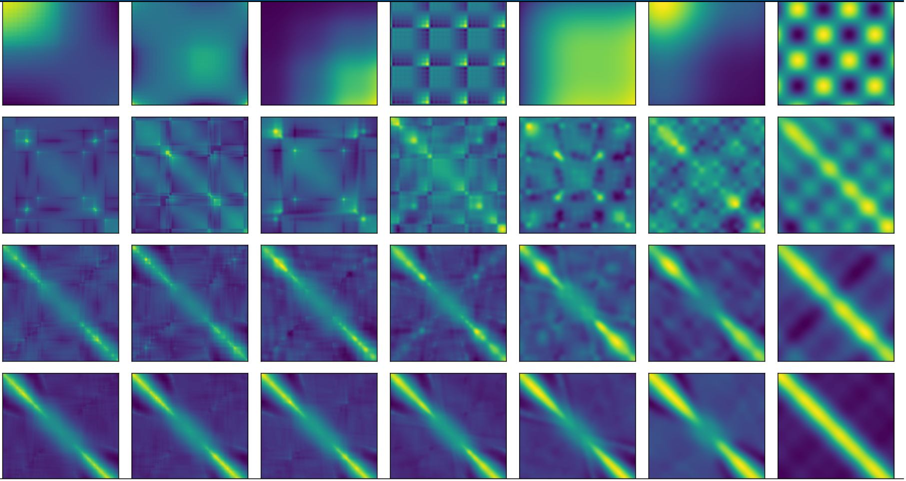
Random Fourier Series Features

Defense “Expert Lab on Robot Learning”



Motivation

Outline

Motivation

Methodology

Evaluation

Conclusion

Deep Learning and Neural Networks

- (Deep) neural networks dominate AI

Deep Learning and Neural Networks

- (Deep) neural networks dominate AI
  - extremely expressive
  - great predictive power

Deep Learning and Neural Networks

- (Deep) neural networks dominate AI
  - extremely expressive
  - great predictive power
  - lack uncertainty estimation

Deep Learning and Neural Networks

TECHNISCHE UNIVERSITÄT DARMSTADT

- (Deep) neural networks dominate AI
  - extremely expressive
  - great predictive power
  - lack uncertainty estimation
- Lead to the development of *Bayesian* neural networks

Deep Learning and Neural Networks

TECHNISCHE UNIVERSITÄT DARMSTADT

- (Deep) neural networks dominate AI
  - extremely expressive
  - great predictive power
  - lack uncertainty estimation
- Lead to the development of *Bayesian* neural networks
  - intractable exact inference
  - complicated training
  - ...

Gaussian Processes (GPs)

- GPs are still the go-to model for reliable uncertainty quantification

Gaussian Processes (GPs)

TECHNISCHE UNIVERSITÄT DARMSTADT

- GPs are still the go-to model for reliable uncertainty quantification
- but performance highly depends on the kernel choice…
  - tackled by kernel learning

Gaussian Processes (GPs)

- GPs are still the go-to model for reliable uncertainty quantification
- but performance highly depends on the kernel choice…
  - tackled by kernel learning
- exact inference complexity is cubic w.r.t. number of data points
  - prohibits online use of GPs

TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation

Methodology

Evaluation

Conclusion

Random Fourier Features

$$\boldsymbol{z}_\omega(\boldsymbol{x}) = \begin{bmatrix} \cos(\langle \omega | \boldsymbol{x} \rangle) \\ \sin(\langle \omega | \boldsymbol{x} \rangle) \end{bmatrix}$$

- resort to “classical” Bayesian regression
- explicit posterior over the weights

Random Fourier Features

$$\boldsymbol{z}_\omega(\boldsymbol{x}) = \begin{bmatrix} \cos(\langle \omega | \boldsymbol{x} \rangle) \\ \sin(\langle \omega | \boldsymbol{x} \rangle) \end{bmatrix}$$

- resort to “classical” Bayesian regression
- explicit posterior over the weights
- approximate every stationary kernel $k(\cdot)$:

$$k(\boldsymbol{x} - \boldsymbol{y}) = \mathbb{E}_{\omega \sim p(\cdot)} \left[\langle \boldsymbol{z}_\omega(\boldsymbol{x}) | \boldsymbol{z}_\omega(\boldsymbol{y}) \rangle \right] \approx \frac{1}{N} \sum_{i=1}^{N} \langle \boldsymbol{z}_{\omega_i}(\boldsymbol{x}) | \boldsymbol{z}_{\omega_i}(\boldsymbol{y}) \rangle , \quad \omega_i \sim p(\cdot)$$

