

# Optimizing Gaussian Processes

## Honours Research Project

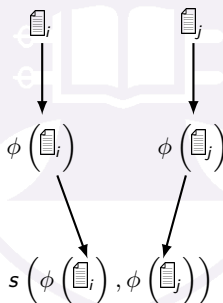
Michael Ciccotosto-Camp - 44302913



THE UNIVERSITY  
OF QUEENSLAND  
AUSTRALIA

# Problem Setting and Motivation

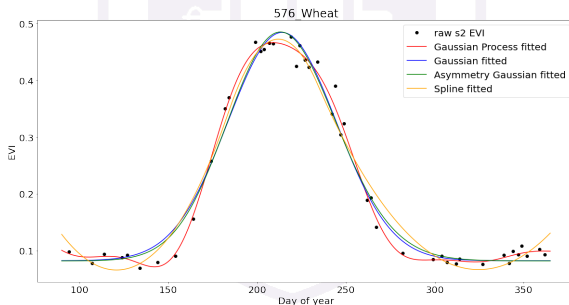
- Suppose we have some inputs  $[\mathbb{I}_1, \dots, \mathbb{I}_n]$  (with their corresponding experimental observations  $[y_1, \dots, y_n]$ ), where  $\mathbb{I}_i$  can take a number of different of form (perhaps a tree data structure or vectors of values).



- The function  $s$  provides us with some notion of similarity between inputs after they've been "transformed" into a nicer form using a *feature map*  $\phi$ .

# Problem Setting and Motivation

- The idea of studying time series prediction came from a research group from the Gatton campus, lead by Dr Potgieter, analysing crop growth from previous seasons to forecast when certain phenological stages will take place in the current harvest.



# Introduction to Gaussian Processes

- A Gaussian Process (GP) is a collection of random variables with index set  $I$ , such that every finite subset of random variables has a joint Gaussian distribution and are completely characterized by a mean function  $m : X \rightarrow \mathbb{R}$  and a kernel  $k : X \times X \rightarrow \mathbb{R}$  (in this context, think of the kernel as a function that provides some notion of similarity between points).

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))].$$

# Predictions

- Using the assumption that our data can be modelled as a Gaussian process, we can write out the new distribution of the observed noisy values along the points at which we wish to test the underlying function as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_\star \end{bmatrix} \sim \mathbb{N} \left( \mathbf{0}, \begin{bmatrix} \mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_n^2 \mathbb{I}_{n \times n} & \mathbf{K}_{\mathbf{X}_\star \mathbf{X}} \\ \mathbf{K}_{\mathbf{X}_\star \mathbf{X}} & \mathbf{K}_{\mathbf{X}_\star \mathbf{X}_\star} \end{bmatrix} \right).$$

(using the notation  $(\mathbf{K}_{\mathbf{W}\mathbf{W}'} )_{i,j} \triangleq k(\mathbf{w}_i, \mathbf{w}'_j)$ )

# Predictions

- Using the assumption that our data can be modelled as a Gaussian process, we can write out the new distribution of the observed noisy values along the points at which we wish to test the underlying function as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_* \end{bmatrix} \sim \mathbb{N} \left( \mathbf{0}, \begin{bmatrix} \mathbf{K}_{\mathbf{xx}} + \sigma_n^2 \mathbb{I}_{n \times n} & \mathbf{K}_{\mathbf{x}_* \mathbf{x}}^\top \\ \mathbf{K}_{\mathbf{x}_* \mathbf{x}} & \mathbf{K}_{\mathbf{x}_* \mathbf{x}_*} \end{bmatrix} \right).$$

(using the notation  $(\mathbf{K}_{\mathbf{ww}'} )_{i,j} \triangleq k(\mathbf{w}_i, \mathbf{w}_j')$ )

- The mean and covariance can then be computed as

$$\begin{aligned} \bar{\mathbf{y}}_* &= \mathbf{K}_{\mathbf{x}_* \mathbf{x}} \left[ \mathbf{K}_{\mathbf{xx}} + \sigma_n^2 \mathbb{I}_{n \times n} \right]^{-1} \mathbf{y} \\ \text{cov}(\mathbf{y}_*) &= \mathbf{K}_{\mathbf{x}_* \mathbf{x}_*} - \mathbf{K}_{\mathbf{x}_* \mathbf{x}} \left[ \mathbf{K}_{\mathbf{xx}} + \sigma_n^2 \mathbb{I}_{n \times n} \right]^{-1} \mathbf{K}_{\mathbf{x}_* \mathbf{x}}^\top. \end{aligned}$$

