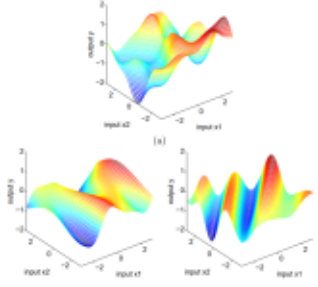
Introduction to Machine Learning

Covariance Functions for GPs



Learning goals

- Covariance functions encode key assumptions about the GP
- Know common covariance functions like squared exponential and Matérn

COVARIANCE FUNCTION OF A GP

The marginalization property of the Gaussian process implies that for any finite set of input values, the corresponding vector of function values is Gaussian:

$$\boldsymbol{f} = \left[f\left(\mathbf{x}^{(1)}\right), ..., f\left(\mathbf{x}^{(n)}\right) \right] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}),$$

- The covariance matrix $\boldsymbol{K}$ is constructed based on the chosen inputs $\left\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\right\}$.
- Entry $\boldsymbol{K}_{ij}$ is computed by $k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)$.
- Technically, for **every** choice of inputs $\left\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\right\}$, $\boldsymbol{K}$ needs to be positive semi-definite in order to be a valid covariance matrix.
- A function $k(.,.)$ satisfying this property is called **positive definite**.

COVARIANCE FUNCTION OF A GP / 2

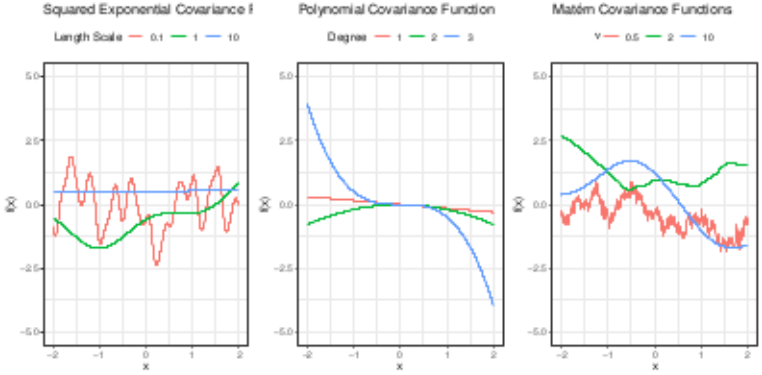
- Recall, the purpose of the covariance function is to control to which degree the following is fulfilled:

  If two points $\mathbf{x}^{(i)}, \mathbf{x}^{(j)}$ are close in $\mathcal{X}$-space, their function values $f(\mathbf{x}^{(i)}), f(\mathbf{x}^{(j)})$ should be close (**correlated**!) in $\mathcal{Y}$-space.

- Closeness of two points $\mathbf{x}^{(i)}, \mathbf{x}^{(j)}$ in input space $\mathcal{X}$ is measured in terms of $\boldsymbol{d} = \mathbf{x}^{(i)} - \mathbf{x}^{(j)}$:

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = k(\boldsymbol{d})$$

COMMONLY USED COVARIANCE FUNCTIONS / 2



- Random functions drawn from Gaussian processes with a Squared Exponential Kernel (left), Polynomial Kernel (middle), and a Matérn Kernel (right, $\ell = 1$).
- The length-scale hyperparameter determines the "wiggliness" of the function.
- For Matérn, the ν parameter determines how differentiable the process is.

CHARACTERISTIC LENGTH-SCALE / 2

For $p \geq 2$ dimensions, the squared exponential can be parameterized:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^\top \boldsymbol{M} (\mathbf{x} - \mathbf{x}')\right)$$

Possible choices for the matrix $\boldsymbol{M}$ include

$$\boldsymbol{M}_1 = \ell^{-2}\boldsymbol{I} \qquad \boldsymbol{M}_2 = \text{diag}(\boldsymbol{\ell})^{-2} \qquad \boldsymbol{M}_3 = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top + \text{diag}(\boldsymbol{\ell})^{-2}$$

where $\boldsymbol{\ell}$ is a p-vector of positive values and $\boldsymbol{\Gamma}$ is a $p \times k$ matrix.

The 2nd (and most important) case can also be written as

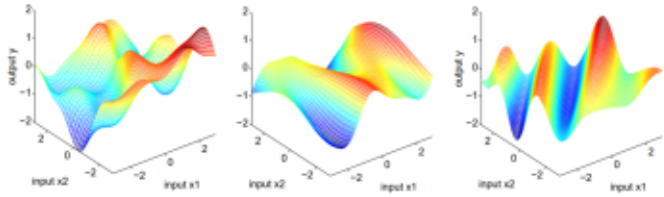
$$k(\mathbf{d}) = \exp\left(-\frac{1}{2}\sum_{j=1}^{p}\frac{d_j^2}{\ell_j^2}\right)$$

CHARACTERISTIC LENGTH-SCALE / 3

What is the benefit of having an individual hyperparameter ℓ_i for each dimension?

- The $\ell_1, \ldots, \ell_p$ hyperparameters play the role of **characteristic length-scales**.
- Loosely speaking, ℓ_i describes how far you need to move along axis i in input space for the function values to be uncorrelated.
- Such a covariance function implements **automatic relevance determination** (ARD), since the inverse of the length-scale ℓ_i determines the relevancy of input feature i to the regression.
- If ℓ_i is very large, the covariance will become almost independent of that input, effectively removing it from inference.
- If the features are on different scales, the data can be automatically **rescaled** by estimating $\ell_1, \ldots, \ell_p$

CHARACTERISTIC LENGTH-SCALE / 4



For the first plot, we have chosen $\boldsymbol{M} = \boldsymbol{I}$: the function varies the same in all directions. The second plot is for $\boldsymbol{M} = \text{diag}(\ell)^{-2}$ and $\ell = (1, 3)$: The function varies less rapidly as a function of x_2 than x_1 as the length-scale for x_1 is less. In the third plot $\boldsymbol{M} = \Gamma\Gamma^T + \text{diag}(\ell)^{-2}$ for $\Gamma = (1, -1)^\top$ and $\ell = (6, 6)^\top$. Here Γ gives the direction of the most rapid variation. (Image from Rasmussen & Williams, 2006)