Tractable Function-Space Variational Inference in Bayesian Neural Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Reliable predictive uncertainty estimation plays an important role in allowing neural networks to be deployed in safety-critical settings. A popular approach for estimating the predictive uncertainty of neural networks is to treat the network parameters as random variables and infer an approximate posterior that can be used to obtain a distribution over network predictions. However, explicit inference over neural network parameters makes it difficult to incorporate meaningful prior information about the data generating process into training. In this paper, we pursue an alternative approach. Noting that stochastic neural networks define distributions over functions induced by distributions over parameters, we follow prior work in framing Bayesian inference in neural networks as inferring a posterior distribution over functions and propose a scalable function-space variational inference method that allows incorporating prior information and encourages reliable predictive uncertainty estimation. We show that the proposed method leads to state-of-the-art uncertainty estimation and predictive performance on a range of prediction problems, and demonstrate that it performs well on a challenging safety-critical medical diagnosis task in which reliable uncertainty estimation is essential.

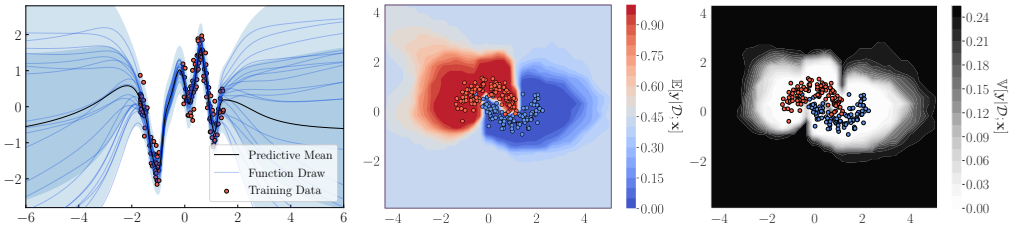
1 Introduction

Machine learning models succeed at an increasingly wide range of narrowly defined tasks [Krizhevsky et al., 2012, Mnih et al., 2013, Silver et al., 2016, Jumper et al., 2021], but fail without warning when used on inputs that are meaningfully different from the data they were trained on [Amodei et al., 2016, Hendrycks et al., 2021]. To deploy machine learning models in safety-critical environments where failures are costly or may endanger human lives, machine learning methods must be reliable and have the ability to 'fail gracefully.' As a promising tool for incorporating fail-safe mechanisms into machine learning systems, predictive uncertainty quantification allows machine learning models to express their confidence in the correctness of their predictions Gal and Ghahramani [2016].

In this paper, we develop a method for obtaining reliable uncertainty estimates in Bayesian neural networks (BNNs, Neal [1996]). While BNNs have promised to combine the advantages of deep learning and Bayesian inference, existing approaches for approximate inference in BNNs fall short of this promise and have been demonstrated to result in approximate posterior predictive distributions that underperform non-Bayesian methods both in terms of predictive accuracy and uncertainty quantification—making them of limited use in practice [Ovadia et al., 2019, Foong et al., 2019, Farquhar et al., 2020a, Band et al., 2021]. A potential reason for this shortcoming is that commonly used parameter-space inference methods make it difficult to define meaningful priors that effectively incorporate information about the data-generating process into inference.

To avoid this limitation, we follow the foundation work by Sun et al. [2019] and consider a variational objective defined explicitly in terms of distributions over *functions* induced by distributions over pa-

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) Predictive Distribution **(b)** Predictive Mean **(c)** Predictive Variance

Figure 1: 1D regression on the *Snelson* dataset and binary classification on the *Two Moons* dataset. The plots show the predictive distributions of a BNNs, obtained via function-space variational inference (FSVI). For further illustrative exampled and comparisons to deep ensembles and BNNs learned via parameter-space variational inference, see Appendix B.

rameters. In contrast to prior works that rely on approximation techniques that prevent such function-space variational objectives to be used with high-dimensional inputs and highly-overparameterized neural networks [Sun et al., 2019], we propose a simple estimator of the Kullback-Leibler divergence between distributions over functions that allows for stochastic variational inference. The proposed estimation procedure allows defining priors that explicitly encourage high predictive uncertainty away from the training data as well as priors that reflect relevant information about the task at hand.

We demonstrate that this approach leads to posterior approximations that exhibit significantly improved predictive uncertainty estimates compared to a wide array of state-of-the-art Bayesian and non-Bayesian methods. Figure 1 shows examples of predictive distributions obtained via function-space variational inference on low-dimensional, easy-to-visualize datasets. As can be seen in the figures, the predictive distributions fit the training data well while also exhibiting a high degree of predictive uncertainty in parts of the input space far away from the training data, as desired. Anonymized code to reproduce our results is available at `https://github.com/fsvibnns/fsvi`.

Contributions. We propose a simple estimation procedure for performing function-space variational inference in BNNs. The variational method allows for the incorporation of meaningful prior information about the data-generating process into the inference and produces reliable predictive uncertainty estimates. We perform a thorough empirical evaluation in which we compare the proposed approach to a wide array of competing methods and show that it consistently results in high predictive performance and reliable predictive uncertainty estimates, outperforming other methods in terms of predictive accuracy, robustness to distribution shifts, and uncertainty-based detection of distributionally-shifted data samples. We evaluate the proposed method on standard benchmarking datasets as well as on a safety-critical medical diagnosis task in which reliable uncertainty estimation is essential.

2 Preliminaries

We consider supervised learning tasks on data $\mathcal{D} \doteq \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N = (\mathbf{X}_\mathcal{D}, \mathbf{y}_\mathcal{D})$ with inputs $\mathbf{x}_n \in \mathcal{X} \subseteq \mathbb{R}^D$ and targets $\mathbf{y}_n \in \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}^Q$ for regression and $\mathcal{Y} \subseteq \{0, 1\}^Q$ for classification tasks. Bayesian neural networks (BNNs) are stochastic neural networks trained using (approximate) Bayesian inference. Denoting the parameters of such a stochastic neural network by the multivariate random variable $\mathbf{\Theta} \in \mathbb{R}^P$ and letting the function mapping defined by a neural network architecture be given by $f : \mathcal{X} \times \mathbb{R}^P \to \mathcal{Y}$, we obtain a random function $f(\cdot\,; \mathbf{\Theta})$. For a parameter realization $\boldsymbol{\theta}$, we obtain a corresponding function realization, $f(\cdot\,; \boldsymbol{\theta})$. When evaluated at a finite collection of points $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^m$, $f(\mathbf{X}; \mathbf{\Theta})$ is a multivariate random variable and $f(\mathbf{X}; \boldsymbol{\theta})$ is a vector.

