University of Amsterdam
Informatics Institute

AMLAB
Amsterdam Machine Learning Lab

Machine Learning 1

Lecture 12.5 - Kernel Methods
Gaussian Processes - Bayesian Regression

Erik Bekkers

(Bishop 6.4.2, 6.4.3)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: KirillM | Getty Images

Regression with GP's

- We have observed $\{(\mathbf{x}_i, f_i)\}_{i=1}^N$ where we assume

$$f_i = f(\mathbf{x}_i) = y(\mathbf{x}_i) + \varepsilon \,, \qquad\qquad \varepsilon \sim \mathcal{N}(0,\beta^{-1})$$

$K(X, X)$

- Assume we have a GP for y(x), so for any

$$\mathbf{y} = \begin{bmatrix} y(\mathbf{x}_1) \\ \vdots \\ y(\mathbf{x}_N) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \dots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \dots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix} \right)$$

- Then $f(\,\cdot\,)$ is also a GP since $\mathbf{f} = \mathbf{y} + \boldsymbol{\varepsilon}$, and the sum of two independent random variables is also Gaussian distributed

non-parametric

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, K(\mathbf{X}, \mathbf{X}) + \beta^{-1}\boldsymbol{I})$$

vs

parametric

$\underline{f} \sim N(\Phi w, \beta^{-1} I)$

$\underline{w} \sim p(\underline{w})$

equivalent kernel

Predictions with GP's

X

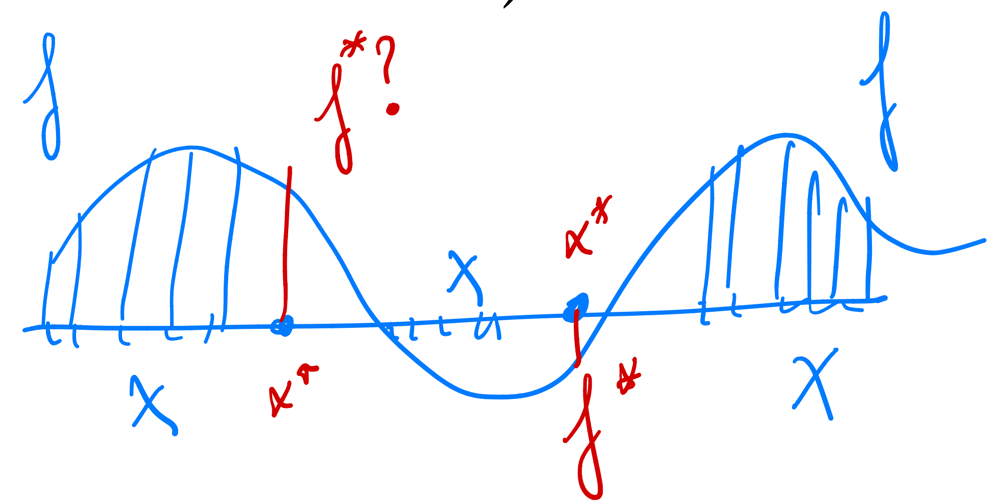
- The joint distribution of test points $\mathbf{f}^*$ (at $\mathbf{X}^*$) and $\mathbf{f}$ (train points), according to our GP, is given by

$$\begin{pmatrix} k(x_1, x_1^*), & k(x_1, x_2^*), & \dots, & k(x_1, x_{N^*}^*) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_N, x_1^*), & k(x_N, x_2^*), & \dots, & k(x_N, x_{N^*}^*) \end{pmatrix}$$

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{X},\mathbf{X}) + \beta^{-1}\mathbf{I} & \mathbf{K}(\mathbf{X},\mathbf{X}^*) \\ \mathbf{K}(\mathbf{X}^*,\mathbf{X}) & \mathbf{K}(\mathbf{X}^*,\mathbf{X}^*) + \beta^{-1}\mathbf{I} \end{bmatrix}\right)$$

Gaussian conditioning property!

- Then



$$p(\mathbf{f}^* \mid \mathbf{X}^*, \mathbf{X}, \mathbf{f}) = \mathcal{N}\left(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*\right)$$

