Structured Inter-domain Inducing Points for Variational Gaussian Processes

Shengyang Sun
University of Toronto

Outline

- Background: Inter-domain Inducing Points & Variational Fourier Features
- Harmonic variational Gaussian Processes
- Neural networks as Inter-domain Inducing Points

Gaussian Processes

- Gaussian processes (GPs) are natural generalizations of multivariate Gaussian distributions,

$$f\left(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right) \sim \mathcal{N}\left(\mu\left(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right), \begin{bmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) \\ \vdots & \ddots & \vdots \\ k(x_n,x_1) & \cdots & k(x_n,x_n) \end{bmatrix}\right)$$

function values $\mathbf{f}_X$ mean kernel matrix $\boldsymbol{K}_{XX}$

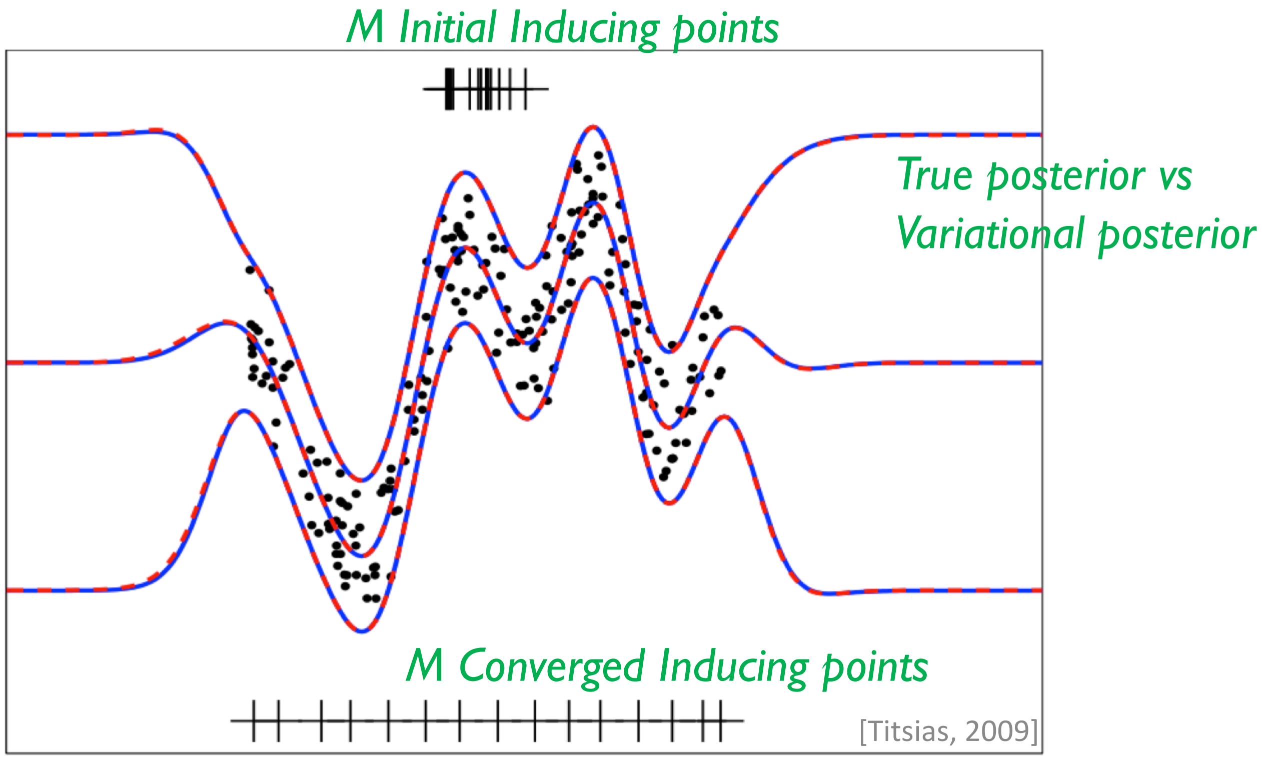
- Under a Gaussian likelihood, the GP posterior has explicit expressions.

$$\mathbf{f}_\star | \mathbf{y} \sim \mathcal{N}\left(\mathbf{K}_{\star\mathbf{X}}\left(\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}, \mathbf{K}_{\star\star} - \mathbf{K}_{\star\mathbf{X}}\left(\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{K}_{\mathbf{X}\star}\right)$$

Cubic computations

Inducing Points

- *Inducing points Z are a small set of points to* summarize the dataset in variational GPs[1] (VGPs),



Computational complexities: $\mathcal{O}(N^3) \rightarrow \mathcal{O}(M^3)$

[1][Titsias, 2009; Hensman et. al., 2015]

Inducing Points

- While the GP model fixes $k(X, X)$, the VGP optimizes Z for approximate posterior.

$$\begin{bmatrix} k(X,X) & k(X,Z) \\ k(Z,X) & k(Z,Z) \end{bmatrix} \qquad \begin{bmatrix} k(X,X) & B \\ B^T & C \end{bmatrix}$$

- VGPs can be done as long as the augmented kernel matrix is PSD.

- *How to design PSD augmented kernels?*

Inter-domain Inducing Points

- A kernel can be characterized as the covariance of a stochastic process

$$k(x, x') \longleftrightarrow Cov(f(x), f(x'))$$

- Given any function $w: \mathcal{X} \rightarrow \mathcal{R}$, an inducing variable[1] is defined as,

$$u_w = \int f(x)w(x)dx$$

- The augmented covariance can be computed as,

$k(X,X)$	$k(X,W)$
$k(W,X)$	$k(W,W)$

$$k(x, w) \longleftrightarrow Cov(f(x), u_w) = \int k(x, x')w(x')dx'$$

$$k(w, w') \longleftrightarrow Cov(u_w, u_{w'}) = \int k(x, x')w(x)w'(x')dxdx'$$

[1]Lázaro-Gredilla & Figueiras-Vidal, 2009

Variational Fourier Features

- A kernel can be characterized as the covariance of a stochastic process

$$k(x, x') \longleftrightarrow Cov(f(x), f(x'))$$

- Given any function $w: \mathcal{X} \rightarrow \mathcal{R}$, an inducing variable[1] is defined as,

$$u_w = < f, w >_{\mathcal{H}}$$

- The augmented covariance can be computed as,

$$\begin{bmatrix} k(X, X) & W(X)^T \\ W(X) & < W, W >_{\mathcal{H}} \end{bmatrix}$$

$$k(x, w) \longleftrightarrow Cov(f(x), u_w) = < k(x, \cdot), w >_{\mathcal{H}} = w(x)$$

$$k(w, w') \longleftrightarrow Cov(u_w, u_{w'}) = < w, w' >_{\mathcal{H}}$$

[1]Hensman et al., 2017

Why Care?

