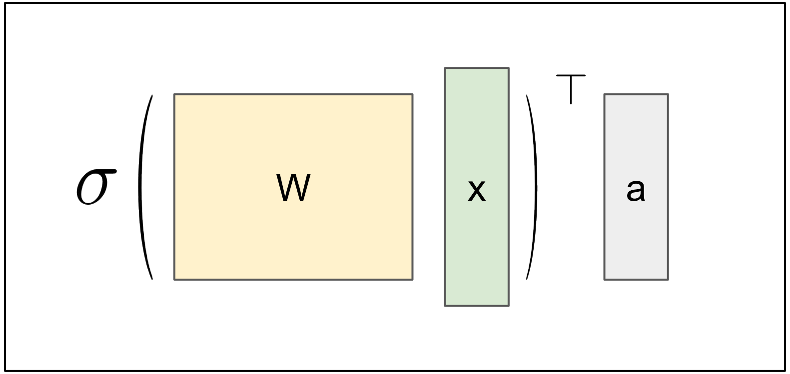
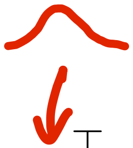
A Probabilistic Perspective on Neural Networks

Neural Networks as Inter-Domain Inducing Points

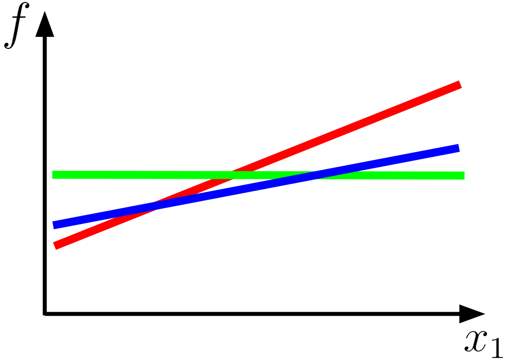
Jiaxin Shi
Microsoft Research New England
jiaxinshi@microsoft.com

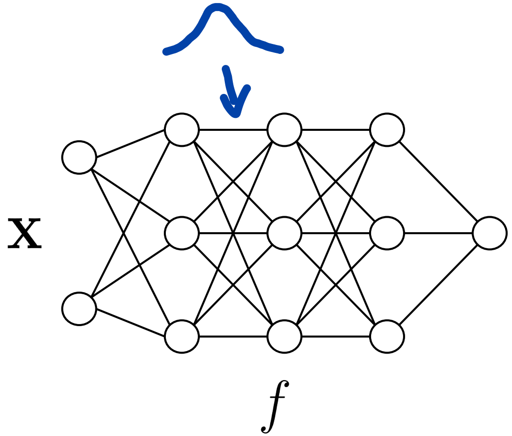


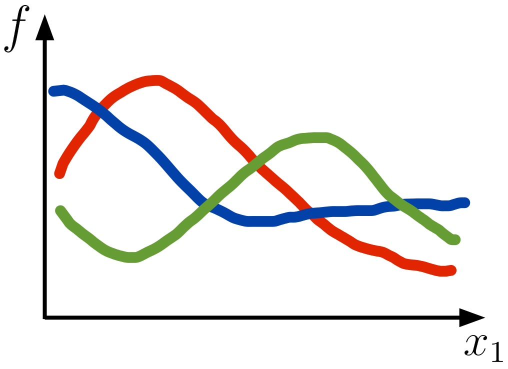
Existing Probabilistic Perspectives on Neural Networks



