Tractable Function-Space Variational Inference in Bayesian Neural Networks

Tim G. J. Rudner[†] [*]   Zonghao Chen[‡]   Yee Whye Teh[†]   Yarin Gal[†]

[†] University of Oxford [‡] University College London [*] Corresponding author: tim.rudner@cs.ox.ac.uk.   @timrudner

Summary

- We propose a function-space approach to variational inference in BNNs and derive a *tractable* function-space variational objective by approximating the BNN's variational and prior distributions via linearization of the function mapping.
- This approach leads to competitive predictive accuracy and significantly improved predictive uncertainty estimates compared to related methods, including deep ensembles, the Laplace approximation, parameter-space variational inference, and Monte Carlo Dropout.

Background

- Consider data $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N = (\mathbf{X}_\mathcal{D}, \mathbf{y}_\mathcal{D})$ 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.
- Consider a function mapping defined by a neural network architecture given by $f : \mathcal{X} \times \mathbb{R}^P \to \mathbb{R}^Q$

Parameter-Space Variational Inference in BNNs

- Goal: Find posterior over parameters $p(\boldsymbol{\theta} \mid \mathcal{D})$.
- Find variationally via $\min_{q(\boldsymbol{\theta}) \in \mathcal{Q}_\Theta} \mathbb{D}_{\mathrm{KL}}(q_\Theta \,\|\, p_{\Theta|\mathcal{D}})$,

$$\Leftrightarrow \max_{q(\boldsymbol{\theta}) \in \mathcal{Q}_\Theta} \{\mathbb{E}_{q_\Theta}[\log p(\mathbf{y} \mid f(\mathbf{X}_\mathcal{D}; \boldsymbol{\theta}))] - \mathbb{D}_{\mathrm{KL}}(q_\Theta \,\|\, p_\Theta)\}$$

Induced Distributions over Functions

- $f(\cdot\,; \Theta)$ is a random function induced by random vector Θ.
- Prior distribution over functions (induced by p_Θ):

$$p_{f(\cdot\,;\Theta)}(f(\cdot\,;\boldsymbol{\theta})) = \int_{\mathbb{R}^P} p_\Theta(\boldsymbol{\theta}')\, \delta(f(\cdot\,;\boldsymbol{\theta}) - f(\cdot\,;\boldsymbol{\theta}'))\, \mathrm{d}\boldsymbol{\theta}' \quad (1)$$

- Posterior distribution over functions (induced by $p_{\Theta|\mathcal{D}}$):

$$p_{f(\cdot\,;\Theta)|\mathcal{D}}(f(\cdot\,;\boldsymbol{\theta}) \mid \mathcal{D}) = \int_{\mathbb{R}^P} p_{\Theta|\mathcal{D}}(\boldsymbol{\theta}' \mid \mathcal{D})\, \delta(f(\cdot\,;\boldsymbol{\theta}) - f(\cdot\,;\boldsymbol{\theta}'))\, \mathrm{d}\boldsymbol{\theta}' \quad (2)$$

Function-Space Variational Inference in BNNs

Function-Space Variational Inference

- Goal: Find posterior over functions $p(f(\cdot\,;\boldsymbol{\theta}) \mid \mathcal{D})$.
- Find variationally via

$$\min_{q_\Theta \in \mathcal{Q}_\theta} \mathbb{D}_{\mathrm{KL}}(q_{f(\cdot\,;\Theta)} \,\|\, p_{f(\cdot\,;\Theta)|\mathcal{D}}), \quad (3)$$

where

$$q_{f(\cdot\,;\Theta)}(f(\cdot\,;\boldsymbol{\theta})) = \int_{\mathbb{R}^P} q_\Theta(\boldsymbol{\theta}')\, \delta(f(\cdot\,;\boldsymbol{\theta}) - f(\cdot\,;\boldsymbol{\theta}'))\, \mathrm{d}\boldsymbol{\theta}' \quad (4)$$

- Data Processing Inequality (Polyanskiy and Wu, 2017):

$$\mathbb{D}_{\mathrm{KL}}(q_{f(\cdot\,;\Theta)} \,\|\, p_{f(\cdot\,;\Theta)}) \leq \mathbb{D}_{\mathrm{KL}}(q_\Theta \,\|\, p_\Theta) \quad (5)$$

- Function-space variational objective:

$$\mathcal{F}(q_\Theta) = \mathbb{E}_{q_{f(\mathbf{X}_\mathcal{D};\Theta)}}[\log p(\mathbf{y}_\mathcal{D} \mid f(\mathbf{X}_\mathcal{D}; \boldsymbol{\theta}))] - \sup_{\mathbf{X} \in \mathcal{X}_\mathbb{N}} \mathbb{D}_{\mathrm{KL}}(q_{f(\mathbf{X};\Theta)} \,\|\, p_{f(\mathbf{X};\Theta)}) \quad (6)$$

where $\mathcal{X}_\mathbb{N} \doteq \bigcup_{n \in \mathbb{N}} \{\mathbf{X} \in \mathcal{X}_n \mid \mathcal{X}_n \subseteq \mathbb{R}^{n \times D}\}$.