Letting $p(\mathbf{y}_\mathcal{D} \mid f(\mathbf{X}_\mathcal{D}; \boldsymbol{\theta}))$ be the likelihood of observing the targets $\mathbf{y}$ under the stochastic function $f(\cdot\,; \mathbf{\Theta})$ evaluated at inputs $\mathbf{X}_\mathcal{D}$ and letting $p_\mathbf{\Theta}$ be a prior distribution over the stochastic network parameters $\mathbf{\Theta}$, we can use Bayesian inference for finding the conditional distribution over parameters given the observed data, $p_{\mathbf{\Theta}|\mathcal{D}}$ [MacKay, 1992, Neal, 1996]. However, since the mapping f is a non-linear function of the stochastic parameters $\mathbf{\Theta}$, exact inference is analytically intractable. Variational inference is an approximate method that seeks to sidestep this intractability by framing posterior inference as a variational optimization problem [Wainwright and Jordan, 2008]. Following this approach,

we can obtain a Bayesian neural network defined in terms of a variational distribution over parameters $q_{\boldsymbol{\Theta}}$ by solving $\min_{q_{\boldsymbol{\Theta}} \in \mathcal{Q}_{q_{\boldsymbol{\Theta}}}} \mathbb{D}_{\mathrm{KL}}(q_{\boldsymbol{\Theta}} \,\|\, p_{\boldsymbol{\Theta}|\mathcal{D}})$, where $\mathcal{Q}_{q_{\boldsymbol{\Theta}}}$ is a variational family of distributions. If $\mathcal{Q}_{q_{\boldsymbol{\Theta}}}$ is the family of mean-field Gaussian distributions and the prior distribution over parameters $p_{\boldsymbol{\Theta}}$ given by a diagonal Gaussian distribution, the resulting variational objective is amenable to stochastic variational inference and can be optimized using gradient-based methods. [Blundell et al., 2015, Graves, 2011, Hinton and van Camp, 1993, Hoffman et al., 2013].

2.1 A Function-Space Perspective on Variational Inference in Bayesian Neural Networks

By framing variational inference in BNNs as computing an approximation to a posterior distribution over *functions*, $p_{f(\cdot\,;\boldsymbol{\Theta})|\mathcal{D}}$, we can incorporate meaningful prior information about the underlying data-generating process into training. Specifically, we can express the problem of finding a posterior distribution over functions variationally as

$$\min_{q_{\boldsymbol{\Theta}} \in \mathcal{Q}_{q_{\boldsymbol{\Theta}}}} \mathbb{D}_{\mathrm{KL}}(q_{f(\cdot\,;\boldsymbol{\Theta})} \,\|\, p_{f(\cdot\,;\boldsymbol{\Theta})|\mathcal{D}}),$$

where $p_{f(\cdot;\boldsymbol{\Theta})}$ the a prior distribution over functions induced by a prior distribution over parameters $p_{\boldsymbol{\Theta}}$, and $q_{f(\cdot\,;\boldsymbol{\Theta})}$ is a variational distribution over functions induced by a variational distribution over parameters $q_{\boldsymbol{\Theta}}$. As discussed by Burt et al. [2021], this variational objective is guraanteed to be well-defined for suitably chosen prior distributions over functions. Specifically, the KL divergence between two distributions over functions generated from different distributions over parameters applied to the same mapping (e.g., the same neural network architecture) is well-defined (i.e., finite) if the KL divergence between the distributions over parameters is finite, since, by the strong data processing inequality [Polyanskiy and Wu, 2017],

$$\mathbb{D}_{\mathrm{KL}}(q_{f(\cdot\,;\boldsymbol{\Theta})} \,\|\, p_{f(\cdot;\boldsymbol{\Theta})}) \leq \mathbb{D}_{\mathrm{KL}}(q_{\boldsymbol{\Theta}} \,\|\, p_{\boldsymbol{\Theta}}). \tag{1}$$

As a result, if $\mathbb{D}_{\mathrm{KL}}(q_{\boldsymbol{\Theta}} \,\|\, p_{\boldsymbol{\Theta}}) < \infty$, which is the case for finite-dimensional parameter vectors $\boldsymbol{\Theta}$ and $q_{\boldsymbol{\Theta}}$ absolutely continuous with respect to $p_{\boldsymbol{\Theta}}$, then the function-space KL divergence is finite and thus well-defined as a variational objective.

Hence, for a likelihood function defined on a finite set of training targets $\mathbf{y}_{\mathcal{D}}$ and a suitably defined prior distribution over functions, we can express the variational problem above equivalently as the well-defined maximization problem $\max_{q_{\boldsymbol{\Theta}} \in \mathcal{Q}_{\boldsymbol{\theta}}} \mathcal{F}(q_{\boldsymbol{\Theta}}, \mathbf{X}_{\mathcal{D}}, \mathbf{y}_{\mathcal{D}})$ with

$$\mathcal{F}(q_{\boldsymbol{\Theta}}, \mathcal{D}) \doteq \mathbb{E}_{q_{f(\mathbf{X}_{\mathcal{D}};\boldsymbol{\Theta})}}[\log p(\mathbf{y}_{\mathcal{D}} \mid f(\mathbf{X}_{\mathcal{D}}; \boldsymbol{\theta}))] - \mathbb{D}_{\mathrm{KL}}(q_{f(\cdot;\boldsymbol{\Theta})} \,\|\, p_{f(\cdot;\boldsymbol{\Theta})}), \tag{2}$$

where $\mathbb{D}_{\mathrm{KL}}(q_{f(\cdot;\boldsymbol{\Theta})} \,\|\, p_{f(\cdot;\boldsymbol{\Theta})})$ is also a KL divergence between distributions over functions.

Unfortunately, evaluating the KL divergence in Equation (2) is intractable in general. To obtain a tractable objective, Sun et al. [2019] show that $\mathbb{D}_{\mathrm{KL}}(q_{f(\cdot;\boldsymbol{\Theta})} \,\|\, p_{f(\cdot;\boldsymbol{\Theta})})$ can be expressed as the supremum of the KL divergence from $q_{f(\cdot;\boldsymbol{\Theta})}$ to $p_{f(\cdot;\boldsymbol{\Theta})}$ over all *finite* sets of evaluation points, resulting in the objective function

$$\mathcal{F}(q_{\boldsymbol{\Theta}}, \mathcal{D}) = \mathbb{E}_{q_{f(\mathbf{X}_{\mathcal{D}};\boldsymbol{\Theta})}}[\log p(\mathbf{y}_{\mathcal{D}} \mid f(\mathbf{X}_{\mathcal{D}}; \boldsymbol{\theta}))] - \sup_{\mathbf{X} \in \mathcal{X}_{\mathbb{N}}} \mathbb{D}_{\mathrm{KL}}(q_{f(\mathbf{X};\boldsymbol{\Theta})} \,\|\, p_{f(\mathbf{X};\boldsymbol{\Theta})}), \tag{3}$$

where $\mathcal{X}_{\mathbb{N}} \doteq \bigcup_{n \in \mathbb{N}} \{\mathbf{X} \in \mathcal{X}_n \mid \mathcal{X}_n \subseteq \mathbb{R}^{n \times D}\}$ is the collection of all finite sets of evaluation points. However, this objective function is still challenging to optimize in practice: The supremum cannot be obtained analytically and—more problematically—the KL divergence term itself is analytically intractable and difficult to estimate—even for a single evaluation point. While the supremum can be estimated via Monte Carlo sampling, the fact that the KL divergence between BNNs cannot be easily estimated poses has prevented the application of function-space variational inference to challenging prediction tasks. For example, Sun et al. [2019] attempted to circumvent estimating the KL divergence between BNNs directly by instead approximating the gradients of the KL divergence with respect to the variational parameters, but this approach does not scale to large neural networks or high-dimensional input data.