- Accurate posterior inference
  - The Nystrom approximation can be more accurate[1].

- Computational benefits
  - The kernel matrix $k(W, W)$ can be structured[2].

- Wider applicable scenarios of kernel methods

[1]Burt et al., 2019; [2] Burt et al., 2020; Dutordoir et al., 2020

Harmonic Variational Gaussian Processes

A simple example

- Given two inputs z_1, z_2, we define two inter-domain inducing functions,

$$w_1 = \frac{1}{2}(\delta_{z_1} + \delta_{-z_1})(\cdot) \qquad w_2 = \frac{1}{2}(\delta_{z_2} - \delta_{-z_2})(\cdot)$$

- The augmented kernel,

$$k(w_1, w_2) = \frac{1}{4}\int k(x, x')\big(\delta_{z_1} + \delta_{-z_1}\big)(x)\big(\delta_{z_2} - \delta_{-z_2}\big)(x')dx\,dx'$$

$$= \frac{1}{4}\big(k(z_1, z_2) - k(z_1, -z_2) + k(-z_1, z_2) - k(-z_1, -z_2)\big)$$

$$= 0$$

If k is invariant to negations: $k(x, x') = k(-x, -x')$

A simple example

- The kernel matrix $k(W, W)$ is 2x2 block diagonal.

m	$k_1(W_1, W_1)$	0
m	0	$k_2(W_2, W_2)$

- Two times of inducing points with only two times of computations: $2m^3$ instead of $8m^3$!

Generalizing the simple example

A negative transformation

$$x \rightleftarrows -x$$

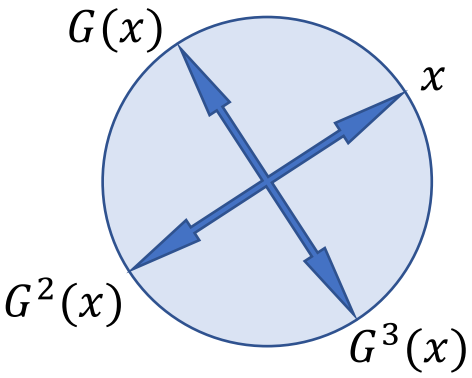
Kernel is invariant to negations

$$k(x, x') = k(-x, -x')$$

2 types of inducing points

$$\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

T-cyclic transformation G



Kernel is invariant to G

$$k(x, x') = k(G(x), G(x'))$$

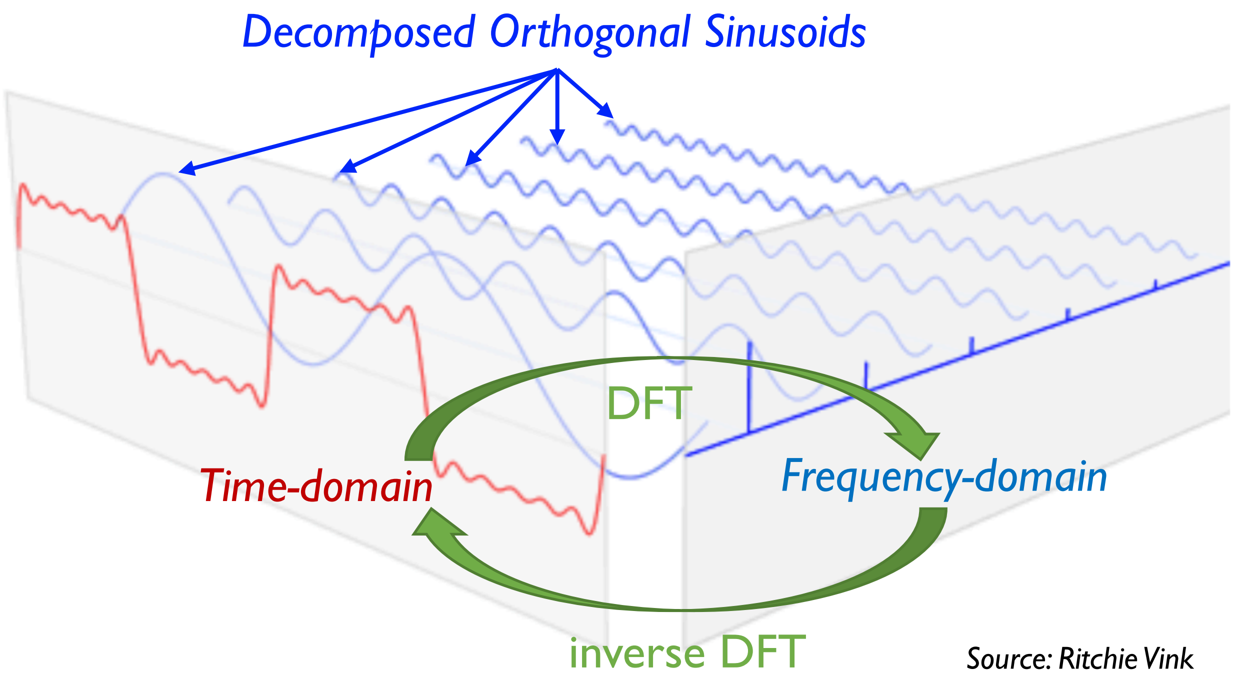
T types of inducing points

$$\frac{1}{T}\left[e^{-i\frac{2\pi ts}{T}}\right]_{t,s=1}^{T}$$

Discrete Fourier Transform (DFT)

Harmonic Kernel Decomposition

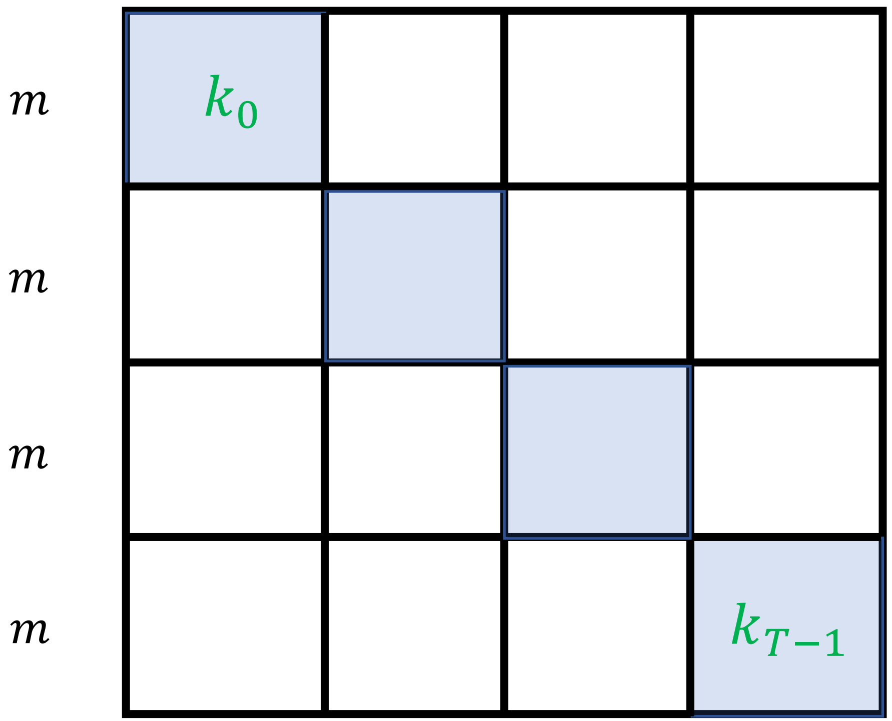
- DFT: "time-domain" representations into "frequency-domain" representations,