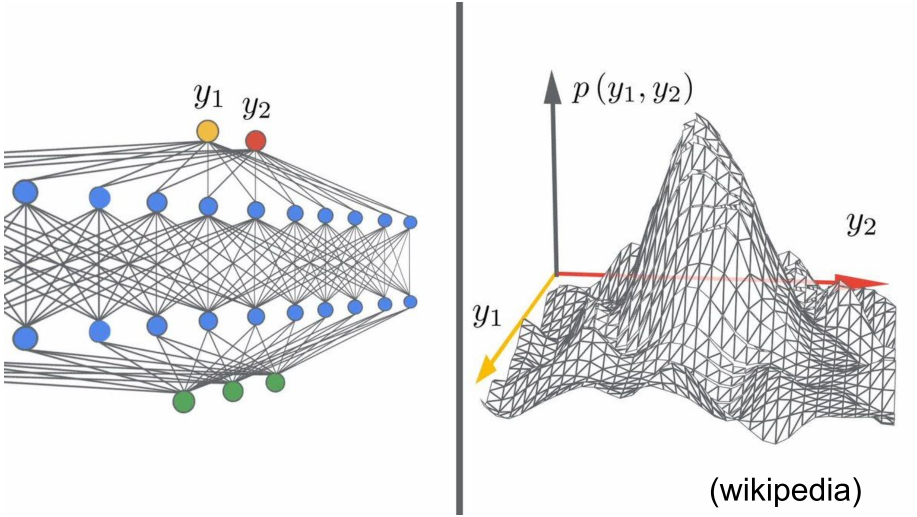
$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$$







Existing Probabilistic Perspectives on Neural Networks



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

$$f(\mathbf{x}) = \sum_{m=1}^{M} a_m \sigma(\mathbf{w}_m^\top \mathbf{x})$$

$$a_m \sim N(0, \sigma_a^2) \quad w_{mj} \sim N(0, \sigma_w^2)$$

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

$$\partial_t f_\theta(x) = (\nabla f_\theta(x))^T \partial_t \theta = \frac{2}{N} \sum_{i=1}^{N} (\nabla f_\theta(x))^T \nabla f_\theta(x_i)(y_i - f_\theta(x_i))$$

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

Existing Probabilistic Perspectives on Neural Networks

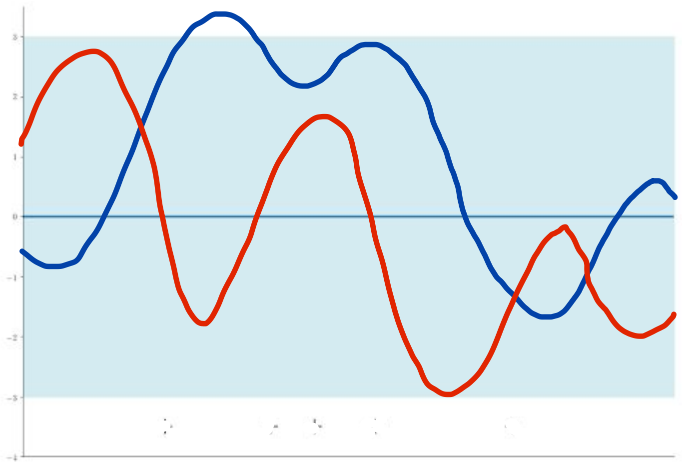
Pros

- Allows a probabilistic pipeline for learning neural networks
- Guidance on initialization (modeling), training (exact inference for GPs), and prediction (predict with uncertainty)

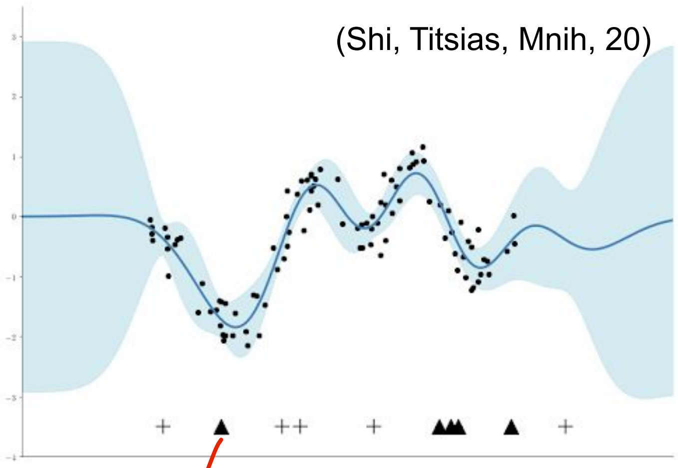
Cons

- Relies heavily on the infinite-width assumption (the CLT & linearization).
- Over-simplification by ignoring the importance of individual weights.
  - Putting a simple distribution over them /+ linearization
- Performance fails to match NNs with standard training.

