UNIVERSITÄT ZU LÜBECK
INSTITUTE FOR ROBOTICS
AND COGNITIVE SYSTEMS

Humanoid Robotics
Summer term 2018
May 16, 2018

Nils Rottmann
rottmann@rob.uni-luebeck.de

Exercise sheet 2 - Movement Representations

Please prepare the following exercises for the upcoming tutorial.

Task 1: Dynamical Systems

a) A parameterized trajectory generator can be defined as

$$\boldsymbol{\tau} = f(\boldsymbol{w}), \tag{1}$$

where $\boldsymbol{\tau}$ is the desired trajectory and $f(\boldsymbol{w})$ the function, which defines the trajectory based on parameters $\boldsymbol{w}$. One approach to define such a parameterized trajectory generator is based on a second order dynamical system. Such a system could be a simple spring-damper system, as depicted in Figure 1. The second order differential equation, which describes the presented system can be determined as

$$m\ddot{x} + d\dot{x} + cx = F. \tag{2}$$

We can now rearrange this equation such that

$$\frac{m}{c}\ddot{x} + \frac{d}{c}\dot{x} = \frac{F}{c} - x = g - x. \tag{3}$$

Hence, the term $\frac{F}{c} = g$ defines now a target position of the system. Further,

$$\begin{aligned} \frac{m}{d}\ddot{x} &= \frac{c}{d}(g - x) - \dot{x} \\ \ddot{x} &= \frac{d}{m}\left(\frac{c}{d}(g - x) - \dot{x}\right). \end{aligned} \tag{4}$$

Generalizing our second-order linear model leads to

$$\ddot{x} = \alpha\left(\beta(g - x) - \dot{x}\right). \tag{5}$$

In order to encode a desired acceleration profile (e.g. drive a circle) we have to add a forcing function $f_w(t)$ to the model

$$\ddot{x} = \alpha\left(\beta(g - x) - \dot{x}\right) + f_{\boldsymbol{w}}(t). \tag{6}$$

This forcing function is defined by parameters $\boldsymbol{w}$. This parameters have to be learned. Furthermore, such a function is generally built on normalized basis functions which are defined in the region $[0, 1]$. Thus, we have to encode a temporal scaling

$$\ddot{x} = \frac{1}{\tau^2}\alpha\left(\beta(g - x) - \dot{x}\tau\right) + f_{\boldsymbol{w}}(z), \tag{7}$$

where z is the phase variable defined in the region $z \in [0, 1]$ and τ is the time resolution. An advantage of such a model is the well-defined behavior and its stability by construction, but on the other hand this comes with only a limited class of possible movements.

b) A normalized radial basis function (NRBF) is defined as

$$\Phi_i(z) = \exp\left(-0.5\frac{z - c_i)^2}{h_i}\right). \tag{8}$$

Using K of these NRBF we can define our forcing function as

$$f_{\boldsymbol{w}}(z) = \frac{\sum_{i=1}^{K}\Phi_i(z)w_i}{\sum_{j=1}^{K}\Phi_j(z)}. \tag{9}$$

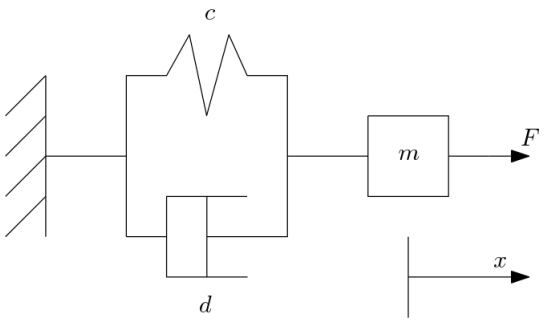

Figure 1 simple spring-damper system

In matrix notation we get

$$f_{\boldsymbol{w}}(z) = \boldsymbol{\Psi}^\top(z)\boldsymbol{w}, \qquad \Psi_i(z) = \frac{\Phi_i(z)}{\sum_{j=1}^{K} \Phi_j(z)}. \tag{10}$$

In order to understand RBF, take a look at Figure 2 where different RBFs are shown. The left figure shows six RBFs with $c_i = 0, 0.2, \ldots, 1$ and $h = 0.1$. The right figure has the same c_i but $h = 0.01$. As seen, the c defines the position of the peak of the RBF and the h defines the width of the RBF.