Random Fourier Features

Approximating the Squared Exponential

For the SE kernel: tractable Fourier transform

Random Fourier Features

Approximating the Squared Exponential

For the SE kernel: tractable Fourier transform

$$k_{\mathrm{SE}}(\boldsymbol{x} - \boldsymbol{y}) = \exp\left\{ -\langle \boldsymbol{x} - \boldsymbol{y} | \boldsymbol{x} - \boldsymbol{y} \rangle / 2 \right\}$$

$$p(\omega) = \left(\mathcal{F} k_{\mathrm{SE}}\right)(\omega) = \mathcal{N}(\omega \mid \mathbf{0}, \mathbf{I})$$

Random Fourier Features
Approximating the Squared Exponential

TECHNISCHE UNIVERSITÄT DARMSTADT

For the SE kernel: tractable Fourier transform

$$k_{\mathrm{SE}}(\boldsymbol{x}-\boldsymbol{y}) = \exp\left\{ -\langle \boldsymbol{x}-\boldsymbol{y} | \boldsymbol{x}-\boldsymbol{y} \rangle / \, 2 \right\}$$

$$p(\omega) = \left(\mathcal{F} k_{\mathrm{SE}}\right)(\omega) = \mathcal{N}(\omega \mid \mathbf{0}, \mathbf{I})$$

$$\begin{aligned}
\mathbb{E}\left[\langle \boldsymbol{z}_\omega(\boldsymbol{x}) | \boldsymbol{z}_\omega(\boldsymbol{y}) \rangle\right] &= \mathbb{E}\left[\left\langle \begin{bmatrix} \cos(\langle \omega | \boldsymbol{x} \rangle) \\ \sin(\langle \omega | \boldsymbol{x} \rangle) \end{bmatrix} \middle| \begin{bmatrix} \cos(\langle \omega | \boldsymbol{y} \rangle) \\ \sin(\langle \omega | \boldsymbol{y} \rangle) \end{bmatrix} \right\rangle\right] \\
&= \mathbb{E}\left[\cos(\langle \omega | \boldsymbol{x} \rangle) \cos(\langle \omega | \boldsymbol{y} \rangle) + \sin(\langle \omega | \boldsymbol{x} \rangle) \sin(\langle \omega | \boldsymbol{y} \rangle)\right] \\
&= \mathbb{E}\left[\cos(\langle \omega | \boldsymbol{x}-\boldsymbol{y} \rangle)\right] = \mathrm{Re}\left(\mathbb{E}\left[\exp\{i \langle \omega | \boldsymbol{x}-\boldsymbol{y} \rangle\}\right]\right) \\
&= \mathrm{Re}\left(\left(\mathcal{F}^{-1} p\right)(\boldsymbol{x}-\boldsymbol{y})\right) = \mathrm{Re}\left(k_{\mathrm{SE}}(\boldsymbol{x}-\boldsymbol{y})\right) = k_{\mathrm{SE}}(\boldsymbol{x}-\boldsymbol{y})
\end{aligned}$$

Random Fourier Features

Approximating the Squared Exponential

For the SE kernel: tractable Fourier transform

$$k_{\mathrm{SE}}(\boldsymbol{x}-\boldsymbol{y}) = \exp\big\{-\langle \boldsymbol{x}-\boldsymbol{y} | \boldsymbol{x}-\boldsymbol{y}\rangle / 2\big\}$$

$$p(\omega) = \big(\mathcal{F}k_{\mathrm{SE}}\big)(\omega) = \mathcal{N}(\omega \,|\, \mathbf{0}, \mathbf{I})$$