- HKD: DFT applied to kernels
  - Orthogonal kernel sum decomposition

Harmonic Variational Gaussian Process

- HVGP: a scalable variational GP approximation



$T{\times}m$: T types of orthogonal inducing points

Similar to SVGP:

- *Large Datasets*
- *High Dimensional Inputs*
- *Trainable Inducing Points*

Better than SVGP:

- *More Inducing Points*
- *Less Computational Costs*
- *Easier Parallelisms*

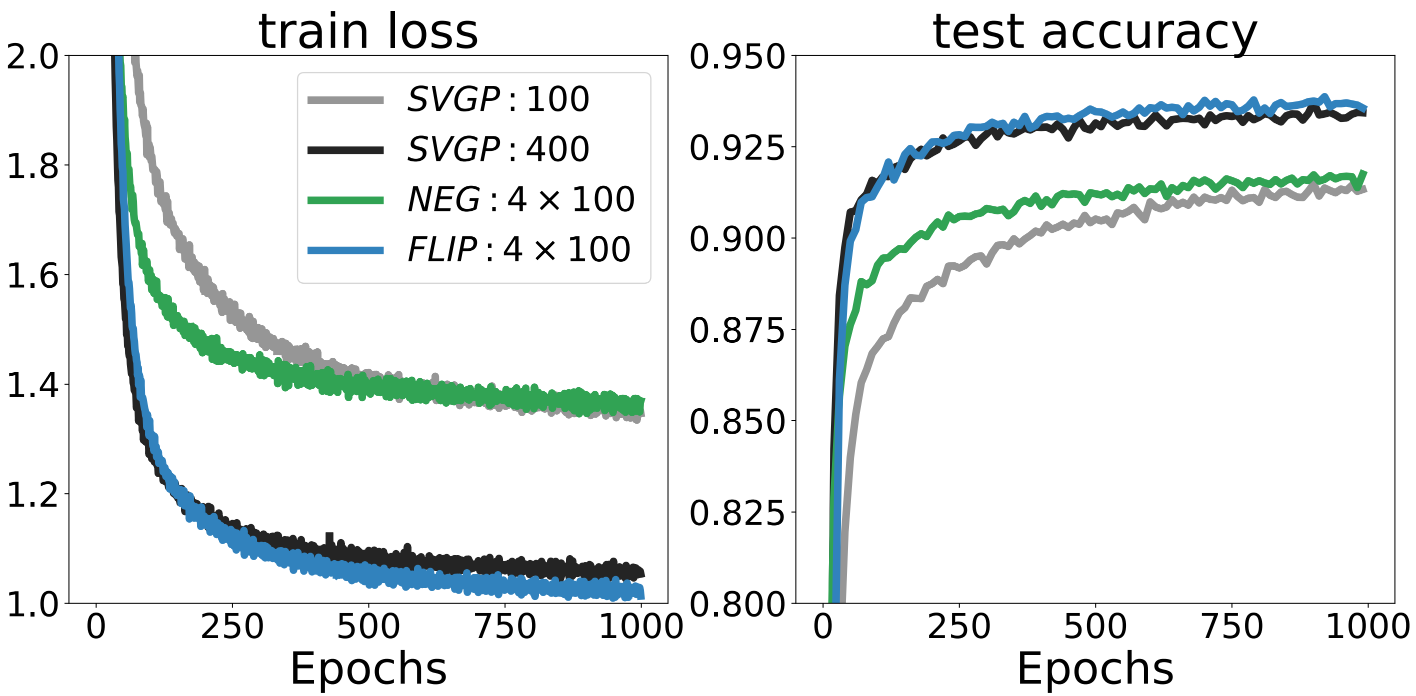
Substantial reduction in terms of computational complexities: $\mathcal{O}(T^3m^3) \rightarrow \mathcal{O}(Tm^3 + T^2m^2)$

Harmonic Variational Gaussian Process

- HVGP: a scalable variational GP approximation

Substantial reduction in terms of computational complexities: $\mathcal{O}(T^3m^3) \rightarrow \mathcal{O}(Tm^3 + T^2m^2)$

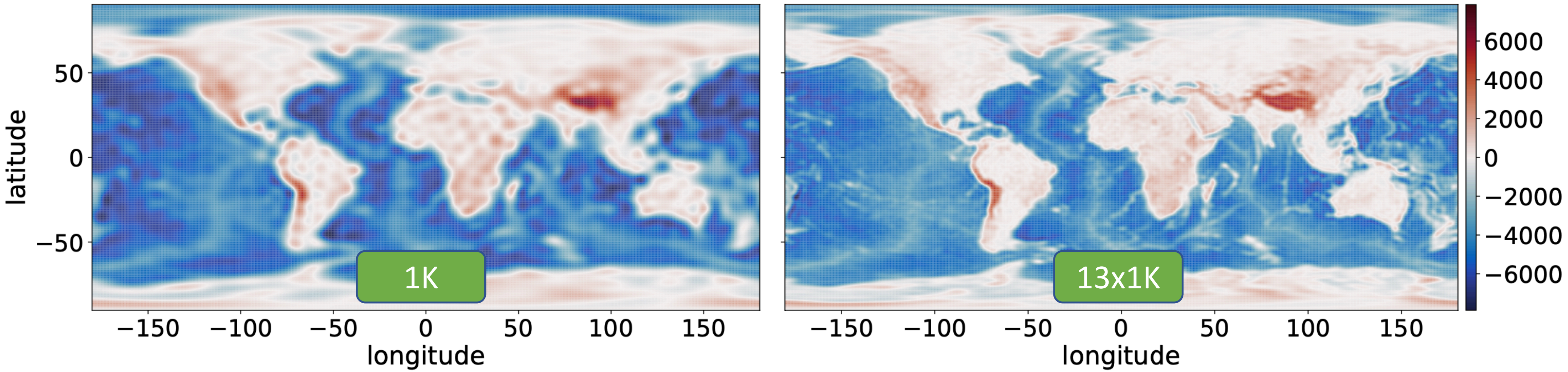
High-fidelity GP approximation if the transformation is properly chosen



