty to be the entropy of the mean of the predictive distribution with respect to samples from the posterior. We evaluate models trained on MNIST using MNIST's test split as ID data and a subset of the training set of Fashion MNIST as OOD data. Likewise, we evaluate models trained on Fashion MNIST using Fashion MNIST's test split as ID data and a subset of the training set of MNIST as OOD data. We use the same preprocessing as for classification, as well as the same baselines with the same hyper-parameters (see Appendix C.4). We first fit a model, then evaluate the extend by which the epistemic uncertainty under the model is predictive of the ID and OOD data using a single threshold obtained by a depth-1 decision tree fit to minimize the classification loss. We estimate mean of the predictive distribution by Monte-Carlo integration using 100 posterior samples. We report the mean and standard error of the accuracy of the threshold to classify OOD from ID data based on epistemic uncertainty for the 5 models trained on different random seeds (see Appendix C.4).

C.6 VARIATIONAL MEASURE EVALUATION

We evaluate our inference method by comparing the samples drawn from the exact posterior over functions with the approximate posterior obtained with our method (GFSVI). We follow the setup by Wilson et al. [2022] and we compute the average Wasserstein-2 metric between 1000 samples drawn from a GP posterior with a RBF kernel evaluated at the test points, and samples from the approximate posterior of GFSVI, sparse GP and FVI evaluated at the same points and with the same prior. We consider the Boston, Concrete, Energy, Wine and Yacht datasets for which the exact GP posterior can be

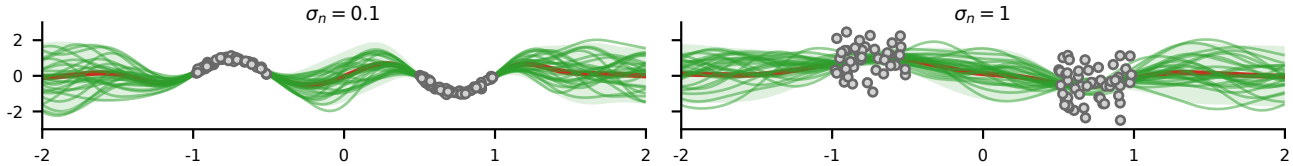


Figure 4: Our method (GFSVI) effectively regularizes functions generated by the Bayesian neural network (BNN) both in settings where the generative process is very noisy ($\sigma_n = 1$) or not ($\sigma_n = 0.1$).

computed and use the same preprocessing as for regression (see Appendix C.3). We report the mean and standard error of the average Wasserstein-2 metric across the 5 folds of cross-validation. The Wasserstein-2 metric is computed using the Python Optimal Transport library [Flamary et al., 2021].

Baseline specification FVI and GFSVI have the same two hidden layer neural network architecture with 100 neurons each and hyperbolic tangent activation. These models are fit with the same learning rate and set of 500 measurement points jointly sampled from a uniform distribution over the feature-space and mini-batch of training samples. We use $\gamma = 10^{-15}$ for the regularized KL divergence. We further consider a sparse GP with 100 inducing points.

C.7 SOFTWARE

We use the JAX [Bradbury et al., 2018] and DM-Haiku [Hennigan et al., 2020] Python libraries to implement our Bayesian neural networks. MFVI, linearized Laplace and TFSVI were implemented based on the information in the papers, and code for FVI was adapted to the JAX library from the implementation provided by the authors. We further use the GPJAX Python library for experiments involving Gaussian processes [Pinder and Dodd, 2022].

C.8 HARDWARE

All models were fit using a single NVIDIA RTX 2080Ti GPU with 11GB of memory.

D ADDITIONAL EXPERIMENTAL RESULTS

In this section, we present additional figures for our qualitative uncertainty evaluation as well as further experimental results on regression, out-of-distribution detection and robustness under input distribution shift tasks. We also provide plots illustrating the eigenvalue decay of different kernels, and figures showing the influence of γ in the regularized KL divergence.

D.1 QUALITATIVE UNCERTAINTY EVALUATION

Regression We further find that our method (GFSVI) provides strong regularization when the data generative process is noisy (see Figure 4) and is more robust than FVI to situations where ones computational budget constrains the number of measurement points M to be small (Figure 8). In contrast to FVI, GFSVI accurately approximates the exact GP posterior under rough (Matérn-1/2) GP priors effectively incorporating prior knowledge defined by the GP prior to the inference process (see Figure 2). Likewise, GFSVI adapts to the variability of the functions specified by the kernel (see Figure 6). We also find that GFSVI requires a larger number of measurement points to capture the behavior of a rougher prior (see Figure 7).

