Rethinking Function-Space Variational Inference in Bayesian Neural Networks

Anonymous Authors
Anonymous Institution

Abstract

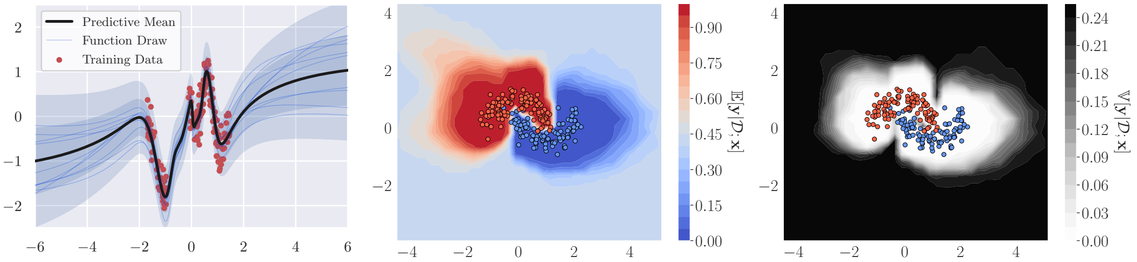
Bayesian neural networks (BNNs) define distributions over functions induced by distributions over parameters. In practice, this model specification makes it difficult to define and use meaningful prior distributions over functions that could aid in training. What's more, previous attempts at defining an explicit function-space variational objective for approximate inference in BNNs require approximations that do not scale to high-dimensional data. We propose a new function-space approach to variational inference in BNNs and derive a *tractable* variational by linearizing the BNN's posterior predictive distribution about its mean parameters, allowing function-space variational inference to be scaled to large and high-dimensional datasets. We evaluate this approach empirically and show that it leads to models with competitive predictive accuracy and significantly improved predictive uncertainty estimates compared to parameter-space variational inference.

1. Introduction

Approximate inference in Bayesian neural networks (BNNs) typically involves performing probabilistic inference directly over a set of stochastic network parameters. Unfortunately, standard approaches for parameter-space inference in BNNs often do not result in approximate posterior predictive distributions that reliably exhibit high predictive uncertainty away from the training data or under distribution shift, making them of limited use in practice.

Instead of *explicitly* performing approximate inference over BNN parameters, we propose a method for tractable and efficient approximate inference in BNNs by inferring an approximate posterior distribution over the network parameters *implicitly* and optimizing a variational objective over the induced distribution over *functions* instead. This way, it is possible to to better control the distribution over functions induced by the network parameters and obtain higher-quality uncertainty estimates than state-of-the-art Bayesian and non-Bayesian methods. We evaluate the resulting approximate posterior predictive distribution empirically on a number of high-dimensional prediction tasks and demonstrate that it outperforms related methods in terms of accuracy, out-of-distribution detection, and calibration.

The main contributions of this paper is the conceptualization and formalization of a new approach to function-space variational inference in BNNs that is more scalable and better performing than previously proposed approaches. We address the conceptual and practical limitations of prior work, perform an extensive empirical evaluation on high-dimensional prediction tasks, and conduct ablation studies on the proposed method.



(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 distribution of a BNN, obtained via function-space variational inference (FSVI) under the local approximation described in Section 4. For further plots, see Appendix 4.

2. Preliminaries

We consider supervised learning tasks on data $\mathcal{D} \overset{\text{def}}{=} \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$ 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 Consider a neural network $f(\mathbf{x}; \boldsymbol{\theta})$ parameterized by stochastic parameters $\boldsymbol{\theta} \in \mathbb{R}^P$ and define a conditional distribution of targets given function values $f(\mathbf{x}; \boldsymbol{\theta})$: $p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}; f)$. For $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$, we thus obtain a joint distribution $p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}; f)\, p(\boldsymbol{\theta})$, where the semicolon denotes a dependency on some non-stochastic quantity (in this case, the architecture of the neural network), and the resulting distribution of the targets under a given architecture and parameter vector is determined by the model $p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}; f)$. For example, for regression and softmax classification tasks, the conditional distribution $p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}; f)$ would be a Gaussian distribution with mean $f(\mathbf{x}; \boldsymbol{\theta})$ (and some variance) or a categorical distribution defined via a softmax model (Bridle, 1990), respectively.

Mean-Field Variational Inference Since BNNs are non-linear in their parameters, exact inference over the network parameters is analytically intractable. Mean-field variational inference is a variational approach for finding an approximate posterior distribution over network parameters. For further details, see Appendix 2.

3. A Function-Space Perspective on Variational Inference

In this section, we present a function-space perspective on variational inference in BNNs and discuss shortcomings of prior approaches to function-space variational inference (FSVI).

Consider again the probabilistic model $p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}; f)\, p(\boldsymbol{\theta})$ defined in Section 2. Instead of defining the probabilistic model explicitly in terms of the parameters, we will instead define it explicitly in terms of the stochastic functions induced by the stochastic parameters $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$. Specifically, we consider a probabilistic model of the targets and a latent random function f distributed according to some prior distribution over functions, $p(f \,|\, \mathbf{x}; \boldsymbol{\theta})$, parameterized by $\boldsymbol{\theta}$. For a model $p(\mathbf{y} \,|\, f(\mathbf{x}; \boldsymbol{\theta}))$, we can then express the joint distribution over targets and latent random functions as $p(\mathbf{y}, f \,|\, \mathbf{x}) = p(\mathbf{y} \,|\, f) p(f \,|\, \mathbf{x}; \boldsymbol{\theta})$ and frame the inference problem of finding a posterior distribution over functions, $p(f|\mathcal{D})$, variationally as minimizing the KL divergence $\mathbb{D}_{\mathrm{KL}}(q(f; \boldsymbol{\theta}) \,\|\, p(f|\mathcal{D}))$, where $q(f; \boldsymbol{\theta})$ and $p(f|\mathcal{D})$ are *distributions over functions* defined on an infinite index set. For a likelihood function defined on a finite set of training targets $\mathbf{y}$, we can express this minimization problem as maximizing the variational objective

$$\mathcal{F}(q(f;\boldsymbol{\theta})) \overset{\text{def}}{=} \mathbb{E}_{q(f_{\mathbf{X}_\mathcal{D}};\boldsymbol{\theta})}[\log p(\mathbf{y} \mid f_{\mathbf{X}_\mathcal{D}})] - \mathbb{D}_{\text{KL}}(q(f;\boldsymbol{\theta}) \,\|\, p(f)), \tag{1}$$

where $\mathbb{D}_{\text{KL}}(q(f;\boldsymbol{\theta}) \,\|\, p(f))$ is again a KL divergence between distributions over functions. For a measure-theoretic derivation of this result, see Appendix 1. Unfortunately, it is not immediately clear how to evaluate such a KL divergence if $q(f;\boldsymbol{\theta})$ and $p(f)$ are BNN posterior and prior predictive distributions. In an effort to make this objective more tractable, Sun et al. (2019) show that $\mathbb{D}_{\text{KL}}(q(f;\boldsymbol{\theta}) \,\|\, p(f))$ can be expressed as the supremum of the KL divergence from $q(f;\boldsymbol{\theta})$ to $p(f)$ over all *finite* sets of evaluation points, $\mathbf{X}_\mathcal{I}$, that is,

$$\mathcal{F}(q(f;\boldsymbol{\theta})) = \mathbb{E}_{q(f_{\mathbf{X}_\mathcal{D}};\boldsymbol{\theta})}[\log p(\mathbf{y} \mid f_{\mathbf{X}_\mathcal{D}})] - \sup_{n\in\mathbb{N},\mathbf{X}_\mathcal{I}\in\mathcal{X}^n} \mathbb{D}_{\text{KL}}(q(f_{\mathbf{X}_\mathcal{I}};\boldsymbol{\theta}) \,\|\, p(f_{\mathbf{X}_\mathcal{I}})). \tag{2}$$