Performances over Flip-MNIST

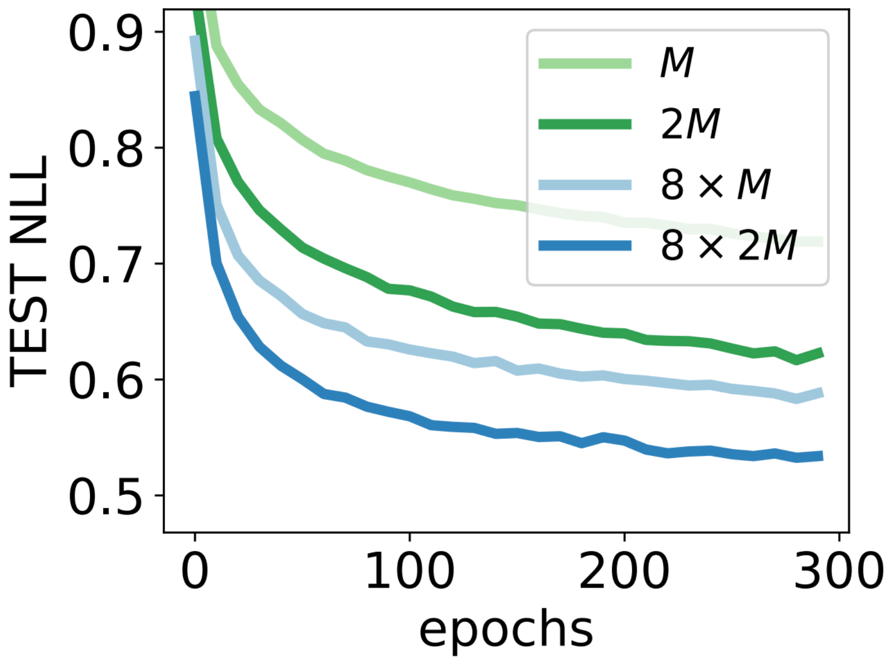
Experimental Results

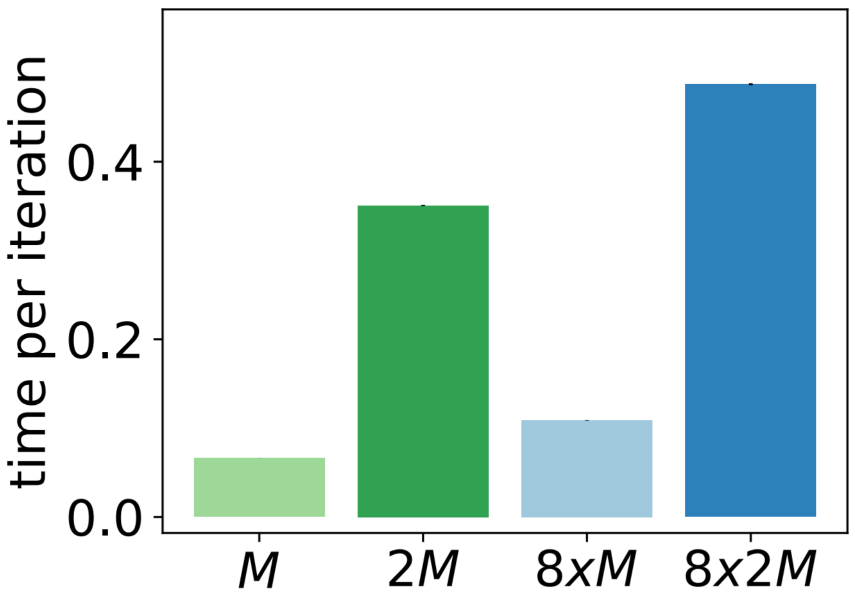
- More inducing points



Experimental Results

- Predictive performances & Parallelisms





Experimental Results

- Flexible model designs

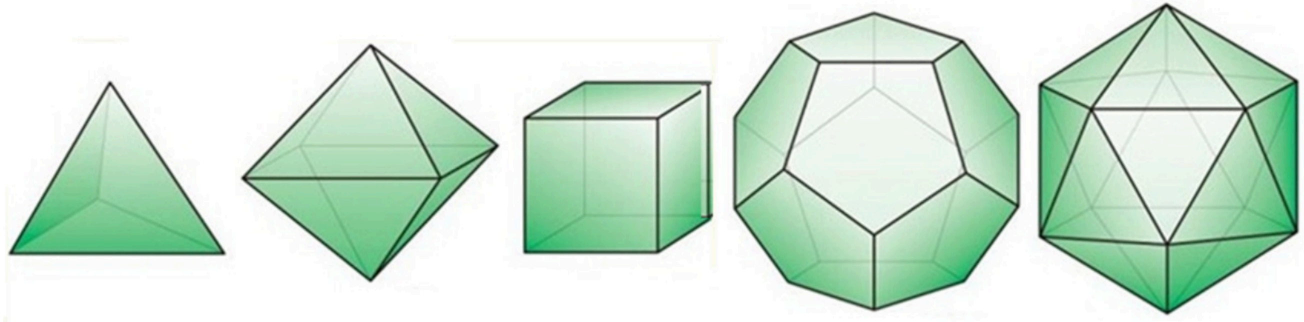
M	Model	ACC	NLL	sec/iter
384x2, 1K	M	79.01±0.11	0.86±0.00	0.17
	2M	80.27±0.04	0.81±0.00	0.52
	M+M	79.98 ±0.21	0.80±0.01	0.46
	2xM	80.04±0.04	0.80±0.00	0.37
	4xM	**80.52±0.20**	**0.75±0.01**	0.37
384x3, 1K	M	82.41±0.08	0.73±0.01	0.40
	2M	-	-	-
	M+M	83.26±0.19	0.69±0.01	1.24
	2xM	**84.97±0.08**	0.60±0.00	0.90
	4xM	**84.85±0.11**	**0.58±0.00**	0.90

CIFAR-10 classification via deep convolutional GPs

Future Directions

- Transformations over adaptive manifolds.

- Transformations beyond cyclic groups.



Source: Ouyang et al., 2017

- Expressive kernel learning.

