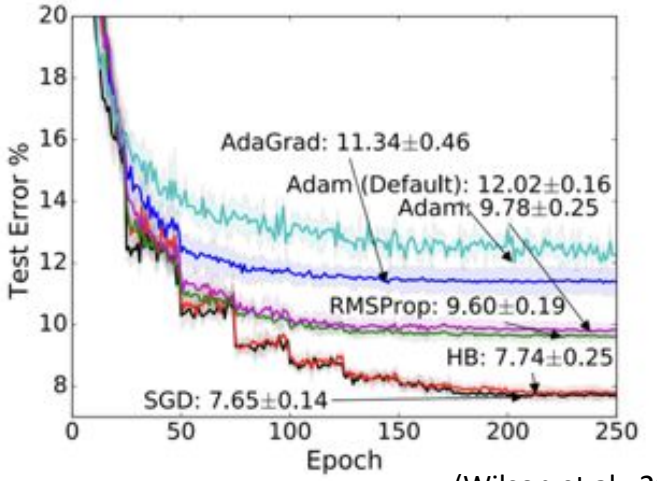
Functional Variational Bayesian Neural Networks

Shengyang Sun*[1], Guodong Zhang*[1], Jiaxin Shi*[2], Roger Grosse[1]

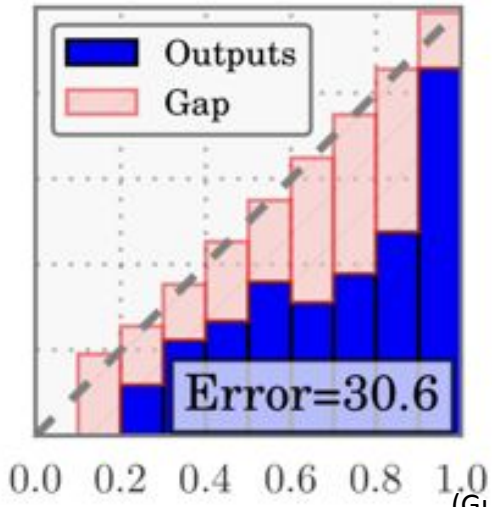
University of Toronto[1] Tsinghua University[2]

Motivation: Why uncertainty in neural nets?



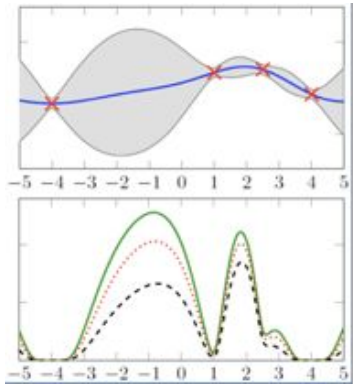
(Wilson et al., 2017)

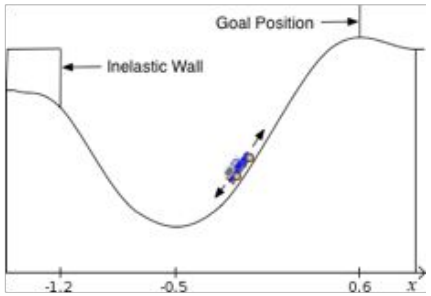
Generalization



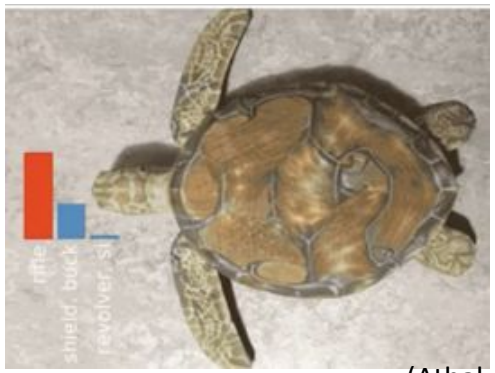
(Guo et al., 2017)

Calibration





Exploration

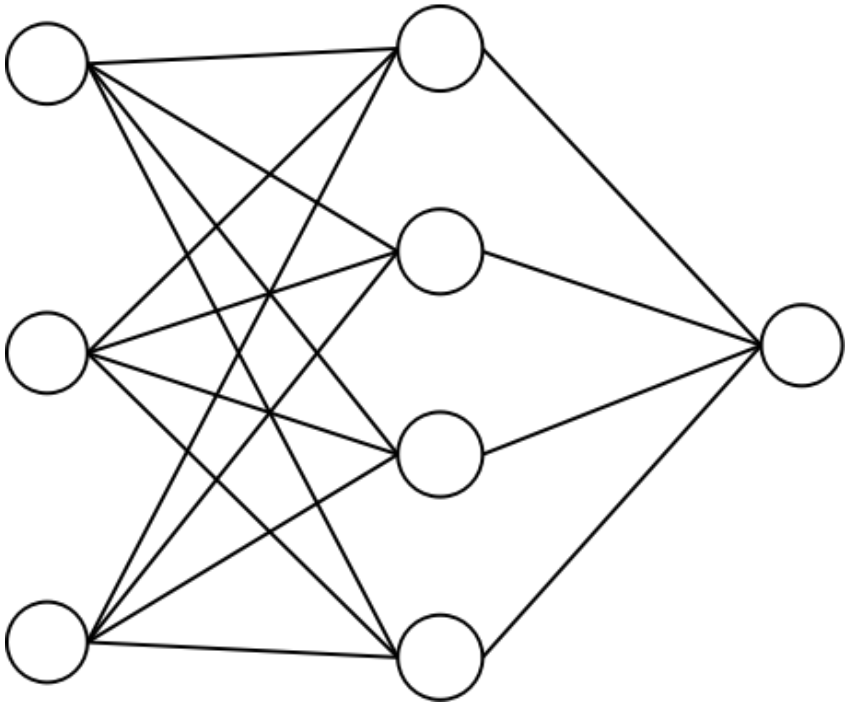


(Athalye et al., 2017)