Classification We find that GFSVI better captures the beliefs induced by the smooth RBF and rough Matérn-1/2 Gaussian process priors compared to FVI (see Figures 9 and 10). Moreover, GFVSI both accurately fits the training data and shows greater uncertainty outside of its support relative to BNNs baselines with weight-space and function-space priors. Unlike for the toy data regression experiments where the GP posterior was the ground truth, the Laplace (approximate) GP posterior in Figures 9 and 10 only represents a possible approximation to the now in-tractable posterior (due to the softmax inverse link function). Thus the GP should not be considered as the ground truth nor as the optimal approximation in the classification

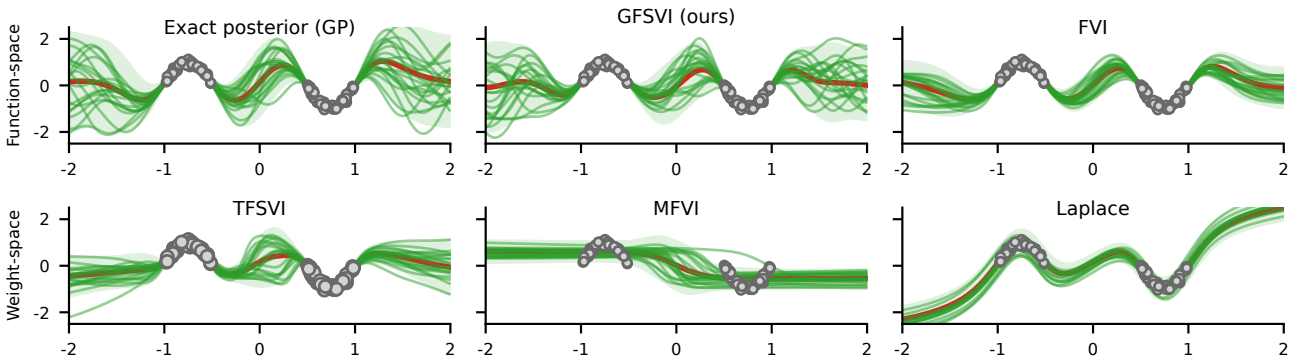


Figure 5: Our method (GFSVI) with an RBF Gaussian process (GP) prior accurately approximates the exact GP posterior unlike the function-space prior baseline (FVI). Weight-space prior baselines do not provide a straight-forward mechanism to incorporate prior assumptions regarding the functions generated by BNNs and underestimate the epistemic uncertainty (MFVI, Laplace). The lower row is identical to the one in Figure 2 in the main text and is reproduced here to make comparison easier.

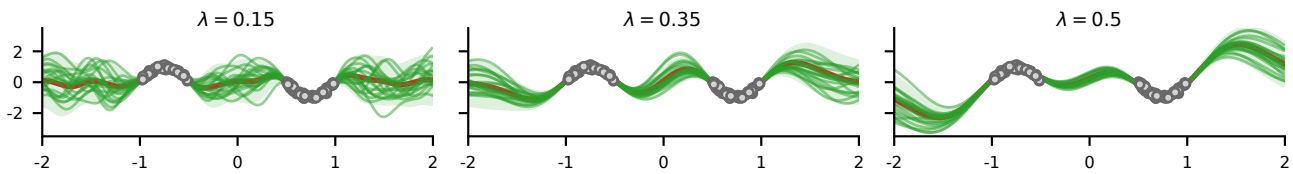


Figure 6: Our method (GFSVI) allows to incorporate prior beliefs in terms of function variability using the characteristic length-scale parameter of the Gaussian process (GP) prior. GFSVI was fit using a GP prior with RBF covariance function.

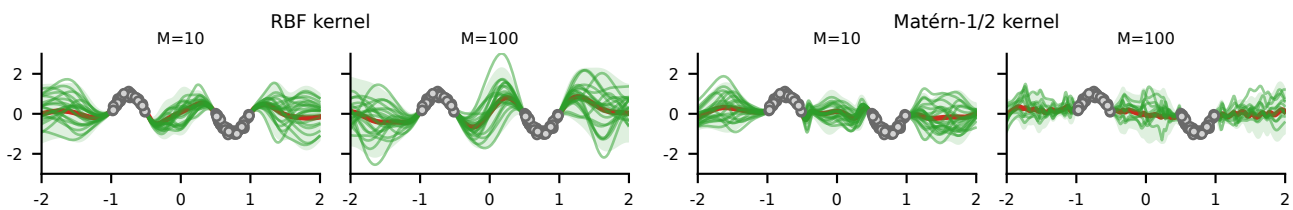


Figure 7: Our method (GFSVI) captures the smooth behavior of a Gaussian process (GP) prior with RBF covariance function even if the number of measurement points is small (M=10). However, in that setting GFSVI fails to reproduce the rough effect of a GP prior with a Matérn-1/2 covariance function, and requires a larger amount of measurement points to do so (M=100).

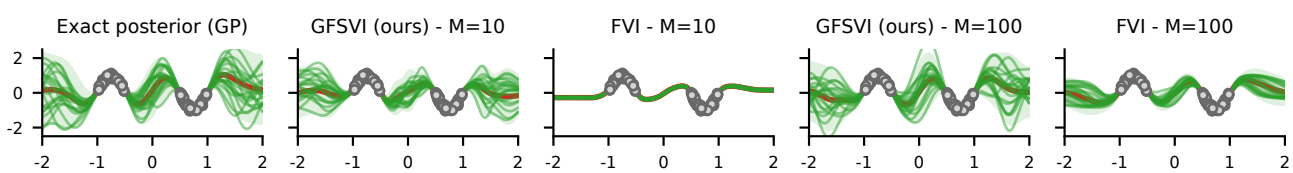


Figure 8: Our method (GFSVI) already provides a reasonable approximation to the exact posterior with small numbers of measurement points (M=10) while function-space baseline FVI requires many more (M=100).

setting, but is nevertheless useful to give a idea of the level of uncertainty a BNN with a GP prior should provide outside of the support of the data.

