An Introduction to Gaussian Process regression

Martin Eigel, David Sommer
Leibniz MMS Summer School
Schloss Dagstuhl 2021

August 21, 2021

What are Gaussian Processes?

Definition [RW06]

A Gaussian Process is a collection of Random Variables, any *finite number* of which have a joint Gaussian distribution.

Formal definition:

Let $\mathcal{X}$ be some index set and $f(\mathbf{x})$ a random variable for every $\mathbf{x} \in \mathcal{X}$. Then $(f(\mathbf{x}))_{\mathbf{x}\in\mathcal{X}}$ is a Gaussian process if and only if for every $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ we have

$$\begin{pmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_n) \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

for some $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$.

Mean and Covariance

A Gaussian process is completely specified by its mean and covariance function:

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

Then we have

$$\begin{pmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_n) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_n) \end{pmatrix}, \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \dots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \dots & k(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} \right).$$

The notation for a GP with mean $m(\mathbf{x})$ and covariance $k(\mathbf{x}, \mathbf{x}')$ is

$$f \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')).$$

Samples from a GP prior

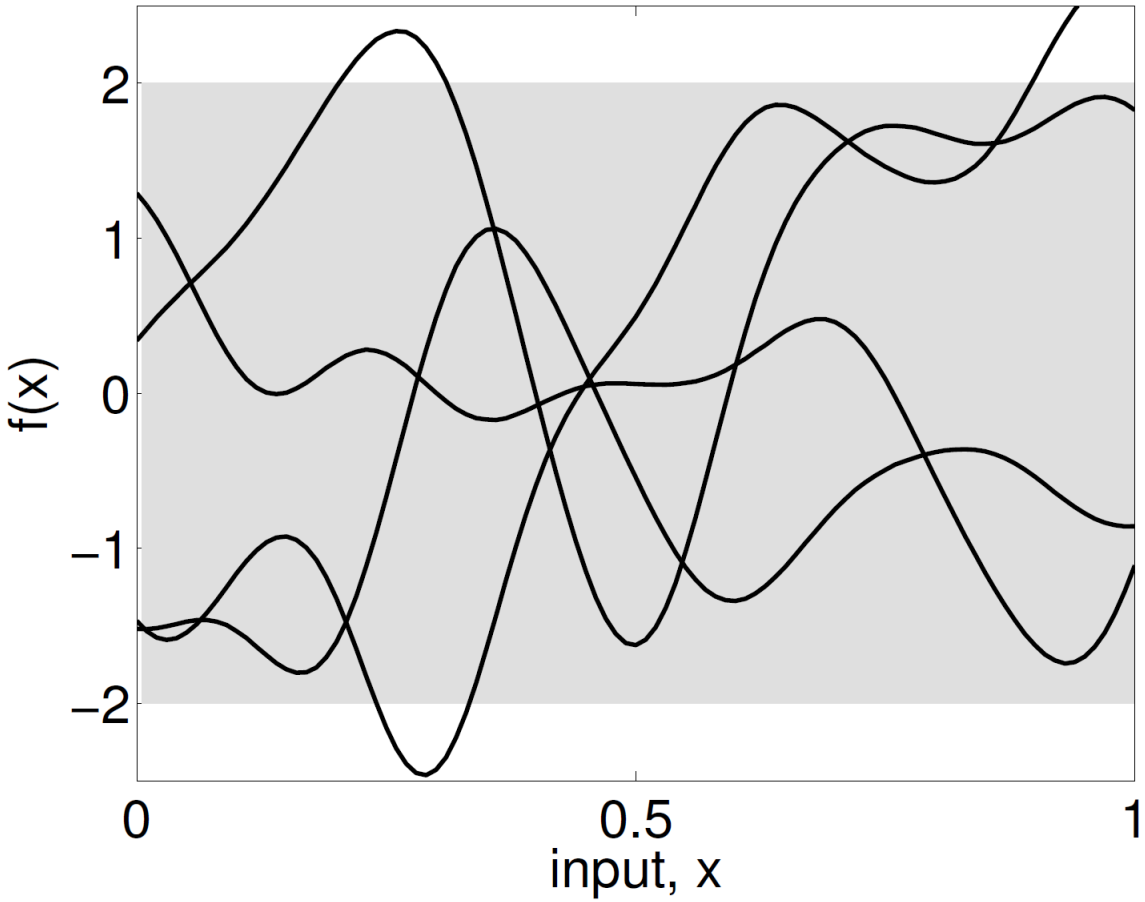

Figure: Samples from a squared exponential GP prior. Shaded region is twice the standard deviation at each input. Taken from [RW06].