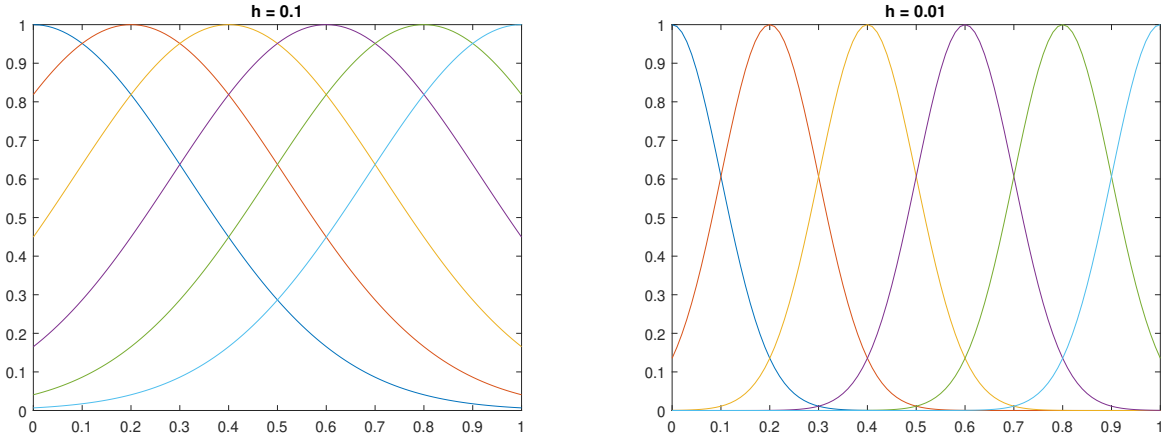

Figure 2 RBF

c) Assume we have some observations $\boldsymbol{y}$, a system matrix $\boldsymbol{X}$ and unknown parameters $\boldsymbol{\beta}$. Since the observations will probably not perfect fit our system we have also an error $\boldsymbol{\epsilon}$. Such a system can be described as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{11}$$

The idea is to determine the parameters $\boldsymbol{\beta}$ such that the squared error $\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}$ becomes minimal. Hence,

$$\begin{aligned} \boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} \end{aligned} \tag{12}$$

Searching for a minimum using the derivative yields to

$$\begin{aligned} \frac{\partial \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}}{\partial \boldsymbol{\beta}} &= -2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = 0 \\ \rightarrow \quad & 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = 2\boldsymbol{X}^\top \boldsymbol{y} \\ \rightarrow \quad & \boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \end{aligned} \tag{13}$$

To calculate the parameters $\boldsymbol{w}$ we have to measure desired trajectories (e.g. using an OptiTrack System). With this measurements we can compute our target values for our forcing function $\boldsymbol{f}_t$. Linear Regression yields

$$\boldsymbol{w} = (\boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{\Psi}^\top \boldsymbol{f}_t. \tag{14}$$

Task 2: Probabilistic Systems

a) The Bayes' Theorem is given as

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}. \tag{15}$$

Here $p(A|B)$ is the conditional probability of event $A = [a_1, a_2, \ldots, a_N]$ under the condition that event $B = [b_1, b_2, \ldots, b_M]$ occured and $p(B|A)$ the conditional probability of event B under the condition that event A occured. The probabilities $p(A)$ and $p(B)$ are the a-priori probabilities of the events A and B. Now, let us recapture some probability calculation. The probability of an event $a_i \in A$ can be written as the sum of the conditional probabilities

$$p(a_i) = \sum_{j=1}^{M} p(a_i|b_j)p(b_j). \tag{16}$$

The joint probability of an event a_i and b_j can be written as

$$p(a_i, b_j) = p(a_i|b_j)p(b_j) = p(b_j|a_i)p(a_i). \tag{17}$$

This is also the basis of the Bayes' Theorem. Furthermore, the sum of probabilities have to sum up to 1

$$\begin{aligned} \sum_{i=1}^{N} p(a_i) &= 1 \\ \sum_{i=1}^{N} p(a_i|b_j) &= 1 \\ \sum_{i=1,j=1}^{N,M} p(a_i, b_j) &= 1. \end{aligned} \tag{18}$$

The Bayes' Theorem allows to make estimates about a probability of an event A, which may not be observed directly using information about an event B which is related to event A and can be observed. Moreover, the Bayes' Theorem lays the foundation of position estimation methods, such as Kalman Filters.

