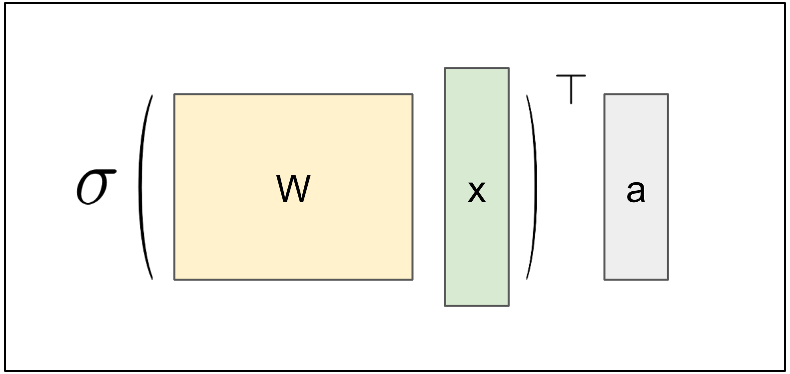
Neural Networks as Inter-Domain Inducing Points

Shengyang Sun*, Jiaxin Shi*, Roger Grosse



Existing Probabilistic Perspectives on Neural Networks

Infinite-width neural networks at initialization are Gaussian processes (Neal 92, Lee et al. 18)

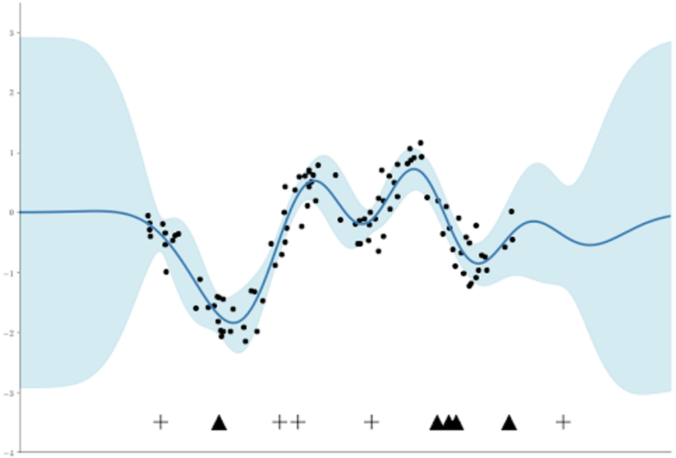
$$f(\mathbf{x}) = \sum_{m=1}^{M} a_m \sigma(\mathbf{w}_m^\top \mathbf{x}) \quad w_{mj} \sim N(0, \sigma_w^2)$$

Infinite-width neural networks at training are Gaussian processes (NTK, Jacot et al. 18)

$$\Theta^{(L)}(x, y) := (\nabla f_\theta(x))^T \nabla f_\theta(y)$$

- Relies heavily on the infinite-width assumption.
- Ignores the importance of individual weights.
- Performance fails to match NNs with standard training.

Sparse Gaussian Processes



Sparse GP predictive distribution

Inducing Points (Z): A small number of inputs summarizing the training data

(Shi, Titsias, Mnih, 20)

Sparse GPs and Two-Layer NNs

Predictive mean of Sparse GP $$\mu(\mathbf{x}) = k(\mathbf{Z}, \mathbf{x})^\top \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{m}$$

Two-layer Neural Networks $$f(\mathbf{x}) = \sigma(\mathbf{W}\mathbf{x})^\top \mathbf{a}$$

Sparse GPs and Two-Layer NNs

Predictive mean of Sparse GP $$\mu(\mathbf{x}) = k(\mathbf{Z}, \mathbf{x})^\top \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{m}$$

Two-layer Neural Networks $$f(\mathbf{x}) = \sigma(\mathbf{W}\mathbf{x})^\top \mathbf{a}$$

Nonlinear Linear

Sparse GPs and Two-Layer NNs

Predictive mean of Sparse GP $$\mu(\mathbf{x}) = k(\mathbf{Z}, \mathbf{x})^\top \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{m}$$

Two-layer Neural Networks $$f(\mathbf{x}) = \sigma(\mathbf{W}\mathbf{x})^\top \mathbf{a}$$

Problem: Activations are not necessarily positive-type functions

Sparse GPs and Two-Layer NNs

Predictive mean of Sparse GP $\mu(\mathbf{x}) = k(\mathbf{Z}, \mathbf{x})^\top \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{m}$

Two-layer Neural Networks $f(\mathbf{x}) = \sigma(\mathbf{W}\mathbf{x})^\top \mathbf{a}$