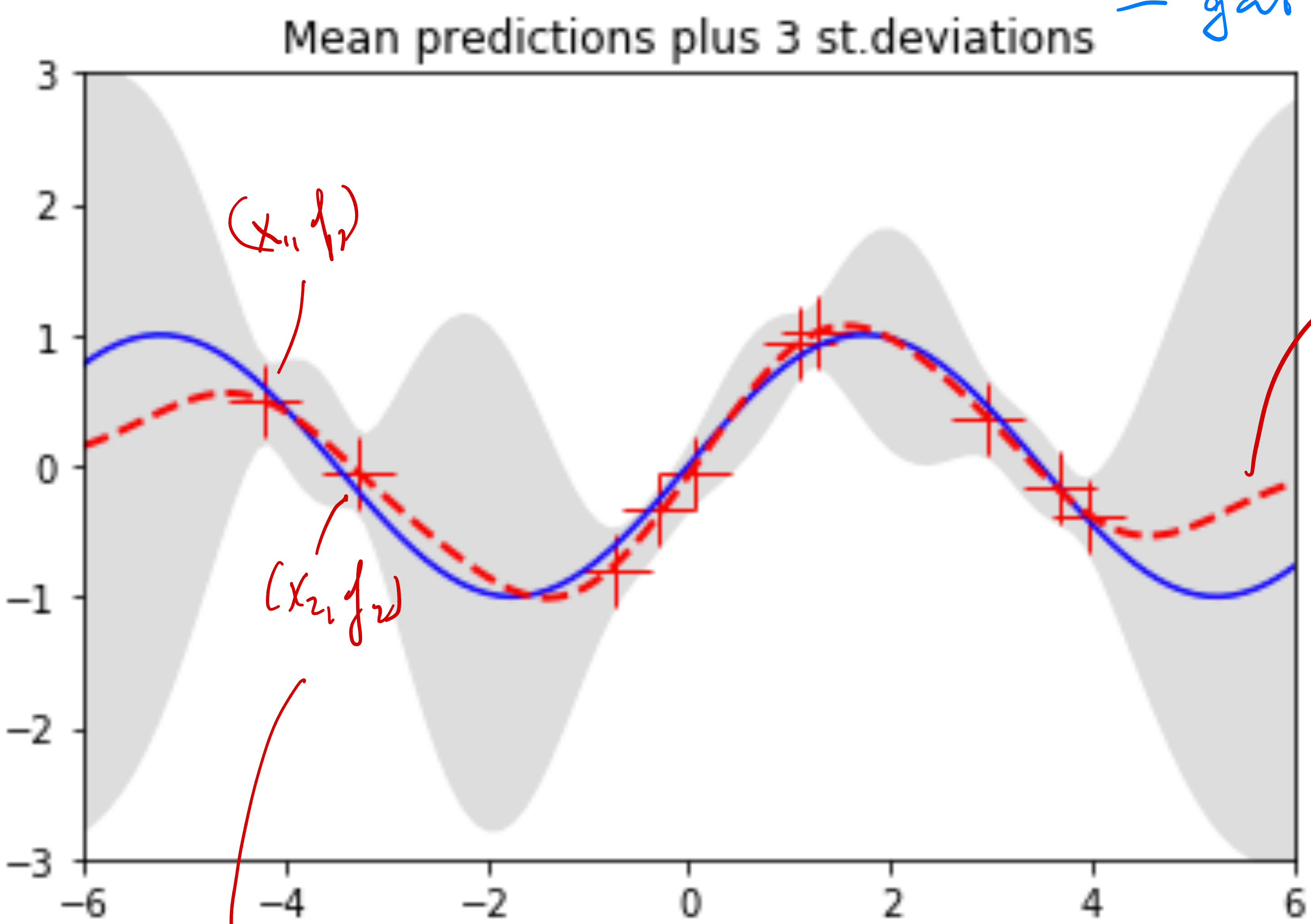
with

$$\boldsymbol{\mu}^* = \mathbf{K}(\mathbf{X}^*,\mathbf{X})\left(\mathbf{K}(\mathbf{X},\mathbf{X}) + \beta^{-1}\mathbf{I}\right)^{-1}\mathbf{f}$$

$$\boldsymbol{\Sigma}^* = \mathbf{K}(\mathbf{X}^*,\mathbf{X}^*) + \beta^{-1}\mathbf{I} - \mathbf{K}(\mathbf{X}^*,\mathbf{X})\left(\mathbf{K}(\mathbf{X},\mathbf{X}) + \beta^{-1}\mathbf{I}\right)^{-1}\mathbf{K}(\mathbf{X},\mathbf{X}^*)$$

Predictions with GP's

active learning!
- identify uncertain regions
- gather more data



(x_1, f_1)