Inductive biases Figure 11 compares GFSVI to the exact posterior across two different priors and three model architectures (details in C.1). We find that the BNN's ability to incorporate the beliefs introduced by the GP prior depends on its size and activation function. When using piece-wise linear activations (ReLU), small models are prone to underfitting for smooth priors (RBF), and to collapsing uncertainty for rough priors (Matérn-1/2). By contrast, when using smooth activations (Tanh), smaller models suffice, and they are compatible with most standard GP priors (the results shown in Figure 11 extend

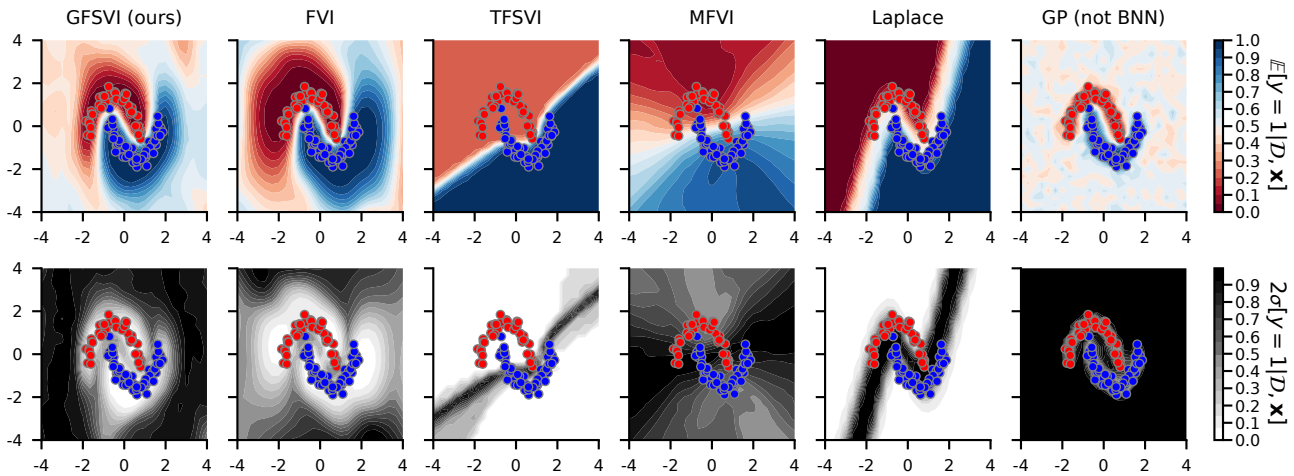


Figure 9: Our method (GFSVI) with a RBF Gaussian process (GP) prior accurately captures the smooth decision boundary induced by the prior and shows high uncertainty outside of the data support. Weight-space baselines do not provide a straight-forward mechanism to incorporate prior assumptions regarding the functions generated by BNNs and underestimate the epistemic uncertainty (TFSVI, Laplace) or underfit the data (MFVI).

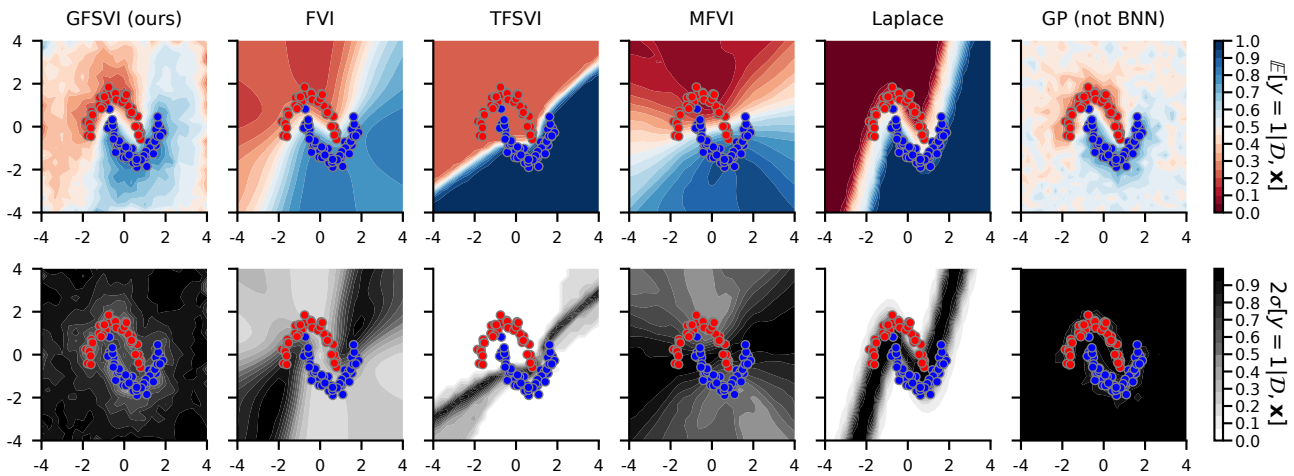


Figure 10: Our method (GFSVI) with a Matérn-1/2 Gaussian process (GP) prior accurately captures the rough decision boundary unlike the function-space baseline (FVI). Weight-space baselines do not provide a straight-forward mechanism to incorporate prior assumptions regarding the functions generated by BNNs and underestimate the epistemic uncertainty (TFSVI, Laplace) or underfit the data (MFVI).

to RBF, Matérn family, and Rational Quadratic in our experiments). We also analyzed how the number M of measurement points affects performance. Figures 7 and 17 show that capturing the properties of rough GP priors and estimating D_{KL}^{γ} with these priors requires larger M.