GP Regression

Consider

$$y = f(\mathbf{x}) + \varepsilon,$$

where

- $y \in \mathbb{R}$
- $\mathbf{x} \in \mathbb{R}^d$
- $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$ is the assumed measurement noise.

Given: training data $(y_i, \mathbf{x}_i)_{i=1}^N$

Wanted: Prediction $\mathbf{f}^* = (f(\mathbf{x}_1^*), \ldots, f(\mathbf{x}_M^*))^T$ at a separate set of input points $(\mathbf{x}_i^*)_{i=1}^M$.

The GP regression method:
Assume that $f \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$ is a GP on the index set $\mathbb{R}^d$ with mean $\mathbf{0}$ and some covariance function k.

GP Regression

Collect matrices of training inputs $X = (\mathbf{x}_1 \dots \mathbf{x}_N) \in \mathbb{R}^{d \times N}$ and test inputs $X^* \in \mathbb{R}^{d \times M}$.
Collect target vector $\mathbf{y} = (y_1, \dots, y_N)^T$.

Define the corresponding covariance matrices $K(X, X), K(X^*, X), K(X, X^*), K(X^*, X^*)$ accordingly.

Then we have

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}^* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K(X, X) + \sigma_n^2 I & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{pmatrix}\right).$$

By conditioning on $\mathbf{y}$ we get the predictive distribution

$$\mathbf{f}^* \sim \mathcal{N}(\bar{\mathbf{f}}^*, \mathrm{cov}(\mathbf{f}^*))$$

with

$$\begin{aligned} \bar{\mathbf{f}}^* &= K(X^*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y}, \\ \mathrm{cov}(\mathbf{f}^*) &= K(X^*, X^*) - K(X^*, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, X^*). \end{aligned}$$