Unfortunately, this objective is still extremely challenging to estimate in practice: The supremum cannot be found analytically, searching for it iteratively may lead to undesirable optimization behavior (Sun et al., 2019), and the KL divergence term itself is intractable as well—even for finite $\mathbf{X}_\mathcal{I}$. What's more, existing approaches to approximating the KL divergence do not scale to high input or target dimensions (Sun et al., 2019).

We propose a fundamentally different approach to function-space variational inference. Starting from Equation (1), we consider a locally accurate approximation to $q(f;\boldsymbol{\theta})$ and $p(f)$ by linearizing them about their mean parameters. Assuming a Gaussian distribution over the network parameters, this approximation turns $q(f;\boldsymbol{\theta})$ and $p(f)$ into Gaussian processes. To evaluate the resulting locally accurate KL divergence, we make a prior conditional matching assumption, which results in a tractable KL divergence evaluated at a *finite* number of evaluation points. We present this approximation in more detail next.

4. Function-Space Variational Inference via Local Linearization

The primary obstacle to making the objective in Equation (1) tractable is the KL divergence from $q(f;\boldsymbol{\theta})$ to $p(f)$. We start by considering the distribution over parameters that gives rise to the distribution over functions $q(f;\boldsymbol{\theta})$. Specifically,

Assumption 1 (Mean-Field Variational Distribution Over Parameters):
Assume a factorized Gaussian variational distribution, $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

We denote the distribution over functions induced by $q(\boldsymbol{\theta})$ as $q(f;\boldsymbol{\theta})$. To obtain a tractable distribution over functions, we make a local approximation about the BNN's mean parameters:

Assumption 2 (Linearization about Mean Parameters):
Linearize the stochastic function f about its mean parameters $\boldsymbol{\mu}$, to obtain the locally accurate approximation $f(\mathbf{x};\boldsymbol{\theta}) \approx \tilde{f}(\mathbf{x};\boldsymbol{\theta}) \equiv f(\mathbf{x};\boldsymbol{\mu}) + \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})(\boldsymbol{\theta} - \boldsymbol{\mu})$, where $\mathcal{J}_{\boldsymbol{\mu}}$ denotes the Jacobian $\frac{\partial f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\mu}}$ and $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Due to local linearity, the approximation $\tilde{f}(\mathbf{x};\boldsymbol{\theta})$ will be accurate for realizations $\hat{\boldsymbol{\theta}}$ close to $\boldsymbol{\mu}$, and hence, the distribution over $\tilde{f}(\mathbf{x};\boldsymbol{\theta})$ (induced by $\boldsymbol{\theta}$) will be close to the distribution over $f(\mathbf{x};\boldsymbol{\theta})$ for small variance parameters $\boldsymbol{\Sigma}$. Under the two assumptions above, we obtain a locally accurate approximate distribution over functions:

Proposition 1 (Predictive Distribution of Linearized BNN) *Consider $\boldsymbol{\theta}$, $f(\mathbf{x};\boldsymbol{\theta})$, and $\tilde{f}(\mathbf{x};\boldsymbol{\theta})$ as defined above. The mean and variance of $\tilde{q}(\tilde{f}(\mathbf{x};\boldsymbol{\theta}))$ are given by*

$$\mathbb{E}_{\tilde{q}(\tilde{f}(\mathbf{x};\boldsymbol{\theta}))}[\tilde{f}(\mathbf{x};\boldsymbol{\theta})] = f(\mathbf{x};\boldsymbol{\mu}) \quad \textit{and} \quad \mathbb{V}(\tilde{f}(\mathbf{x};\boldsymbol{\theta})) = \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})\boldsymbol{\Sigma}\mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x}')^\top$$

and the predictive distribution $\tilde{q}$ over $\tilde{f}(\mathbf{x};\boldsymbol{\theta})$ is a Gaussian process given by

$$\tilde{q}(\tilde{f}(\mathbf{x});\boldsymbol{\theta}) = \mathcal{GP}(\tilde{f} \mid f(\mathbf{x};\boldsymbol{\mu}), \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})\boldsymbol{\Sigma}\mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x}')^\top). \tag{3}$$

Proof See Appendix 1. ■

Under the linearization about $\boldsymbol{\mu}$, we obtain a local approximation to the objective:

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

where $\tilde{p}(\tilde{f})$ is a local approximation to a prior distribution over functions induced by *some* Gaussian prior distribution over the network parameters. The approximate variational objective in Equation (4) includes a KL divergence between two Gaussian processes $\tilde{q}(\tilde{f})$ and $\tilde{p}(\tilde{f})$. Unfortunately, in the absence of additional assumptions about $\tilde{q}(\tilde{f})$ and $\tilde{p}(\tilde{f})$, this KL divergence is still intractable (de G. Matthews et al., 2016).

To obtain a tractable variational objective, we make an assumption about how the variational predictive distribution $\tilde{q}(\tilde{f};\boldsymbol{\theta}) = \tilde{q}(\tilde{f}_{\mathbf{X}_*}, \tilde{f}_{\mathbf{X}_\mathcal{D}}, \tilde{f}_{\mathbf{X}_\mathcal{I}};\boldsymbol{\theta})$ factorizes, where $\mathbf{X}_\mathcal{D}$ is a finite set of training inputs, $\mathbf{X}_\mathcal{I}$ is a finite set of so-called inducing points, $\mathbf{X}_* \overset{\text{def}}{=} \mathcal{X}\backslash\{\mathbf{X}_\mathcal{D}, \mathbf{X}_\mathcal{I}\}$ is an infinite set of evaluation points containing all points in the data space except for $\mathbf{X}_\mathcal{D}$ and $\mathbf{X}_\mathcal{I}$, and $\tilde{f}_\mathbf{X}$ are function values at evaluation points $\mathbf{X}$. Specifically, we assume prior conditional matching, that is:

Assumption 3 (Prior Conditional Matching):
Let the variational distribution factorize as

$$\tilde{q}(\tilde{f}_{\mathbf{X}_*}, \tilde{f}_{\mathbf{X}_\mathcal{D}}, \tilde{f}_{\mathbf{X}_\mathcal{I}};\boldsymbol{\theta}) \overset{\text{def}}{=} \tilde{p}(\tilde{f}_{\mathbf{X}_*} \mid \tilde{f}_{\mathbf{X}_\mathcal{D}})\tilde{p}(\tilde{f}_{\mathbf{X}_\mathcal{D}} \mid \tilde{f}_{\mathbf{X}_\mathcal{I}})\tilde{q}(\tilde{f}_{\mathbf{X}_\mathcal{I}};\boldsymbol{\theta}). \tag{5}$$

Under Assumption 3, we can now simplify the function-space variational objective as follows:

Proposition 2 *Under Assumptions 1, 2, and 3, we obtain the variational objective*

$$\mathcal{F}(q(f;\boldsymbol{\theta})) = \mathbb{E}_{q(f_{\mathbf{X}_\mathcal{D}};\boldsymbol{\theta})}[\log p(\mathbf{y} \mid f_{\mathbf{X}_\mathcal{D}})] - \mathbb{D}_{\mathrm{KL}}(\tilde{q}(\tilde{f}_{\mathbf{X}_\mathcal{I}};\boldsymbol{\theta}) \,\|\, \tilde{p}(\tilde{f}_{\mathbf{X}_\mathcal{I}})), \tag{6}$$

where $\mathbb{D}_{\mathrm{KL}}(\tilde{q}(\tilde{f}_{\mathbf{X}_\mathcal{I}};\boldsymbol{\theta}) \,\|\, \tilde{p}(\tilde{f}_{\mathbf{X}_\mathcal{I}}))$ is analytically tractable.

Proof Sketch We can then express the variational objective in Equation (4) as

$$\mathcal{F}(q(f;\boldsymbol{\theta})) = \mathbb{E}_{q(f_{\mathbf{X}_\mathcal{D}};\boldsymbol{\theta})}[\log p(\mathbf{y} \mid f_{\mathbf{X}_\mathcal{D}})] - \mathbb{D}_{\mathrm{KL}}(\tilde{q}(\tilde{f}_{\mathbf{X}_*}, \tilde{f}_{\mathbf{X}_\mathcal{D}}, \tilde{f}_{\mathbf{X}_\mathcal{I}};\boldsymbol{\theta}) \,\|\, \tilde{p}(\tilde{f}_{\mathbf{X}_*}, \tilde{f}_{\mathbf{X}_\mathcal{D}}, \tilde{f}_{\mathbf{X}_\mathcal{I}})), \tag{7}$$

and under prior conditional matching, this objectives becomes

$$\mathbb{D}_{\mathrm{KL}}(\tilde{p}(\tilde{f}_{\mathbf{X}_*} \mid \tilde{f}_{\mathbf{X}_\mathcal{D}})\tilde{p}(\tilde{f}_{\mathbf{X}_\mathcal{D}} \mid \tilde{f}_{\mathbf{X}_\mathcal{I}})\tilde{q}(\tilde{f}_{\mathbf{X}_\mathcal{I}}) \,\|\, \tilde{p}(\tilde{f}_{\mathbf{X}_*} \mid \tilde{f}_{\mathbf{X}_\mathcal{D}})\tilde{p}(\tilde{f}_{\mathbf{X}_\mathcal{D}} \mid \tilde{f}_{\mathbf{X}_\mathcal{I}})\tilde{p}(\tilde{f}_{\mathbf{X}_\mathcal{I}})), \tag{8}$$

which simplifies to $\mathbb{D}_{\mathrm{KL}}(\tilde{q}(\tilde{f}_{\mathbf{X}_\mathcal{I}};\boldsymbol{\theta}) \,\|\, \tilde{p}(\tilde{f}_{\mathbf{X}_\mathcal{I}}))$, which is a KL divergence between multivariate Gaussian distributions and can be expressed analytically. For a full, measure-theoretic proof, see de G. Matthews et al. (2016, Sections 3.2 and 3.3) ■

Algorithm 1 FSVI: Function-Space Variational Inference

```
Input: data $\mathcal{D}$, size $|\mathbf{X}_{\mathcal{I}}|$, learning rate $\eta$, prior mean $\boldsymbol{\mu}_0$, prior variance $\boldsymbol{\Sigma}_0$;
Initialization: $\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}_0}$, $\mathcal{I} \sim p_{\mathcal{I}}$; while $\ell(q(f;\boldsymbol{\theta})) - \mathbb{D}_{\mathrm{KL}}(\tilde{q}(\tilde{f}_{\mathbf{X}_{\mathcal{I}}};\boldsymbol{\theta}) \,\|\, \tilde{p}(\tilde{f}_{\mathbf{X}_{\mathcal{I}}}))$ not converged do
    $\mathbf{X}_{\mathcal{I}} \subset \mathcal{I}$, $\mathcal{B} \subset \mathcal{D}$
    $\boldsymbol{\Theta}(\mathbf{X}_{\mathcal{I}}, \mathbf{X}_{\mathcal{I}}) = \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{X}_{\mathcal{I}}) \boldsymbol{\Sigma} \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{X}_{\mathcal{I}})^\top$
    $\ell(q(f;\boldsymbol{\theta})) = \frac{1}{S} \sum_{i=1}^{S} \sum_{(\mathbf{X}_{\mathcal{B}}, \mathbf{y}_{\mathcal{B}})} \log p(\mathbf{y}_{\mathcal{B}} \,|\, f(\mathbf{X}_{\mathcal{B}}, \boldsymbol{\epsilon}_i; \boldsymbol{\theta}))$, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{\epsilon} \,|\, \mathbf{0}, \mathbf{I})$
    $\mathbb{D}_{\mathrm{KL}}(\tilde{q}(\tilde{f}_{\mathbf{X}_{\mathcal{I}}};\boldsymbol{\theta}) \,\|\, \tilde{p}(\tilde{f}_{\mathbf{X}_{\mathcal{I}}})) = \sum_{j=1}^{|\mathbf{X}_{\mathcal{I}}|} -\frac{1}{2} \log \frac{\sqrt{[\boldsymbol{\Sigma}_0]_{jj}}}{\sqrt{[\boldsymbol{\Theta}(\mathbf{X}_{\mathcal{I}}, \mathbf{X}_{\mathcal{I}})]_{jj}}} + \frac{[\boldsymbol{\Theta}(\mathbf{X}_{\mathcal{I}}, \mathbf{X}_{\mathcal{I}})]_{jj} + (f(\mathbf{X}_{\mathcal{I}}; \boldsymbol{\mu}) - \boldsymbol{\mu}_0)^2}{2[\boldsymbol{\Sigma}_0]_{jj}}$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \left( \ell(q(f;\boldsymbol{\theta})) - \mathbb{D}_{\mathrm{KL}}(\tilde{q}(\tilde{f}_{\mathbf{X}_{\mathcal{I}}};\boldsymbol{\theta}) \,\|\, \tilde{p}(\tilde{f}_{\mathbf{X}_{\mathcal{I}}})) \right)$
```

5. Empirical Evaluation

We evaluate the proposed function-space variational inference (FSVI) method on illustrative regression and classification tasks as well as on high-dimensional classification tasks prior work (Sun et al., 2019) was unable to scale to. 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 uncertainty estimation.