D.2 REGRESSION ON TABULAR DATA

We present additional regression results reporting the mean square error (MSE) of evaluated methods across the considered baselines, see Table 6. We find that GFSVI also performs competitively in terms of MSE compared to baselines and obtains the best mean rank, matching best the performing methods on nearly all datasets. In particular, we find that using GP priors in the linearized BNN setup with GFSVI yields improvements over the weight-space priors used in TFSVI and that GFSVI performs slightly better than FVI. Function-space VI methods (TFSVI, GFSVI, FVI) significantly improves over weight-space VI mostly performing similarly to the linearized Laplace approximation. Further improvement over baselines are obtained when considering GP priors with GFSVI and FVI. Finally, GFSVI compares favorably to the GP and sparse GP.

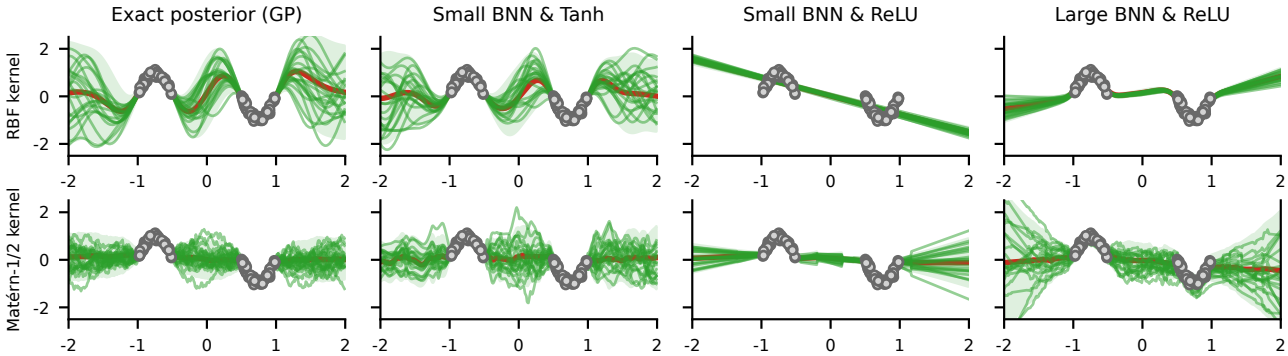


Figure 11: Our method requires that the Bayesian neural network (BNN) and Gaussian process (GP) prior share similar inductive biases to provide an accurate approximation to the exact posterior.

Table 6: Test mean square error (MSE) of evaluated methods on regression datasets. We find that GFSVI (ours) also performs competitively in terms of MSE compared to baselines and obtains the best mean rank, matching best the performing methods on nearly all datasets.

Dataset	Function-space priors		Weight-space priors				Gaussian Processes (Gold Standards)		
	GFSVI (ours)	FVI	TFSVI	MFVI	VIP	Laplace	GWI	Sparse GP	GP
Boston	**0.123 ± 0.021**	**0.136 ± 0.022**	0.995 ± 0.092	0.532 ± 0.072	0.201 ± 0.056	0.203 ± 0.047	0.273 ± 0.069	0.122 ± 0.014	0.115 ± 0.020
Concrete	**0.114 ± 0.008**	**0.116 ± 0.004**	0.389 ± 0.015	0.698 ± 0.046	**0.109 ± 0.008**	**0.116 ± 0.007**	0.145 ± 0.017	0.399 ± 0.020	0.116 ± 0.007
Energy	0.003 ± 0.000	0.003 ± 0.000	0.003 ± 0.000	0.152 ± 0.024	0.043 ± 0.036	**0.002 ± 0.000**	0.003 ± 0.001	0.087 ± 0.005	0.087 ± 0.004
Kin8nm	0.071 ± 0.001	0.075 ± 0.003	0.073 ± 0.001	0.290 ± 0.111	**0.068 ± 0.002**	0.083 ± 0.001	0.071 ± 0.001	0.088 ± 0.002	*(infeasible)*
Naval	**0.000 ± 0.000**	**0.001 ± 0.001**	**0.000 ± 0.000**	0.007 ± 0.003	0.002 ± 0.000	**0.000 ± 0.000**	0.197 ± 0.174	0.000 ± 0.000	*(infeasible)*
Power	**0.052 ± 0.001**	**0.054 ± 0.002**	0.054 ± 0.001	0.058 ± 0.002	**0.054 ± 0.002**	**0.054 ± 0.002**	0.052 ± 0.001	0.071 ± 0.001	*(infeasible)*
Protein	0.459 ± 0.005	0.466 ± 0.004	**0.429 ± 0.004**	0.537 ± 0.008	**0.421 ± 0.005**	0.446 ± 0.006	0.425 ± 0.003	0.408 ± 0.002	*(infeasible)*
Wine	**0.652 ± 0.022**	0.663 ± 0.009	1.297 ± 0.093	**0.655 ± 0.023**	**0.627 ± 0.013**	0.637 ± 0.031	0.682 ± 0.048	0.607 ± 0.033	0.585 ± 0.032
Yacht	**0.003 ± 0.001**	0.004 ± 0.001	0.221 ± 0.037	0.682 ± 0.140	0.004 ± 0.001	**0.002 ± 0.001**	0.008 ± 0.003	0.399 ± 0.064	0.355 ± 0.030
Wave	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**	0.001 ± 0.001	0.000 ± 0.000	*(infeasible)*
Denmark	**0.155 ± 0.004**	0.287 ± 0.003	0.163 ± 0.004	0.225 ± 0.003	0.189 ± 0.008	0.194 ± 0.003	0.197 ± 0.004	0.260 ± 0.001	*(infeasible)*
Mean rank	1.364	2.000	2.182	3.182	1.636	1.727	-	-	-