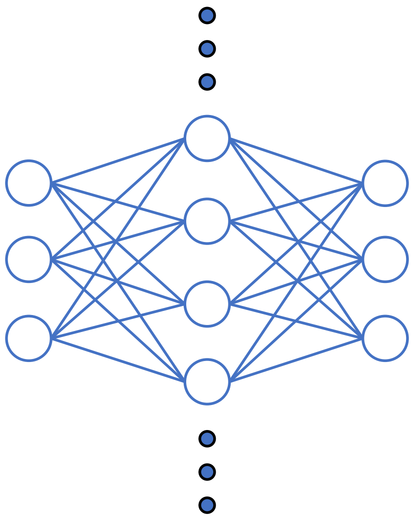
HVGP: Orthogonal Inter-domain Inducing Points for Substantial Computational Improvements

Neural Networks as Inter-domain Inducing Points

Existing Kernel Perspectives on Neural Networks

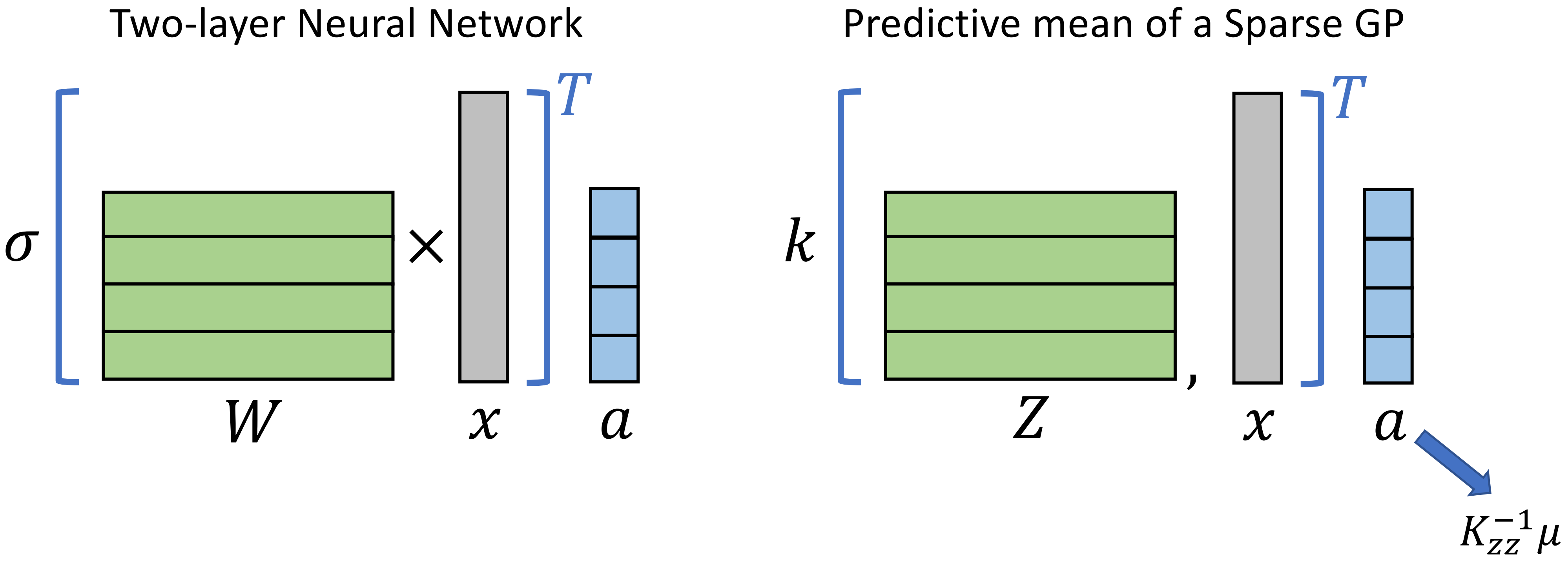
Infinite-width neural networks at initialization are Gaussian processes (Neal 92, Lee et al. 18)

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



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

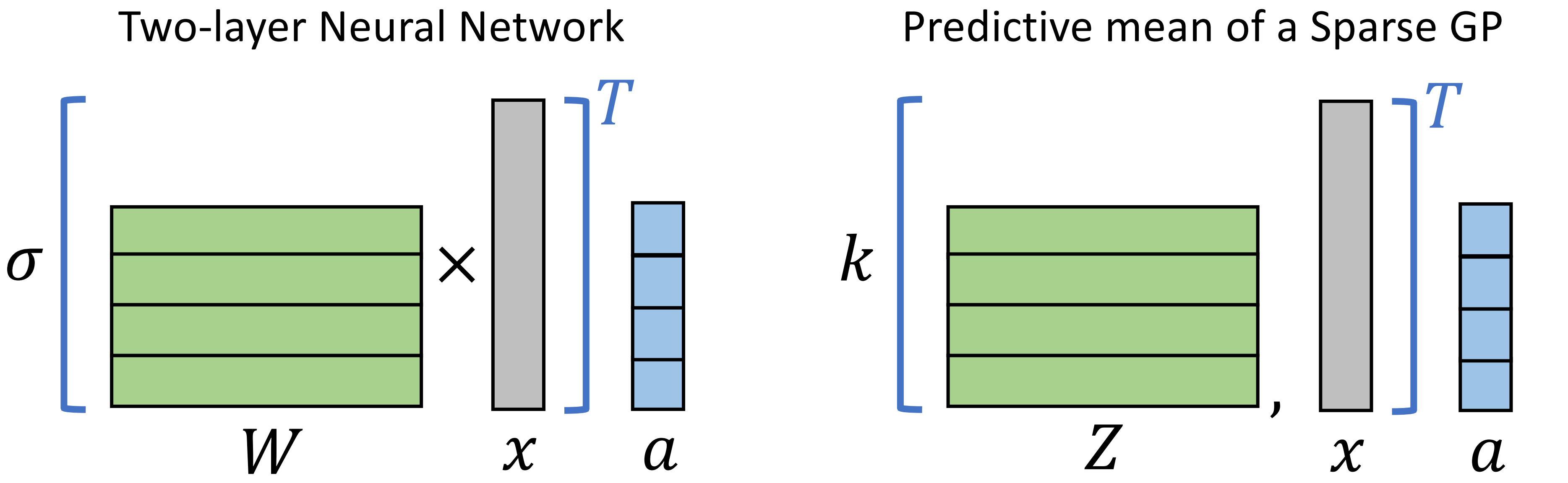
Neural Networks as Inter-domain Inducing Points