In the next section, we propose an estimation procedure that allows scaling function-space variational inference to large networks and high-dimensional input data.

3 A Tractable Function-Space Variational Objective

The primary obstacle to making the objective in Equation (2) tractable is the KL divergence from $q_{f(\cdot;\boldsymbol{\Theta})}$ to $p_{f(\cdot;\boldsymbol{\Theta})}$. In this section, we describe how to obtain an estimator of the KL divergence amenable to gradient-based optimization.

3.1 Approximating the Function-Space Kullback-Leibler Divergence via Local Linearization

First, we note that the KL divergence in Equation (3) is intractable for two reasons: (**i**) For BNNs or other non-linear models, we do not have access to the probability density functions of the multivariate distributions $q_{f(\mathbf{X};\boldsymbol{\Theta})}$ and $p_{f(\mathbf{X};\boldsymbol{\Theta})}$; (**ii**) For all but extremely simple input spaces, we are unable to compute the supremum over all possible finite sets of evaluation points. In the remainder of this section, we outline a simple approach for obtaining an estimator of the KL divergence that allows for scalable gradient-based optimization of Equation (3).

We first approach the problem of computing the KL divergence between two BNNs evaluated at a finite set of points. To obtain an estimator of KL divergence $\mathbb{D}_{\mathrm{KL}}(q_{f(\mathbf{X};\boldsymbol{\Theta})} \,\|\, p_{f(\mathbf{X};\boldsymbol{\Theta})})$, we approximate $q_{f(\mathbf{X};\boldsymbol{\Theta})}$ and $p_{f(\mathbf{X};\boldsymbol{\Theta})}$ by computing the first-order Taylor expansion of the *mapping* f about the mean parameters of $q_{\boldsymbol{\Theta}}$ and $p_{\boldsymbol{\Theta}}$, respectively. To obtain locally accurate estimates of the probability densities of $q_{f(\mathbf{X};\boldsymbol{\Theta})}$ and $p_{f(\mathbf{X};\boldsymbol{\Theta})}$, we use the following result:

Proposition 1 (Distribution under Linearized Mapping)**.** *For a stochastic function $f(\cdot\,;\boldsymbol{\Theta})$ defined in terms of stochastic parameters $\boldsymbol{\Theta} \sim g_{\boldsymbol{\Theta}}$, and $g_{\boldsymbol{\Theta}}$ with $\mathbf{m} \doteq \mathbb{E}_{g_{\boldsymbol{\Theta}}}[\boldsymbol{\Theta}]$, and Jacobian $\mathcal{J}(\cdot, \mathbf{m}) \doteq \frac{\partial f(\cdot\,;\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}|_{\boldsymbol{\Theta}=\mathbf{m}}$, denote the linearization of the stochastic function $f(\cdot\,;\boldsymbol{\Theta})$ about $\mathbf{m}$ by*

$$f(\cdot\,;\boldsymbol{\Theta}) \approx \tilde{f}(\cdot\,;\boldsymbol{\Theta}) \doteq f(\cdot\,;\mathbf{m}) + \mathcal{J}(\cdot, \mathbf{m})(\boldsymbol{\Theta} - \mathbf{m}).$$

Then the mean and covariance of the distribution over the linearized mapping $\tilde{f}$ at $\mathbf{X}, \mathbf{X}' \in \mathcal{X}$ are given by

$$\mathbb{E}[\tilde{f}(\mathbf{X};\boldsymbol{\Theta})] = f(\mathbf{X};\mathbf{m}) \quad \textit{and} \quad \mathrm{Cov}[\tilde{f}(\mathbf{X};\boldsymbol{\Theta}), \tilde{f}(\mathbf{X}';\boldsymbol{\Theta})] = \mathcal{J}(\mathbf{X}, \mathbf{m})\mathbf{S}\mathcal{J}(\mathbf{X}', \mathbf{m})^{\top}.$$

Proof. See Appendix A. □

Corollary 1. *If $g_{\boldsymbol{\Theta}}$ is a multivariate Gaussian distribution with mean $\mathbf{m}$ and diagonal co-variance $\mathbf{S}$, then the distribution $\tilde{g}$ over $\tilde{f}(\mathbf{X}\,;\boldsymbol{\Theta})$ is given by $\tilde{g}_{\tilde{f}(\mathbf{X}\,;\boldsymbol{\Theta})} = \mathcal{N}(f(\mathbf{X}\,;\mathbf{m}), \mathcal{J}(\mathbf{X}, \mathbf{m})\mathbf{S}\mathcal{J}(\mathbf{X}', \mathbf{m})^{\top})$.*

Proof. The result follows from the fact that Gaussianity is preserved under affine transformations. □

Proposition 1 tells us that if $q_{\boldsymbol{\Theta}}$ and $p_{\boldsymbol{\Theta}}$ are both Gaussian distributions, then the induced distributions under the linearized mapping $\tilde{f}$ evaluated at a finite set of evaluation points will be Gaussian as well. This means that for Gaussian variational or prior distributions over $\boldsymbol{\Theta}$, we can obtain a locally accurate estimate of the probability density function of the induced multivariate random variable $f(\mathbf{X};\boldsymbol{\Theta})$.[1] We can use this insight to estimate the KL divergence in the variational objective by $\mathbb{D}_{\mathrm{KL}}(\tilde{q}_{\tilde{f}(\mathbf{X};\boldsymbol{\Theta})} \,\|\, \tilde{p}_{\tilde{f}(\mathbf{X};\boldsymbol{\Theta})})$ for several different variational and prior distributions over parameters. For example, for an isotropic Gaussian prior and a mean-field Gaussian variational distribution, $\mathbb{D}_{\mathrm{KL}}(\tilde{q}_{\tilde{f}(\mathbf{X};\boldsymbol{\Theta})} \,\|\, \tilde{p}_{\tilde{f}(\mathbf{X};\boldsymbol{\Theta})})$ is a KL divergence between two multivariate Gaussians and can be obtained analytically.

