The practical effectiveness of BNNs is limited by our ability to specify meaningful prior distributions and by the intractability of posterior inference. Choosing a meaningful prior distribution over network weights is difficult because the weights have a complicated relationship to the function computed by the network. Stochastic variational inference is appealing because the update rules resemble ordinary backprop [1, 2], but fitting accurate posterior distribution is difficult due to strong and complicated posterior dependencies [3, 4, 5].
To tackle this problem, they introduce functional Bayesian neural networks (fBNNs), which maximize an Evidence Lower BOund (ELBO) defined directly on the stochastic processes, i.e. distributions over functions. A BNN is trained to produce a distribution of functions with small KL divergence to the true posterior over functions. They show that fBNNs can specify priors entailing rich structures, including Gaussian processes and implicit stochastic processes. Empirically, they find fBNNs extrapolate well using various structured priors, provide reliable uncertainty estimates, and scale to large datasets.

1 Functional Evidence Lower Bound (fELBO)

Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, a Bayesian neural network (BNN) is defined in terms of a prior $p(\mathbf{w})$ on the weights, as well as the likelihood $p(\mathcal{D}|\mathbf{w})$. Variational Bayesian methods [6, 1, 2] attempt to fit an approximate posterior $q(\mathbf{w})$ to maximize the ELBO:

$$\mathcal{L}_q = \mathbb{E}_q[\log\ p(\mathcal{D}|\mathbf{w})] - \mathrm{KL}[q(\mathbf{w})||p(\mathbf{w})]. \tag{1}$$

Functional variational inference maximizes the fELBO, akin to the weight space ELBO in (1), except that the distributions are over functions rather than weights.

$$\mathcal{L}_q = \mathbb{E}_q[\log\ p(\mathcal{D}|f)] - \mathrm{KL}[q||p]. \tag{2}$$

Here p is a stochastic process prior over functions $f : \mathcal{X} \to \mathcal{Y}$, such as a GP. $\mathrm{KL}[q||p]$ is the KL divergence between two stochastic processes, but it does not have a convenient form as $\int \log \frac{q(f)}{p(f)} q(f)\,df$ due to there is no infinite-dimensional Lebesgue measure [7, 8]. Therefore, they prove that it equals the supremum of marginal KL divergences over all finite sets of inputs. Working on this, they introduce a training objective which approximates the fELBO using finite measurement sets and the Spectral Stein gradient estimator.

Theorem 1. *For two stochastic processes* P *and* Q*,*

$$\mathrm{KL}[Q|P] = \sup_{n\in\mathbb{N}, \boldsymbol{X}\in\mathcal{X}^n} \mathrm{KL}[Q_{\boldsymbol{X}}||P_{\boldsymbol{X}}]$$

fELBO becomes

$$\mathcal{L}_q = \mathbb{E}_q[\log\ p(\mathcal{D}|f)] - \sup_{n\in\mathbb{N}, \boldsymbol{X}\in\mathcal{X}^n} \mathrm{KL}[q(\mathbf{f}^{\boldsymbol{X}})||p(\mathbf{f}^{\boldsymbol{X}})] \tag{3}$$

2 KL Divergence Gradients

For implicit distributions whose density are intractable but are defined through a tractable sample process, the Spectral Stein gradient estimator (SSGE) [9] is a recently proposed method for estimating the log density derivative function of an implicit distribution, only require samples from the distributions. Specifically, given a continuous differentiable density $q(\mathbf{x})$, and a positive definite kernel $k(\mathbf{x}, \mathbf{x}')$ in the Stein class [10] of q, they show

$$\nabla_{x_i} \log\ q(\mathbf{x}) = -\sum_{j=1}^{\infty} [\mathbb{E}_q \nabla_{x_i} \psi_j(\mathbf{x})] \psi_j(\mathbf{x}),$$

where $\{\psi_j\}_{j \geqslant 1}$ is a series of eigenfunctions of k given by Mercer's theorem: $k(\mathbf{x}, \mathbf{x}') = \sum_j \mu_j \psi_j(\mathbf{x}) \psi_j(\mathbf{x}')$. While the likelihood term of the fELBO is tractable, the KL divergence term remains intractable because we don't have an explicit formula for the variational posterior density $q_\phi(\mathbf{f}^{\mathbf{X}})$. To derive an approximation, first observe that