Adversarial Robustness

Background: Bayesian neural networks

Combining Bayesian methods with deep learning models



$$W \sim \mathcal{N}(0, \eta^2)$$ Prior

$$f(x) = \mathrm{NN}(x; W)$$ Neural nets

$$y = \mathcal{P}(f(x); \theta)$$ Likelihood

Commonly-used Likelihoods

Regression: $y = f(x) + \epsilon^2, \epsilon \sim \mathcal{N}(0, \sigma^2)$

Classification: $y = \mathrm{softmax}(f(x))$

Background: Bayesian neural networks

- Variational Inference fits a $q(w)$ to maximize the Evidence Lower Bound (ELBO)

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(y|x,w)] - \mathrm{KL}[q(w)||p(w)]$$

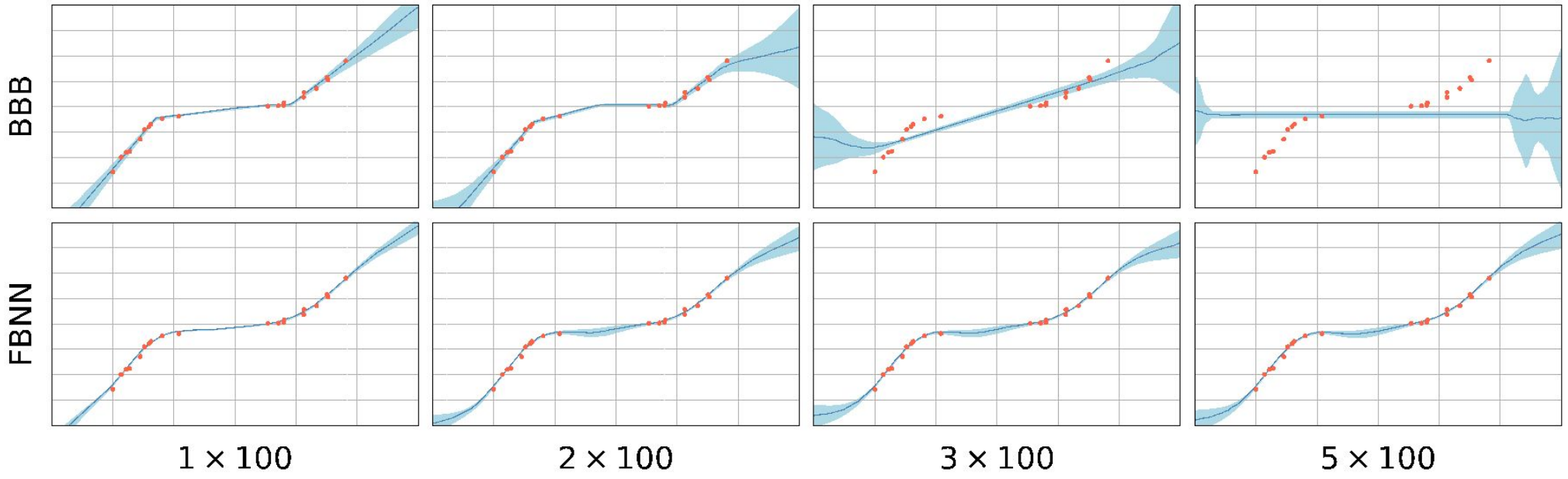
  - Fast, stochastic training
  - Noisy Backpropagation
  - Compact representation of posterior uncertainty

Background: Bayesian neural networks

- Recent Attempts Towards Expressive Posteriors
  - Factorized Gaussian (Blundell et al., 2015)
  - Matrix variate Gaussian (Louizos & Welling, 2016; Sun et al., 2017; Zhang et al., 2017)
  - Normalizing Flow (Louizos & Welling, 2017)
  - Implicit distribution (Shi et al., 2017)
  - SGLD (Wang et al., 2018)

- Is variational posterior the key problem for BNNs?

Mis-Specified Weight-Space Prior



Predictive results for BBB, HMC and FBNN, when varying network sizes.

This paper

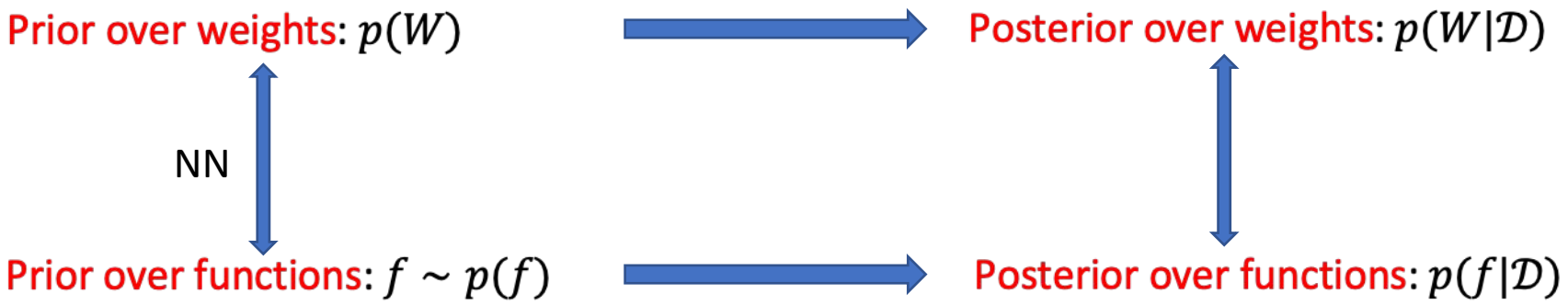
distribution over functions

- A theorem on KL divergence between two stochastic processes
- Two algorithms for functional variational inference
  - Adversarial
  - Sampling-based
- Functional variational BNNs
  - Reliable uncertainty
  - Sensible Extrapolation
  - Scalability
  - Model-Agnostic