# Unoptimized GPR

---

## Algorithm 1: Unoptimized GPR

---

**input** : Observations  $\mathbf{X}, \mathbf{y}$  and a test input  $\mathbf{x}_*$ .

**output**: A prediction  $\bar{f}_*$  with its corresponding variance  $\mathbb{V}[f_*]$ .

- 1  $\mathbf{L} = \text{cholesky}(\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_n^2 \mathbb{I}_{n \times n})$
  - 2  $\boldsymbol{\alpha} = \text{lin-solve}(\mathbf{L}^\top, \text{lin-solve}(\mathbf{L}, \mathbf{y}))$
  - 3  $\bar{\mathbf{y}}_* = \mathbf{K}_{\mathbf{x}_* \mathbf{X}} \boldsymbol{\alpha}$
  - 4  $\mathbf{v} = \text{lin-solve}(\mathbf{L}, \mathbf{K}_{\mathbf{x}_* \mathbf{X}})$
  - 5  $\mathbb{V}[f_*] = \mathbf{K}_{\mathbf{x}_* \mathbf{x}_*} - \mathbf{v}^\top \mathbf{v}$
  - 6 **return**  $\bar{f}_*, \mathbb{V}[f_*]$
-

# Problems with Unoptimized GPR

---

## Algorithm 2: Unoptimized GPR

---

**input** : Observations  $\mathbf{X}, \mathbf{y}$  and a prediction inputs  $\mathbf{x}_*$ .

**output**: A prediction  $\bar{f}_*$  with its corresponding variance  $\mathbb{V}[f_*]$ .

---

```

1  $\mathbf{L} = \text{cholesky}(\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_n^2 \mathbb{I}_{n \times n})$ 
2  $\boldsymbol{\alpha} = \text{lin-solve}(\mathbf{L}^\top, \text{lin-solve}(\mathbf{L}, \mathbf{y}))$ 
3  $\bar{f}_* = \mathbf{K}_{\mathbf{x}_* \mathbf{X}} \boldsymbol{\alpha}$ 
4  $\mathbf{v} = \text{lin-solve}(\mathbf{L}, \mathbf{K}_{\mathbf{x}_* \mathbf{x}_*})$ 
5  $\mathbb{V}[f_*] = \mathbf{K}_{\mathbf{x}_* \mathbf{x}_*} - \mathbf{v}^\top \mathbf{v}$ 
6 return  $\bar{f}_*, \mathbb{V}[f_*]$ 

```

---

- Lines 1, 2 and 4 can be incredibly slow as computing  $\mathbf{K}_{\mathbf{X}\mathbf{X}}$  doing a Cholesky decomposition and performing linear solves scale poorly as the number of inputs,  $n$ , grows.



# Nystrom Approximation

- The Nystrom method we seek a matrix  $\mathbf{Q} \in \mathbb{R}^{n \times k}$  that satisfies  $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^* \mathbf{A}\|_F \leq \varepsilon$ , where  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is positive semi definite matrix, to form the rank- $k$  approximation

# Nystrom Approximation

- The Nystrom method we seek a matrix  $\mathbf{Q} \in \mathbb{R}^{n \times k}$  that satisfies  $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^* \mathbf{A}\|_F \leq \varepsilon$ , where  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is positive semi definite matrix, to form the rank- $k$  approximation

$$\mathbf{A} \simeq \mathbf{Q}\mathbf{Q}^* \mathbf{A}$$

# Nystrom Approximation

- The Nystrom method we seek a matrix  $\mathbf{Q} \in \mathbb{R}^{n \times k}$  that satisfies  $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^* \mathbf{A}\|_F \leq \varepsilon$ , where  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is positive semi definite matrix, to form the rank- $k$  approximation

$$\begin{aligned}\mathbf{A} &\simeq \mathbf{Q}\mathbf{Q}^* \mathbf{A} \\ &\simeq \mathbf{Q}(\mathbf{Q}^* \mathbf{A} \mathbf{Q}) \mathbf{Q}^*\end{aligned}$$

# Nystrom Approximation

- The Nystrom method we seek a matrix  $\mathbf{Q} \in \mathbb{R}^{n \times k}$  that satisfies  $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^* \mathbf{A}\|_F \leq \varepsilon$ , where  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is positive semi definite matrix, to form the rank- $k$  approximation

$$\begin{aligned}\mathbf{A} &\simeq \mathbf{Q}\mathbf{Q}^* \mathbf{A} \\ &\simeq \mathbf{Q} (\mathbf{Q}^* \mathbf{A} \mathbf{Q}) \mathbf{Q}^* \\ &= \mathbf{Q} (\mathbf{Q}^* \mathbf{A} \mathbf{Q}) (\mathbf{Q}^* \mathbf{A} \mathbf{Q})^\dagger (\mathbf{Q}^* \mathbf{A} \mathbf{Q}) \mathbf{Q}^*\end{aligned}$$