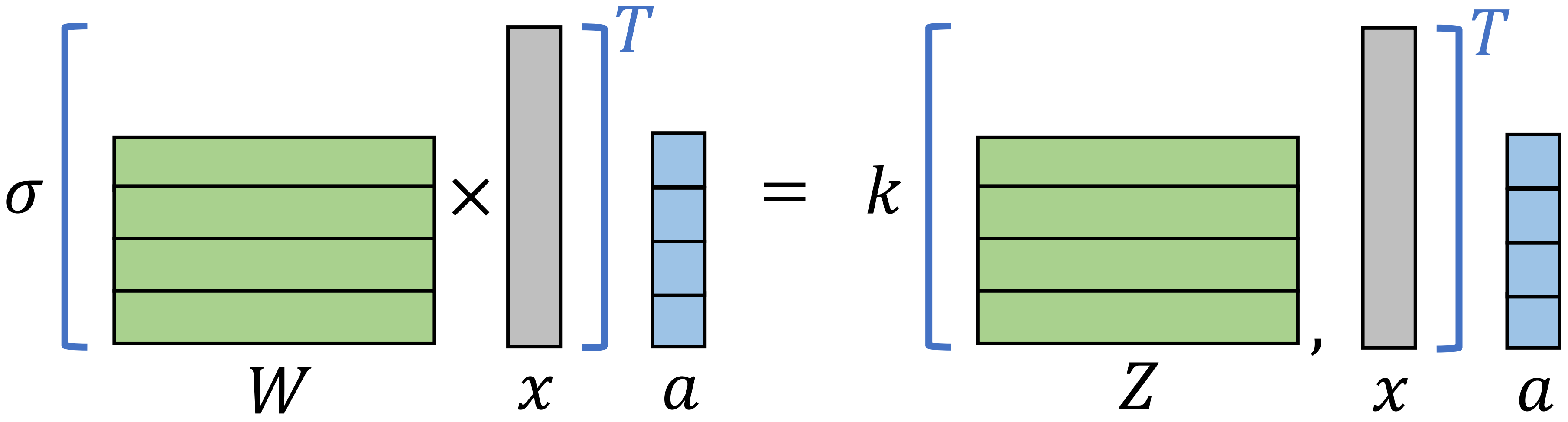
Neural Networks as Inter-domain Inducing Points

- We define the variational Fourier feature z_i, $z_i(x) = \sigma(w_i^T x)$. Then,

$$k(x, z_i) = z_i(x) = \sigma\left(w_i^T x\right), \qquad k\left(z_i, z_j\right) =< \sigma\left(w_i^T \cdot\right), \sigma\left(w_j^T \cdot\right) >_{\mathcal{H}}$$



Neural Networks as Inter-domain Inducing Points



A New Interpretation of finite-width NN:

- Each activation function $\sigma(\cdot; w)$ can be seen as an inter-domain inducing point $k(\cdot; z)$.
- The number of hidden units equals to the number of inducing points.
- A two-layer NN becomes equivalent to the predictive mean of a variational GP.

The variational GP: $f(x) \sim \mathcal{N}(\mathrm{NN}(x), \sigma^2(x))$

- Performance matches the standard NN

Numerical Experiments

Direct Uncertainty from post-trained NNs

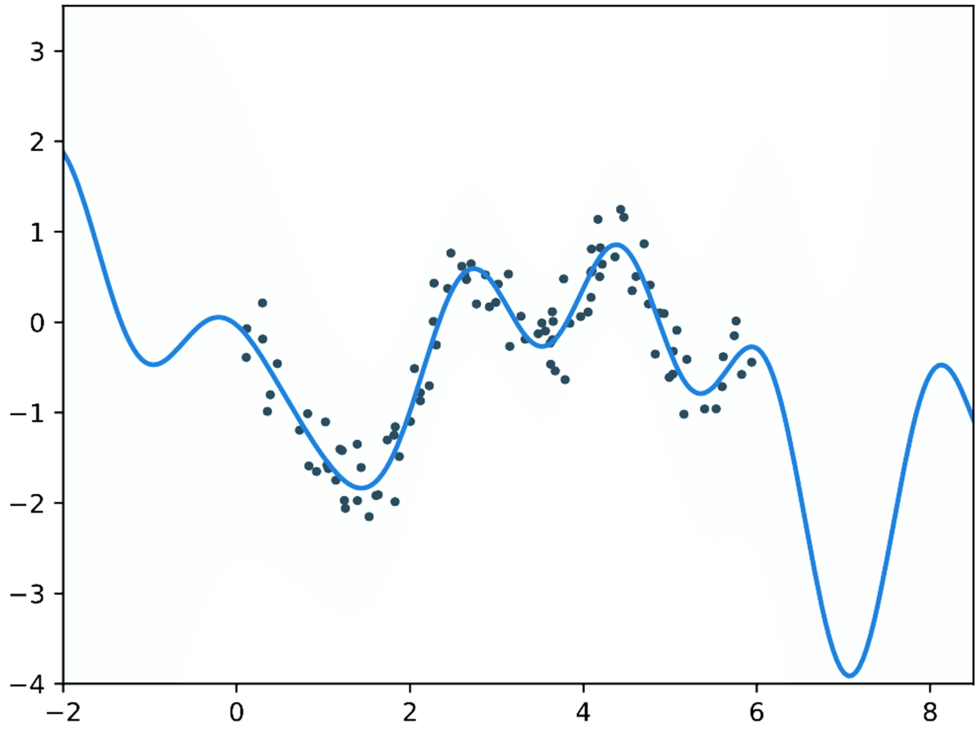
1. Train a neural network by standard backprop.
2. After training, each hidden unit is an inter-domain inducing point.
3. Compute (approximate) predictive variance of the corresponding sparse GP:

Nystrom Approximation Error — Inducing Variable Variance

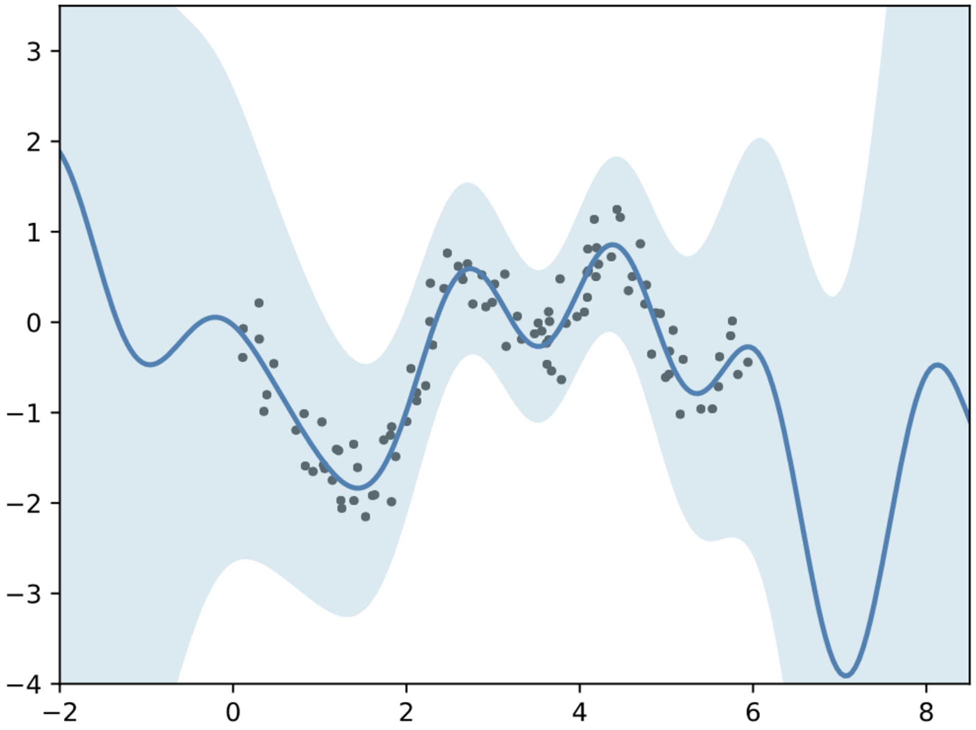
$$\sigma^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_{\mathbf{zx}}^\top \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{k}_{\mathbf{zx}} + \mathbf{k}_{\mathbf{zx}}^\top \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{S} \mathbf{K}_{\mathbf{zz}}^{-1} \mathbf{k}_{\mathbf{zx}}$$
$$\approx k(\mathbf{x}, \mathbf{x})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

- We derived analytic expressions of $k(z_i, z_j)$ for two-layer neural networks.

Numerical Experiments

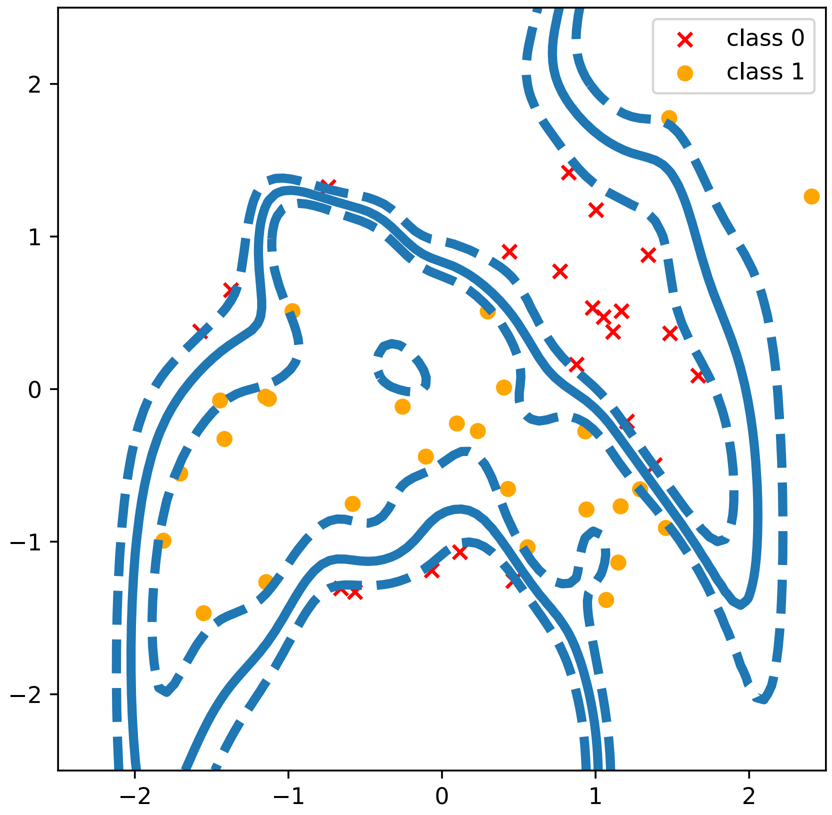


Two-Layer Cosine Network



Uncertainty from post-trained NNs

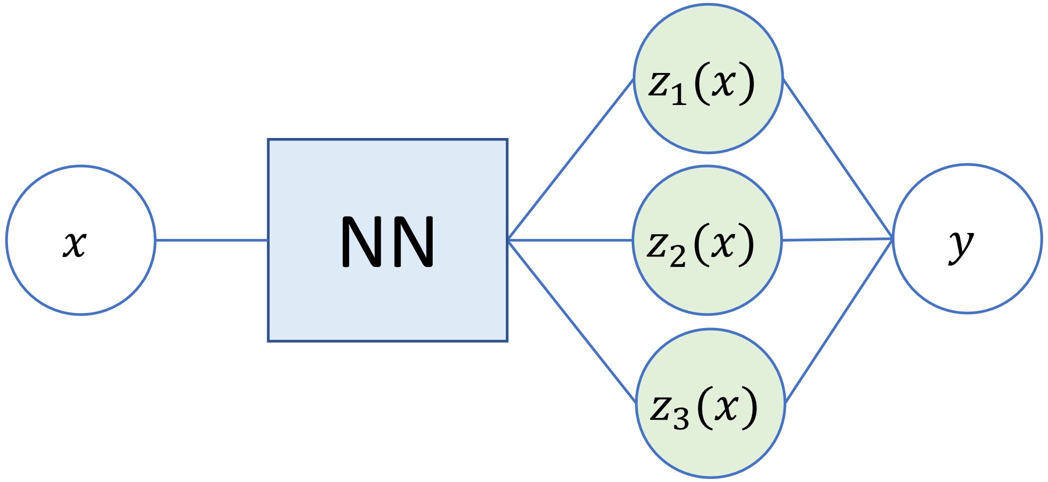
Numerical Experiments



Uncertainty from post-trained Two-Layer Erf NNs

Deep Neural Networks

- Argument[1]: A deep neural network also corresponds to a variational GP.
- Each hidden unit at the second-last layer is an inter-domain inducing point.



- Caveat: The analytic expression of $k(z_i, z_j)$ is generally intractable for deep networks.

[1]This is in contrast to Dutordoir et al., 2021

Future Directions

- Direct uncertainty from post-trained deep neural networks.
- Generalization bounds for NNs.
- Alternative regularizations in NN training.

$$\left| \frac{1}{n} \sum_i l(f(\boldsymbol{x}_i), y_i) - \mathbb{E}_P[l(f(\boldsymbol{x}), y)] \right| \leq C_1 + C_2 \frac{||f||_{\mathcal{H}}^{\alpha}}{n^{\beta}}, \quad C_1, C_2, \alpha, \beta \geq 0$$

Source: Belkin et al., 2018

Obstacle: accurate & efficient approximations of the kernel $k(z_i, z_j)$.