5.1. Illustrative Examples

Figure 1 shows the posterior predictive distribution obtained via FSVI on a 1D regression and a binary classification problem. On the regression problem (Figure 1(*a*)), the posterior predictive distribution is certain about the training data, becomes somewhat uncertain when interpolating between datapoints (in the interval $[-1, 0]$), and grows very uncertain away from the training data, as desired. Figures 1(*b*) and 1(*c*) show the posterior predictive mean and variance obtained via FSVI on the *Two Moons* classification task. As can be seen in Figure 1(*b*), the predictive mean (in the form of binary class probabilities) is highly confident around the data manifold and converges to 0.5, the maximum level of uncertainty, further away from it. Similarly, the epistemic uncertainty over the class probabilities shown in Figure 1(*c*), is low on and close to the data manifold and increases further away from it.

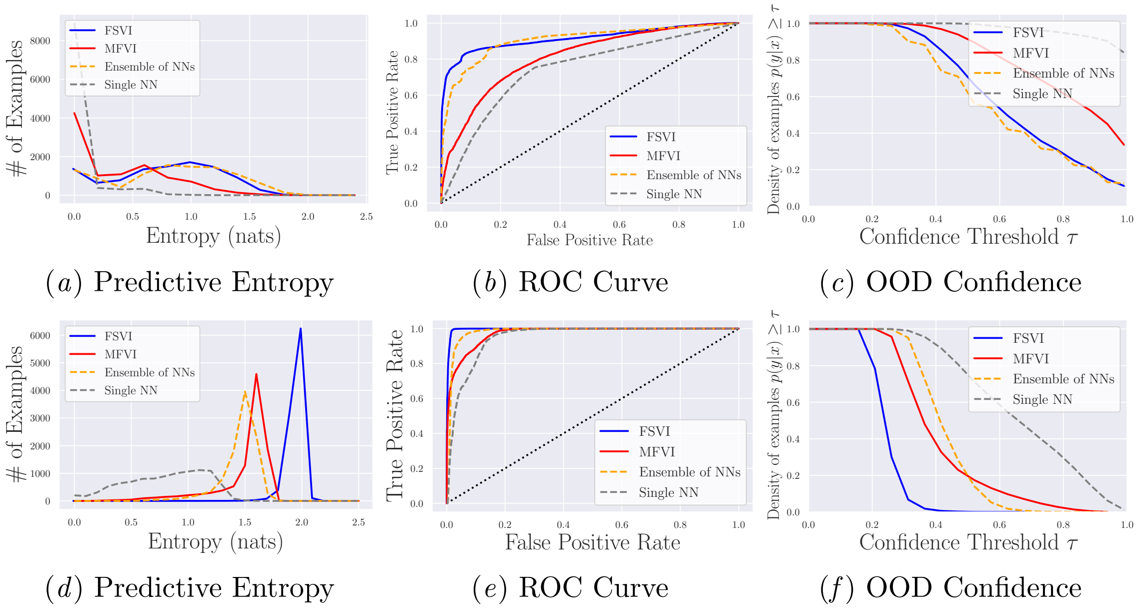
5.2. Evaluation of In- and Out-of-Distribution Performance

To evaluate the predictive performance of FSVI, we consider a selection of widely used high-dimensional classification datasets to which prior approaches to function-space variational inference were unable to scale.

Table 1 and Figure 2 show that FSVI consistently either performs on par with related methods or outperforms them. More specifically, Figure 2, shows that FSVI has as a predictive entropy on out-of-distribution inputs that is as high or higher than that of ensembles or BNNs with MFVI, indicating high uncertainty under distribution shift, as we would like. This high level of uncertainty is reflected in the corresponding receiver operating characteristic (ROC) curve, which shows that FSVI outperforms other methods at distinguishing in- from out-of-distribution inputs. Finally, as can be seen in the rightmost column of Figure 2,

Table 1: Comparison of in- and out-of-distribution performance metrics. AUROC: area under ROC curve. ECE: expected calibration error. [1]Computed from mutual information scores. [2]Computed from predictive entropy scores. [3]BNN trained with MFVI (Blundell et al., 2015). [4]BNN MAP estimate obtained by training a deterministic neural network with weight regularization.

	fashionMNIST/MNIST				CIFAR-10/SVHN			
Model	Accuracy	ECE	AUROC[1]	AUROC[2]	Accuracy	ECE	AUROC[1]	AUROC[2]
Ours: FSVI	91%	**0.02**	**91%**	86%	85%	**0.03**	54%	**99%**
MFVI[3]	91%	0.04	85%	84%	85%	0.04	68%	97%
MAP[4]	91%	0.07	72%	76%	86%	0.09	72%	90%
Ensemble	**93%**	**0.02**	**91%**	90%	**91%**	**0.03**	88%	98%



(a) Predictive Entropy

(b) ROC Curve

(c) OOD Confidence

(d) Predictive Entropy

(e) ROC Curve

(f) OOD Confidence

Figure 2: Uncertainty evaluation metrics for in- and out-of-distribution (OOD) prediction. Top row: results for models trained on fashionMNIST, with MNIST images as OOD inputs; Bottom row: results for models trained on CIFAR-10, with SVHN images as OOD inputs. Left column: closer to diagonal is better; Center column: closer to top left corner is better; Right column: closer to bottom left corner is better. For further details, see Appendix 4.

FSVI exhibits low confidence on out-of-distribution inputs, as desired. Table 1 further corroborates these observations and shows that FSVI leads to high predictive accuracy *and* good uncertainty estimates. In Appendix 4, we provide an ablation study on the effect of the number of inducing points on a BNN's predictive accuracy and the quality of its predictive uncertainty estimates.

6. Conclusion

We proposed a fundamentally new approach to variational inference in BNNs, where the parameters are inferred *indirectly* by performing inference over the induced distribution over functions. We showed that FSVI exhibits an in- and out-of-distribution predictive performance on par or better than related state-of-the-art Bayesian and non-Bayesian approaches.

References

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.

John S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In Françoise Fogelman Soulié and Jeanny Hérault, editors, *Neurocomputing*, pages 227–236, Berlin, Heidelberg, 1990. Springer Berlin Heidelberg. ISBN 978-3-642-76153-9.

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.

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.

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.

