Modern Gaussian Processes: Scalable Inference and Novel Applications

(Part II-a) Model Approximations

Edwin V. Bonilla and **Maurizio Filippone**

CSIRO's Data61, Sydney, Australia and EURECOM, Sophia Antipolis, France

July 14th, 2019

Outline

Random Feature Expansions

Bochner's theorem

- Continuous shift-invariant covariance function

$$k(\mathbf{x}_i - \mathbf{x}_j|\boldsymbol{\theta}) = \sigma^2 \int p(\boldsymbol{\omega}|\boldsymbol{\theta}) \exp\left(\iota(\mathbf{x}_i - \mathbf{x}_j)^\top \boldsymbol{\omega}\right) d\boldsymbol{\omega}$$

Bochner's theorem

- Continuous shift-invariant covariance function

$$k(\mathbf{x}_i - \mathbf{x}_j|\boldsymbol{\theta}) = \sigma^2 \int p(\boldsymbol{\omega}|\boldsymbol{\theta}) \exp\left(\iota(\mathbf{x}_i - \mathbf{x}_j)^\top \boldsymbol{\omega}\right) d\boldsymbol{\omega}$$

- Monte Carlo estimate

$$k(\mathbf{x}_i - \mathbf{x}_j|\boldsymbol{\theta}) \approx \frac{\sigma^2}{N_{\mathrm{RF}}} \sum_{r=1}^{N_{\mathrm{RF}}} \mathbf{z}(\mathbf{x}_i|\tilde{\boldsymbol{\omega}}_r)^\top \mathbf{z}(\mathbf{x}_j|\tilde{\boldsymbol{\omega}}_r)$$

with

$$\tilde{\boldsymbol{\omega}}_r \sim p(\boldsymbol{\omega}|\boldsymbol{\theta})$$

$$\mathbf{z}(\mathbf{x}|\boldsymbol{\omega}) = [\cos(\mathbf{x}^\top \boldsymbol{\omega}), \sin(\mathbf{x}^\top \boldsymbol{\omega})]^\top$$

Rahimi and Recht, *NIPS*, 2008 - Lázaro-Gredilla et al., *JMLR*, 2010

GPs with Random Fourier Features

- Define

$$\boldsymbol{\Phi} = \sqrt{\frac{\sigma^2}{N_{\mathrm{RF}}}} \left[\cos\left(\mathbf{X}\boldsymbol{\Omega}\right), \sin\left(\mathbf{X}\boldsymbol{\Omega}\right)\right]$$

and

$$\mathbf{f} = \boldsymbol{\Phi}\mathbf{w}$$

- GPs become Bayesian linear models with

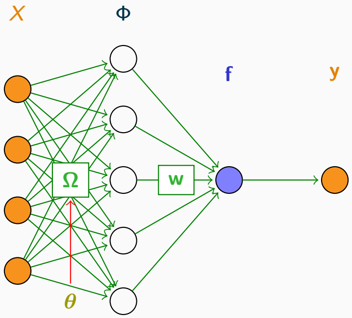
$$p\left(\mathbf{w}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$$

- Low-rank approximation of $\mathbf{K}$

$$\mathrm{cov}(\mathbf{f}) = \mathbb{E}[\boldsymbol{\Phi}\mathbf{w}\mathbf{w}^\top\boldsymbol{\Phi}^\top] = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top \approx \mathbf{K}$$

GPs with Random Features become Bayesian Linear Models

- Neural Network-like diagram



Low-Rank Approximations

Low-Rank Approximations

- Marginal likelihood GP regression:

$$-\frac{1}{2}\log|\mathbf{K_y}| - \frac{1}{2}\mathbf{y}^\top \mathbf{K_y}^{-1}\mathbf{y} + \mathrm{const.}$$

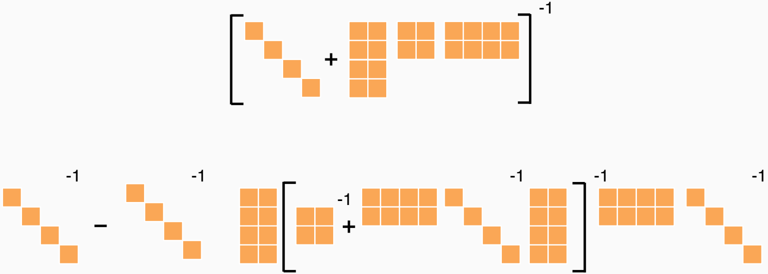
- Most GP approximations aim to form a low-rank approximation to the covariance matrix

$$\mathbf{K_y} = \mathbf{K} + \sigma^2\mathbf{I} \approx \mathbf{UCV} + \sigma^2\mathbf{I}$$