Finite-width neural networks are variational GPs with inter-domain inducing points

End Remarks

- This talk covers,
  - Scalable Variational Gaussian Processes via Harmonic Kernel Decomposition (Sun et al., ICML 2021)
  - Neural Networks as Inter-Domain Inducing Points (Sun et al., AABI 2020)
- Careful design of inter-domain inducing points can bring substantial computational savings.
- Inter-domain inducing points provide a promising direction to understand finite neural networks.

References

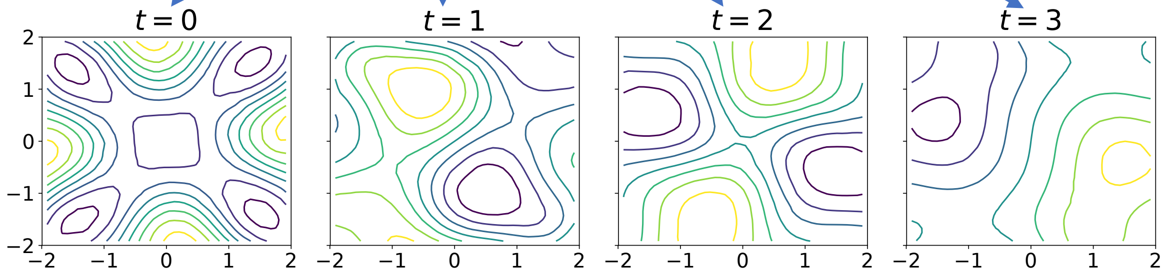
- **Sun, Shengyang**, et al. "Scalable Variational Gaussian Processes via Harmonic Kernel Decomposition." *International Conference on Machine Learning*. PMLR, 2021.
- **Sun, Shengyang**, Jiaxin Shi, and Roger Baker Grosse. "Neural Networks as Inter-Domain Inducing Points." Third Symposium on Advances in Approximate Bayesian Inference. 2020.
- Titsias, Michalis. "Variational learning of inducing variables in sparse Gaussian processes." Artificial intelligence and statistics. PMLR, 2009.
- Hensman, James, Alexander Matthews, and Zoubin Ghahramani. "Scalable variational Gaussian process classification." Artificial Intelligence and Statistics. PMLR, 2015.
- Lázaro-Gredilla, Miguel, and Anibal Figueiras-Vidal. "Inter-domain Gaussian processes for sparse inference using inducing features." Advances in Neural Information Processing Systems 22 (2009).
- Hensman, James, Nicolas Durrande, and Arno Solin. "Variational Fourier Features for Gaussian Processes." *J. Mach. Learn. Res.* 18.1 (2017): 5537-5588.
- Burt, David, Carl Edward Rasmussen, and Mark Van Der Wilk. "Rates of convergence for sparse variational Gaussian process regression." *International Conference on Machine Learning*. PMLR, 2019.
- Dutordoir, Vincent, Nicolas Durrande, and James Hensman. "Sparse Gaussian processes with spherical harmonic features." *International Conference on Machine Learning*. PMLR, 2020.
- Burt, David R., Carl Edward Rasmussen, and Mark van der Wilk. "Variational orthogonal features." *arXiv preprint arXiv:2006.13170* (2020).
- Dutordoir, Vincent, et al. "Deep neural networks as point estimates for deep Gaussian processes." Advances in Neural Information Processing Systems 34 (2021).
- Neal, Radford M. Bayesian learning for neural networks. Vol. 118. Springer Science & Business Media, 2012.
- Lee, Jaehoon, et al. "Deep neural networks as gaussian processes." arXiv preprint arXiv:1711.00165 (2017).
- Jacot, Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." Advances in neural information processing systems 31 (2018).

Harmonic Kernel Decomposition

$$k(x, x')$$

HKD

$$[k_0(x, x'), k_1(x, x'), k_2(x, x'), k_3(x, x')]$$



Theorem: orthogonal kernel decomposition

$$\mathrm{k}(x, x') = k_0(x, x') + k_1(x, x') + k_2(x, x') + k_3(x, x')$$

Harmonic Kernel Decomposition

- The HKD is an orthogonal decomposition of kernels and RKHSs,

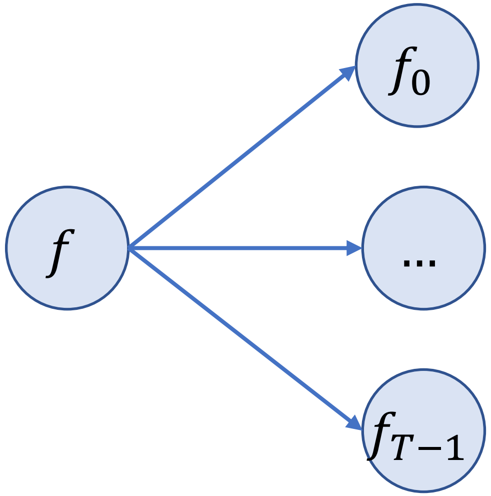
$$k(\mathbf{x}, \mathbf{x}') = \sum_{t=0}^{T-1} k_t(\mathbf{x}, \mathbf{x}') \qquad \mathcal{H}_k = \bigoplus_{t=0}^{T-1} \mathcal{H}_{k_t}$$

- The HKD is widely applicable to many kernels: RBF, Matérn, polynomial, periodic, …

Kernels k	Inner-Product	Stationary	Stationary
Input Space $\mathcal{X}$	Complex, Real	Real	Torus
Transformation G	Rotation, Reflection	Negation	Translation

Harmonic Variational Gaussian Process

- HVGP: a scalable variational GP approximation



$u_0 = f_0(z_0)$

…

$u_{T-1} = f_{T-1}(z_{T-1})$

HKD *Orthogonal inducing points*

Harmonic Variational Gaussian Process

- From kernel decomposition to GP decomposition:

$$f = \sum_{t=0}^{T-1} f_t, \; f_t \sim \mathcal{GP}(0, k_t)$$

- The HVGP introduces an independent variational posterior for each component GP,

$$f = \sum_{t=0}^{T-1} f_t, \; q_t(f_t, \mathbf{u}_t) = p_t(f_t|\mathbf{u}_t)q_t(\mathbf{u}_t)$$

- The variational posterior can be optimized by maximizing the ELBO,

$$\mathbb{E}_{q(f_0,\ldots,f_{T-1})}\left[\log p\left(\mathbf{y}\middle|\sum_{t=0}^{T-1} f_t, \mathbf{X}\right)\right] - \sum_{t=0}^{T-1} \mathrm{KL}\left(q_t(\mathbf{u}_t)\|p_t(\mathbf{u}_t)\right)$$