


# Gaussian Processes

**Mark van der Wilk**

Department of Computing  
Imperial College London

@markvanderwilk  
m.vdwilk@imperial.ac.uk

January 27, 2023

# Recap: Model

- ▶ Specify prior on weights  $p(\mathbf{w})$
- ▶ Defines distribution on functions through  $f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{w}$
- ▶ Observe data through likelihood
$$p(\mathbf{y}|f(X)) = \prod_{n=1}^N \mathcal{N}(y_n; f(\mathbf{x}_n), \sigma^2)$$

## Recap: Inference

- ▶ Find posterior on weights  $p(\mathbf{w}|\mathbf{y})$
- ▶ Combine this with  $p(\mathbf{y}^*|\mathbf{w})$  to find  $p(\mathbf{y}^*|\mathbf{y})$ :

$$p(\mathbf{y}_*|\mathbf{y}) = \mathcal{N}\left(\boldsymbol{\theta}; \quad \boldsymbol{\phi}(\mathbf{x}_*)^\top [\mathbf{I}_M + \sigma^{-2}\Phi(\mathbf{X})^\top \Phi(\mathbf{X})]^{-1} \sigma^{-2} \Phi(\mathbf{X})^\top \mathbf{y} \right. \\ \left. \boldsymbol{\phi}(\mathbf{x}_*)^\top [\mathbf{I}_M + \sigma^{-2}\Phi(\mathbf{X})^\top \Phi(\mathbf{X})]^{-1} \boldsymbol{\phi}(\mathbf{x}_*) + \sigma^2 \mathbf{I}_N \right)$$

- ▶ Apply Woodbury to go from  $O(NM^2 + M^3) \rightarrow O(N^3)$ :

$$p(y_*|\mathbf{y}) = \mathcal{N}\left(y_*; \quad \boldsymbol{\phi}(\mathbf{x}_*)^\top \Phi(\mathbf{X})^\top [\Phi(\mathbf{X})\Phi(\mathbf{X})^\top + \sigma^2 \mathbf{I}_N]^{-1} \mathbf{y}, \right. \\ \left. \boldsymbol{\phi}(\mathbf{x}_*)^\top \boldsymbol{\phi}(\mathbf{x}_*) + \sigma^2 \right. \\ \left. - \boldsymbol{\phi}(\mathbf{x}_*)^\top \Phi(\mathbf{X})^\top [\Phi(\mathbf{X})\Phi(\mathbf{X})^\top + \sigma^2 \mathbf{I}_N]^{-1} \Phi(\mathbf{X}) \boldsymbol{\phi}(\mathbf{x}_*) \right)$$

- ▶ Apply Kernel trick  $\boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}')$

$$p(y_*|\mathbf{y}) = \mathcal{N}\left(y_*; \quad k(\mathbf{x}_*, \mathbf{X}) [k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}_N]^{-1} \mathbf{y}, \right.$$

# Today

- ▶ Develop interpretation of the maths we got from Woodbury
- ▶ This is a way of specifying distributions on functions
- ▶ But without parameters!

# How do we get rid of parameters?

Model:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; 0, \mathbf{I}_M) \quad (1)$$

$$f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{w} \quad (2)$$

$$p(\mathbf{y}|f(X)) = \prod_{n=1}^N \mathcal{N}(y_n; f(\mathbf{x}_n), \sigma^2) \quad (3)$$

Observation: Likelihood only depends on function value

⇒ Can we ignore the distribution on weights, and work directly with function values?

All we are **really** interested in, is:

- ▶ Predicting data  $p(\mathbf{y}^*|\mathbf{y})$
- ▶ Predicting function values  $p(f(X^*))|\mathbf{y})$

# Distribution on Function Values

Let's start by analysing distribution on function values  $p(f(X))$  (board)

- ▶ Function values are linear transformation of Gaussian RV  
For arbitrary  $N$  inputs arranged in a matrix  $X \in \mathbb{R}^{N \times D}$ :

$$f(X) = \Phi(X)\mathbf{w} \quad (4)$$

- ▶ As usual  $\implies$  Gaussian distributed, and can find mean+var

$$\mathbb{E}_{\mathbf{w}}[f(X)] = \mathbb{E}_{\mathbf{w}}[\Phi(X)\mathbf{w}] = 0 \quad (5)$$

$$\mathbb{V}_{\mathbf{w}}[f(X)] = \mathbb{E}_{\mathbf{w}}[\Phi(X)\mathbf{w}\mathbf{w}^T\Phi(X)^T] = \Phi(X)\Phi(X)^T \quad (6)$$

$$\implies p(f(X)) = \mathcal{N}(f(X); 0, \Phi(X)\Phi(X)^T) \quad (7)$$

- ▶ All function values are correlated
- ▶ Kernel trick applies!  $[\Phi(X)\Phi(X)^T]_{ij} = \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)$