Using this approximation, we obtain an estimator of the variational objective given by

$$\tilde{\mathcal{F}}(q_{\boldsymbol{\Theta}}, \mathcal{D}) \doteq \mathbb{E}_{q_{f(\mathbf{X}_{\mathcal{D}};\boldsymbol{\Theta})}}[\log p(\mathbf{y}_{\mathcal{D}} \mid f(\mathbf{X}_{\mathcal{D}};\boldsymbol{\theta}))] - \sup_{\mathbf{X} \in \mathcal{X}_{\mathbb{N}}} \mathbb{D}_{\mathrm{KL}}(\tilde{q}_{\tilde{f}(\mathbf{X};\boldsymbol{\Theta})} \,\|\, \tilde{p}_{\tilde{f}(\mathbf{X};\boldsymbol{\Theta})}), \tag{4}$$

where the arguments of the KL divergence have been replaced by the (locally accurate) approximations to the variational and prior distributions over functions evaluated at $\mathbf{X}$, respectively. Since the stochastic functions $\tilde{f}(\cdot\,;\boldsymbol{\Theta})$ induced by $q_{\boldsymbol{\Theta}}$ and $p_{\boldsymbol{\Theta}}$ under the linearized mapping will be closer to the stochastic function under f the smaller the variance of $q_{\boldsymbol{\Theta}}$ and $p_{\boldsymbol{\Theta}}$, respectively, the approximation to the KL divergence will be more accurate the smaller the variance of $q_{\boldsymbol{\Theta}}$ and $p_{\boldsymbol{\Theta}}$.

Next, we turn to computing the supremum. Unlike Sun et al. [2019], who consider the supremum as a separate optimization problem, we do not seek to compute the supremum by searching over points $\mathbf{X} \in \mathcal{X}_{\mathbb{N}}$ but instead propose to estimate the supremum at every gradient step via a simple finite-sample estimator. Specifically, letting $I(\mathbf{X}) \doteq \mathbb{D}_{\mathrm{KL}}(\tilde{q}_{\tilde{f}(\mathbf{X};\boldsymbol{\Theta})} \,\|\, \tilde{p}_{\tilde{f}(\mathbf{X};\boldsymbol{\Theta})})$, we estimate $G = \sup_{\mathbf{X} \in \mathcal{X}_{\mathbb{N}}} I(\mathbf{X})$ using the Monte Carlo estimator

$$\hat{G}(\mathcal{X}_{\mathcal{C}}^{S}) = \max_{\mathbf{X} \in \mathcal{X}_{\mathcal{C}}^{S}} I(\mathbf{X}), \tag{5}$$

[1] Although the result in Corollary 1 extends to all distributions in the location-scale family of distributions, such as Laplace and Cauchy distributions, we leave an empirical exploration of non-Gaussian variational distributions over parameters for future work.

where $\mathcal{X}_{\mathcal{C}}^S \doteq \{\mathbf{X}_{\mathcal{C}}^{(i)}\}_{i=1}^S$ is a collection of S sets of *context points* $\mathbf{X}_{\mathcal{C}}^{(i)} \doteq \{\mathbf{x}^{(j)}\}_{j=1}^K$ jointly sampled from a context distribution $p_{\mathcal{X}_{\mathcal{C}}}$. Each set of context points $\mathbf{X}_{\mathcal{C}}^{(i)}$ can be viewed as a single Monte Carlo sample from the input space so that the estimator $\hat{G}(\mathcal{X}_{\mathcal{C}}^S)$ provides an S-sample Monte Carlo estimate of the supremum. While this estimator is crude and only provides a rough approximation to the true supremum, it encourages the variational distribution over functions to match the prior distribution over functions on the sets of context points. The choice of the context distribution $p_{\mathcal{X}_{\mathcal{C}}}$ can be informed by knowledge about the prediction task and should be viewed as a problem-specific modeling choice. Similarly, the numbers of samples S and K are hyperparameters to be optimized with a validation set. For details on how $p_{\mathcal{X}_{\mathcal{C}}}$ is chosen in the empirical evaluation in Section 5, see Appendix F. For an ablation study on different choices of $p_{\mathcal{X}_{\mathcal{C}}}$, see Appendix D.

3.2 Stochastic Estimation of the Function-Space Variational Objective

Let $q_{\boldsymbol{\Theta}}$ be a Gaussian mean-field variational distribution, let $p_{\boldsymbol{\Theta}}$ be an isotropic Gaussian prior, let $(\mathbf{X}_{\mathcal{B}}, \mathbf{y}_{\mathcal{B}})$ be a mini-batch of the training data, and reparameterize $\boldsymbol{\Theta}$ as $\hat{\boldsymbol{\Theta}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\epsilon}^{(j)}) \doteq \boldsymbol{\mu} + \boldsymbol{\Sigma} \odot \boldsymbol{\epsilon}^{(j)}$. Using the estimator $\hat{G}(\mathcal{X}_{\mathcal{C}}^S)$ defined above and estimating the expected log-likelihood via Monte Carlo sampling, we obtain a Monte Carlo estimator for the function-space variational objective:

$$\bar{\mathcal{F}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathcal{X}_{\mathcal{C}}^S, \mathcal{D}) = \frac{1}{M} \sum_{j=1}^{M} \log p(\mathbf{y}_{\mathcal{B}} \mid f(\mathbf{X}_{\mathcal{B}}; \hat{\boldsymbol{\Theta}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\epsilon}^{(j)}))) - \max_{\mathbf{X} \in \mathcal{X}_{\mathcal{C}}^S} \mathbb{D}_{\mathrm{KL}}(\tilde{q}_{\tilde{f}(\mathbf{X}; \hat{\boldsymbol{\Theta}})} \,\|\, \tilde{p}_{\tilde{f}(\mathbf{X}; \hat{\boldsymbol{\Theta}})})$$

with $\boldsymbol{\epsilon}^{(j)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_P)$ and $\mathcal{X}_{\mathcal{C}}^S$ as defined above. Crucially, to avoid having to perform expensive matrix operations as part of the KL computation, we make a diagonalization approximation to the covariance matrices of $\tilde{q}$ and $\tilde{p}$, respectively.

Selection of Prior. For all experiments that involve uncertainty quantification, we chose a prior distribution over parameters that induces a prior distribution over functions $p_{f(\cdot;\boldsymbol{\Theta})}$ that, under the likelihood, exhibits a high degree of predictive uncertainty. Such a prior will encourage the BNN to have high uncertainty at evaluation points from regions in input space where $p_{\mathcal{X}_{\mathcal{C}}}$ has non-zero support and, under smoothness constraints, on evaluation points in nearby regions. For settings where prior information is encoded in data—for example, in the form of expert demonstrations of robotic manipulation tasks, or in the form of pre-trained networks in continual or transfer learning—an empirical prior that reflect this information can be specified. For further details, see Appendix F, and for ablation studies on the effect of different prior choices on performance, see Appendix D.