Supplementary Materials

1. Theoretical Results

1.1. Linearization of Bayesian Neural Network Predictive Distribution

Proposition 1 (Predictive Distribution of Linearized BNN):
Consider $\boldsymbol{\theta}$, $f(\mathbf{x};\boldsymbol{\theta})$, and $\tilde{f}(\mathbf{x};\boldsymbol{\theta})$ as defined above. The mean and variance of $\tilde{q}(\tilde{f}(\mathbf{x};\boldsymbol{\theta}))$ are given by

$$\mathbb{E}_{\tilde{q}(\tilde{f}(\mathbf{x};\boldsymbol{\theta}))}[\tilde{f}(\mathbf{x};\boldsymbol{\theta})] = f(\mathbf{x};\boldsymbol{\mu}) \quad \text{and} \quad \mathbb{V}(\tilde{f}(\mathbf{x};\boldsymbol{\theta})) = \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})\boldsymbol{\Sigma}\mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x}')^\top$$

and the predictive distribution $\tilde{q}$ over $\tilde{f}(\mathbf{x};\boldsymbol{\theta})$ is given by

$$\tilde{q}(\tilde{f}(\mathbf{x});\boldsymbol{\theta}) = \mathcal{GP}(\tilde{f} \mid f(\mathbf{x};\boldsymbol{\mu}), \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})\boldsymbol{\Sigma}\mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x}')^\top). \tag{1.1}$$

Proof Since $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\Sigma})$, and $\tilde{f}(\mathbf{x};\boldsymbol{\theta}) = f(\mathbf{x};\boldsymbol{\mu}) + \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})(\boldsymbol{\theta}-\boldsymbol{\mu})$ is a linear transformation of $\boldsymbol{\theta}$, $\tilde{f}(\mathbf{x};\boldsymbol{\theta})$ is a Gaussian process

$$\tilde{q}(\tilde{f}(\mathbf{x});\boldsymbol{\theta}) = \mathcal{GP}(\tilde{f}|m(\mathbf{x}), S(\mathbf{x},\mathbf{x}')) \tag{1.2}$$

with some predictive mean $m(\mathbf{x})$ and predictive covariance $S(\mathbf{x},\mathbf{x}')$. To find $\tilde{q}(\tilde{f}(\mathbf{x};\boldsymbol{\theta}))$, we need to find the predictive mean $m(\mathbf{x})$ and the predictive covariance $S(\mathbf{x},\mathbf{x}')$, which, by definition, we can write as:

$$m(\mathbf{x}) = \mathbb{E}[\tilde{f}(\mathbf{x};\boldsymbol{\theta})] \tag{1.3}$$

and

$$\begin{aligned} S(\mathbf{x},\mathbf{x}') &= \mathrm{Cov}(\tilde{f}(\mathbf{x};\boldsymbol{\theta}), \tilde{f}(\mathbf{x}';\boldsymbol{\theta})) && (1.4)\\ &= \mathbb{E}[(\tilde{f}(\mathbf{x};\boldsymbol{\theta}) - \mathbb{E}[\tilde{f}(\mathbf{x};\boldsymbol{\theta})])\,(\tilde{f}(\mathbf{x}';\boldsymbol{\theta}) - \mathbb{E}[\tilde{f}(\mathbf{x}';\boldsymbol{\theta})])^\top]. && (1.5)\end{aligned}$$

To see that $m(\mathbf{x}) = \mathbb{E}[\tilde{f}(\mathbf{x};\boldsymbol{\theta})] = f(\mathbf{x};\boldsymbol{\mu})$, note that, by linearity of expectation, we have

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[\tilde{f}(\mathbf{x};\boldsymbol{\theta})] && (1.6)\\ &= \mathbb{E}[f(\mathbf{x};\boldsymbol{\mu}) + \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})(\boldsymbol{\theta}-\boldsymbol{\mu})] && (1.7)\\ &= f(\mathbf{x};\boldsymbol{\mu}) + \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})(\mathbb{E}[\boldsymbol{\theta}]-\boldsymbol{\mu}) && (1.8)\\ &= f(\mathbf{x};\boldsymbol{\mu}). && (1.9)\end{aligned}$$

To see that $S(\mathbf{x},\mathbf{x}') = \mathrm{Cov}(\tilde{f}(\mathbf{x};\boldsymbol{\theta}), \tilde{f}(\mathbf{x}';\boldsymbol{\theta})) = \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})\boldsymbol{\Sigma}\mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x}')^\top$, note that in general, $\mathrm{Cov}(\mathbf{X},\mathbf{X}) = \mathbb{E}[\mathbf{X}\mathbf{X}^\top] + \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^\top$, and hence,

$$\mathrm{Cov}(\tilde{f}(\mathbf{x};\boldsymbol{\theta}), \tilde{f}(\mathbf{x}';\boldsymbol{\theta})) = \mathbb{E}[\tilde{f}(\mathbf{x};\boldsymbol{\theta})\tilde{f}(\mathbf{x}';\boldsymbol{\theta})^\top] - \mathbb{E}[\tilde{f}(\mathbf{x};\boldsymbol{\theta})]\mathbb{E}[\tilde{f}(\mathbf{x}';\boldsymbol{\theta})]^\top. \tag{1.10}$$

We already know that $\mathbb{E}[\tilde{f}(\mathbf{x};\boldsymbol{\theta})] = f_{\boldsymbol{\mu}}(\mathbf{x})$, so we only need to find $\mathbb{E}[\tilde{f}(\mathbf{x};\boldsymbol{\theta})\tilde{f}(\mathbf{x}';\boldsymbol{\theta})^\top]$:

$$\begin{aligned} &\mathbb{E}_{q(\boldsymbol{\theta})}[\tilde{f}(\mathbf{x};\boldsymbol{\theta})\tilde{f}(\mathbf{x}';\boldsymbol{\theta})^\top]\\ &\quad=\mathbb{E}_{q(\boldsymbol{\theta})}[(f(\mathbf{x};\boldsymbol{\mu}) + \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})(\boldsymbol{\theta}-\boldsymbol{\mu}))(f(\mathbf{x}';\boldsymbol{\mu}) + \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x}')(\boldsymbol{\theta}-\boldsymbol{\mu}))^\top] && (1.11)\\ &\quad=\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0},\mathbf{I})}[(f_{\boldsymbol{\mu}}(\mathbf{x}) + \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})(\boldsymbol{\mu}+\boldsymbol{\epsilon}\odot\sqrt{\boldsymbol{\Sigma}}-\boldsymbol{\mu}))(f_{\boldsymbol{\mu}}(\mathbf{x}) + \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x}')(\boldsymbol{\mu}+\boldsymbol{\epsilon}\odot\sqrt{\boldsymbol{\Sigma}}-\boldsymbol{\mu}))^\top] && (1.12)\\ &\quad=\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0},\mathbf{I})}[(f(\mathbf{x};\boldsymbol{\mu}) + \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})(\boldsymbol{\epsilon}\odot\sqrt{\boldsymbol{\Sigma}}))(f(\mathbf{x}';\boldsymbol{\mu}) + \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x}')(\boldsymbol{\epsilon}\odot\sqrt{\boldsymbol{\Sigma}}))^\top], && (1.13)\end{aligned}$$

where the reparameterization is possible by Assumption 1. With some algebra, this expression can be further simplified to

$$\mathbb{E}_{q(\boldsymbol{\theta})}[\tilde{f}(\mathbf{x};\boldsymbol{\theta})\tilde{f}(\mathbf{x}';\boldsymbol{\theta})^\top] = f(\mathbf{x};\boldsymbol{\mu})f(\mathbf{x}';\boldsymbol{\mu})^\top + \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})\boldsymbol{\Sigma}\mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x}')^\top. \tag{1.14}$$