D.3 VARIATIONAL MEASURE EVALUATION

Table 7 evaluates our inference method by comparing samples drawn from the exact posterior (where computationally feasible) with the approximate posterior obtained with our method (GFSVI). We follow the setup by Wilson et al. [2022] and we compute the average per-sample Wasserstein-2 metric samples drawn from a GP posterior with RBF kernel evaluated at the test points, and samples from the approximate posterior of GFSVI, sparse GP and FVI evaluated at the same points and with the same prior. More details are provided in Appendix C.6. We find that GFSVI approximates the exact posterior more accurately that FVI, obtaining a higher mean rank, but worse than the gold standard sparse GP, which demonstrates to be most accurate.

Table 7: Average point-wise Wasserstein-2 distance (lower is better) between exact and approximate posterior of reported methods. GFSVI (ours) provides a more accurate approximation than FVI.

Dataset	Boston	Concrete	Energy	Wine	Yacht	Mean rank
GFSVI (ours)	**0.0259 ± 0.0040**	**0.0499 ± 0.0029**	**0.0035 ± 0.0004**	**0.0571 ± 0.0097**	**0.0036 ± 0.0006**	1.0
FVI	0.0469 ± 0.0044	0.0652 ± 0.0037	**0.0037 ± 0.0004**	0.1224 ± 0.0167	**0.0052 ± 0.0013**	1.6
GP sparse	0.0074 ± 0.0022	0.0125 ± 0.0016	0.0042 ± 0.0003	0.0170 ± 0.0035	0.0035 ± 0.0008	-

D.4 OUT-OF-DISTRIBUTION DETECTION WITH IMAGE DATA

We here show an additional plot from our out-of-distribution detection experiment with image data (details in C.5). Figure 12 shows the (normalized) histograms of the entropy of the mean prediction produced by each model on the in-distribution (blue) and out-of-distribution (red) data sets considered in our OOD detection experiment. Methods which estimate the (regularized) KL-divergence in function-space (GFSVI, FVI and TFSVI) use the KMNIST measurement distribution. We find that the entropy produced by GFSVI on in-distribution data highly peaks around 0 while the entropy produced from

out-of-distribution data strongly concentrates around its maximum $\ln(10)$. GFSVI best partitions ID and OOD data based on predictive entropy improving over the function-space prior (FVI) and weight-space prior (TFSVI, MFVI, Laplace) BNN baselines (see Table 2).

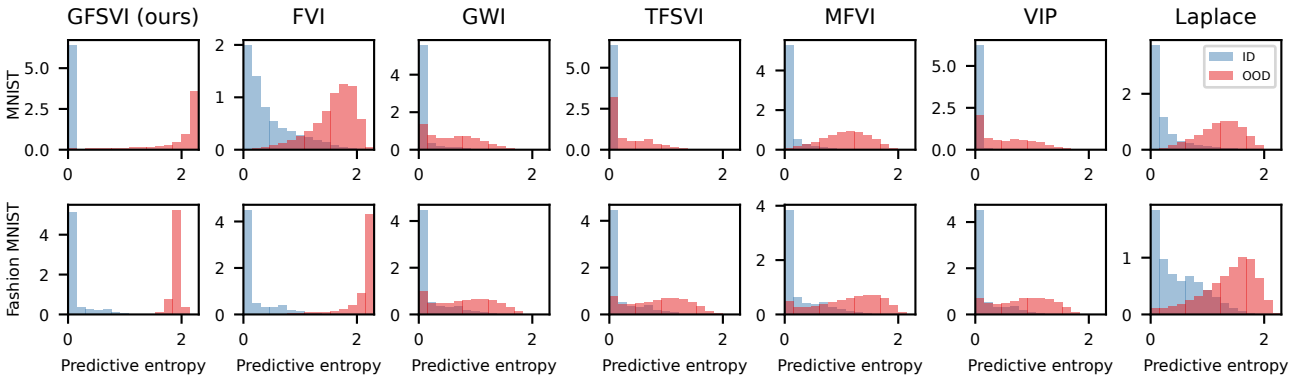


Figure 12: Histograms of the entropy of the mean predictive distribution produced by evaluated methods in the out-of-distribution detection with image data experiment. GFSVI (ours) best partitions in-distribution and out-of-distribution data based on the entropy of its mean predictive distribution.

D.5 INFLUENCE OF MEASUREMENT POINT DISTRIBUTION FOR IMAGE DATA

We present additional results evaluating the influence of the measurement point distribution ρ on the the performance of function-space inference methods when using high-dimensional image data. The measurement point distribution are described in Appendix C.4. Just like in Rudner et al. [2022b], we find that the choice of measurement point distribution may highly influence the OOD detection accuracy. While the expected log-likelihood, accuracy and expected calibration error (ECE) of a model generally remains comparable across measurement point distributions, the OOD accuracy of GFSVI is greatly improved by using samples from KMNIST to evaluate the (regularized) KL divergence. The measurement point distribution determines where the BNN is regularized and thus should be carefully selected especially for high dimensional data.