Selection of Context Distribution. The distribution $p_{\mathcal{X}_{\mathcal{C}}}$ allows us to incorporate information about the data-generating process into training. This way, $p_{\mathcal{X}_{\mathcal{C}}}$ allows us obtain gradient estimates that encourage the variational distribution to match the prior over functions in more relevant parts of the input space. By taking advantage of the abundance of data available in real-world settings, we can construct context distributions from large datasets like ImageNet [Krizhevsky et al., 2012] or from small but diverse datasets like CIFAR-100. In our experiments, we choose two types of context distributions. One distribution constructed from the training data and only containing randomly sampled monochrome images, and one constructed from a related real-world dataset generated from a similar but different data distribution. For further details, see Appendix F, and for ablation studies on the effect of different context distributions on performance, see Appendix D.

Posterior Predictive Distribution. After optimizing the variational objective with respect to the parameters of the variational distribution $q_{\boldsymbol{\Theta}}$, we obtain an approximate posterior predictive distribution

$$q(\mathbf{y}_* \mid \mathbf{x}_*) = \int p(\mathbf{y}_* \mid f(\mathbf{x}_*; \boldsymbol{\theta}))\, q_{f(\mathbf{x}_*;\boldsymbol{\Theta})}\, df(\mathbf{x}_*; \boldsymbol{\theta}) \approx \frac{1}{M_*} \sum_{j=1}^{M_*} p(\mathbf{y}_* \mid f(\mathbf{x}_*; \boldsymbol{\Theta}^{(j)})), \boldsymbol{\Theta}^{(j)} \sim q_{\boldsymbol{\Theta}}, \quad (6)$$

where M_* is the number of Monte Carlo samples used to estimate the predictive distribution.

Computational Complexity. Under the diagonal approximation used to compute $\bar{\mathcal{F}}$, the time and space complexity for stochastic gradient estimation scale linearly in the mini-batch size $n_{\mathcal{B}}$ and the number forward passes through the network per input point (i.e., the number of Monte Carlo samples M used to estimate the expected log-likelihood plus the number of points on which the KL divergence is evaluated, SM), yielding a time and space complexity of $\mathcal{O}(n_{\mathcal{B}} p(M + SK))$. In comparison, parameter-space mean-field variational inference has complexity $\mathcal{O}(n_{\mathcal{B}} p M)$. The time complexity of a prediction on n_* points with M_* Monte Carlo samples has a complexity of $\mathcal{O}(n_* M_*)$.

4 Related Work

There is a growing body of work on function-space approaches to inference in BNNs, deep learning, and applications such as continual learning [Benjamin et al., 2019, Burt et al., 2021, Pan et al., 2020, Sun et al., 2019, Titsias et al., 2020, Ma and Hernández-Lobato, 2021].

Function-Space Inference in Bayesian Neural Networks. Previously proposed methods for FSVI in BNNs are based on approximate gradient estimators and either replace the supremum in Equation (3) with an expectation [Sun et al., 2019] or do not define an explicit variational objective [Wang et al., 2019]. More recent work has attempted to circumvent the intractability of the variational objective in Equation (2) by proposing alternative objectives for function-space inference in BNNs [Ma et al., 2019, Ober and Aitchison, 2020, Ma and Hernández-Lobato, 2021].

Linear Models. Immer et al. [2020] take a similar approach as Khan et al. [2019] and show that approximating BNN posterior distribution via the Laplace and Generalized-Gauss-Newton approximation corresponds to linearizations of different models. Crucially, unlike in our approach, they use a Laplace approximation and do not perform variational inference. As such, they do not optimize the variance parameters. While they also use a linearization similar to the first-order Taylor expansion in Proposition 1, they use it to make predictions, whereas we only use the linearization to obtain an estimator for the KL divergence between distributions over functions. In our approach, predictions and expected log-likelihood estimates are obtained from the unlinearized BNN.

Pathologies of Variational Inference in Bayesian Neural Networks. Burt et al. [2021] consider the function-space variational objective in Equation (2) and show that the KL divergence between BNNs with different networks architectures are not well-defined. A parallel line of research showed that posterior predictive distributions of shallow BNNs with mean-field variational distributions have a limited ability to represent complex covariance structures in function space [Foong et al., 2019, 2020], but that deep BNNs do not suffer from this limitation [Farquhar et al., 2020b].

5 Empirical Evaluation

The seminal work of Sun et al. [2019] introduced function-space variational inference as a tool for approximate inference in BNNs but the approach proposed in Sun et al. [2019] is limited to small neural networks and low-dimensional input data [Ma and Hernández-Lobato, 2021]. In this section, we evaluate FSVI on high-dimensional classification tasks that were previously out of reach under the approach proposed by Sun et al. [2019] and compare FSVI to several well-established and state-of-the-art methods. We show that FSVI (sometimes *significantly*) outperforms existing Bayesian and non-Bayesian methods in terms of their in-distribution uncertainty calibration and out-of-distribution predictive uncertainty estimation. For a details on models, training and validation procedures, and datasets used, see Appendix F. Our code can be accessed at `https://github.com/fsvibnns/fsvi`.

5.1 Predictive Performance, Uncertainty Estimation, and Distribution Shift Detection

In this set of experiments, we assess the reliability of the uncertainty estimates generated by FSVI. If a BNN trained via FSVI is able to perform reliable uncertainty estimation, its predictions will be significantly higher on input points that were generated according to a different data-generating distribution than the training data. For models trained on the FashionMNIST dataset, we use the MNIST and NotMNIST datasets as out-of-distribution evaluation points, while for models trained on the CIFAR-10 dataset, we use the SVHN dataset as out-of-distribution evaluation points.

For models trained on either FashionMNIST or CIFAR-10, we evaluate their in-distribution performance in terms of test accuracy, test log-likelihood, and test calibration. To evaluate the quality of different models' uncertainty estimates, we compute uncertainty estimates for the pairs FashionMNIST/MNIST, FashionMNIST/NotMNIST, and CIFAR-10/SVHN to and measure for a range of thresholds how well the datasets in each pair can be separated solely based on the uncertainty estimates. This experiment setup follows prior work by van Amersfoort et al. [2020] and Immer et al. [2020]. We report the area under the receiver operating characteristic (ROC) curve in Tables 1 and 2.

Predictive Performance and Calibration. To assess in-distribution predictive performance and calibration, we report the test accuracy, negative log-likelihood (NLL), and expected calibration error (ECE) for models trained on FashionMNIST and CIFAR-10 in Tables 1 and 2. On both

Table 1: Comparison of in- and out-of-distribution performance metrics on FashionMNIST (mean ± standard error over ten random seeds). The last two columns shoow the AUROC for binary in- vs. out-of-distribution detection on MNIST (M) and NotMNIST (NM). MNIST and NotMNIST are used as out-of-distribution datasets. Best overall results for single and ensemble models are printed in boldface with gray shading. Results within a 95% confidence interval of the best overall result are printed in boldface only. All methods use the same four-layer CNN architecture. For further details about model architectures and training and evaluation protocols, see Appendix F.