Gaussian Processes and Sparse GPs



The GP model: prior distribution of functions



Sparse GP predictive distribution

Inducing Points (Z): Summarize the training data by function values at these locations.

Deriving Two-Layer NNs From Sparse GPs

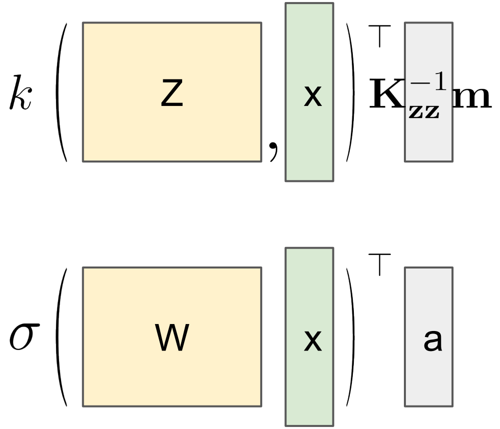
nonlinear linear

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

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

$\mathcal{N}(\mathbf{m}, \mathbf{S})$ are distributions of $f(\mathbf{Z})$

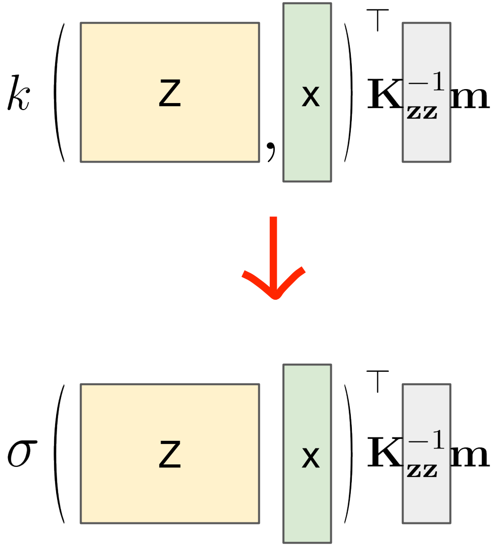
Predictive distribution of sparse GPs



Comparison with two-layer NNs

Neural Networks as Inter-domain Inducing Points (Sun, Shi, Grosse, 20)
https://openreview.net/pdf?id=NgqYp7sAW6t

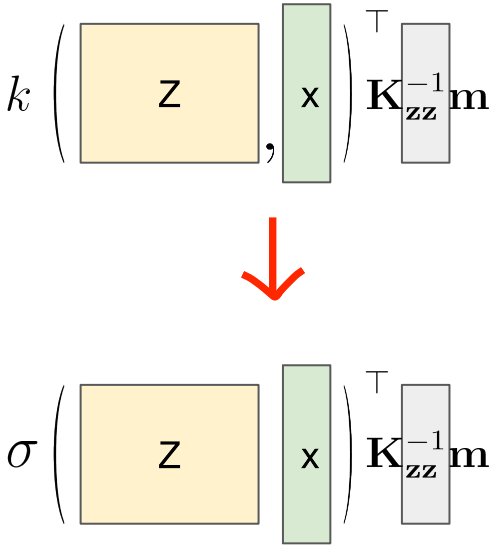
Deriving Two-Layer NNs From Sparse GPs



$$k(\mathbf{x}, \mathbf{z}_i) = \mathbb{E}\left[f(\mathbf{x})f(\mathbf{z}_i)\right]$$

Neural Networks as Inter-domain Inducing Points (Sun, Shi, Grosse, 20)
https://openreview.net/pdf?id=NgqYp7sAW6t