b) We have the information, that the AIDS test is 99.9% sensitive and 99.7% specific. Let $A \in$ [infected, non-infected] the event, which defines if a person is infected or not and $B = [+, -]$ the event, which defines the result

of the AIDS test. From the above descriptions we can then define our conditional probabilities as follows

$$\begin{aligned} &p(+|\text{infected}) = 0.999 \\ &p(-|\text{non-infected}) = 0.997 \\ &p(-|\text{infected}) = 1 - p(+|\text{infected}) = 1 - 0.999 = 0.001 \\ &p(+|\text{non-infected}) = 1 - p(-|\text{non-infected}) = 1 - 0.997 = 0.003 \end{aligned} \tag{19}$$

and the a-priori probabilities as

$$\begin{aligned} &p(\text{infected}) = 0.001 \\ &p(\text{non-infected}) = 1 - p(\text{infected}) = 1 - 0.001 = 0.999 \\ &p(+) = p(+|\text{infected})p(\text{infected}) + p(+|\text{non-infected})p(\text{non-infected}) = 0.004 \end{aligned} \tag{20}$$

Using the Bayes' Theorem we get

$$p(\text{infected}|+) = \frac{p(+|\text{infected})p(\text{infected})}{p(+)} = \frac{0.999 \cdot 0.001}{0.004} = 0.2498 \approx 25\%. \tag{21}$$

Hence, if the test is positive, with a probability of 25% the person has AIDS.

c) In order to understand Gaussian Processes first consider a multivariate normal distribution with $\boldsymbol{x}' = \{x'_1, \ldots, x'_k\}$

$$f(\boldsymbol{x}'|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}' - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}' - \boldsymbol{\mu})\right), \tag{22}$$

which can be written as

$$\boldsymbol{x}' \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{23}$$

Figure 3 shows such a normal distribution. By partition the Gaussian random vector $\boldsymbol{x}'$ into $\boldsymbol{x}$ and $\boldsymbol{y}$, where both are jointly Gaussian random vectors, the term (23) becomes

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{C}^\top & \boldsymbol{B} \end{bmatrix}\right). \tag{24}$$

The marginal distribution of $\boldsymbol{x}$ is

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{A}) \tag{25}$$

and the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$ is

$$\boldsymbol{x}|\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{C}\boldsymbol{B}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top). \tag{26}$$

Thus, the conditional expectation and the covariance matrix can be written as

$$\begin{aligned} \mathrm{E}(\boldsymbol{x}|\boldsymbol{y}) &= \boldsymbol{\mu}_x + \boldsymbol{C}\boldsymbol{B}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y) \\ \mathrm{var}(\boldsymbol{x}|\boldsymbol{y}) &= \boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top. \end{aligned} \tag{27}$$

Proof. Define $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{M}\boldsymbol{y}$ with $\boldsymbol{M} = -\boldsymbol{C}\boldsymbol{B}^{-1}$. There $\boldsymbol{z}$ and $\boldsymbol{y}$ are uncorrelated and, since they are jointly normal, they are independent. This can be shown by

$$\begin{aligned} \mathrm{cov}(\boldsymbol{z}, \boldsymbol{y}) &= \mathrm{cov}(\boldsymbol{x}, \boldsymbol{y}) + \mathrm{cov}(\boldsymbol{M}\boldsymbol{y}, \boldsymbol{y}) \\ &= \boldsymbol{C} + \boldsymbol{M}\mathrm{var}(\boldsymbol{y}) \\ &= \boldsymbol{C} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{B} \\ &= 0 \end{aligned}$$

The expectation value of $\boldsymbol{x}|\boldsymbol{y}$ is calculated to proof the first part of equation (26), using the fact that $\mathrm{E}(\boldsymbol{z}|\boldsymbol{y}) = \mathrm{E}(\boldsymbol{z}) = \boldsymbol{\mu}_x + \boldsymbol{M}\boldsymbol{\mu}_y$.

$$\begin{aligned}
\mathrm{E}(\boldsymbol{x}|\boldsymbol{y}) &= \mathrm{E}(\boldsymbol{z} - \boldsymbol{M}\boldsymbol{y}|\boldsymbol{y}) \\
&= \mathrm{E}(\boldsymbol{z}|\boldsymbol{y}) - \mathrm{E}(\boldsymbol{M}\boldsymbol{y}|\boldsymbol{y}) \\
&= \mathrm{E}(\boldsymbol{z}) - \boldsymbol{M}\boldsymbol{y} \\
&= \boldsymbol{\mu}_x + \boldsymbol{M}(\boldsymbol{\mu}_y - \boldsymbol{y}) \\
&= \boldsymbol{\mu}_x + \boldsymbol{C}\boldsymbol{B}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y)
\end{aligned}$$