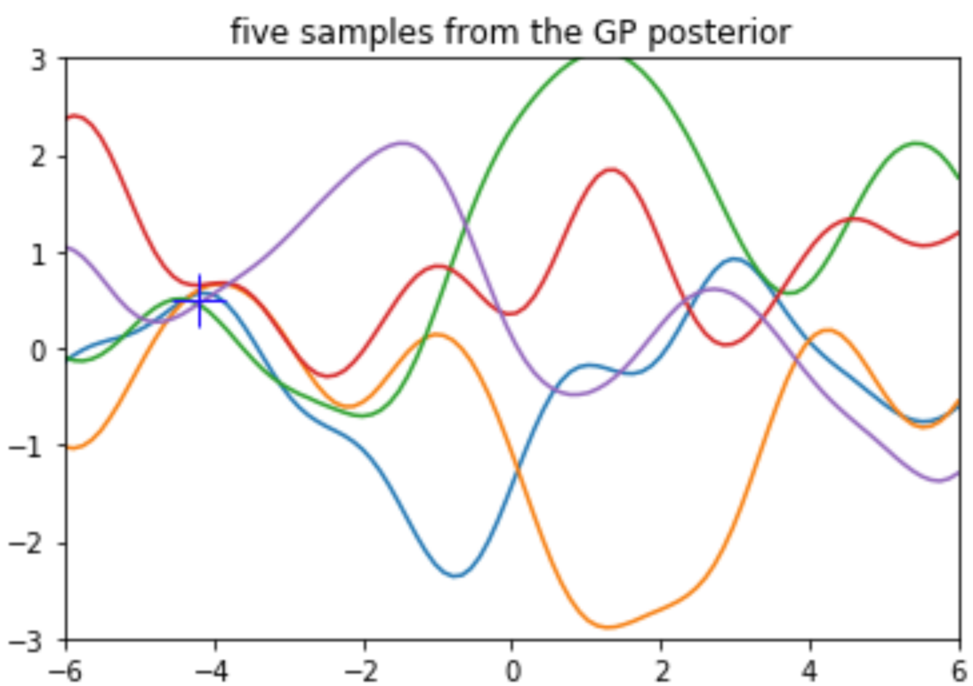
(x_2, f_2)