Deriving Two-Layer NNs From Sparse GPs

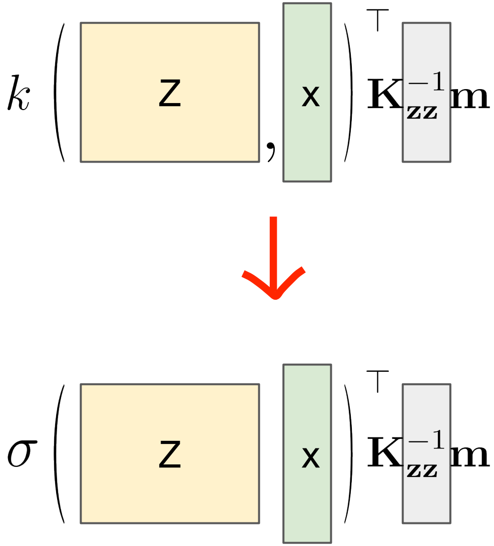


$$g(\mathbf{z}_i) = \langle f, \sigma(\langle \mathbf{z}_i, \cdot \rangle) \rangle_{\mathcal{H}}$$

$$k(\mathbf{x}, \mathbf{z}_i) = \mathbb{E}\left[f(\mathbf{x}) g(\mathbf{z}_i)\right]$$

Neural Networks as Inter-domain Inducing Points (Sun, Shi, Grosse, 20)
https://openreview.net/pdf?id=NgqYp7sAW6t

Deriving Two-Layer NNs From Sparse GPs



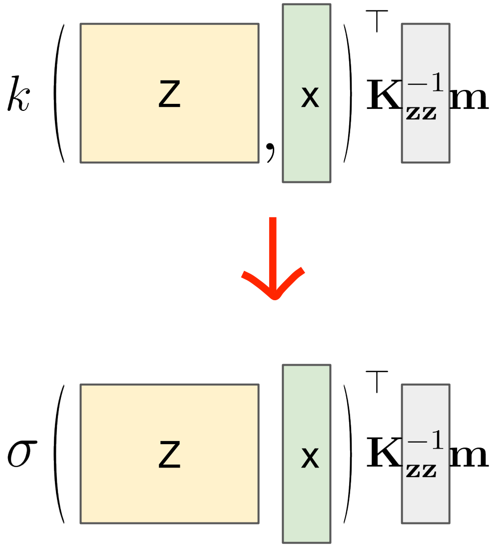
- Assumption: activation function σ in RKHS $\mathcal{H}$
- Inter-domain inducing points

$$g(\mathbf{z}_i) = \langle f, \sigma(\langle \mathbf{z}_i, \cdot \rangle) \rangle_{\mathcal{H}}$$

$$k(\mathbf{x}, \mathbf{z}_i) = \mathbb{E}\left[f(\mathbf{x}) g(\mathbf{z}_i) \right]$$

Neural Networks as Inter-domain Inducing Points (Sun, Shi, Grosse, 20)
https://openreview.net/pdf?id=NgqYp7sAW6t

Deriving Two-Layer NNs From Sparse GPs



- Assumption: activation function σ in RKHS $\mathcal{H}$
- Inter-domain inducing points

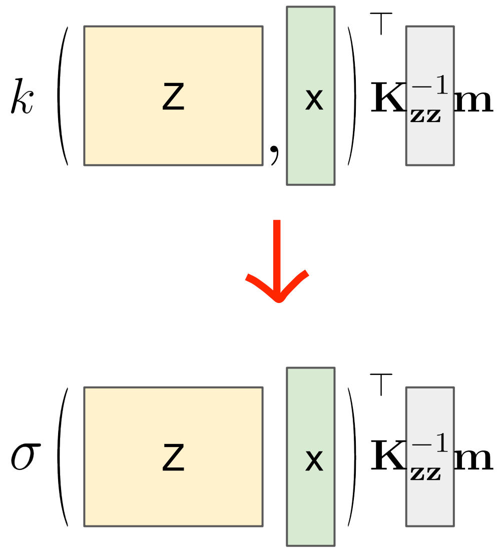
$$g(\mathbf{z}_i) = \langle f, \sigma(\langle \mathbf{z}_i, \cdot \rangle) \rangle_{\mathcal{H}}$$

$$k(\mathbf{x}, \mathbf{z}_i) = \mathbb{E}\left[f(\mathbf{x}) g(\mathbf{z}_i)\right]$$

$$= \langle \mathbb{E}\left[f(\mathbf{x}) \;\; f(\cdot)\right], \sigma(\langle \mathbf{z}_i, \cdot \rangle) \rangle_{\mathcal{H}}$$

Neural Networks as Inter-domain Inducing Points (Sun, Shi, Grosse, 20)
https://openreview.net/pdf?id=NgqYp7sAW6t