$$\nabla_\phi \mathrm{KL}[q_\phi(\mathbf{f}^{\mathbf{X}}) \| p(\mathbf{f}^{\mathbf{X}})] = \mathbb{E}_q[\nabla_\phi \log\ q_\phi(\mathbf{f}^{\mathbf{X}})] + \mathbb{E}_\xi[\nabla_\phi \mathbf{f}^{\mathbf{X}} (\nabla_{\mathbf{f}} \log\ q(\mathbf{f}^{\mathbf{X}}) - \nabla_{\mathbf{f}} \log\ p(\mathbf{f}^{\mathbf{X}}))].$$

The first term (expected score function) is zero, and so can be discarded. The Jacobian $\nabla_\phi \mathbf{f}^{\mathbf{X}}$ can be exactly multiplied by a vector using backpropagation. Therefore, it remains to estimate the log-density derivative $\nabla_{\mathbf{f}} \log\ q(\mathbf{f}^{\mathbf{X}})$ and $\nabla_{\mathbf{f}} \log\ p(\mathbf{f}^{\mathbf{X}})$.

3 The Algorithm

Algorithm 1 Functional Variational Bayesian Neural Networks (fBNNs)

Require: Dataset $\mathcal{D}$, variational posterior $g(\cdot)$, prior p (explicit or implicit), KL weight λ.
Require: Sampling distribution c for random measurement points.
while ϕ not converged **do**
  $\mathbf{X}^M \sim c$; $D_S \subset \mathcal{D}$ ▷ sample measurement points
  $\mathbf{f}_i = g([\mathbf{X}^M, \mathbf{X}^{D_S}], \xi_i; \phi),\ i = 1 \cdots k.$ ▷ sample k function values
  $\Delta_1 = \frac{1}{k} \frac{1}{|D_s|} \sum_i \sum_{(x,y)} \nabla_\phi \log p(y|\mathbf{f}_i(x))$ ▷ compute log likelihood gradients
  $\Delta_2 = \mathrm{SSGE}(p, \mathbf{f}_{1:k})$ ▷ estimate KL gradients
  $\phi \leftarrow \mathrm{Optimizer}(\phi, \Delta_1 - \lambda \Delta_2)$ ▷ update the parameters
end while

The whole algorithm is as presented. In each iteration, the measurement points include a mini-batch $\mathcal{D}$ from the training data and random points $\mathbf{X}^M$ from a distribution c. $\mathbf{X}^{\mathcal{D}}$ and $\mathbf{X}^M$ were forwarded together through the network $g(\cdot; \phi)$ which defines the variational posterior q_ϕ. Then the algorithm tries to maximize the following objective correpsonding to fELBO.

$$\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}, y) \in \mathcal{D}} \mathbb{E}_{q_\phi}[\log\ p(y|f(\mathbf{x}))] - \lambda \mathrm{KL}[q(\mathbf{f}^{\mathcal{D}}, \mathbf{f}^M) \| p(\mathbf{f}^{\mathcal{D}}, \mathbf{f}^M)],$$

where λ is a regularization hyperparameter.

4 Toy Example

In this experiment, consider 20 inputs randomly sampled from the interval $[-4, -1] \cup [1, 4]$, and targets y which are noisy observations of a periodic function: $y = x^3 + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 9)$. Here they use GP prior,

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}' + exp(-\frac{(\mathbf{x} - \mathbf{x}')^\top (\mathbf{x} - \mathbf{x}')}{2}),$$