Approximations

1. Linearize mapping:

$$f(\cdot; \Theta) \approx \tilde{f}(\cdot; \Theta) \doteq f(\cdot; \mathbf{m}) + \mathcal{J}_\mathbf{m}(\cdot)(\Theta - \mathbf{m}) \quad (7)$$

with $\mathcal{J}_\mathbf{m}(\cdot) \doteq \left.\frac{\partial f(\cdot;\Theta)}{\partial \Theta}\right|_{\Theta=\mathbf{m}}$ to get

$$\tilde{q}_{\tilde{f}(\cdot\,;\Theta)}(\tilde{f}(\cdot\,;\boldsymbol{\theta})) \approx q_{f(\cdot\,;\Theta)}(f(\cdot\,;\boldsymbol{\theta})) \quad (8)$$

$$\tilde{p}_{\tilde{f}(\cdot\,;\Theta)}(\tilde{f}(\cdot\,;\boldsymbol{\theta})) \approx p_{f(\cdot\,;\Theta)}(f(\cdot\,;\boldsymbol{\theta})) \quad (9)$$

2. Estimate supremum via maximum over finite sample:

$$\max_{\mathbf{X} \in \mathcal{X}_\mathcal{C}^S} I(\mathbf{X}) \approx \sup_{\mathbf{X} \in \mathcal{X}_\mathbb{N}} I(\mathbf{X})), \text{ where } \mathcal{X}_\mathcal{C}^S \doteq \{\mathbf{X}_\mathcal{C}^{(i)}\}_{i=1}^S \quad (10)$$

Approximate Function-Space Variational Objective

- For $q_\Theta(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, maximize

$$\bar{\mathcal{F}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{M} \sum_{j=1}^M \log p(\mathbf{y}_\mathcal{B} \mid f(\mathbf{X}_\mathcal{B}; \hat{\Theta}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\epsilon}^{(j)}))) - \max_{\mathbf{X} \in \mathcal{X}_\mathcal{C}^S} \mathbb{D}_{\mathrm{KL}}(\tilde{q}_{\tilde{f}(\mathbf{X};\hat{\Theta})} \,\|\, \tilde{p}_{\tilde{f}(\mathbf{X};\hat{\Theta})}) \quad (11)$$

where $\boldsymbol{\epsilon}^{(j)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_P)$.

Empirical Evaluation

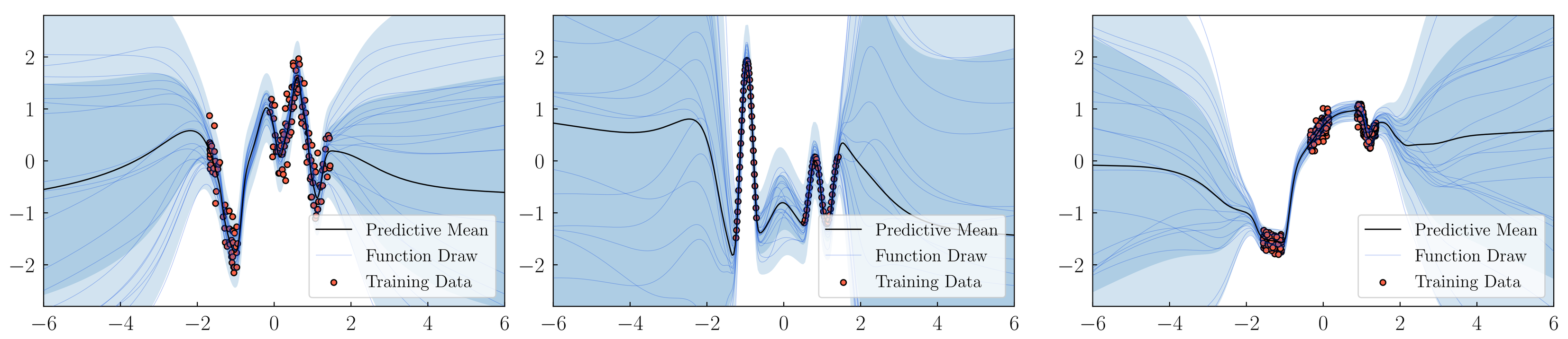

Figure 1: FSVI Posterior predictive distributions on 1D regression datasets.