# Nystrom Approximation

- The Nystrom method we seek a matrix  $\mathbf{Q} \in \mathbb{R}^{n \times k}$  that satisfies  $\|\mathbf{A} - \mathbf{Q}\mathbf{Q}^* \mathbf{A}\|_F \leq \varepsilon$ , where  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is positive semi definite matrix, to form the rank- $k$  approximation

$$\begin{aligned}\mathbf{A} &\simeq \mathbf{Q}\mathbf{Q}^* \mathbf{A} \\ &\simeq \mathbf{Q} (\mathbf{Q}^* \mathbf{A} \mathbf{Q}) \mathbf{Q}^* \\ &= \mathbf{Q} (\mathbf{Q}^* \mathbf{A} \mathbf{Q}) (\mathbf{Q}^* \mathbf{A} \mathbf{Q})^\dagger (\mathbf{Q}^* \mathbf{A} \mathbf{Q}) \mathbf{Q}^* \\ &\simeq (\mathbf{A} \mathbf{Q}) (\mathbf{Q}^* \mathbf{A} \mathbf{Q})^\dagger (\mathbf{Q}^* \mathbf{A}).\end{aligned}$$

# Random Fourier Feature Approximation

- The RFF technique hinges on Bochner's theorem which characterises positive definite functions (namely kernels) and states that any positive definite functions can be represented as

$$k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y}) = \int_{\mathbb{C}^d} \exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle) \mu_k(d\boldsymbol{\omega})$$

where  $\mu_k$  is a positive finite measure on the frequencies of  $\boldsymbol{\omega}$ .

# Random Fourier Feature Approximation

- The RFF technique hinges on Bochners theorem which characterises positive definite functions (namely kernels) and states that any positive definite functions can be represented as

$$k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y}) = \int_{\mathbb{C}^d} \exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle) \mu_k(d\boldsymbol{\omega})$$

where  $\mu_k$  is a positive finite measure on the frequencies of  $\boldsymbol{\omega}$ .

- This integral can then be approximated via the following Monte Carlo estimate

$$k(\mathbf{x} - \mathbf{y}) = \int_{\mathbb{C}^d} \exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle) p(\boldsymbol{\omega}) d\boldsymbol{\omega}$$

# Random Fourier Feature Approximation

- The RFF technique hinges on Bochners theorem which characterises positive definite functions (namely kernels) and states that any positive definite functions can be represented as

$$k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y}) = \int_{\mathbb{C}^d} \exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle) \mu_k(d\boldsymbol{\omega})$$

where  $\mu_k$  is a positive finite measure on the frequencies of  $\boldsymbol{\omega}$ .

- This integral can then be approximated via the following Monte Carlo estimate

$$\begin{aligned} k(\mathbf{x} - \mathbf{y}) &= \int_{\mathbb{C}^d} \exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle) p(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &= \mathbb{E}_{\boldsymbol{\omega} \sim p(\cdot)} (\exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle)) \end{aligned}$$



# Random Fourier Feature Approximation

- The RFF technique hinges on Bochners theorem which characterises positive definite functions (namely kernels) and states that any positive definite functions can be represented as

$$k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y}) = \int_{\mathbb{C}^d} \exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle) \mu_k(d\boldsymbol{\omega})$$

where  $\mu_k$  is a positive finite measure on the frequencies of  $\boldsymbol{\omega}$ .

- This integral can then be approximated via the following Monte Carlo estimate

$$\begin{aligned} k(\mathbf{x} - \mathbf{y}) &= \int_{\mathbb{C}^d} \exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle) p(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &= \mathbb{E}_{\boldsymbol{\omega} \sim p(\cdot)} (\exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle)) \\ &\simeq \frac{1}{D} \sum_{j=1}^D \exp(i\langle \boldsymbol{\omega}_j, \mathbf{x} - \mathbf{y} \rangle) \end{aligned}$$

# Random Fourier Feature Approximation

- The RFF technique hinges on Bochners theorem which characterises positive definite functions (namely kernels) and states that any positive definite functions can be represented as

$$k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y}) = \int_{\mathbb{C}^d} \exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle) \mu_k(d\boldsymbol{\omega})$$

where  $\mu_k$  is a positive finite measure on the frequencies of  $\boldsymbol{\omega}$ .