Method	Accuracy ↑	NLL ↓	ECE ↓	AUROC M ↑	AUROC NM ↑
MAP	91.73±0.08	0.288±0.003	0.037±0.001	87.00±0.30	74.85±1.31
MFVI [Blundell et al., 2015]	91.03±0.04	0.354±0.003	0.038±0.001	93.10±0.34	88.88±0.74
MFVI (tempered)	91.38±0.05	0.519±0.005	0.058±0.001	86.30±0.29	80.78±0.68
MFVI (radial) [Farquhar et al., 2020a]	90.31±0.11	0.340±0.001	0.035±0.001	84.40±0.68	82.11±1.15
MC DROPOUT	90.55±0.04	0.230±0.001	**0.012**±0.001	88.46±0.57	80.02±1.04
SWAG [Maddox et al., 2019]	92.56±0.05	0.300±0.000	0.043±0.001	85.18±0.35	80.31±0.30
DUQ [van Amersfoort et al., 2020]	92.40±0.20	–	–	95.50±0.70	94.60±1.80
BNN-LAPLACE [Immer et al., 2020]	92.25±0.10	0.244±0.003	**0.012**±0.003	95.55±0.60	–
SPG [Ma and Hernández-Lobato, 2021]	91.60±0.14	0.254±0.050	–	95.60±6.00	–
FSVI ($p_{\mathbf{X}_\mathcal{I}}$ = random monochrome)	**93.15**±0.13	0.203±0.004	**0.014**±0.002	96.55±0.41	95.27±0.63
FSVI ($p_{\mathbf{X}_\mathcal{I}}$ = KMNIST)	**93.18**±0.14	**0.188**±0.005	**0.013**±0.002	**99.75**±0.17	**98.64**±0.23
Deep Ensemble	92.49±0.01	0.242±0.001	**0.019**±0.000	89.22±0.09	83.17±0.91
FSVI Ensemble	**94.44**±0.07	**0.181**±0.001	0.020±0.001	**97.85**±0.15	**96.95**±0.20

Table 2: Comparison of in- and out-of-distribution performance metrics on CIFAR-10 (mean ± standard error over ten random seeds). SVHN and corrupted CIFAR-10 (C-CIFAR) are used as an out-of-distribution datasets. The penultimate column shows the AUROC for binary in- vs. out-of-distribution detection on SVHN. Best overall results for single and ensemble models are printed in boldface with gray shading. Results within a 95% confidence interval of the best overall result are printed in boldface only. All methods use a ResNet-18 architecture. For further details about model architectures and training and evaluation protocols, see Appendix F.

Method	Accuracy↑	NLL↓	ECE↓	OOD-AUROC↑	C-CIFAR Acc↑
MAP	92.19±0.15	0.307±0.006	0.046±0.001	95.17±0.40	78.55±1.01
MFVI [Blundell et al., 2015]	89.98±0.09	0.340±0.006	0.040±0.001	92.14±0.34	79.36±1.35
MFVI (tempered)	90.87±0.11	0.360±0.003	0.048±0.001	91.82±0.90	79.86±1.32
MC DROPOUT	91.32±0.06	0.310±0.004	0.041±0.001	90.32±0.57	80.19±1.44
SWAG [Maddox et al., 2019]	93.13±0.14	0.480±0.001	0.067±0.002	89.79±0.50	76.12±0.51
VOGN [Osawa et al., 2019]	84.27±0.20	0.477±0.006	0.040±0.002	87.60±0.20	–
DUQ [van Amersfoort et al., 2020]	**94.10**±0.20	–	–	92.70±1.30	–
SPG [Ma and Hernández-Lobato, 2021]	77.69±0.64	0.675±0.030	–	88.30±4.00	–
FSVI ($p_{\mathbf{X}_\mathcal{I}}$ = random monochrome)	92.55±0.04	**0.295**±0.003	0.034±0.001	94.69±0.19	**80.88**±0.47
FSVI ($p_{\mathbf{X}_\mathcal{I}}$ = CIFAR-100)	92.57±0.04	**0.293**±0.003	**0.029**±0.001	**98.09**±0.09	**81.21**±0.47
Deep Ensemble	95.13±0.06	0.158±0.001	0.019±0.001	**98.04**±0.07	**81.22**±0.37
FSVI Ensemble	**95.29**±0.03	**0.150**±0.003	**0.011**±0.001	**99.42**±0.36	**81.44**±0.43

FashionMNIST and CIFAR-10, FSVI achieves the lowest NLL and either the best or second-best predictive accuracy and ECE, respectively, across all methods. Notably, FSVI significantly outperforms SPG [Ma and Hernández-Lobato, 2021], an alternative function-space variational inference method.

Predictive Uncertainty under Distribution Shift. In Tables 1 and 2, we report evaluation metrics that elucidate the reliability of different methods' predictive uncertainty under distribution shift. FSVI exhibits reliable predictive uncertainty estimates that allow distinguishing between in- and out-of-distribution inputs with high accuracy. As would be expected, we observe that using context distributions that reflect our knowledge about the data-generating process can significantly improve uncertainty quantification under FSVI. For the FashionMNIST experiment, we used the KMNIST dataset, which contains grayscale images of Kuzushiji letters, and for the CIFAR-10 experiment, we used the CIFAR-100 dataset, which contains RGB images of 100 classes. Both KMNIST and CIFAR-100 differ from the OOD datasets (MNIST and NotMNIST and SVHN, respectively) used to compute OOD-AUROC metrics in Tables 1 and 2, but using them as context distributions significantly increased the ability of BNNs trained via FSVI to identify distributionally shifted samples. Since the variational objective encourages matching the prior (which we chose to have high variance) on samples from the context distribution can improve uncertainty estimation in regions of the input space far from the training data.

5.2 Generalization and Reliability of Predictive Uncertainty under Distribution Shift

To assess the reliability of predictive models in deep learning, Ovadia et al. [2019] propose the following desiderata: In order for a model to be considered reliable, it ought to (i) exhibit low predictive uncertainty on training data and high predictive uncertainty on out-of-distribution inputs, (ii) generate predictive uncertainty estimates that allow distinguishing in- from out-of-distribution data points, and (iii) if possible, maintain high predictive accuracy even under distribution shift. Models that satisfy these desiderata are less likely to make poor, high-confidence predictions and are more amenable for use in safety-critical downstream tasks.