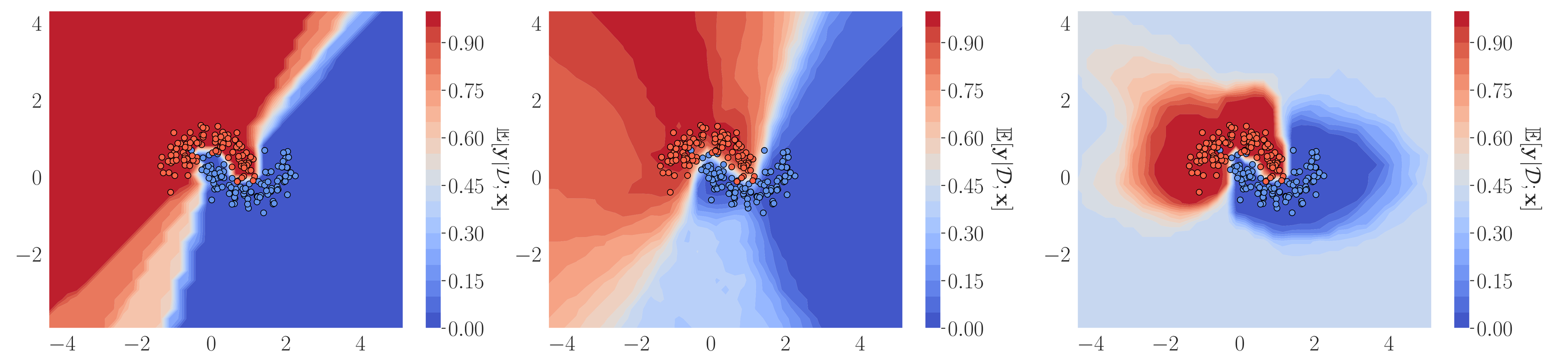

Figure 2: FSVI Posterior predictive distributions on the *Two Moons* dataset.

Table 1. Comparison of in- and out-of-distribution performance metrics (mean ± standard error over ten random seeds).

Method	Accuracy ↑	ECE ↓	AUROC M ↑	AUROC NM ↑
MAP	91.73±0.08	0.037±0.001	87.00±0.30	74.85±1.31
MFVI	91.03±0.04	0.038±0.001	93.10±0.34	88.88±0.74
MFVI (tempered)	91.38±0.05	0.058±0.001	86.30±0.29	80.78±0.68
MFVI (radial)	90.31±0.11	0.035±0.001	84.40±0.68	82.11±1.15
MC DROPOUT	90.55±0.04	**0.012**±0.001	88.46±0.57	80.02±1.04
SWAG	92.56±0.05	0.043±0.001	85.18±0.35	80.31±0.30
DUQ	92.40±0.20	–	95.50±0.70	94.60±1.80
BNN-LAPLACE	92.25±0.10	**0.012**±0.003	95.55±0.60	–
SPG	91.60±0.14	–	95.60±6.00	–
FSVI ($p_{\mathbf{X}_\mathcal{C}}$ = random monochrome)	**92.52**±0.13	**0.014**±0.002	96.55±0.41	95.15±0.71
FSVI ($p_{\mathbf{X}_\mathcal{C}}$ = KMNIST)	**92.67**±0.15	**0.012**±0.002	**99.81**±0.19	**97.44**±0.24
Deep Ensemble	92.49±0.01	**0.019**±0.000	89.22±0.09	83.17±0.91
FSVI Ensemble	**94.44**±0.07	0.020±0.001	**97.85**±0.15	**96.95**±0.20

Method	Accuracy↑	ECE↓	OOD-AUROC↑	C-CIFAR Acc↑
MAP	92.19±0.15	0.046±0.001	95.17±0.40	78.55±1.01
MFVI	89.98±0.09	0.040±0.001	92.14±0.34	79.36±1.35
MFVI (tempered)	90.87±0.11	0.048±0.001	91.82±0.90	79.86±1.32
MC DROPOUT	91.32±0.06	0.041±0.001	90.32±0.57	80.19±1.44
SWAG	93.13±0.14	0.067±0.002	89.79±0.50	76.12±0.51
VOGN	84.27±0.20	0.040±0.002	87.60±0.20	–
DUQ	**94.10**±0.20	–	92.70±1.30	–
SPG	77.69±0.64	–	88.30±4.00	–
FSVI ($p_{\mathbf{X}_\mathcal{C}}$ = random monochrome)	92.21±0.04	0.035±0.001	94.57±0.24	**80.76**±0.48
FSVI ($p_{\mathbf{X}_\mathcal{C}}$ = CIFAR-100)	92.27±0.04	**0.028**±0.001	**98.02**±0.10	**81.03**±0.49
Deep Ensemble	95.13±0.06	0.019±0.001	**98.04**±0.07	**81.22**±0.37
FSVI Ensemble	**95.19**±0.03	**0.013**±0.001	**99.19**±0.41	**81.35**±0.48

Full paper: https://timrudner.com/fsvi