The covariance matrix can be derived from

$$\begin{aligned}
\mathrm{var}(\boldsymbol{x}|\boldsymbol{y}) &= \mathrm{var}(\boldsymbol{z} - \boldsymbol{M}\boldsymbol{y}|\boldsymbol{y}) \\
&= \mathrm{var}(\boldsymbol{z}|\boldsymbol{y}) - \mathrm{var}(\boldsymbol{M}\boldsymbol{y}|\boldsymbol{y}) - \boldsymbol{M}\mathrm{cov}(\boldsymbol{z}, -\boldsymbol{y}) - \mathrm{cov}(\boldsymbol{z}, -\boldsymbol{y})\boldsymbol{M}^\top \\
&= \mathrm{var}(\boldsymbol{z}|\boldsymbol{y}) \\
&= \mathrm{var}(\boldsymbol{z}) \\
&= \mathrm{var}(\boldsymbol{x} + \boldsymbol{M}\boldsymbol{y}) \\
&= \mathrm{var}(\boldsymbol{x}) + \boldsymbol{M}\mathrm{var}(\boldsymbol{y})\boldsymbol{M}^\top + \boldsymbol{M}\mathrm{cov}(\boldsymbol{x}, \boldsymbol{y}) + \mathrm{cov}(\boldsymbol{y}, \boldsymbol{x})\boldsymbol{M}^\top \\
&= \boldsymbol{A} + \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{B}^{-1}\boldsymbol{C}^\top - 2\boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top \\
&= \boldsymbol{A} + \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top - 2\boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top \\
&= \boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top
\end{aligned}$$

□

Processes without Noise Gaussian Processes are probabilistic models which can be used to estimate, on the basis of known data, a mean and a variance for an unknown data point. They use the relation given in equation (27).

From now on a more convenient nomenclature is used. Consider a training set $T = \{(\boldsymbol{x}_i, y_i)\} = (X, \boldsymbol{y})$ where $\boldsymbol{x}_i$ denotes an input vector of dimension D and y_i denotes a scalar output. Here, X contains all the input data and $\boldsymbol{y}$ all the output (target) data of the training set. In order to determine expectation values for $\boldsymbol{y}_*$, the values $\boldsymbol{y}$ at the states $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ have to be known. Rewriting equation (24) with the predefined training set yields

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}_* \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}(X) \\ \boldsymbol{\mu}(X_*) \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(X, X) & \boldsymbol{K}(X, X_*) \\ \boldsymbol{K}(X_*, X) & \boldsymbol{K}(X_*, X_*) \end{bmatrix} \right), \tag{28}$$

where $\boldsymbol{\mu}(\cdot)$ are the expectation values given a certain input set and $\boldsymbol{K}(\cdot, \cdot)$ a covariance matrix determined through two input sets. The expectation values and the variance of the unknown values $\boldsymbol{y}_*$ can be derived with equation (27), as

$$\begin{aligned}
\mathrm{E}(\boldsymbol{y}_*|\boldsymbol{y}, X, X_*) &= \boldsymbol{\mu}_* + \boldsymbol{K}_*^\top \boldsymbol{K}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \\
\mathrm{var}(\boldsymbol{y}_*|\boldsymbol{y}, X, X_*) &= \boldsymbol{K}_{**} - \boldsymbol{K}_*^\top \boldsymbol{K}^{-1}\boldsymbol{K}_*,
\end{aligned} \tag{29}$$

where $\boldsymbol{\mu}(X) = \boldsymbol{\mu}$, $\boldsymbol{\mu}(X_*) = \boldsymbol{\mu}_*$, $\boldsymbol{K}(X, X) = \boldsymbol{K}$, $\boldsymbol{K}(X, X_*) = \boldsymbol{K}_*$ and $\boldsymbol{K}(X_*, X_*) = \boldsymbol{K}_{**}$. The mean function $\boldsymbol{\mu}(\cdot)$ can be generated using proper information of the target values y_i given the input $\boldsymbol{x}_i$ or, if non such information is available, can be set to zero. The covariance function (kernel function) defines nearness or similarity and the covariance matrix $\boldsymbol{K}$ has to be positive definite. Possible types of such kernels are