GP Regression

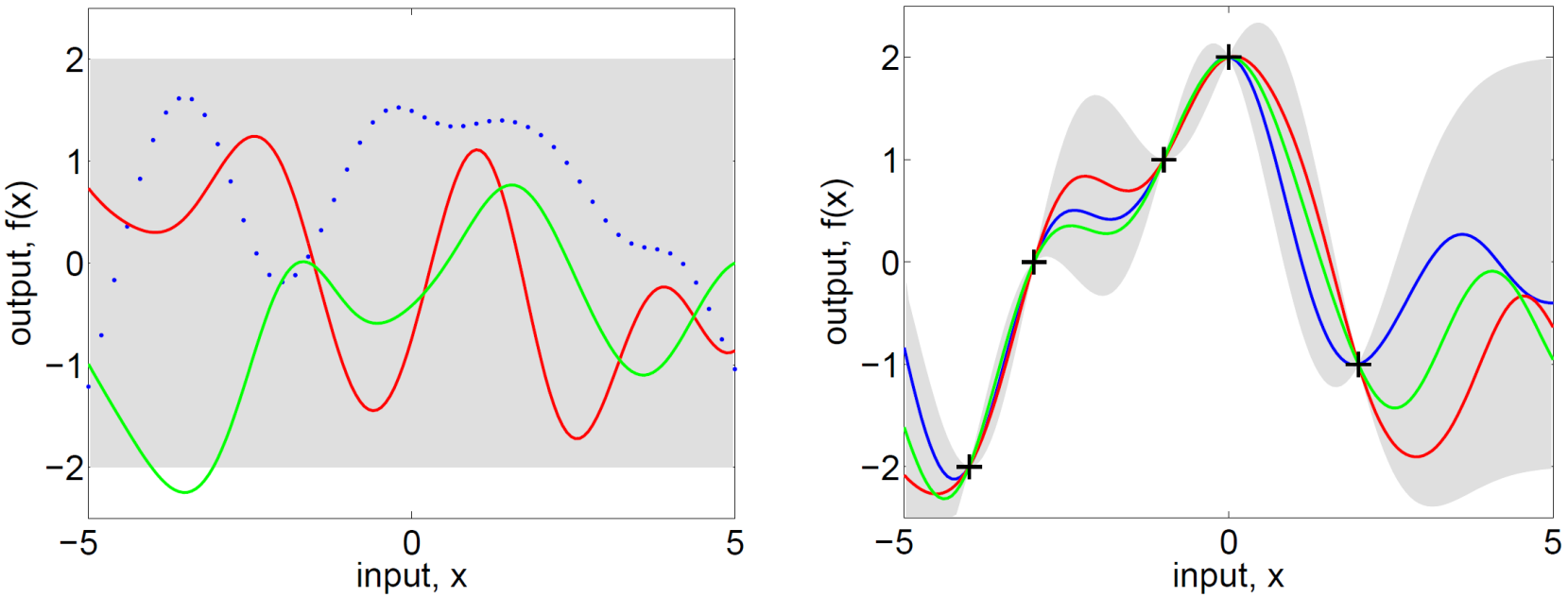

Figure: Samples from a prior (left) and a posterior (right). Taken from [RW06].

Kernel Choice

A popular choice for k is the RBF kernel

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-(\mathbf{x} - \mathbf{x}')\Lambda^{-1}(\mathbf{x} - \mathbf{x}')\right)$$

with

- signal variance σ_f^2
- lengthscales $\Lambda = \text{diag}(l_1^2, \ldots, l_N^2)$

Choice of the hyperparameters σ_f, σ_n, Λ?

Usually by *evidence maximization*, i.e. ML-II (*Type II Maximum Likelihood*):
For a kernel $k_{\boldsymbol{\theta}}$ with hyperparameters $\boldsymbol{\theta}$ choose

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathbf{y}|\boldsymbol{\theta}).$$

This determination of the hyperparameters is the actual training of the GP.

Hyperparameter optimization

$\log p(\mathbf{y}|X, \boldsymbol{\theta})$ is the **log likelihood function** for the Gaussian process. It is given by

$$\ln p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{1}{2}\ln|C_N| - \frac{1}{2}\mathbf{y}^\intercal C_N^{-1}\mathbf{y} - \frac{N}{2}\ln(2\pi),$$

where

$$C_N = K(X, X) + \sigma_n^2 I \quad \in \mathbb{R}^{N\times N}$$

is the **covariance matrix** and its inverse C_N^{-1} is called the **precision matrix**.
We optimize it by gradient ascend, i.e. successively updating

$$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} + \alpha\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{y}|\boldsymbol{\theta})$$

with suitable (adaptive) step size α.
The partial derivatives we have to compute are given by

$$\frac{\partial}{\partial\theta_i}\ln p(\mathbf{y}|\theta) = -\frac{1}{2}\mathsf{Tr}\left(C_N^{-1}\frac{\partial C_N}{\partial\theta_i}\right) + \frac{1}{2}\mathbf{y}^\intercal C_N^{-1}\frac{\partial C_N}{\partial\theta_i}C_N^{-1}\mathbf{y}.$$

[RW06] CE. Rasmussen and CKI. Williams, *Gaussian processes for machine learning*, Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, USA, January 2006.