Functional Variational Inference

- Reinterpreting BNNs as stochastic processes

- Consider the the space of functions $f: \mathcal{X} \rightarrow \mathcal{Y}$



Functional Variational Inference

- Given dataset $\mathcal{D} = (\mathrm{X}^{\mathcal{D}}, y^{\mathcal{D}})$, consider variational inference in f, we get functional ELBO:

$$\mathcal{L}(q) \coloneqq \mathbb{E}_q\left[\log p\left(y^{\mathcal{D}}\middle|f\right)\right] - \mathrm{KL}[q(f)||p(f)]$$

$p(f)$: prior process

- Gaussian Process
- Student-t's Process
- Piecewise Constant functions
- …

$\mathrm{q}(f)$: variational stochastic process

- Gaussian Process
- NN with weight stochasticity (BNN)
- NN with node stochasticity (GAN GEN)
- …

Functional Variational Inference

- **Theorem [Functional KL Divergence].** *For two stochastic processes P, Q, the KL divergence is the supremum of marginal KL divergences over all finite subset of inputs $x_{1:n}$*

$$\mathrm{KL}[P\|Q] = \sup_{n,\, \mathbf{x}_{1:n}} \mathrm{KL}[P_{\mathbf{x}_{1:n}} \| Q_{\mathbf{x}_{1:n}}].$$

- **Example [KL between conditional processes].**

$$\begin{aligned}
\mathrm{KL}[p(f|\mathcal{D}_1)\|p(f|\mathcal{D}_2)] &= \sup_{n,\mathbf{x}_{1:n}} \mathrm{KL}[p(\mathbf{f}^{\mathbf{x}}|\mathcal{D}_1)\|p(\mathbf{f}^{\mathbf{x}}|\mathcal{D}_2)] \\
&= \sup_{n,\mathbf{x}_{1:n}} \mathbb{E}_{p(\mathbf{f}^{\mathbf{x}},\mathbf{f}^{D_1\cup D_2}|\mathcal{D}_1)} \log \frac{p(\mathbf{f}^{\mathbf{x}},\mathbf{f}^{D_1\cup D_2}|\mathcal{D}_1)}{p(\mathbf{f}^{\mathbf{x}},\mathbf{f}^{D_1\cup D_2}|\mathcal{D}_2)} \\
&= \sup_{n,\mathbf{x}_{1:n}} \mathbb{E}_{p(\mathbf{f}^{\mathbf{x}},\mathbf{f}^{D_1\cup D_2}|\mathcal{D}_1)} \log \frac{p(\mathbf{f}^{D_1\cup D_2}|\mathcal{D}_1)p(\mathbf{f}^{\mathbf{x}}|\mathbf{f}^{D_1\cup D_2},\mathcal{D}_1)}{p(\mathbf{f}^{D_1\cup D_2}|\mathcal{D}_2)p(\mathbf{f}^{\mathbf{x}}|\mathbf{f}^{D_1\cup D_2},\mathcal{D}_2)} \\
&= \sup_{n,\mathbf{x}_{1:n}} \mathbb{E}_{p(\mathbf{f}^{\mathbf{x}},\mathbf{f}^{D_1\cup D_2}|\mathcal{D}_1)} \log \frac{p(\mathbf{f}^{D_1\cup D_2}|\mathcal{D}_1)p(\mathbf{f}^{\mathbf{x}}|\mathbf{f}^{D_1\cup D_2})}{p(\mathbf{f}^{D_1\cup D_2}|\mathcal{D}_2)p(\mathbf{f}^{\mathbf{x}}|\mathbf{f}^{D_1\cup D_2})} \\
&= \mathrm{KL}[p(\mathbf{f}^{D_1\cup D_2}|\mathcal{D}_1)\|p(\mathbf{f}^{D_1\cup D_2}|\mathcal{D}_2)]
\end{aligned}$$

Functional Variational Inference

- functional ELBO

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q[\log p(\mathbf{y}^D|f)] - \sup_{\mathbf{X}} \mathrm{KL}[q(\mathbf{f}^{\mathbf{X}})||p(\mathbf{f}^{\mathbf{X}})] \\ &= \inf_{\mathbf{X}} \sum_{(\mathbf{x}^D, y^D)\in\mathcal{D}} \mathbb{E}_q[\log p(y^D|f(\mathbf{x}^D))] - \mathrm{KL}[q(\mathbf{f}^{\mathbf{X}})||p(\mathbf{f}^{\mathbf{X}})] \\ &:= \inf_{\mathbf{X}} \mathcal{L}_{\mathbf{X}}(q).\end{aligned}$$