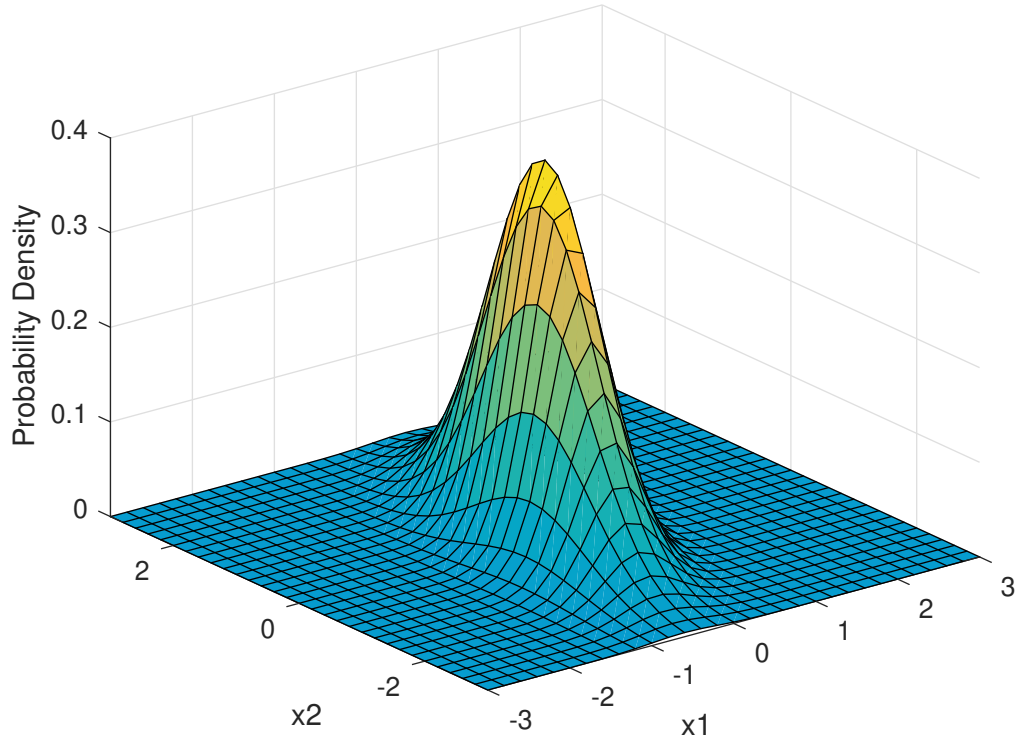

Figure 3 The figure shows a two dimensional Gaussian Distribution created using Matlab with $\boldsymbol{\mu} = [0\ 0]$ and $\boldsymbol{\Sigma} = [0.2\ 0.2;\ 0.2\ 1.0]$.

- $k = k(\boldsymbol{x} - \boldsymbol{x}')$ (stationary, invariant to translation in the input space)
- $k = k(||\boldsymbol{x} - \boldsymbol{x}'||)$ (isotropic, invariant to all rigid motions)
- $k = k(\boldsymbol{x} \cdot \boldsymbol{x}')$ (dot product, invariant to rotation)

A commonly used kernel function is the squared-exponential covariance function $k_{\text{se}}(\tau) = \sigma_f^2 \cdot \exp(-\frac{\tau^2}{2l^2})$ with $\tau = ||\boldsymbol{x} - \boldsymbol{x}'||$. The variables σ_f and l are hyperparameters. This kernel is smooth because the function is infinitely differentiable. The properties of a kernel around $\mathbf{0}$ determine the smoothness of the stationary process. A one dimensional example of a Gaussian Process is shown in figure 4.