To illustrate these desiderata, we follow Ovadia et al. [2019] and consider the rotated MNIST task, where a model is trained on MNIST and evaluated on rotated MNIST digits. The goal is to maintain a high level of predictive accuracy (measured in terms of Brier scores) while exhibiting an increasing level of predictive uncertainty on distribution shifts of increasing magnitude. Figure 2 shows Brier scores (lower is better) and predictive entropy estimates (higher means more uncertain) of four different models. As rotating the MNIST digits gradually shifts the data distributions, we would expect Brier scores to increase (corresponding to worse predictive accuracy) as the rotation angle increases. A model with reliable predictive entropy estimates would only experience a small decrease under distribution shift while exhibiting a large increase in predictive uncertainty. As can be seen in the plot, the Brier scores of FSVI decreases the least, while FSVI's uncertainty is significantly higher than other models'. To assess the reliability of different uncertainty quantification methods on a more challenging distribution-shift task, we consider corrupted CIFAR-10 inputs under the second-mildest corruption level used in [Ovadia et al., 2019] and report our results in Table 2. Consistent with the rotated MNIST results, FSVI achieves the highest accuracy on the corrupted data.

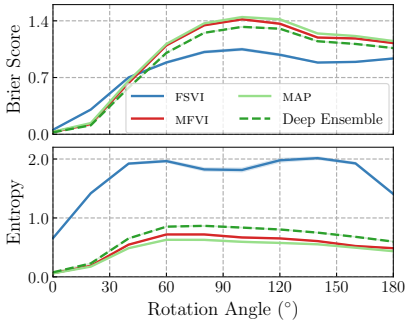


Figure 2: Predictive uncertainty and accuracy on rotated MNIST. Models with reliable uncertainty estimates would exhibit higher predictive uncertainty the more the digits are rotated. Ideally, such models would maintain high predictive accuracy (low Brier score).

5.3 Safety-Critical Uncertainty-Aware Selective Prediction: Diabetic Retinopathy Diagnosis

To evaluate the reliability of the predictive uncertainty of FSVI in a real-world safety-critical setting, we consider the task of diagnosing diabetic retinopathy (DR), a medical condition that can lead to impaired vision, from retina scans [Leibig et al., 2017, Filos et al., 2019, Band et al., 2021]. We use two publicly available datasets, EyePACS [2015] and APTOS [2019], each containing RGB images of a human retina graded by a medical expert on the following scale: 0 (no DR), 1 (mild DR), 2 (moderate DR), 3 (severe DR), and 4 (proliferative DR). The Kaggle dataset was collected from patients in the United States, while the APTOS dataset was collected from patients in India using cheaper but more modern scanning devices. We follow Leibig et al. [2017], Filos et al. [2019], and Band et al. [2021] and binarize all examples from both the EyePACS and APTOS datasets by dividing the classes up into sight-threatening diabetic retinopathy—defined as moderate diabetic retinopathy or worse (classes $\{2, 3, 4\}$)—and non-sight-threatening diabetic retinopathy—defined as no or mild diabetic retinopathy (classes $\{0, 1\}$). This results in a binary prediction task.

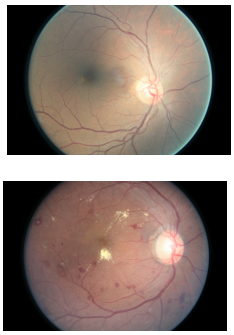
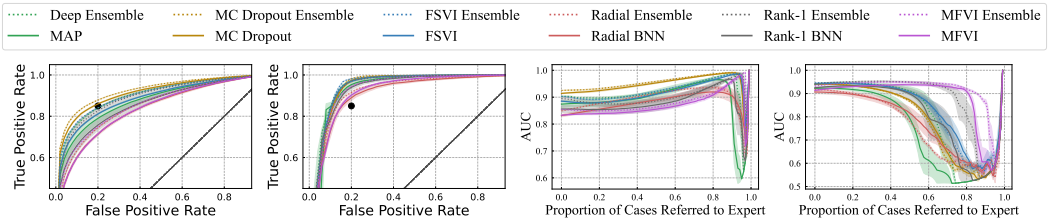


Figure 3: Retina scan examples. **Top:** healthy. **Bottom:** unhealthy.

To assess the reliability of predictive models when medical training and test data are obtained from different patient populations or collected with the same medical equipment, we follow Band et al. [2021] and use the Kaggle dataset for training and the distributionally shifted APTOS dataset for evaluation. The results are shown in Figure 4, which plot the ROC curves for the binary prediction problems as well as the area under the ROC curve for an uncertainty aware selective prediction task. For further details about the uncertainty-aware selective prediction evaluation protocol, see Appendix F.4. Figure 4 shows that FSVI performs well on all four tasks and is only outperformed by MC DROPOUT. For full tabular results, see Appendix B.1.

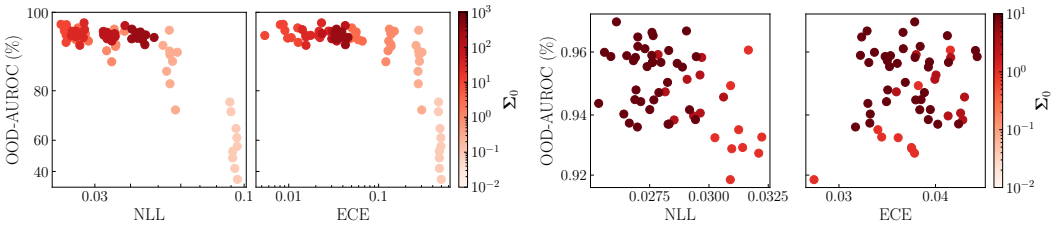


(a) ROC: In-Domain **(b)** ROC: Country Shift **(c)** Selective Prediction AUC: In-Domain **(d)** Selective Prediction AUC: Country Shift

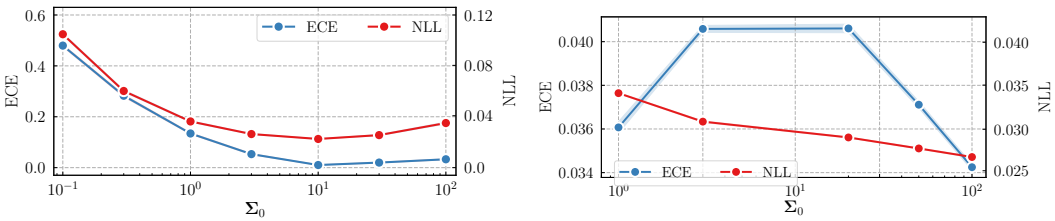
Figure 4: We jointly assess model predictive performance and uncertainty quantification on both in-domain and distributionally shifted data. **Left:** The *receiver operating characteristic curve* (ROC) for in-population diagnosis on the (**a**) EyePACS [2015] test set and for (**b**) changing medical equipment and patient populations on the APTOS [2019] test set. The dot in **black** denotes the NHS-recommended 85% sensitivity and 80% specificity ratios [Widdowson, 2016]. **Right:** *Selective prediction* on AUC in the (**c**) EyePACS [2015] and the (**d**) APTOS [2019] settings. Shading denotes standard error computed over six random seeds.