Since the $f_{\boldsymbol{\mu}}(\mathbf{x})f_{\boldsymbol{\mu}}(\mathbf{x}')^\top$ terms cancel out, we obtain the covariance function

$$S(\mathbf{x},\mathbf{x}') \overset{\text{def}}{=} \operatorname{Cov}(\tilde{f}(\mathbf{x};\boldsymbol{\theta}),\tilde{f}(\mathbf{x}';\boldsymbol{\theta})) = \mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x})\boldsymbol{\Sigma}\mathcal{J}_{\boldsymbol{\mu}}(\mathbf{x}')^\top, \tag{1.15}$$

which concludes the proof. ■

1.2. Function-Space Variational Objective

This proof follows steps from de G. Matthews et al. (2016). Consider measures $\hat{P}$ and P both of which define distributions over some function f, indexed by an infinite index set X. Let $\mathcal{D}$ be a dataset and let $\mathbf{X}_\mathcal{D}$ denote a set of inputs and $\mathbf{y}_\mathcal{D}$ a set of targets. Consider the measure-theoretic version of Bayes' Theorem (Schervish, 1995):

$$\frac{d\hat{P}}{dP}(f) = \frac{p_X(Y\,|\,f)}{p(Y)}, \tag{1.16}$$

where $p_X(Y\,|\,f)$ is the likelihood and $p(Y) = \int_{\mathbb{R}^X} p_X(Y\,|\,f)dP(f)$ is the marginal likelihood. We assume that the likelihood function is evaluated at a finite subset of the index set X. Denote by $\pi_C : \mathbb{R}^X \to \mathbb{R}^C$ a projection function that takes a function and returns the same function, evaluated at a finite set of points C, so we can write

$$\frac{d\hat{P}}{dP}(f) = \frac{d\hat{P}_{\mathbf{X}_\mathcal{D}}}{dP_{\mathbf{X}_\mathcal{D}}}(\pi_{\mathbf{X}_\mathcal{D}}(f)) = \frac{p(\mathbf{y}_\mathcal{D}\,|\,\pi_{\mathbf{X}_\mathcal{D}}(f))}{p(\mathbf{y}_\mathcal{D})}, \tag{1.17}$$

and similarly, the marginal likelihood becomes $p(\mathbf{y}_\mathcal{D}) = \int_{\mathbf{X}_\mathcal{D}} p(\mathbf{y}_\mathcal{D}\,|\,f_{\mathbf{X}_\mathcal{D}})dP_{\mathbf{X}_\mathcal{D}}(f_{\mathbf{X}_\mathcal{D}})$. Now, considering the measure-theoretic version of the KL divergence between an approximating stochastic process Q and a posterior stochastic process $\hat{P}$, we can write

$$\mathbb{D}_{\mathrm{KL}}(Q\,\|\,\hat{P}) = \int_{\mathbb{R}^X} \log\frac{dQ}{dP}(f)dQ(f) - \int_{\mathbb{R}^X} \log\frac{d\hat{P}}{dP}(f)dQ(f), \tag{1.18}$$

where P is some prior stochastic process. Now, considering the second term, we can apply the measure-theoretic Bayes' Theorem to obtain

$$\begin{aligned}\int_{\mathbb{R}^X} \log\frac{d\hat{P}}{dP}(f)dQ(f) &= \int_{\mathbb{R}^{\mathbf{X}_\mathcal{D}}} \log\frac{d\hat{P}_{\mathbf{X}_\mathcal{D}}}{dP_{\mathbf{X}_\mathcal{D}}}(f_{\mathbf{X}_\mathcal{D}})dQ_{\mathbf{X}_\mathcal{D}}(f_{\mathbf{X}_\mathcal{D}}) \qquad (1.19)\\ &= \mathbb{E}_{Q_{\mathbf{X}_\mathcal{D}}}[\log p(\mathbf{y}_\mathcal{D}\,|\,f_{\mathbf{X}_\mathcal{D}})] - \log p(\mathbf{y}_\mathcal{D}), \qquad (1.20)\end{aligned}$$

giving us

$$\mathbb{D}_{\mathrm{KL}}(Q\,\|\,\hat{P}) = \int_{\mathbb{R}^X} \log\frac{dQ}{dP}(f)dQ(f) - \mathbb{E}_{Q_{\mathbf{X}_\mathcal{D}}}[\log p(\mathbf{y}_\mathcal{D}\,|\,f_{\mathbf{X}_\mathcal{D}})] + \log p(\mathbf{y}_\mathcal{D}). \tag{1.21}$$

Rearranging, we can get

$$p(\mathbf{y}_{\mathcal{D}}) = \mathbb{E}_{Q_{\mathbf{X}_{\mathcal{D}}}}\left[\log p\left(\mathbf{y}_{\mathcal{D}} \mid f_{\mathbf{X}_{\mathcal{D}}}\right)\right] - \int_{\mathbb{R}^X} \log \frac{dQ}{dP}(f) dQ(f) + \mathbb{D}_{\mathrm{KL}}(Q \,\|\, P]) \tag{1.22}$$

$$\geq \mathbb{E}_{Q_{\mathbf{X}_{\mathcal{D}}}}\left[\log p\left(\mathbf{y}_{\mathcal{D}} \mid f_{\mathbf{X}_{\mathcal{D}}}\right)\right] - \int_{\mathbb{R}^X} \log \frac{dQ}{dP}(f) dQ(f). \tag{1.23}$$

By the measure-theoretic definition of the KL divergence, we can thus write

$$p(\mathbf{y}_{\mathcal{D}}) \geq \mathbb{E}_{Q_{\mathbf{X}_{\mathcal{D}}}}\left[\log p\left(\mathbf{y}_{\mathcal{D}} \mid f_{\mathbf{X}_{\mathcal{D}}}\right)\right] - \int_{\mathbb{R}^X} \log \frac{dQ}{dP}(f) dQ(f) \tag{1.24}$$

$$= \mathbb{E}_{Q_{\mathbf{X}_{\mathcal{D}}}}\left[\log p\left(\mathbf{y}_{\mathcal{D}} \mid f_{\mathbf{X}_{\mathcal{D}}}\right)\right] - \mathbb{D}_{\mathrm{KL}}(Q \,\|\, P), \tag{1.25}$$

which corresponds to the expression for the function-space variational objective in Section 3.