# Predicting

Let's focus on  $p(f(X^*)|\mathbf{y})$ .

$$p(f(X^*)|\mathbf{y}) \stackrel{\text{AT}}{=} \frac{\int p(\mathbf{y}, f(X), f(X^*)) \, df(X)}{p(\mathbf{y})}$$
$$\stackrel{\text{MA}}{=} \frac{\int p(\mathbf{y}|f(X)) p(f(X), f(X^*)) \, df(X)}{p(\mathbf{y})}$$

$$p(f(X), f(X^*)) = \mathcal{N}\left(\begin{bmatrix} f(X) \\ f(X^*) \end{bmatrix}; 0, \begin{bmatrix} \Phi(X)\Phi(X)^\top & \Phi(X)\Phi(X^*)^\top \\ \Phi(X^*)\Phi(X)^\top & \Phi(X^*)\Phi(X^*)^\top \end{bmatrix}\right)$$

$$p(f(X), f(X^*)) = \mathcal{N}\left(\begin{bmatrix} f(X) \\ f(X^*) \end{bmatrix}; 0, \begin{bmatrix} k(X, X) & k(X, X^*) \\ k(X^*, X) & k(X^*, X^*) \end{bmatrix}\right)$$

Easiest way: Find joint, Gaussian conditioning (board)

$$\mathcal{N}\left(\begin{bmatrix} f(X^*) \\ \mathbf{y} \end{bmatrix}; 0, \begin{bmatrix} k(X^*, X^*) & k(X^*, X) \\ k(X, X^*) & k(X, X) + \sigma^2 \mathbf{I}_N \end{bmatrix}\right) \quad (8)$$

# What is a Gaussian Process?

- ▶ Same as what we get from BLR + Woodbury + kernel trick!
- ▶ No need for parameters, only kernel  $k$ !

Who needs parameters?

$\implies$  Can answer any prediction question  
using only distribution on function *values*.



# What is a Gaussian Process?

A (possibly infinite) collection of Random Variables such that each finite collection has a Gaussian distribution.

## Properties

- ▶ I will index this collection with  $x$ .
- ▶ For this to be a valid collection of RVs, sum rule must hold:

$$p(f(x_1), f(x_2)) = \int p(f(x_1), f(x_2), f(x_3)) \mathrm{d}f(x_3) \quad (9)$$

# Specifying Gaussian Processes

Can specify the function value densities  $p(f(X))$  using:

- ▶ Mean function  $\mu : \mathcal{X} \rightarrow \mathbb{R}$
- ▶ Covariance function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$

$$p(f(X)) = \mathcal{N}(f(X); \mu(X), k(X, X))$$

$$\mu(X) \in \mathbb{R}^N \qquad k(X, X) \in \mathbb{R}^{D \times D}$$

$$[\mu(X)]_i = \mu(\mathbf{x}_i) \qquad [k(X, X)]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

Covariance function  $k(\cdot, \cdot)$  must be a positive definite function.  
I.e.  $k(X, X)$  is PSD for any choice of  $X$ .

# BLR Specifies a GP

- ▶ BLR specifies a density of a collection of RVs
- ▶ Collection of random variables is **function values** at all locations

$$p(f(X)) = \mathcal{N}(f(X); 0, \Phi(X)\Phi(X)^\top) \quad (10)$$

$$= \mathcal{N}(f(X); 0, k(X, X)) \quad (11)$$

- ▶  $\implies$  BLR specifies a GP and a kernel
- ▶ Directly specifying a kernel, also specifies a GP
- ▶ We viewed BLR as specifying a distribution on functions

# GPs as distributions on functions

Can we view a GP as a distribution on functions?

▶▶ Yes! Kolmogorov Extension Theorem (not examined).

# Conclusion

- ▶ Covariance functions / kernels specify GPs
- ▶ GPs specify distributions on function values directly
- ▶ To make predictions, we only need distributions on function values
- ▶ So who needs parameters?
- ▶ BLR specifies a GP

# Recommended reading

- ▶ Rasmussen and Williams (2006) §2.1 + §2.2

# References I

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*. MIT press, Cambridge, MA, USA.