- Inter-domain inducing point $z : \mathcal{X} \to \mathbb{R}$
- Variational Fourier Features[1] (VFF) generalizes the kernel function

$$k(z, \mathbf{x}) = z(\mathbf{x}), k(z, z') = \langle z, z' \rangle_{\mathcal{H}}$$

[1] (Hensman et al, 2017)

Sparse GPs and Two-Layer NNs

Predictive mean of Sparse GP

$$\mu(\mathbf{x}) = k(\mathbf{Z}, \mathbf{x})^\top \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{m}$$

$$z_i(\mathbf{x}) = \sigma(\mathbf{w}_i^\top \mathbf{x}) = k(\mathbf{z}_i, \mathbf{x})$$

Two-layer Neural Networks

$$f(\mathbf{x}) = \sigma(\mathbf{W}\mathbf{x})^\top \mathbf{a}$$

- Inter-domain inducing point $z : \mathcal{X} \to \mathbb{R}$
- Variational Fourier Features[1] (VFF) generalizes the kernel function

$$k(z, \mathbf{x}) = z(\mathbf{x}), k(z, z') = \langle z, z' \rangle_{\mathcal{H}}$$

[1] (Hensman et al, 2017)

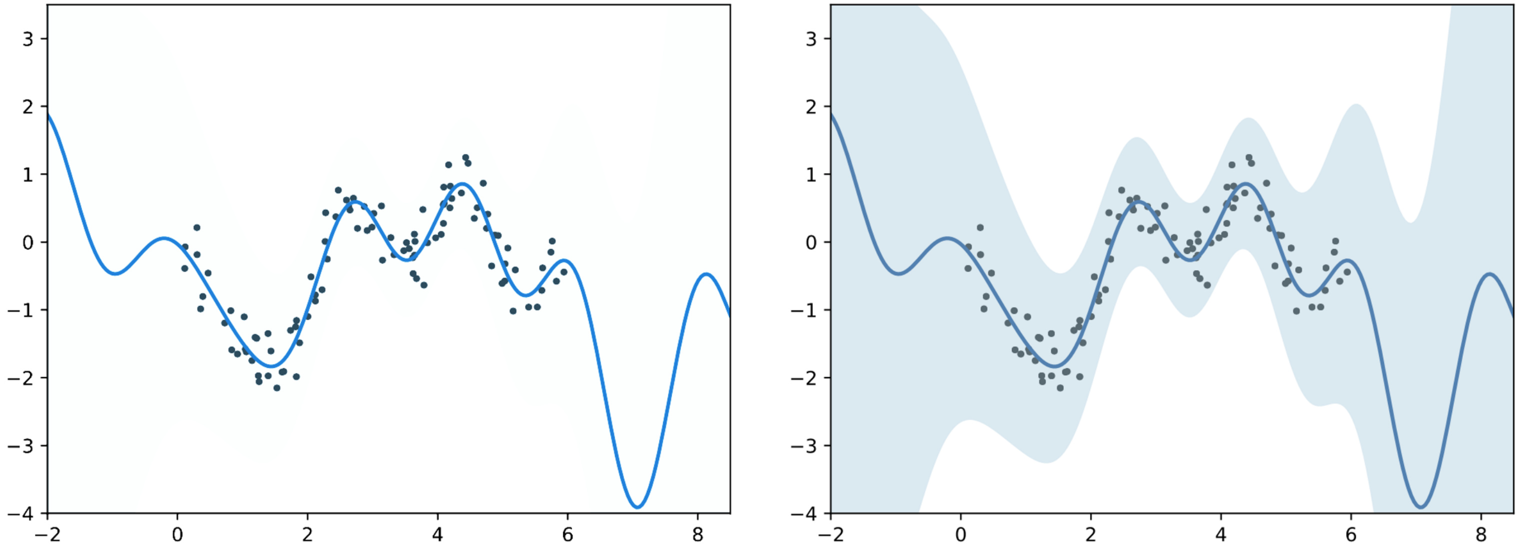
Numerical Experiments

Uncertainty from post-trained NNs

1. Train a two-layer neural network by standard backprop.
2. After training, each hidden unit is an inter-domain inducing point.
3. Compute (approximate) predictive variance of the corresponding sparse GP:

$$
\begin{aligned}
\sigma^2(\mathbf{x}) &= k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_{\mathbf{zx}}^\top \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{k}_{\mathbf{zx}} + \mathbf{k}_{\mathbf{zx}}^\top \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{S} \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{k}_{\mathbf{zx}} \\
&\approx k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_{\mathbf{zx}}^\top \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{k}_{\mathbf{zx}}
\end{aligned}
$$

Numerical Experiments



Uncertainty from post-trained NNs

Future Work

- Multi-layer neural networks
- Convolutional, Recurrent Structures
- How does this framework help us understand neural networks ?

Thanks