- This integral can then be approximated via the following Monte Carlo estimate

$$\begin{aligned} k(\mathbf{x} - \mathbf{y}) &= \int_{\mathbb{C}^d} \exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle) p(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &= \mathbb{E}_{\boldsymbol{\omega} \sim p(\cdot)} (\exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle)) \\ &\simeq \frac{1}{D} \sum_{j=1}^D \exp(i\langle \boldsymbol{\omega}_j, \mathbf{x} - \mathbf{y} \rangle) \\ &= \sum_{j=1}^D \left( \frac{1}{\sqrt{D}} \exp(i\langle \boldsymbol{\omega}_j, \mathbf{x} \rangle) \right) \overline{\left( \frac{1}{\sqrt{D}} \exp(i\langle \boldsymbol{\omega}_j, \mathbf{y} \rangle) \right)} \end{aligned}$$

# Random Fourier Feature Approximation

- The RFF technique hinges on Bochners theorem which characterises positive definite functions (namely kernels) and states that any positive definite functions can be represented as

$$k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y}) = \int_{\mathbb{C}^d} \exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle) \mu_k(d\boldsymbol{\omega})$$

where  $\mu_k$  is a positive finite measure on the frequencies of  $\boldsymbol{\omega}$ .

- This integral can then be approximated via the following Monte Carlo estimate

$$\begin{aligned} k(\mathbf{x} - \mathbf{y}) &= \int_{\mathbb{C}^d} \exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle) p(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &= \mathbb{E}_{\boldsymbol{\omega} \sim p(\cdot)} (\exp(i\langle \boldsymbol{\omega}, \mathbf{x} - \mathbf{y} \rangle)) \\ &\simeq \frac{1}{D} \sum_{j=1}^D \exp(i\langle \boldsymbol{\omega}_j, \mathbf{x} - \mathbf{y} \rangle) \\ &= \sum_{j=1}^D \left( \frac{1}{\sqrt{D}} \exp(i\langle \boldsymbol{\omega}_j, \mathbf{x} \rangle) \right) \overline{\left( \frac{1}{\sqrt{D}} \exp(i\langle \boldsymbol{\omega}_j, \mathbf{y} \rangle) \right)} \\ &= \langle \boldsymbol{\varphi}(\mathbf{x}), \boldsymbol{\varphi}(\mathbf{y}) \rangle_{\mathbb{C}^D} \end{aligned}$$

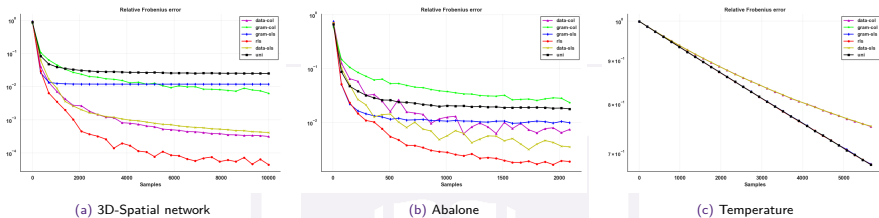


Figure: Comparison of Nystrom methods for various datasets.

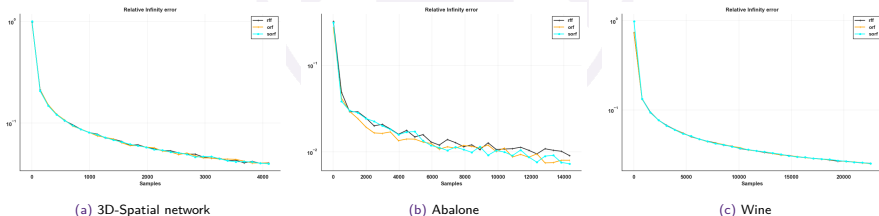


Figure: Comparison of RFF methods for various datasets.

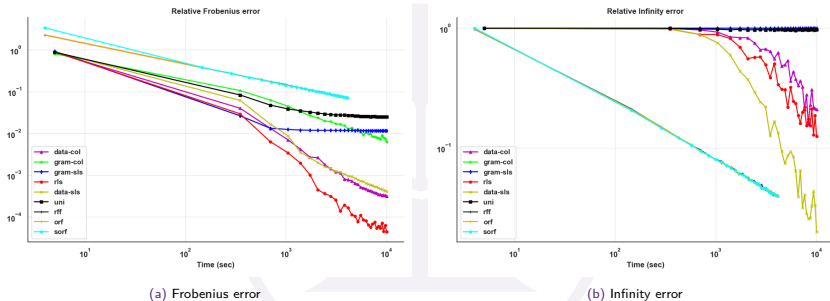


Figure: Comparison between Nystrom and RFF approximations for the 3D-Spatial network data.

# Moving Forward

- We saw that the Nystrom technique is better at producing approximations of the Gram matrix,  $\mathbf{K}_{XX}$ , with smaller *relative Frobenius errors* while the RFF technique is better at producing approximations with smaller *relative infinity errors*. Which is better for GP prediction?
- Recall, the other bottle neck in the GPR algorithm was the Cholesky decomposition. We can employ faster linear system solvers, namely the *Conjugate Gradient* (CG) and *Minimum Residual* (MINRES) method.