$$\begin{aligned}
\mathbb{E}\big[\langle \boldsymbol{z}_\omega(\boldsymbol{x}) | \boldsymbol{z}_\omega(\boldsymbol{y})\rangle\big] &= \mathbb{E}\left[\left\langle \begin{bmatrix}\cos(\langle\omega|\boldsymbol{x}\rangle) \\ \sin(\langle\omega|\boldsymbol{x}\rangle)\end{bmatrix} \middle| \begin{bmatrix}\cos(\langle\omega|\boldsymbol{y}\rangle) \\ \sin(\langle\omega|\boldsymbol{y}\rangle)\end{bmatrix} \right\rangle\right] \\
&= \mathbb{E}\big[\cos(\langle\omega|\boldsymbol{x}\rangle)\cos(\langle\omega|\boldsymbol{y}\rangle) + \sin(\langle\omega|\boldsymbol{x}\rangle)\sin(\langle\omega|\boldsymbol{y}\rangle)\big] \\
&= \mathbb{E}\big[\cos(\langle\omega|\boldsymbol{x}-\boldsymbol{y}\rangle)\big] = \mathrm{Re}\Big(\mathbb{E}\big[\exp\{i\,\langle\omega|\boldsymbol{x}-\boldsymbol{y}\rangle\}\big]\Big) \\
&= \mathrm{Re}\Big(\big(\mathcal{F}^{-1}p\big)(\boldsymbol{x}-\boldsymbol{y})\Big) = \mathrm{Re}\big(k_{\mathrm{SE}}(\boldsymbol{x}-\boldsymbol{y})\big) = k_{\mathrm{SE}}(\boldsymbol{x}-\boldsymbol{y})
\end{aligned}$$

- SE kernel is extremely smooth (Stein, 1999)

Random Fourier *Series* Features

TECHNISCHE UNIVERSITÄT DARMSTADT

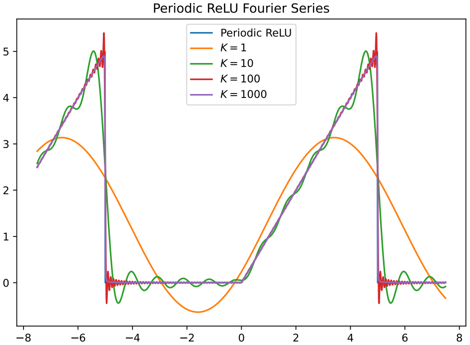
We extend random Fourier features:

$$\boldsymbol{z}_\omega(\boldsymbol{x}) = \begin{bmatrix} \cos(\langle \omega | \boldsymbol{x} \rangle) \\ \sin(\langle \omega | \boldsymbol{x} \rangle) \end{bmatrix} \longrightarrow \begin{aligned} \boldsymbol{z}_\omega(\boldsymbol{x}) &= \sum_{k=1}^{K} \boldsymbol{z}_\omega^{(k)}(\boldsymbol{x}), \\ \boldsymbol{z}_\omega^{(k)}(\boldsymbol{x}) &= \begin{bmatrix} a_k \cos\left(\pi \widetilde{T}^{-1} k \left\langle \omega | \boldsymbol{\Lambda}^{-1} | \boldsymbol{x} \right\rangle\right) \\ b_k \sin\left(\pi \widetilde{T}^{-1} k \left\langle \omega | \boldsymbol{\Lambda}^{-1} | \boldsymbol{x} \right\rangle\right) \end{bmatrix} \end{aligned}$$

- similar to the sine-cosine formulation of Fourier series

Fourier Series
Sine-Cosine Formulation

$$\widehat{f}_K(x) = \frac{a_0}{2} + \sum_{k=1}^{K} a_k \cos\left(\pi \widetilde{T}^{-1} k x\right) + b_k \sin\left(\pi \widetilde{T}^{-1} k x\right)$$



Motivation

Methodology

Evaluation

Conclusion

Hypothesis

TECHNISCHE UNIVERSITÄT DARMSTADT

Central Hypothesis

Random Fourier series features outperform random Fourier features.