Low-Rank Approximations

- Woodbury identity for the inverse

$$(\mathbf{A}+\mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1}+\mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$$

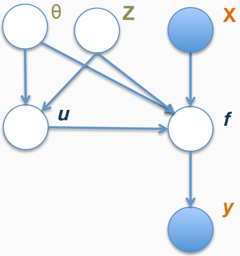


- Similar for the log-determinant
- This reduces complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(M^3)+\mathcal{O}(NM^2)$ with $M \ll N$

Sparse GPs with Nyström Approximation

- Introduce M pseudo-inputs collected in Z ...
- ... and corresponding inducing variables $\mathbf{u}$
- Nyström approximation

$$\mathbf{K} \approx \mathbf{K}_{\mathbf{XZ}}\mathbf{K}_{\mathbf{ZZ}}^{-1}\mathbf{K}_{\mathbf{ZX}}$$

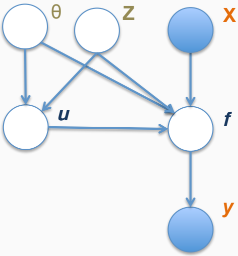


Snelson and Ghahramani, *NIPS*, 2005 – Quinonero-Candela and Rasmussen, *JMLR*, 2005

Sparse GPs with Nyström Approximation

- Introduce M pseudo-inputs collected in Z ...
- ... and corresponding inducing variables $\mathbf{u}$
- Nyström approximations with diagonal correction

$$\mathbf{K} \approx \mathrm{diag}(\mathbf{K} - \mathbf{K}_{\mathbf{XZ}}\mathbf{K}_{\mathbf{ZZ}}^{-1}\mathbf{K}_{\mathbf{ZX}}) + \mathbf{K}_{\mathbf{XZ}}\mathbf{K}_{\mathbf{ZZ}}^{-1}\mathbf{K}_{\mathbf{ZX}}$$



Snelson and Ghahramani, *NIPS*, 2005 – Quinonero-Candela and Rasmussen, *JMLR*, 2005

Structured Inputs

- When inputs lie on a regular 1D grid and $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \kappa(\mathbf{x}_i - \mathbf{x}_j)$
- $\mathbf{K}$ is Toeplitz

$$\mathbf{K} = \begin{pmatrix} a & b & c & d \\ b & a & b & c \\ c & b & a & b \\ d & c & b & a \end{pmatrix}$$

- Solving $\mathbf{K}$ exactly costs $\mathcal{O}(N \log N)$ time!

Saatçi, *Ph.D. Thesis*, 2011

Structured Inputs

- When inputs lie on a regular grid and $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \kappa(\mathbf{x}_i - \mathbf{x}_j)$
- $\mathbf{K}$ can be decomposed is Toeplitz

$$\mathbf{K} = \mathbf{K}_1 \otimes \ldots \otimes \mathbf{K}_d$$

where $\mathbf{K}_p$ has entries $\kappa(\mathbf{x}_i^{(p)}, \mathbf{x}_j^{(p)})$

- Algebraic operations for $\mathbf{K}$ are based on faster ones for each factor $\mathbf{K}_p$ in the Kronecker product

Saatçi, *Ph.D. Thesis*, 2011

Structured Inducing Points

- Consider a sparse GP:

$$\mathbf{K} \approx \mathbf{K}_{\mathbf{XZ}} \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbf{K}_{\mathbf{ZX}}$$

- $\mathbf{Z}$ on a grid makes the inverse fast (Toeplitz)!
- Can afford $M \gg N$
- Still expensive to deal with $\mathbf{K}_{\mathbf{ZX}}$ ... $\mathcal{O}(NM^2)$

Structured Inducing Points

- Consider a sparse GP:

$$\mathbf{K} \approx \mathbf{K}_{\mathbf{XZ}} \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbf{K}_{\mathbf{ZX}}$$

- Kernel Interpolation (KISS-GP)

$$\mathbf{K}_{\mathbf{XZ}} \approx \mathbf{W} \mathbf{K}_{\mathbf{ZZ}}$$

with $\mathbf{W}$ a sparse "interpolation" matrix, so that

$$\mathbf{K} \approx \mathbf{K}_{\mathbf{XZ}} \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbf{K}_{\mathbf{ZX}} \approx \mathbf{W} \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbf{W}^{\top}$$

- All products/inverses are fast even if $M \gg N$!

Wilson, *NIPS*, 2015