- where $\mathrm{X} = x_{1:n}$ ranges over all finite sets of input locations, termed as *measurement points*.
- **Theorem [Lower Bound].** *If X contains all training input locations $\mathrm{X}^{\mathcal{D}}$,*

$$\mathcal{L}_{\mathbf{X}}(q) = \log p(\mathcal{D}) - \mathrm{KL}[q(\mathbf{f}^{\mathbf{X}})\|p(\mathbf{f}^{\mathbf{X}}|\mathcal{D})] \leq \log p(\mathcal{D}).$$

Functional Variational Inference

- Adversarial Functional Variational Inference
  - Fixed-size measurement points

$$\max_{q \in \mathcal{Q}} \quad \min_{|\mathbf{X}|=N_m} \quad \mathcal{L}_{\mathbf{X}}(q).$$

  - Analogy to GANs. Generator Discriminator

Functional Variational Inference

- Sampling-Based Functional Variational Inference
  - Auxiliary distribution c. $\mathrm{X}^{M} \sim c$.
  - Subset of training set $\mathrm{X}^{\mathcal{D}_S}$ to prevent overfitting.
  - measurement points $\mathrm{X} = \mathrm{X}^{M} \cup \mathrm{X}^{\mathcal{D}_S}$

$$\max_{q \in \mathcal{Q}} \mathbb{E}_{\mathbf{X}^{M} \sim C} \mathcal{L}_{\mathbf{X}^{M}, \mathbf{X}^{D_S}}(q).$$

- **Theorem [Consistency]:** *When both p and q are both Gaussian processes, $\mathcal{Q}$ is expressive enough, $\mathcal{D}_S = \mathcal{D}, M > 1$, then we have equivalence:*
  - *Sampling-based functional variational inference with* supp(c) = X *attains its optimum.*
  - *The variational posterior process and the true posterior process are identical.*

Background: Spectral Stein Gradient Estimators (Shi et al., 2018)

- Implicit distributions
- Stein's Equality

$$\mathbb{E}_q[\mathbf{h}(\mathbf{x})\nabla_{\mathbf{x}} \log q(\mathbf{x})^\top + \nabla_{\mathbf{x}}\mathbf{h}(\mathbf{x})] = \mathbf{0}.$$

- Log density derivative

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

- The eigenfunctions $\psi_j(x)$ and their derivatives are approximated by Nystrom method.

Functional Variational Inference

- Estimate KL Divergence Derivatives

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

- The first term is zero.
- SSGE estimates the log-density derivatives $\nabla_{\mathbf{f}} \log q(\mathbf{f}^{\mathbf{X}})$, $\nabla_{\mathbf{f}} \log p(\mathbf{f}^{\mathbf{X}})$
- $\nabla_{\mathbf{f}} \log p(\mathbf{f}^{\mathbf{X}})$ is tractable for explicit priors:
  - Gaussian processes
  - Student-t processes

The algorithm

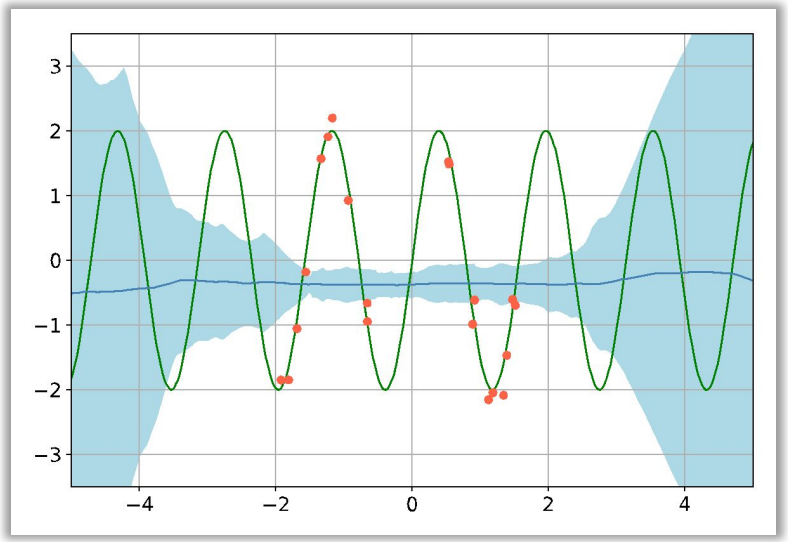
- Objective $$\sum_{(\mathbf{x},y)\in\mathcal{D}_s} \mathrm{E}_{q_\phi}\left[\log p(y|f(x))\right] - \mathrm{KL}[q(\mathbf{f}^M, \mathbf{f}^{\mathcal{D}_s}) \| p(\mathbf{f}^{\mathcal{D}_s}, \mathbf{f}^M)]$$

Algorithm 1 Functional Variational Bayesian Neural Networks (fBNNs)