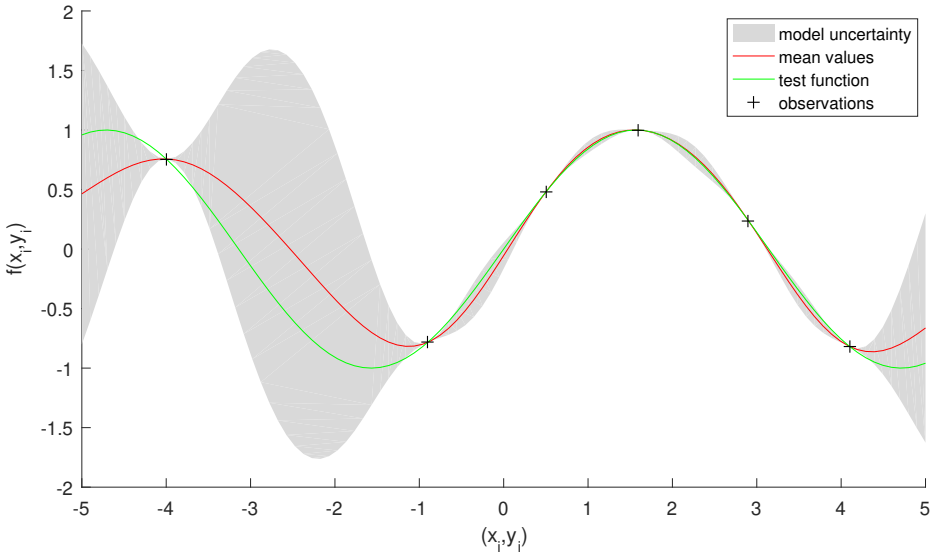

Figure 4 Example for a one dimensional Gaussian Process. The test function is a sinus; Six observation points have been used to determine an estimation result (mean values) and the model uncertainty (variance).

Noisy Observations Are the given data, thus the training set $T = \{(\boldsymbol{x}_i, y_i)\} = (X, \boldsymbol{y})$, corrupted with white noise, e.g. $y_i = z(\boldsymbol{x}_i) + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma_n^2)$, it can be taken into account by adding $\sigma_n^2 \boldsymbol{I}$ to the covariance matrix $\boldsymbol{K}$. The matrix $\boldsymbol{I}$ describes an identity matrix with convenient size. Equation (28) becomes

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}_* \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}(X) \\ \boldsymbol{\mu}(X_*) \end{bmatrix}, \begin{bmatrix} \boldsymbol{K} + \sigma_n^2 \boldsymbol{I} & \boldsymbol{K}_* \\ \boldsymbol{K}_*^\top & \boldsymbol{K}_{**} \end{bmatrix} \right) \tag{30}$$

with

$$\begin{aligned} \mathrm{E}(\boldsymbol{y}_* | \boldsymbol{y}, X, X_*) &= \boldsymbol{\mu}_* + \boldsymbol{K}_*^\top (\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \\ \mathrm{var}(\boldsymbol{y}_* | \boldsymbol{y}, X, X_*) &= \boldsymbol{K}_{**} - \boldsymbol{K}_*^\top (\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} \boldsymbol{K}_*. \end{aligned} \tag{31}$$

Parameter Adjustment Assuming zero mean ($\boldsymbol{\mu}(\cdot) = 0$), in equation (31) only the parameters of the kernel function $\boldsymbol{\theta}$ are unknown. Either one can choose these parameters by trial and error or use Bayesian Optimization to find optimal parameters for a certain problem. The aim of Bayesian Optimization is to find parameters $\boldsymbol{\theta}$ which maximize equation (22). The logarithm of the likelihood,

$$\log(f(\boldsymbol{y} | X, \boldsymbol{\theta})) = -\frac{1}{2} \boldsymbol{y}^\top \boldsymbol{K}^{-1} \boldsymbol{y} - \frac{1}{2} \log||\boldsymbol{K}|| - \frac{n}{2} \log(2\pi), \tag{32}$$

is used for maximization. In order to use optimization methods, such as gradient descent, the derivation of the log likelihood (32), defined as

$$\frac{\partial \log(f(\boldsymbol{y}|X,\boldsymbol{\theta}))}{\partial \theta_j} = \frac{1}{2}\mathrm{tr}\left((\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \boldsymbol{K}^{-1})\frac{\partial \boldsymbol{K}}{\partial \theta_j}\right), \tag{33}$$

with $\boldsymbol{\alpha} = \boldsymbol{K}^{-1}\boldsymbol{y}$ and n the number of data points in the training set, is required.

d) The Gaussian Processes can be used to find a smooth trajectory by using example data. Therefore we have to normalize the data in regard to the time such that one loop (one ∞) is scaled to $t = [0, 1]$. The normalized data can then be used to estimate the mean and variances for the x, y and z coordinates accordingly to the above description of the GP regression. The estimated mean data are the desired values of our trajectory which can be defined as

$$\tau = [\mu_{x,1:T}, \mu_{y,1:T}, \mu_{z,1:T}]^\top. \tag{34}$$