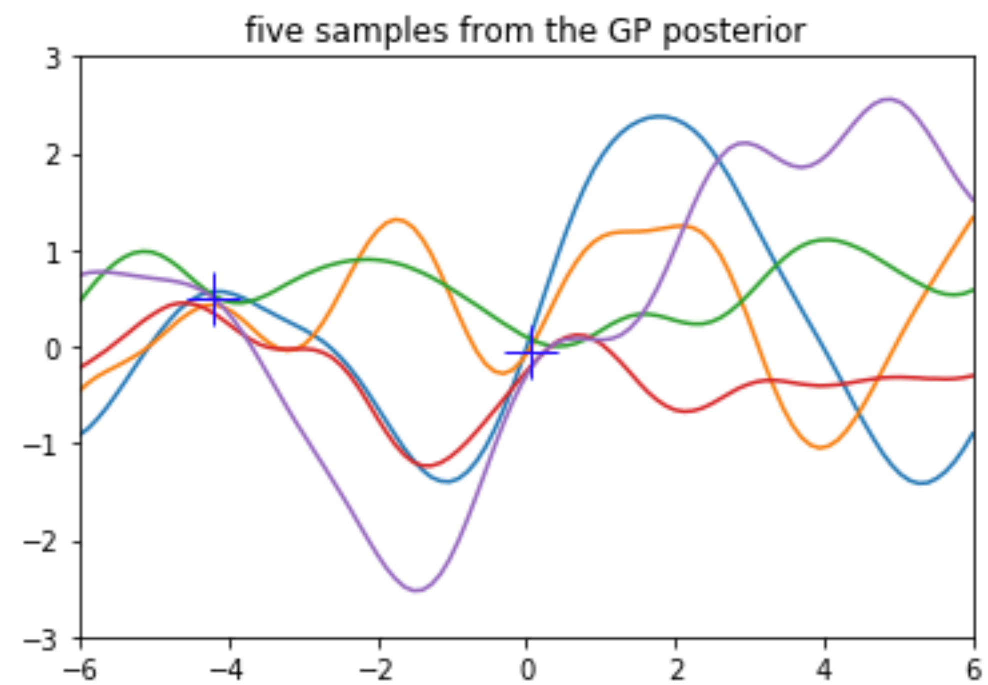
X, f

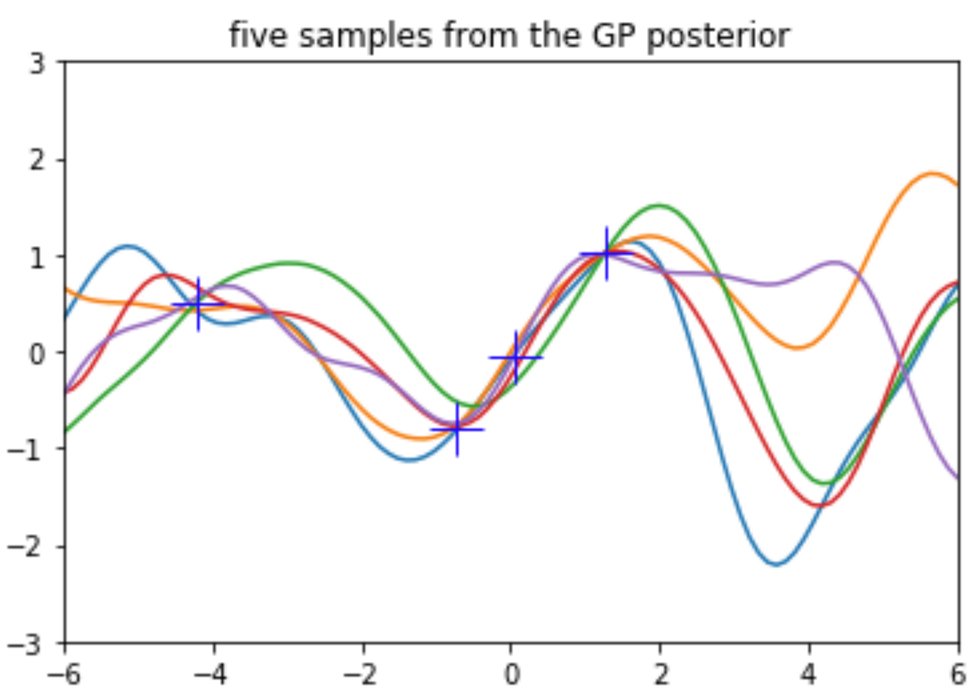
x^*, f^*

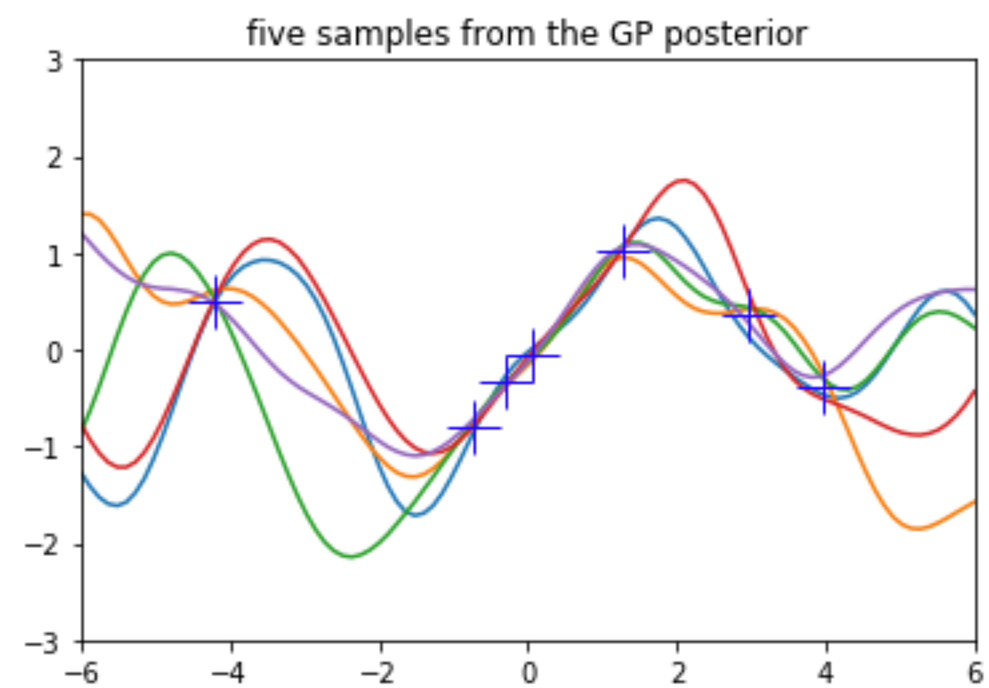
large error

Drawing functions from GP posterior









How to choose kernel parameters?

- The kernel parameters $\theta_0, \theta_1, \theta_2, \theta_3$ are hyperparameters

- Simplest approach: take training observations, for which we know

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, C_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{X})) = \frac{1}{(2\pi)^{N/2} |C_{\boldsymbol{\theta}}|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{f}^T \mathbf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{f}\right)$$

with $C_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{X}) = K_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{X}) + \beta^{-1}\boldsymbol{I}$

- Make a maximum likelihood estimate

$$\max_{\boldsymbol{\theta}} \ln p(\mathbf{f} \,|\, \mathbf{X}, \boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} -\frac{1}{2}\ln|\mathbf{C}_{\boldsymbol{\theta}}| - \frac{1}{2}\mathbf{f}^T\mathbf{C}_{\boldsymbol{\theta}}^{-1}\mathbf{f} - \frac{N}{2}\ln 2\pi$$

- Solve numerically for $\boldsymbol{\theta}$