Evaluation

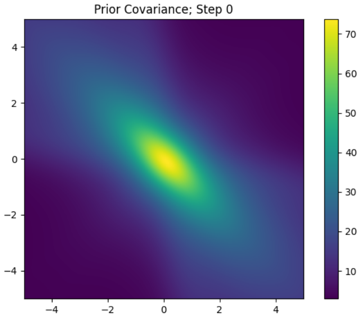
TECHNISCHE UNIVERSITÄT DARMSTADT

- Datasets:
  - Synthetic Data (Cosine, Heaviside, Heavi-Cosine, Gap-Cosine)
  - UCI (Boston, Concrete, Power, Yacht, Energy, Kin8nm, Naval, Protein, Wine)
  - Cartpole

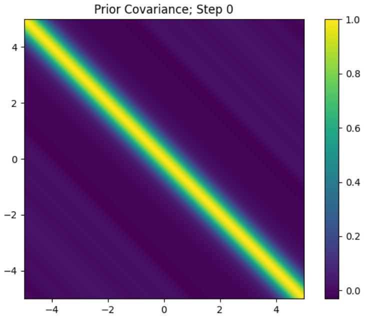
Evaluation

- Datasets:
  - Synthetic Data (Cosine, Heaviside, Heavi-Cosine, Gap-Cosine)
  - UCI (Boston, Concrete, Power, Yacht, Energy, Kin8nm, Naval, Protein, Wine)
  - Cartpole
- Different RFSF Initializations:
  - Random
  - ReLU
  - Single Harmonic (SH)

How the Kernel Learns

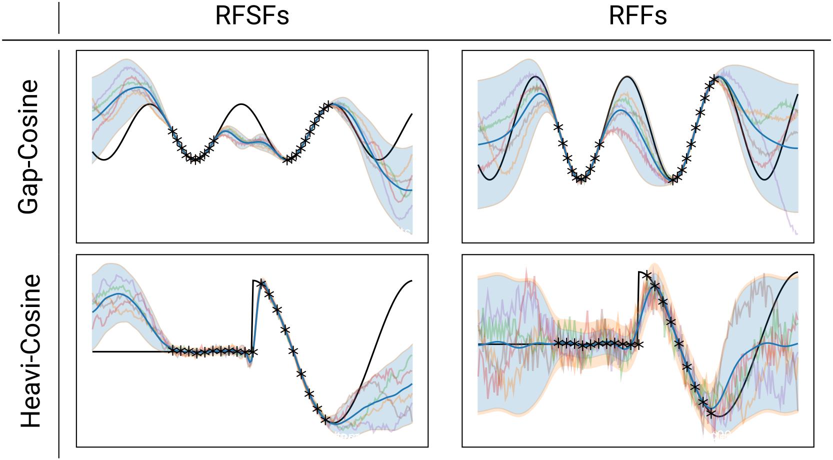


RFSF on Gap-Cosine



RFFs on Gap-Cosine

Results on the Synthetic Data



Quantified Results

Synthetic Data Sets and Cartpole

	Model		Data Set				
			Cosine	Heaviside	Heavi-Cosine	Gap-Cosine	Cartpole
Log-Lik.	*RFSF*	*Random*	***2.43***	*0.11*	*-1.66*	*1.27*	*-9.88 ± 1.86*
		ReLU	*2.34*	***0.80***	*-0.90*	*1.50*	*-12.30 ± 2.31*
		SH	*2.37*	***0.21***	*-1.23*	*1.52*	*-9.73 ± 2.10*
	GP	SE	**2.44**	0.73	**0.77**	**2.58**	**-3.21 ± 1.64**
		RFF	**2.44**	0.73	**0.78**	**2.59**	-7.38 ± 1.94

Quantified Results

UCI Data Sets

TECHNISCHE UNIVERSITÄT DARMSTADT

	Model		Data Set Boston	Concrete	Power	Yacht
Log-Lik.	*RFSF*	*Random*	*−2.40 ± 0.05*	***−2.94 ± 0.05***	*−2.78 ± 0.01*	*−0.80 ± 0.02*
		ReLU	*−2.39 ± 0.05*	***−2.93 ± 0.04***	*−2.80 ± 0.01*	*−0.86 ± 0.02*
		SH	*−2.44 ± 0.06*	***−2.94 ± 0.05***	*−2.78 ± 0.01*	*−0.83 ± 0.02*
	GP	SE	**−2.38 ± 0.05**	**−2.98 ± 0.06**	−2.82 ± 0.01	−0.80 ± 0.02
		RFF	−2.40 ± 0.06	**−3.01 ± 0.05**	−2.84 ± 0.01	−0.80 ± 0.02
	GBLL[1]	Leaky ReLU	−2.90 ± 0.05	−3.09 ± 0.03	−2.77 ± 0.01	−1.67 ± 0.11
		Tanh	−3.06 ± 0.03	−3.21 ± 0.03	−2.83 ± 0.01	−0.70 ± 0.10
	Ensemble[1]	Leaky ReLU	−2.48 ± 0.09	**−3.04 ± 0.08**	**−2.70 ± 0.01**	−0.35 ± 0.07
		Tanh	−2.48 ± 0.08	**−3.03 ± 0.07**	**−2.72 ± 0.01**	**−0.03 ± 0.05**
	MAP[1]	Leaky ReLU	−2.60 ± 0.07	−3.04 ± 0.04	−2.77 ± 0.01	−5.14 ± 1.62
		Tanh	−2.59 ± 0.06	−3.11 ± 0.04	−2.76 ± 0.01	−1.77 ± 0.53