5.4 Ablation Study and Discussion

To investigate the effect of different prior distributions over functions on predictive predictive performance and uncertainty metrics, we performed an ablation study on the effect of changing the variance of the induced prior distribution over functions using the FashionMNIST and CIFAR-10 datasets. In Figure 5, we relate the effect of changing the prior variance $\boldsymbol{\Sigma}_0$ on uncertainty-based distribution shift detection ("OOD-AUROC", higher is better) and the test negative log-likelihood and calibration error (lower is better for both) and find that the prior variance range that results in the best in- vs. shifted distribution uncertainty (i.e., high OOD-AUROC) also results the most accurate prediction and calibration (i.e., low NLL and ECE). This correlation suggests that optimizing for low ECE and NLL, which can be achieved with an in-distribution held-out validation set, will result in reliable uncertainty estimation both on in-distribution data as well as under distribution shift. For an extensive set of ablation studies on all relevant hyperparameters, see Appendix D.



(a) Effect of Prior Variance $\boldsymbol{\Sigma}_0$ on In- and Out-of-Distribution Uncertainty Estimation.



(b) Effect of Prior Variance $\boldsymbol{\Sigma}_0$ on ECE and NLL.

Figure 5: Effect of Prior Variance on Uncertainty Estimation. **Left:** FashionMNIST (Training) & MNIST (OOD). **Right:** CIFAR-10 (Training) & SVHN (OOD). For further ablation studies, see Appendix B.

6 Conclusion

The paper proposed a scalable and effective approach to function-space variational inference in BNNs. We demonstrated that the proposed estimator of the function-space variational objective can be scaled up to high-dimensional data and large neural network architectures and that FSVI exhibits consistently reliable in- and out-of-distribution predictive performance on a wide range of datasets when compared to well-established and state-of-the-art uncertainty quantification methods. We hope that this work will lead to further research into function-space variational inference and the development of more sophisticated data-driven prior distributions over functions.

References

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25:*, pages 1106–1114, 2012.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*. 2013.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. 529, 2016.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zidek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 (7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML 2016, pages 1050–1059, 2016.

Radford M Neal. Bayesian Learning for Neural Networks. 1996.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems 32*. 2019.

Andrew Y. K. Foong, Yingzhen Li, José Miguel Hernández-Lobato, and Richard E. Turner. 'in-between' uncertainty in bayesian neural networks, 2019.

Sebastian Farquhar, Michael A. Osborne, and Yarin Gal. Radial bayesian neural networks: Beyond discrete support in large-scale bayesian deep learning. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1352–1362. PMLR, 26–28 Aug 2020a.

Neil Band, Tim G. J. Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W. Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. Benchmarking Bayesian Deep Learning on Diabetic Retinopathy Detection Tasks. 2021.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger B. Grosse. Functional variational bayesian neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rkxacs0qY7.

David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Comput.*, 4(3):448–472, May 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.448. URL https://doi.org/10.1162/neco.1992.4.3.448.

Martin J Wainwright and Michael I Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008. ISBN 1601981848.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/blundell15.html.

Alex Graves. Practical variational inference for neural networks. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 2348–2356, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, page 5–13, New York, NY, USA, 1993. Association for Computing Machinery. ISBN 0897916115.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, May 2013. ISSN 1532-4435.

David R. Burt, Sebastian W. Ober, Adrià Garriga-Alonso, and Mark van der Wilk. Understanding variational inference in function-space. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021. URL https://openreview.net/forum?id=7P9y3sRa5Mk.

Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and bayesian networks. In Eric Carlen, Mokshay Madiman, and Elisabeth M. Werner, editors, *Convexity and Concentration*, pages 211–249, New York, NY, 2017. Springer New York. ISBN 978-1-4939-7005-6.

Ari Benjamin, David Rolnick, and Konrad Kording. Measuring and regularizing networks in function space. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkMwpiR9Y7.

Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard E. Turner, and Mohammad Emtiyaz Khan. Continual deep learning by functional regularisation of memorable past. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2f3bbb9730639e9ea48f309d9a79ff01-Abstract.html.

Michalis K. Titsias, Jonathan Schwarz, Alexander G. de G. Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxCzeHFDB.

Chao Ma and José Miguel Hernández-Lobato. Functional variational inference based on stochastic process generators. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Ziyu Wang, Tongzheng Ren, Jun Zhu, and Bo Zhang. Function space particle optimization for bayesian neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BkgtDsCcKQ.

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 http://proceedings.mlr.press/v97/ma19b.html.

Sebastian W. Ober and Laurence Aitchison. Global inducing point variational posteriors for bayesian neural networks and deep gaussian processes, 2020.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural networks via local linearization, 2020.

Mohammad Emtiyaz E 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 32*, pages 3094–3104. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8573-approximate-inference-turns-deep-networks-into-gaussian-processes.pdf.

Andrew Y. K. Foong, David R. Burt, Yingzhen Li, and Richard E. Turner. On the expressiveness of approximate inference in bayesian neural networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b6dfd41875bc090bd31d0b1740eb5b1b-Abstract.html.

Sebastian Farquhar, Lewis Smith, and Yarin Gal. Liberty or depth: Deep bayesian neural nets do not need complex weight posterior approximations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/2dfe1946b3003933b7f8ddd71f24dbb1-Abstract.html.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, 2020.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, pages 13153–13164, 2019.

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E 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, pages 4287–4299. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/b53477c2821c1bf0da5d40e57b870d35-Paper.pdf.

Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging Uncertainty Information From Deep Neural Networks for Disease Detection. *Nature Scientific Reports*, 7(1):17816, 2017.

Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks, 2019.

EyePACS. Diabetic Retinopathy Detection Dataset, 2015.

APTOS. APTOS 2019 Blindness Detection Dataset, 2019.

D. T. S. Widdowson. The Management of Grading Quality: Good Practice in the Quality Assurance of Grading. *Tech. Rep.*, 2016.

Alexander G. de G. Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On sparse variational methods and the kullback-leibler divergence between stochastic processes. volume 51 of *Proceedings of Machine Learning Research*, pages 231–239, Cadiz, Spain, 09–11 May 2016. PMLR. URL http://proceedings.mlr.press/v51/matthews16.html.

M. J. Schervish. *Theory of Statistics*. Springer-Verlag, New York, NY, 1995.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1257–1264. MIT Press, 2006. URL http://papers.nips.cc/paper/2857-sparse-gaussian-processes-using-pseudo-inputs.pdf.

Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. Variational deterministic uncertainty quantification, 2021. URL https://openreview.net/forum?id=8W7LTo_zxdE.

Pavel Izmailov, Wesley J. Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Subspace inference for bayesian deep learning. 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 1169–1179. PMLR, 22–25 Jul 2020. URL http://proceedings.mlr.press/v115/izmailov20a.html.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 5.
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Appendix A.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[TODO]**
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix F.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Table 2, Table 2, and Table 3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix F.9.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See `https://github.com/fsvibnns/fsvi`.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]