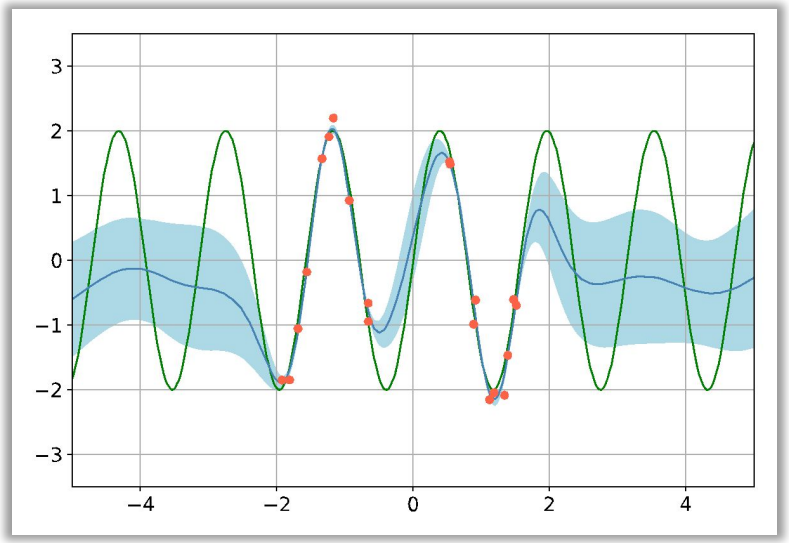
Require: Dataset $\mathcal{D}$, sampling distribution c, variational posterior $g(\cdot)$, prior p (explicit or implicit).

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
  $\mathbf{g}_1 = \frac{1}{k}\sum_i \sum_{(x,y)} \nabla_\phi \log p(y|\mathbf{f}_i(x))$ ▷ compute log likelihood gradients
  $\mathbf{g}_2 = \mathrm{SSGE}(p, \mathbf{f}_{1:k})$ ▷ estimate KL gradients
  $\phi \leftarrow \mathrm{Optimizer}(\phi, \mathbf{g}_1 - \mathbf{g}_2)$ ▷ update the parameters
end while