2. Further Background

Mean-Field Variational Inference Since BNNs are non-linear in their parameters, exact inference over the network parameters is analytically intractable. Mean-field variational inference is a variational approach for finding an approximate posterior distribution over network parameters and and use it to draw samples from a BNN's approximate posterior predictive distribution. Under a mean-field assumption, the joint distribution over the stochastic network parameters has a diagonal covariance, rendering the variational parameters independent of one another. Furthermore, to obtain a tractable variational objective, prior works assume the prior and variational distributions over the networ parameters to be Gaussian (Blundell et al., 2015). This approximation results in a tractable and scalable variational objective, given by

$$\mathcal{L}(q(\boldsymbol{\theta})) \overset{\text{def}}{=} \mathbb{E}_{q(f_{\mathbf{X}_{\mathcal{D}}};\boldsymbol{\theta})}[\log p(\mathbf{y} \,|\, f_{\mathbf{X}_{\mathcal{D}}})] - \mathbb{D}_{\mathrm{KL}}(q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta})), \tag{2.26}$$

where $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \,|\, \mathbf{0}, \mathbf{I})$, and $f_{\mathbf{X}_{\mathcal{D}}}$ are function values at the training inputs $\mathbf{X}_{\mathcal{D}}$.

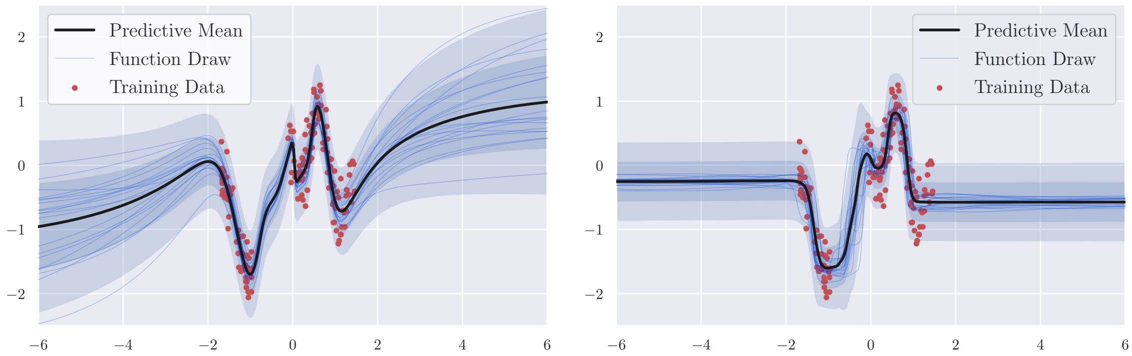
Gaussian Processes Gaussian process (GP) models define distributions over functions. Unlike in BNNs where a prior distribution over parameters *implicitly* induces a prior distribution over functions, Gaussian processes *explicitly* define distributions over functions by specifying a covariance function over possible function realizations. A GP prior $p(f \,|\, \mathbf{x}) = \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ is completely specified by its mean and covariance function, $m(\cdot)$ and $k(\cdot, \cdot)$.

3. Model Details

For the experiments presented in this paper, we diagonalized the covariance of the linearized BNN in the KL divergence. While this simplification is not necessary, it speeds up training larger numbers of inducing points. Furthermore, we assumed a BNN prior that is locally equal to a GP with zero mean and a diagonal covariance scaled by 1e6 on all evaluation points. We chose the set of inducing points by uniformly sampling within some range, e.g., within the range of admissible pixel values for prediction tasks with image inputs.

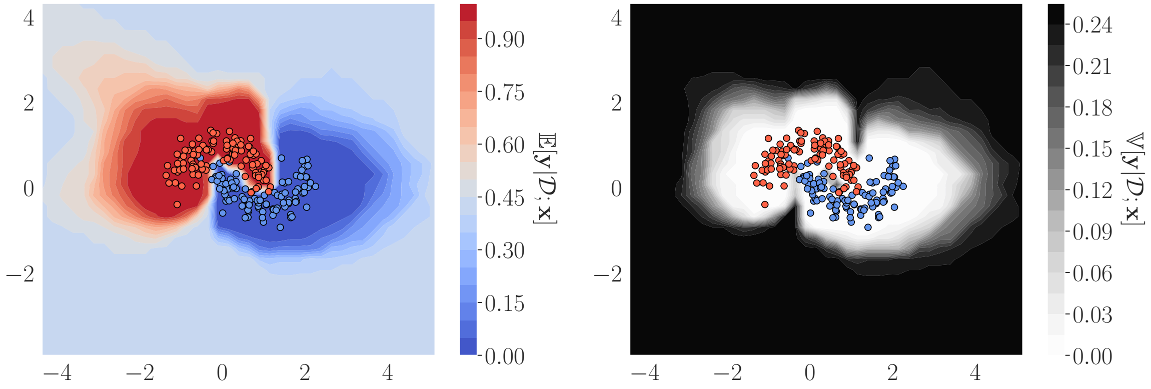
For the *Two Moons* and *Snelson* experiments, we use an fully-connected neural network with two hidden layers and 100 hidden units per layer. For the fashionMNIST/MNIST experiments, we use a three-layer convolutional neural network without batch normalization. For the CIFAR-10/SVHN experiments, we use a seven layer convolutional neural network without batch normalization. We use the Adam optimizer for all experiments with learning rates between η = 1e-4 and η = 1e-3 (depending on which yielded the best performance). The deterministic neural networks that were used for the ensemble were trained with a weight decay of λ = 1e-1. We used early stopping when training BNNs with MFVI to avoid overfitting. We did not use early stopping to train BNNs with FSVI.

4. Further Empirical Results



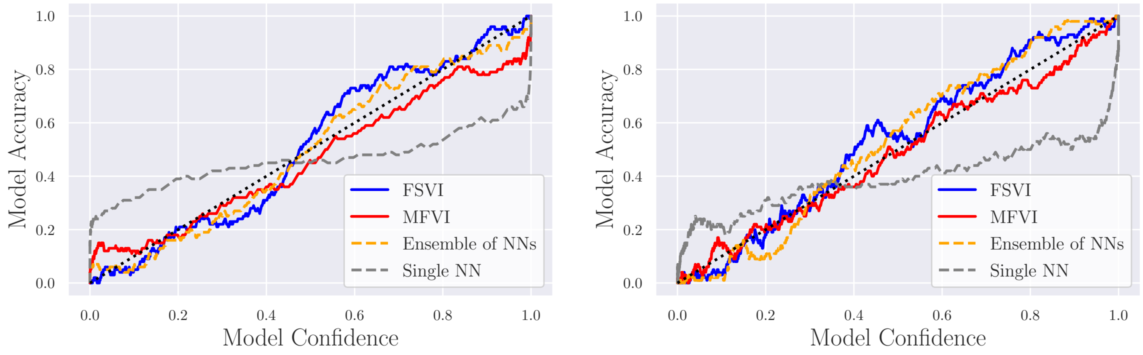
(a) FSVI Posterior Predictive Distribution (b) MFVI Posterior Predictive Distribution

Figure 3: 1D Regression on the *Snelson* datasets. The plots show the predictive distribution of a BNN obtained via function-space variational inference (FSVI) under the local approximation described in Section 4 (Figure 3(a)) and obtained via mean-field variational inference (MFVI) (Figure 3(b)). The plots show noisy data (in **red**), the posterior predictive means (in **black**), ten function draws from the BNNs (in **blue**), and two standard deviations of the empirical distribution over functions.