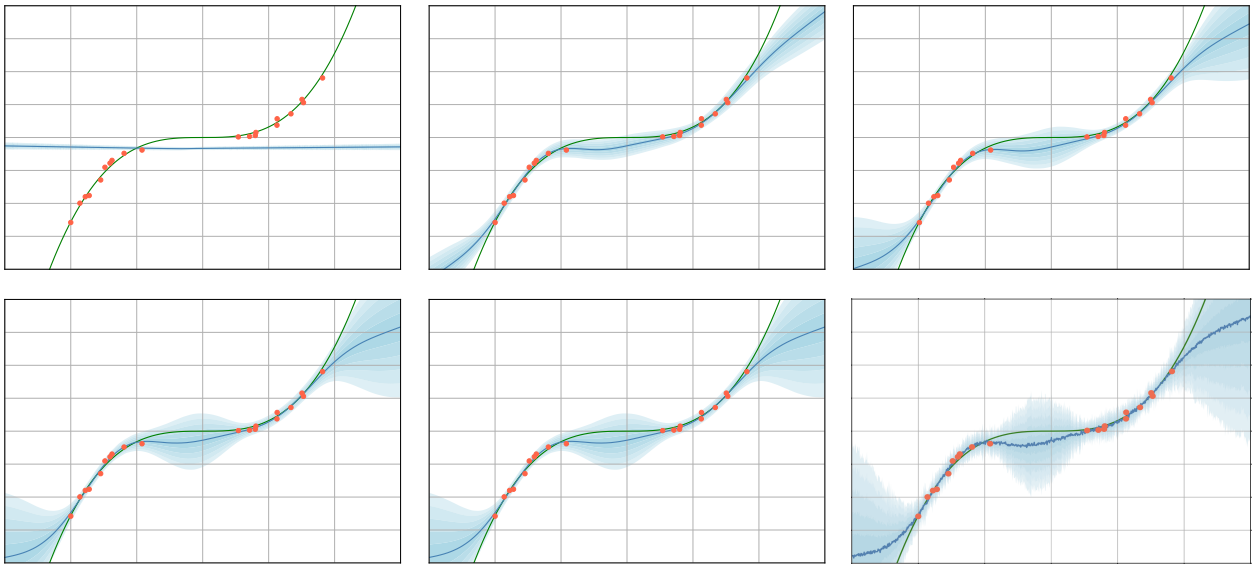

Figure 1: Prediction on the toy function $y = x^3$ using 2 hidden layers of 100 units. Here six subplots represent 0, 2000, 4000, 6000 and 8000 epochs, as well as GP sampling results. Red dots are 20 training points. The green and blue lines represent ground truth and mean prediction, respectively. Shaded areas correspond to standard deviations.

and the variational posterior is represented as a stochastic neural network with independent Gaussian distributions over the weights, i.e. $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \mathtt{diag}(\boldsymbol{\sigma}^2))$.

References

[1] Alex Graves. Practical Variational Inference for Neural Networks. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2348–2356. Curran Associates, Inc., 2011.

[2] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. *arXiv:1505.05424 [cs, stat]*, May 2015. arXiv: 1505.05424.

[3] Christos Louizos and Max Welling. Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors. *arXiv:1603.04733 [cs, stat]*, March 2016. arXiv: 1603.04733.

[4] Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. Noisy Natural Gradient as Variational Inference. *arXiv:1712.02390 [cs, stat]*, December 2017. arXiv: 1712.02390.

[5] Jiaxin Shi, Shengyang Sun, and Jun Zhu. Kernel Implicit Variational Inference. *arXiv:1705.10119 [cs, stat]*, May 2017. arXiv: 1705.10119.

[6] 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, pages 5–13, New York, NY, USA, 1993. ACM. event-place: Santa Cruz, California, USA.

[7] Alexander G. de G. Matthews, James Hensman, Richard E. Turner, and Zoubin Ghahramani. On Sparse variational methods and the Kullback-Leibler divergence between stochastic processes. *arXiv:1504.07027 [stat]*, April 2015. arXiv: 1504.07027.

[8] Nathaniel Eldredge. Analysis and Probability on Infinite-Dimensional Spaces. *arXiv:1607.03591 [math]*, July 2016. arXiv: 1607.03591.

[9] Jiaxin Shi, Shengyang Sun, and Jun Zhu. A Spectral Approach to Gradient Estimation for Implicit Distributions. *arXiv:1806.02925 [cs, stat]*, June 2018. arXiv: 1806.02925.

[10] Qiang Liu, Jason D. Lee, and Michael I. Jordan. A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation. *arXiv:1602.03253 [stat]*, February 2016. arXiv: 1602.03253.