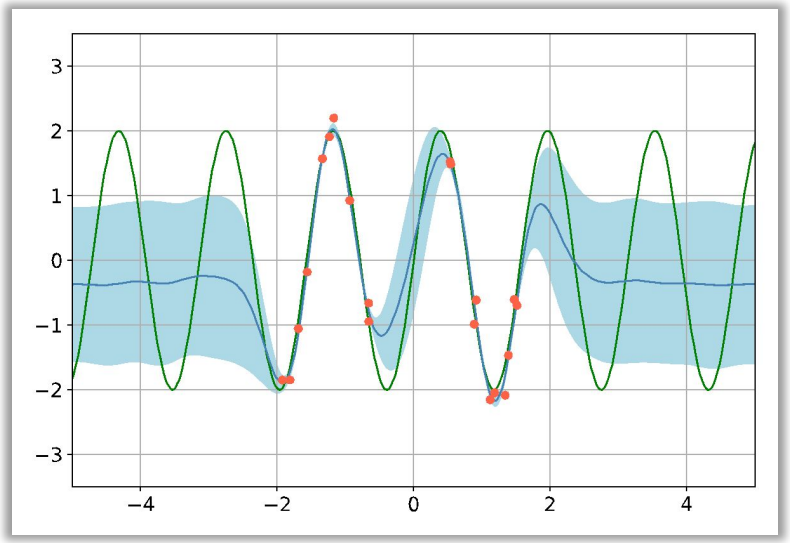
Extrapolating Periodic Structures



BBB

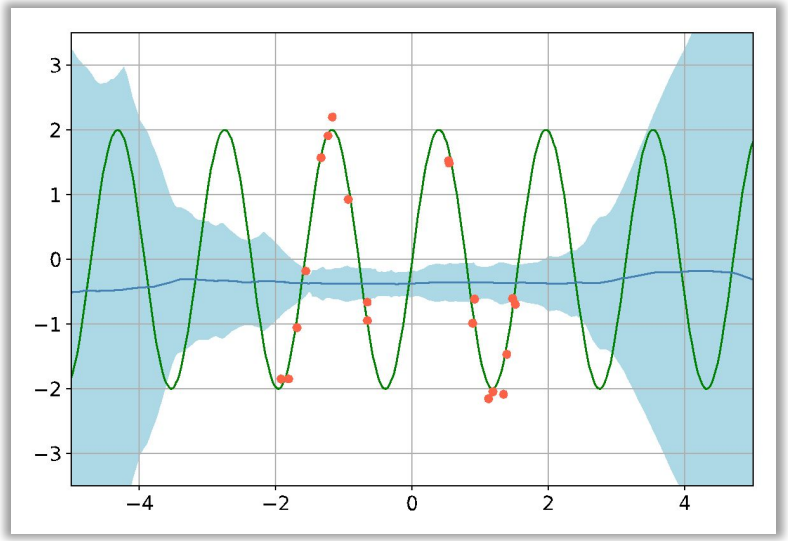


FBNN-RBF

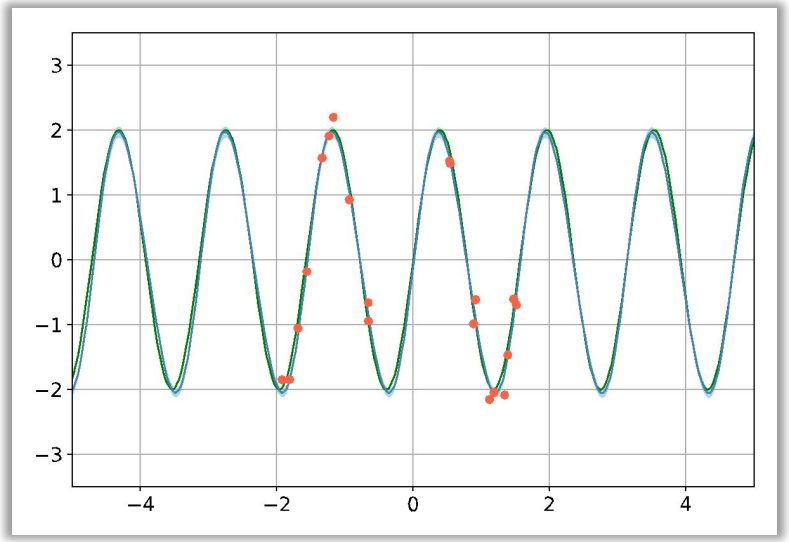


GP-RBF

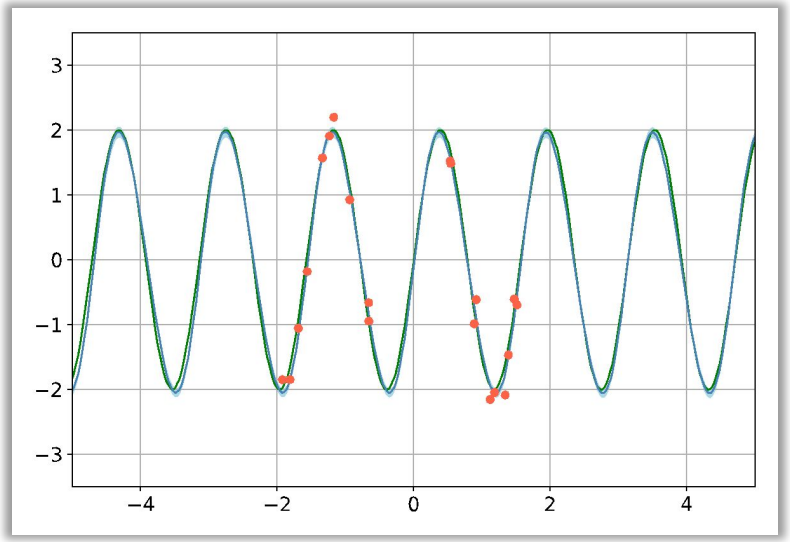
Extrapolating Periodic Structures



BBB



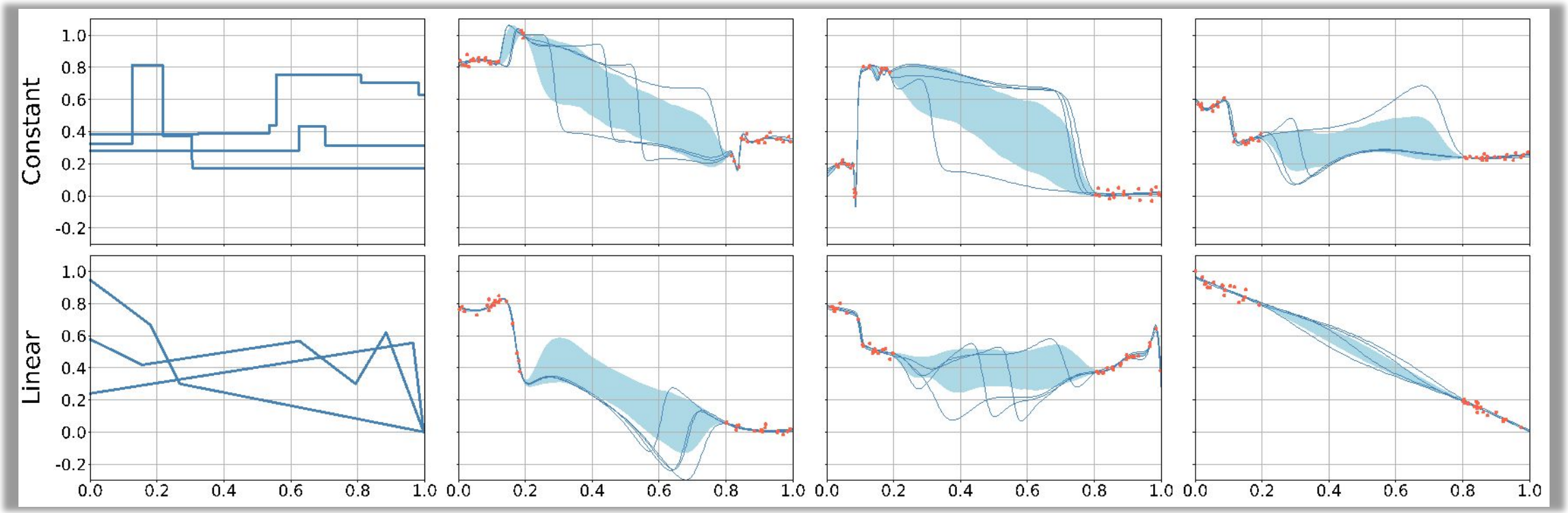
FBNN-PER



GP-PER

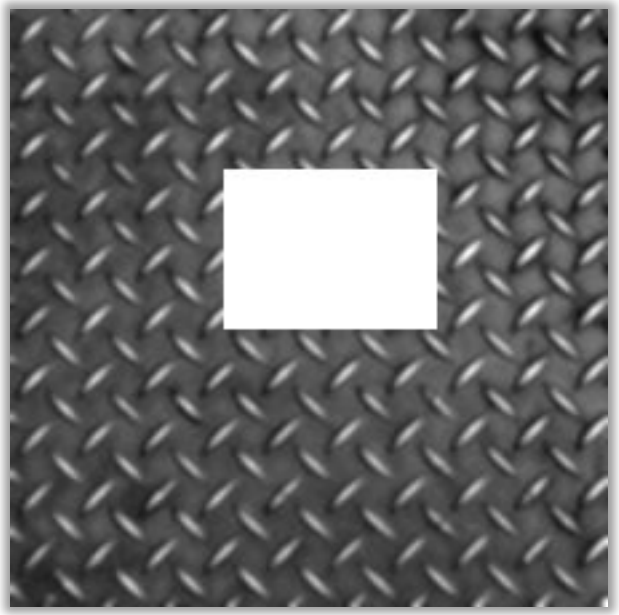
Extrapolating Piecewise Functions

- Implicit Priors

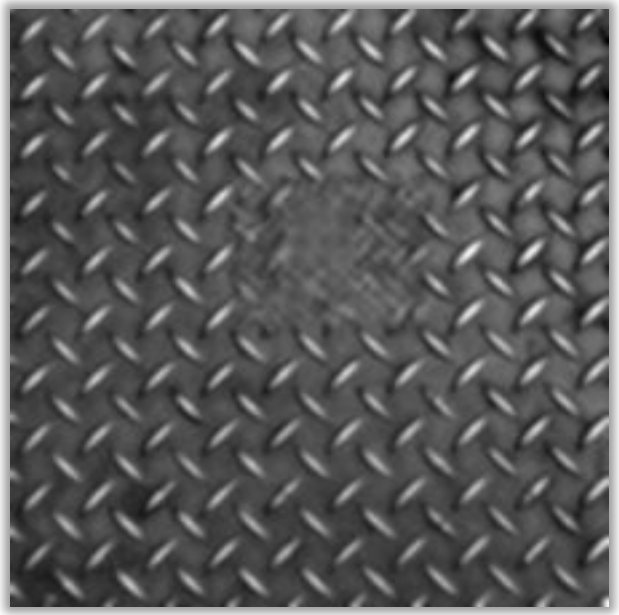


Extrapolating Textures

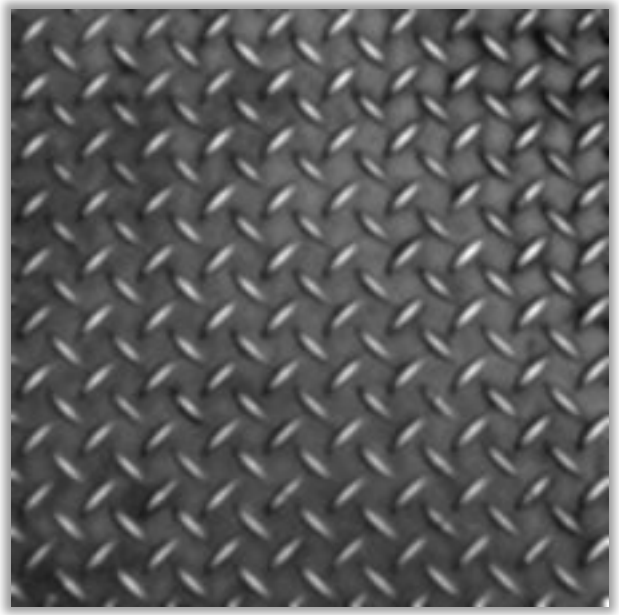
- Large Scale Dataset
- CNN as Generator



Training



FBNN w.o. missing



FBNN w. missing