(a) Posterior Predictive Mean (b) Posterior Predictive Variance

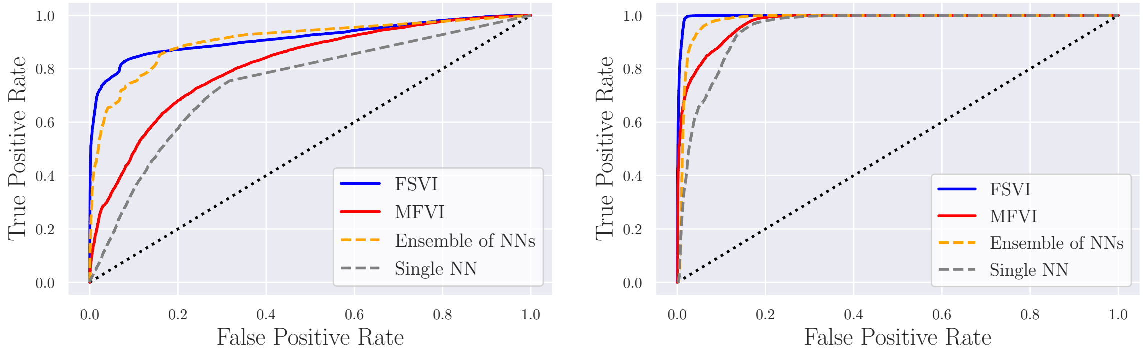
Figure 4: Binary classification on the *Two Moons* dataset. The plots show the posterior predictive mean (Figure 4(a)) and variance (Figure 4(b)) of a BNN obtained via FSVI. They represent the expected class probabilities and the model's epistemic uncertainty over the class probabilities, respectively. The predictive distribution is able to faithfully capture the geometry of the data manifold and exhibits high uncertainty over the class probabilities in areas of the data space of which the data is not informative. In contrast, related methods, such as ensembles, are unable to accurately capture the geometry of the data manifold only exhibit high uncertainty around the decision boundary (van Amersfoort et al., 2020).



(a) fashionMNIST

(b) CIFAR-10

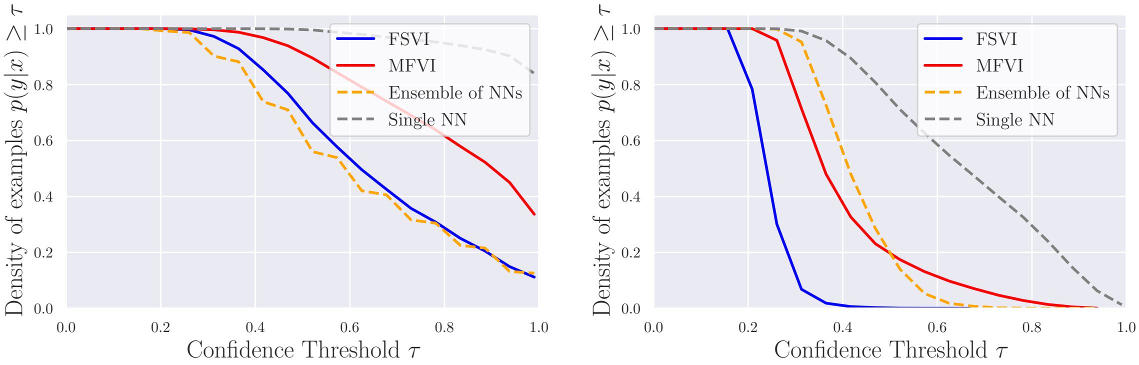
Figure 5: Reliability Diagram. Figure 5(a) shows the reliability of different models, expressed as a plot of model accuracy against model confidence for models trained on fashionMNIST and evaluated on a fashionMNIST test set. Figure 5(b) shows the reliability of different models, expressed as a plot of model accuracy against model confidence for models trained on CIFAR-10 and evaluated on a CIFAR-10 test set.



(a) fashionMNIST/MNIST

(b) CIFAR-10/SVHN

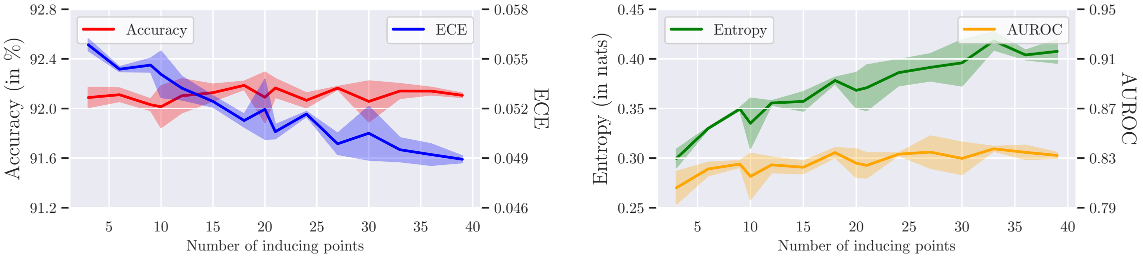
Figure 6: Receiver operating characteristic (ROC) curve. Figure 6(a) shows the ROC for different predictive distributions for the binary classification problem of distinguishing in-distribution inputs (fashionMNIST) from out-of-distribution inputs (MNIST). Figure 6(b) shows the ROC for different predictive distributions for the binary classification problem of distinguishing in-distribution inputs (CIFAR-10) from out-of-distribution inputs (SVHN).



(a) fashionMNIST/MNIST

(b) CIFAR-10/SVHN

Figure 7: Confidence on Out-of-Distribution Inputs. Figure 7(a) shows the confidence of different predictive means of models trained on fashionMNIST, evaluated on out-of-distribution inputs (MNIST). Figure 7(b) shows the confidence of different predictive means of models trained on CIFAR-10, evaluated on out-of-distribution inputs (SVHN). Curves further to the left are better, as they indicate that a model assigns low confidence to a higher number of out-of-distribution inputs. Curves that cover a larger area (i.e., that are further to the top left) are better, as they indicate a higher true than false positive rate.



(a) Effect on In-Distribution Predictions

(b) Effect on Out-of-Distribution Predictions

Figure 8: Figure 8(a) shows the effect of increasing the number of inducing points on in-distribution predictions for BNNs trained viaFSVI on fashionMNIST. Increasing the number of inducing points does not affect the test accuracy, but does increase the predictive mean's expected calibration error. Figure 8(b) shows that increasing the the number of inducing points also increases the predictive entropy on out-of-distribution inputs as well as the area under the receiver operating characteristic curve computed from it.