Deriving Two-Layer NNs From Sparse GPs



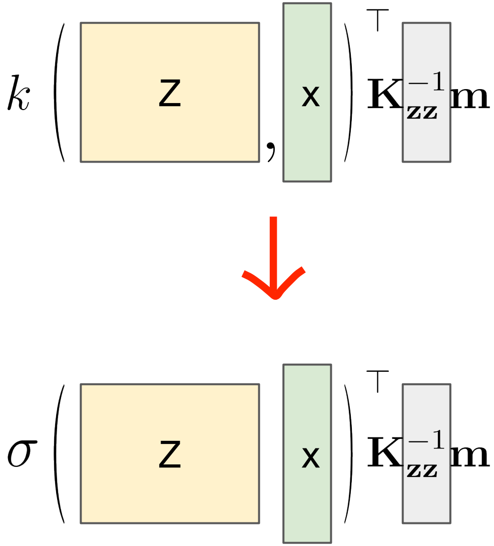
- Assumption: activation function σ in RKHS $\mathcal{H}$
- Inter-domain inducing points

$$g(\mathbf{z}_i) = \langle f, \sigma(\langle \mathbf{z}_i, \cdot \rangle) \rangle_{\mathcal{H}}$$

$$\begin{aligned} k(\mathbf{x}, \mathbf{z}_i) &= \mathbb{E}\left[f(\mathbf{x}) g(\mathbf{z}_i) \right] \\ &= \langle \mathbb{E}\left[f(\mathbf{x}) \;\; f(\cdot) \right], \sigma(\langle \mathbf{z}_i, \cdot \rangle) \rangle_{\mathcal{H}} \\ &= \langle k(\mathbf{x}, \cdot), \sigma(\langle \mathbf{z}_i, \cdot \rangle) \rangle_{\mathcal{H}} \end{aligned}$$

Neural Networks as Inter-domain Inducing Points (Sun, Shi, Grosse, 20)
https://openreview.net/pdf?id=NgqYp7sAW6t

Deriving Two-Layer NNs From Sparse GPs



- Assumption: activation function σ in RKHS $\mathcal{H}$
- Inter-domain inducing points

$$
\begin{aligned}
g(\mathbf{z}_i) &= \langle f, \sigma(\langle \mathbf{z}_i, \cdot \rangle) \rangle_{\mathcal{H}} \\
k(\mathbf{x}, \mathbf{z}_i) &= \mathbb{E}\left[f(\mathbf{x}) g(\mathbf{z}_i) \right] \\
&= \langle \mathbb{E}\left[f(\mathbf{x}) \;\; f(\cdot) \right], \sigma(\langle \mathbf{z}_i, \cdot \rangle) \rangle_{\mathcal{H}} \\
&= \langle k(\mathbf{x}, \cdot), \sigma(\langle \mathbf{z}_i, \cdot \rangle) \rangle_{\mathcal{H}} \\
&= \sigma(\mathbf{z}_i^\top \mathbf{x})
\end{aligned}
$$

Neural Networks as Inter-domain Inducing Points (Sun, Shi, Grosse, 20)
https://openreview.net/pdf?id=NgqYp7sAW6t