Predictive Performance

	Validation RMSE			Validation log-likelihood		
Dataset	BBB	NNG	Ours	BBB	NNG	Ours
Boston	3.171±0.149	2.742±0.125	**2.378±0.104**	-2.602±0.031	-2.446±0.029	**-2.301±0.038**
Concrete	5.678±0.087	5.019±0.127	**4.935±0.180**	-3.149±0.018	**-3.039±0.025**	-3.096±0.016
Energy	0.565±0.018	0.485±0.023	**0.412±0.017**	-1.500±0.006	-1.421±0.005	**-0.684±0.020**
Wine	0.643±0.012	**0.637±0.011**	0.673±0.014	-0.977±0.017	**-0.969±0.014**	-1.040±0.013
Yacht	1.174±0.086	0.979±0.077	**0.607±0.068**	-2.408±0.007	-2.316±0.006	**-1.033±0.033**

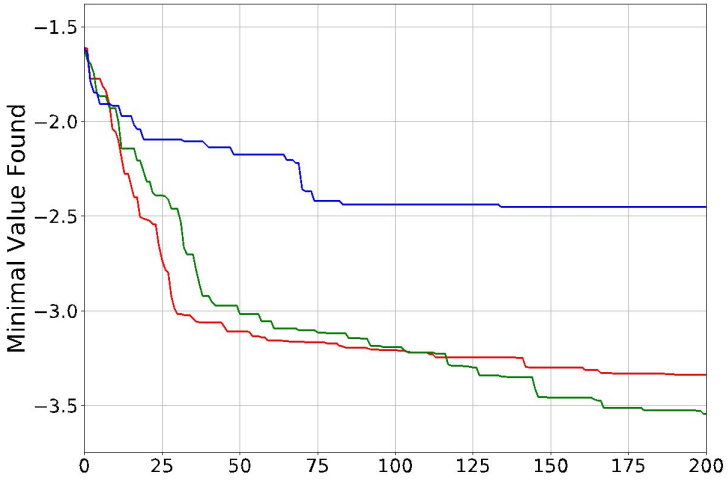
Small Scale regression benchmarks.