Table 8: Influence of the measurement point distribution ρ on expected log-likelihood (log-like.), accuracy (acc.), expected calibration error (ECE) and out-of-distribution detection accuracy (OOD acc.). ρ determines where the BNN will be regularized and strongly influences the out-of-distribution performance of the BNN.

Data	Metric	GFSVI			FVI			TFSVI		
		Random	Random Pixel	KMNIST	Random	Random Pixel	KMNIST	Random	Random Pixel	KMNIST
MNIST	Log-like. (↑)	**-0.033 ± 0.000**	-0.034 ± 0.000	-0.041 ± 0.000	-0.145 ± 0.005	-0.038 ± 0.000	-0.238 ± 0.006	-0.047 ± 0.003	**-0.032 ± 0.001**	-0.041 ± 0.001
	Acc. (↑)	**0.992 ± 0.000**	0.989 ± 0.000	0.991 ± 0.000	0.976 ± 0.001	0.988 ± 0.000	0.943 ± 0.001	0.989 ± 0.000	0.989 ± 0.000	0.989 ± 0.000
	ECE (↓)	**0.002 ± 0.000**	0.004 ± 0.000	0.006 ± 0.000	0.064 ± 0.001	0.003 ± 0.000	0.073 ± 0.003	0.007 ± 0.000	0.003 ± 0.000	0.006 ± 0.000
	OOD acc. (↑)	0.921 ± 0.008	0.868 ± 0.010	**0.980 ± 0.004**	0.894 ± 0.010	0.863 ± 0.003	0.891 ± 0.006	0.887 ± 0.011	0.861 ± 0.008	0.893 ± 0.005
FMNIST	Log-like. (↑)	**-0.260 ± 0.003**	**-0.258 ± 0.002**	-0.294 ± 0.006	-0.300 ± 0.002	-0.293 ± 0.003	-0.311 ± 0.005	-0.261 ± 0.001	**-0.258 ± 0.001**	-0.261 ± 0.002
	Acc. (↑)	**0.910 ± 0.001**	0.908 ± 0.001	**0.909 ± 0.001**	**0.910 ± 0.002**	0.900 ± 0.001	0.906 ± 0.002	**0.909 ± 0.001**	0.908 ± 0.001	0.907 ± 0.001
	ECE (↓)	**0.020 ± 0.003**	0.022 ± 0.001	0.042 ± 0.002	0.027 ± 0.005	**0.018 ± 0.002**	0.024 ± 0.002	0.022 ± 0.002	**0.018 ± 0.001**	**0.021 ± 0.002**
	OOD acc. (↑)	0.853 ± 0.005	0.867 ± 0.005	**0.997 ± 0.001**	0.925 ± 0.005	0.842 ± 0.006	0.975 ± 0.002	0.802 ± 0.006	0.800 ± 0.007	0.779 ± 0.010

D.6 INPUT DISTRIBUTION SHIFT WITH ROTATED IMAGE DATA

We here provide an experiment evaluating our method's (GFSVI) robustness in detecting input distribution shift. We expect the predictive uncertainty of a well-calibrated Bayesian model to be low for in-distribution data and to gradually increase as the input distribution shifts further away from the training data distribution. To test this property, we follow the setup by Sensoy et al. [2018], Rudner et al. [2022b] and assume like the related work that increasing the rotation angle of images gradually increases the level of input "distribution shift". We report the mean and standard-deviation of the average mean predictive entropy of models fit on MNIST [LeCun et al., 2010] and Fashion MNIST [Xiao et al., 2017] for increasingly large angles of rotation of their respective test data partition. We find that GFSVI is confident (low predictive entropy) for images with small rotation angles, and that its predictive entropy increases with the angle. GFSVI therefore exhibits the

expected behavior of a well-calibrated Bayesian model. We note that FVI, Laplace and MFVI tend to be under-confident (high predictive entropy) for small rotation angles, which might be a symptom of underfitting further supported by the results in Table 2. Also, with the exception of TFSVI, the predictive entropy of baselines across different rotation angles is generally higher than the one produced by GFSVI.

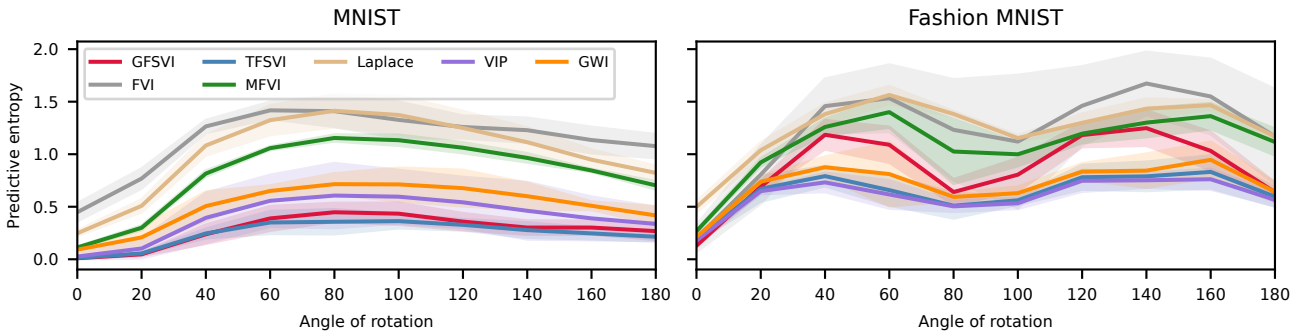


Figure 13: Average predictive entropy of models trained on MNIST and Fashion MNIST and evaluated for different rotation angles of their respective test data partitions. We see that our method (GFSVI) exhibits the behavior of a well-calibrated Bayesian model.