Numerical Experiments

Uncertainty from post-trained NNs

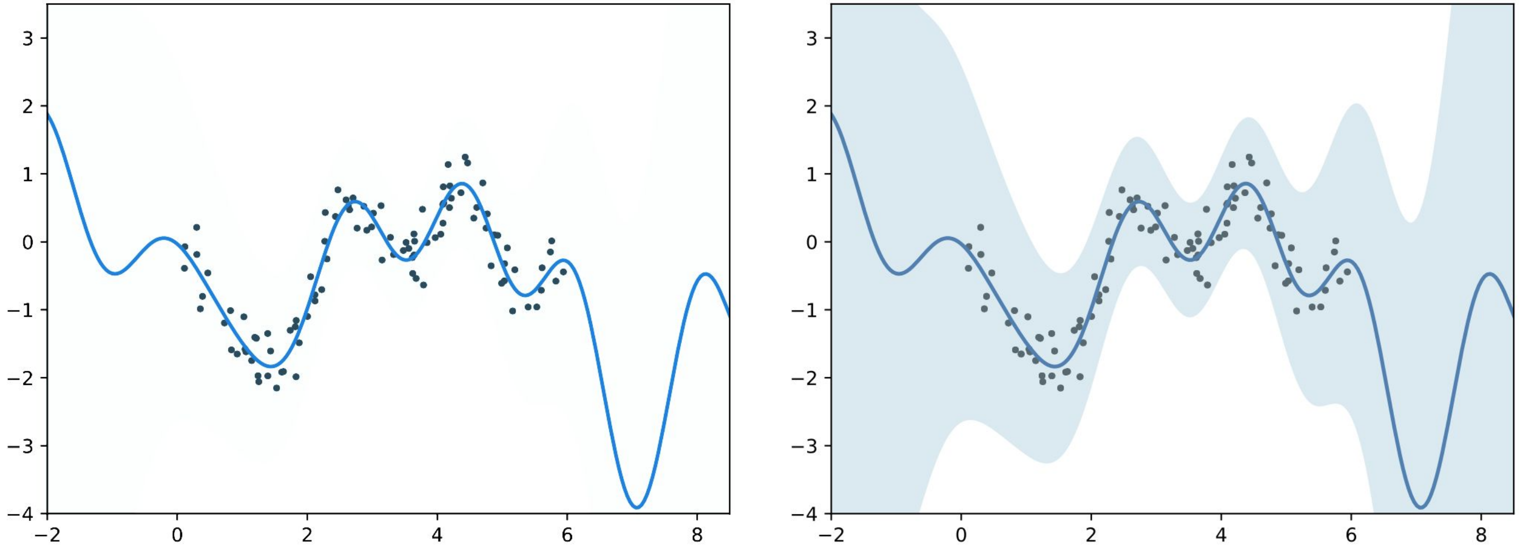
1. Train a two-layer neural network by standard backprop.
2. After training, extract the first-layer weights Z (inter-domain inducing points).
3. Compute (approximate) predictive variance of the sparse GP:

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

0

Neural Networks as Inter-domain Inducing Points (Sun, Shi, Grosse, 20)
https://openreview.net/pdf?id=NgqYp7sAW6t

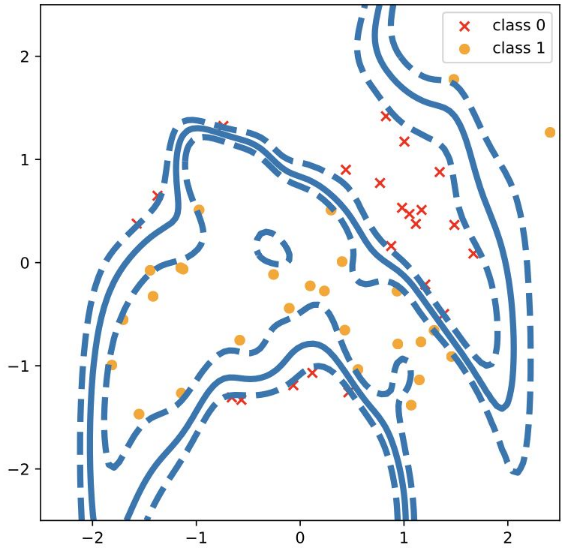
Numerical Experiments



Uncertainty from post-trained NNs

Neural Networks as Inter-domain Inducing Points (Sun, Shi, Grosse, 20)
https://openreview.net/pdf?id=NgqYp7sAW6t

Numerical Experiments



Uncertainty from post-trained NNs

Neural Networks as Inter-domain Inducing Points (Sun, Shi, Grosse, 20)
https://openreview.net/pdf?id=NgqYp7sAW6t

Future Work

- Extend the results to multi-layer NNs
  - What are the second, third, fourth … layer of weights?
  - Convolutional structures?
- How does this help us understand neural networks?
  - Approximation, optimization & generalization

Neural Networks as Inter-domain Inducing Points (Sun, Shi, Grosse, 20)
https://openreview.net/pdf?id=NgqYp7sAW6t