[1] Results taken from Watson et al. (2021), "Latent Derivative Bayesian Last Layer Networks."

Motivation

Methodology

Evaluation

Conclusion

Conclusion

TECHNISCHE UNIVERSITÄT DARMSTADT

Central Hypothesis

Random Fourier series features outperform random Fourier features.

Conclusion

TECHNISCHE UNIVERSITÄT DARMSTADT

Central Hypothesis

Random Fourier series features outperform random Fourier features.

- we compared to RFFs, SE, and BNN methods

Conclusion

TECHNISCHE UNIVERSITÄT DARMSTADT

Central Hypothesis

Random Fourier series features outperform random Fourier features.

- we compared to RFFs, SE, and BNN methods
- advantage of RFSFs is not consistent
- no performance gain
- also true for the SH initialization

Future Work

TECHNISCHE UNIVERSITÄT DARMSTADT

- theoretical analysis what RFSFs approximate
- better understanding of the half-period initialization

TECHNISCHE UNIVERSITÄT DARMSTADT

Methodology

Evaluation

Hyper-Parameter Optimization

- Hyper-Parameters
  - $\boldsymbol{a}_{1:K}$ (sine coefficients)
  - $\boldsymbol{b}_{1:K}$ (cosine coefficients)
  - $\boldsymbol{\Lambda}$ (length-scales)
  - $\widetilde{T}$ (half-period)
  - σ_n^2 (aleatoric noise variance)
- maximization of the marginal log-likelihood
- using the empirical Bayes approximation

$$\boldsymbol{z}_\omega(\boldsymbol{x}) = \sum_{k=1}^{K} \boldsymbol{z}_\omega^{(k)}(\boldsymbol{x}),$$

$$\boldsymbol{z}_\omega^{(k)}(\boldsymbol{x}) = \begin{bmatrix} a_k \cos\left(\pi \widetilde{T}^{-1} k \left\langle \omega \middle| \boldsymbol{\Lambda}^{-1} \middle| \boldsymbol{x} \right\rangle\right) \\ b_k \sin\left(\pi \widetilde{T}^{-1} k \left\langle \omega \middle| \boldsymbol{\Lambda}^{-1} \middle| \boldsymbol{x} \right\rangle\right) \end{bmatrix}$$

Evaluation

Outline

Methodology

Evaluation

Quantified Results

UCI Data Sets; Cont.

		Model	Energy	Kin8nm	Naval	Protein	Wine
			Data Set				
Log-Lik.	*RFSF*	*Random*	***-0.70 ± 0.02***	*0.68 ± 0.05*	*-78.19 ± 69.72*	*-2.94 ± 0.03*	*-0.11 ± 0.07*
		ReLU	*-0.74 ± 0.02*	***0.97 ± 0.03***	*-172.57 ± 104.83*	*-629.05 ± 384.60*	*-0.11 ± 0.06*
		SH	*-0.74 ± 0.02*	*0.52 ± 0.07*	*-62.69 ± 55.40*	*-2.96 ± 0.03*	***0.01 ± 0.06***
	GP	SE	**-0.68 ± 0.02**	-0.22 ± 0.24	**6.91 ± 0.15**	**-2.89 ± 0.00**	-0.84 ± 0.05
		RFF	**-0.69 ± 0.02**	0.75 ± 0.04	-1941.56 ± 248.64	-2.90 ± 0.00	-0.89 ± 0.04