D.7 EXAMPLE OF MODEL MISSPECIFICATION WITH Wild et al. [2022b]

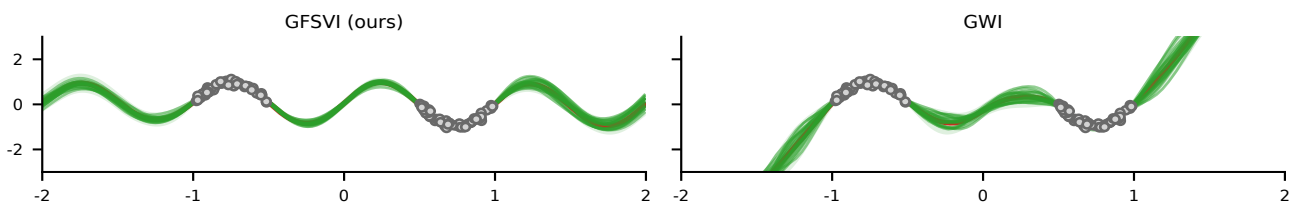


Figure 14: Example of model misspecification when using a periodic GP prior with baseline GWI [Wild et al., 2022b] that does not occur with our method (GFSVI). In GWI, only the posterior covariance is periodic, while the neural network parameterizing the posterior mean results in a function that does not capture the beliefs carried by the (periodic) prior. In contrast, our method accurately captures the GP prior's beliefs and yields a (locally) periodic function.

Figure 14 shows an example of model misspecification when using a periodic GP prior with the baseline GWI [Wild et al., 2022b]. As can be seen in the left panel of the figure, this problem does not occur with our method (GFSVI). While the posterior covariance in GWI reflects the periodicity of the prior, the neural network parametrizing the posterior mean does not result in a periodic function, i.e., the mean does not capture the beliefs specified by the periodic GP prior. In contrast, our method accurately captures the GP prior's beliefs and yields a (locally) periodic function.

D.8 CONVERGENCE SPEED ON UCI DATA

Table 9 shows the training time of our method and baselines MFVI [Blundell et al., 2015] and TFSVI [Rudner et al., 2022b] on the boston dataset using $M = 100$ context points averaged over 5 cross-validation splits, as well as Figure 15 showing the convergence of the validation expected log-likelihood on the boston dataset. Our method converges in more steps than the TFSVI. GFSVI typically takes more time/steps to train that TFSVI as it additionally needs to adapt its features to the beliefs specified by the Gaussian process prior.

Table 9: Training time of our method GFSVI and baselines MFVI [Blundell et al., 2015] and TFSVI [Rudner et al., 2022b] on the boston UCI dataset.

	GFSVI (OURS)	TFSVI	MFVI
TIME (S)	44.15 ± 1.56	36.36 ± 0.90	38.38 ± 10.80

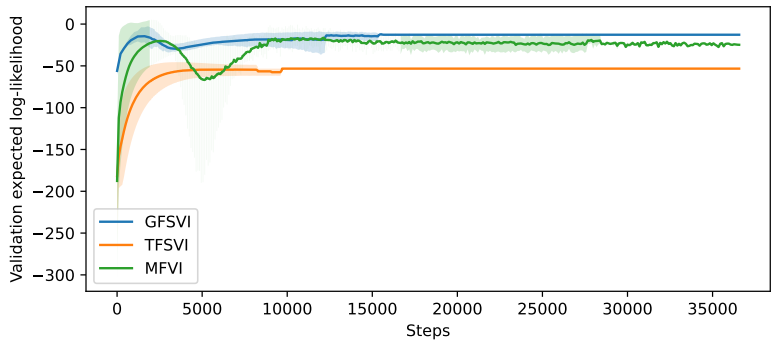


Figure 15: Validation expected log-likelihood of our method (GFSVI) and baselines TFSVI and MFVI. GFSVI (ours) converges on the boston dataset in slightly more steps than TFSVI but in fewer than MFVI.