		Test RMSE		Test log-likelihood	
Dataset	N	BBB	FBNN	BBB	FBNN
Naval	11934	1.6E-4±0.000	**1.2E-4±0.000**	6.950±0.052	**7.130±0.024**
Protein	45730	4.331±0.033	**4.326±0.019**	-2.892±0.007	**-2.892±0.004**
Video Memory	68784	1.879±0.265	**1.858±0.036**	**-1.999±0.054**	-2.038±0.021
Video Time	68784	3.632±1.974	**3.007±0.127**	**-2.390±0.040**	-2.471±0.018
GPU	241600	21.886±0.673	**19.50±0.171**	-4.505±0.031	**-4.400±0.009**

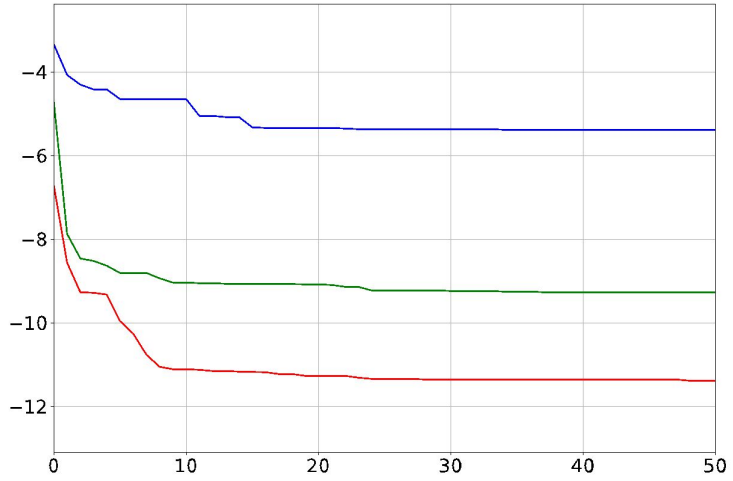
Large Scale regression benchmarks.

Bayesian Optimization

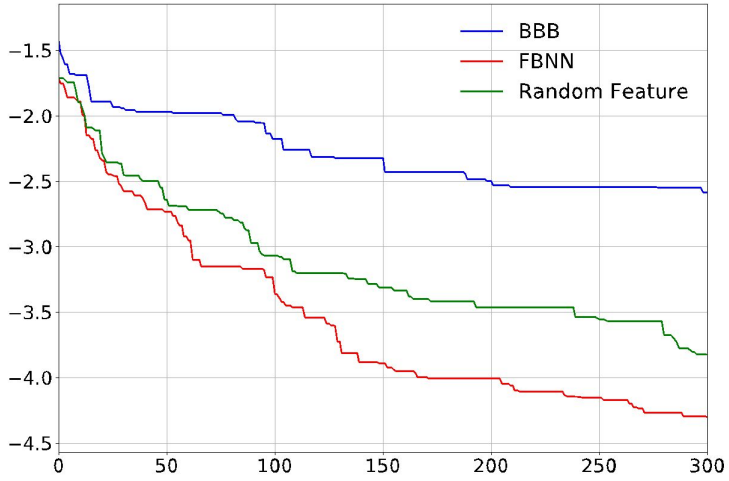
- Max-value Entropy Search (Wang & Jegelka, 2017)
- Functions sampled from GPs, varying kernels.
- Compare BBB, RBF Random Feature, FBNN with true kernel
- Requires parametric function samples



RBF



Arccosine



Matern 12

Contextual Bandits

- Benchmark on evaluating uncertainty estimation (Riquelme et al., 2018).
- Averaged over 8 datasets, 10 runs each.
- Thompson Sampling
- State-of-art

	FBNN 1×50	FBNN 2×50	FBNN 1×500	FBNN 2×500	MULTITASKGP	BBB 1×50	BBB 1×500
M. RANK	5.875	7.125	**4.875**	**5.0**	5.875	11.5	13.375
M. VALUE	46.0	47.0	**45.3**	**44.2**	46.5	56.6	68.1
	BBALPHADIV	PARAMNOISE	NEURALLINEAR	LINFULLPOST	DROPOUT	RMS	UNIFORM
M. RANK	16.0	10.125	10.375	9.25	7.625	8.875	16.75
M. VALUE	87.4	53.0	52.3	NAN	48.3	53.0	100

Thanks

Function-Space Inference versus Weight-Space Inference

- What if having a perfect weight prior in accord to a functional prior?
  - It's still better to perform functional inference!
- Deep linear networks: Equivalent $p(w)$ and $p(f)$, Equivalent $q(w)$ and $q(f)$

$$\mathcal{L}_w = \mathbb{E}_{q_w} \log p(\mathbf{y}^{\mathcal{D}} | \mathbf{X}^{\mathcal{D}}, \prod_{l=0}^{L-1} \mathbf{W}_l) - \mathrm{KL}[q([\mathbf{W}_0, \cdots, \mathbf{W}_{L-1}]) \| p([\mathbf{W}_0, \cdots, \mathbf{W}_{L-1}])]$$

$$\mathcal{L}_f = \mathbb{E}_{q_f} \log p(\mathbf{y}^{\mathcal{D}} | \mathbf{X}^{\mathcal{D}}, f) - \mathrm{KL}[q_f(f) \| p_f(f)]$$

$$= \mathbb{E}_{q_w} \log p(\mathbf{y}^{\mathcal{D}} | \mathbf{X}^{\mathcal{D}}, \prod_{l=0}^{L-1} \mathbf{W}_l) - \mathrm{KL}[q_f(f) \| p_f(f)]$$

Theorem [ELBO Ordering]:

$$\mathcal{L}_f \geq \mathcal{L}_w$$