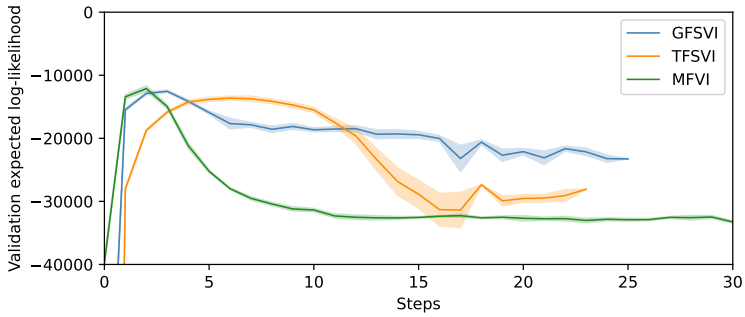
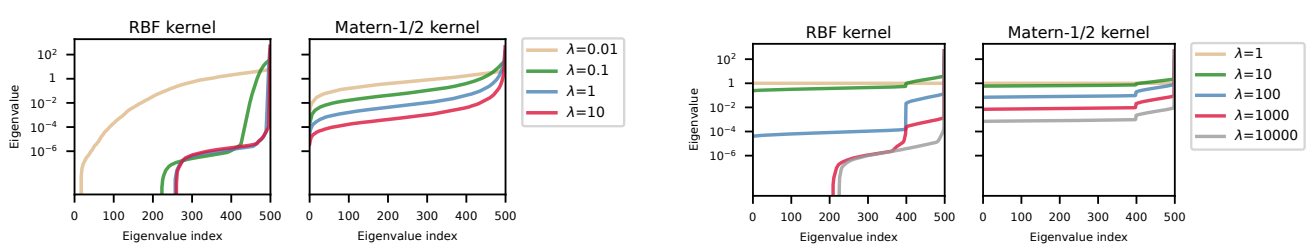


Figure 16: Validation expected log-likelihood of our method (GFSVI) and baselines TFSVI and MFVI on MNIST. GFSVI (ours) converges in slightly more steps than TFSVI but in fewer than MFVI.

D.9 ADDITIONAL PLOTS FOR KERNEL EIGENVALUE DECAY

Figure 17 shows a plot demonstrating the decay rate of the eigenvalues of RBF and Matérn-1/2 kernels evaluated at points sampled uniformly over $\mathcal{X}$. The rate of decay of covariance operator's eigenvalues gives important information about the smoothness of stationary kernels [Williams and Rasmussen, 2006] and that increased smoothness of the kernel leads to faster decay of eigenvalues Santin and Schaback [2016]. For instance, RBF covariance operator eigenvalues decay at near exponential rate independent of the underlying measure [Belkin, 2018] and Matérn kernels eigenvalues decay polynomialy [Chen et al., 2022]. We find that the kernel evaluated at points sampled from a uniform distribution over $\mathcal{X}$ share this same behavior (see Figure 17).



(a) $D = 1$ (b) $D = 100$

Figure 17: Mean eigenvalues of the Gram matrix obtained for different kernels and for varying length-scales over 10 draws from a uniform distribution on $[-2, 2]^D$. The mean eigenvalues are arranged in increasing order. The eigenvalues of the Gram matrix associated with the smooth RBF kernel decays much faster than those of the Matérn-1/2. Furthermore, the eigenvalues decay at a slower rate in high dimensions (D=100).

D.10 ADDITIONAL PLOTS FOR CHOOSING γ IN D_{KL}^{γ}

The γ parameter controls the magnitude of the regularized KL divergence (see Figure 20) and adjusts the relative weight of the regularized KL divergence and expected log-likelihood term in the training objective (see Figure 18). Furthermore, γ also acts as "jitter" preventing numerical errors. We recommend choosing γ large enough to avoid numerical errors while remaining small enough to provide strong regularization.

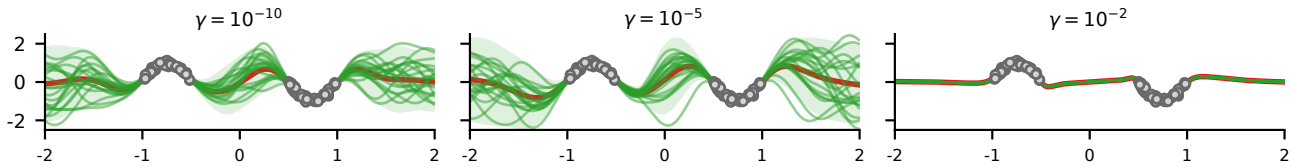


Figure 18: The γ parameter of the regularized KL divergence controls the magnitude of the regularizer in the objective and should be small enough to provide strong regularization.

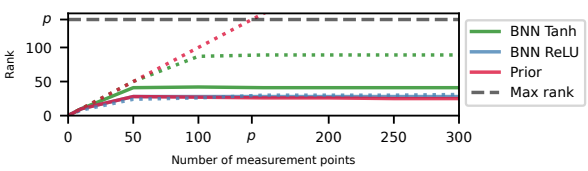


Figure 19: The BNN's covariance adaptation to the prior's covariance rank depends on its activation function. BNNs fit with a RBF prior (full) show lower rank than with a Matérn-1/2 (dotted).

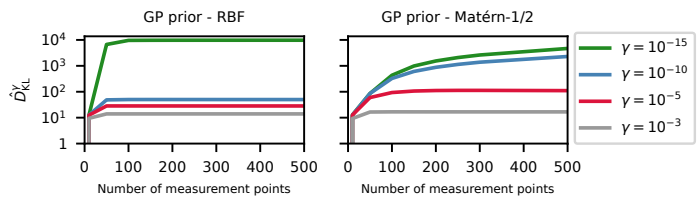


Figure 20: γ explicitly controls the magnitude of the regularized KL-divergence D_{KL}^{γ}. Rougher priors (Matérn-1/2) require more measurement points to accurately estimate D_{KL}^{γ} than smooth priors (RBF).