Mean Square Prediction Error of Misspecified Gaussian Process Models

Thomas Beckers, Jonas Umlauft and Sandra Hirche

***Abstract*— Nonparametric modeling approaches show very promising results in the area of system identification and control. A naturally provided model confidence is highly relevant for system-theoretical considerations to provide guarantees for application scenarios. Gaussian process regression represents one approach which provides such an indicator for the model confidence. However, this measure is only valid if the covariance function and its hyperparameters fit the underlying data generating process. In this paper, we derive an upper bound for the mean square prediction error of misspecified Gaussian process models based on a pseudo-concave optimization problem. We present application scenarios and a simulation to compare the derived upper bound with the true mean square error.**

I. Introduction

Nonparametric or so-called data-driven models are an uprising modeling approach for the identification and control of systems with unknown dynamics. In contrast to classical parametric techniques, the idea is to let the data speak for itself without assuming an underlying, parametric model structure [1]. Nonparametric models require only a minimum of prior knowledge for the regression of complex functions since the complexity of the model scales with the amount of training data [2]. Once a model of a system is learned from data, standard control laws such as model predictive control or feedback linerarization can be sucessfully applied [3], [4]. A general problem of data-driven models is the estimation of the model accuracy which is usually necessary for robust control design and stability considerations [5]. For that reason, Gaussian process (GP) models are a promising nonparametric approach for control because they provide not only a mean prediction, but also a variance as uncertainty measure of the model. Specifically, a GP assigns to every point of an input space a normally distributed random variable. Any finite group of those random variables follows a multivariate Gaussian distribution, and in consequence there exists an analytic solution for the predicted mean and variance of a new test point.

The variance of the prediction is exploited in many different kinds of control approaches [6]–[8]. However, the variance as prediction error measure is only valid if the GP model fits the data generating process, see Fig. 1. A GP model is fully described by a mean function, which is often set to zero [2], and a covariance function. Although GPs with universal covariance functions often produce satisfactory results, the selection of a suitable covariance function is a nontrivial problem [9], [10]. In general, the problem is that only a finite data set is available to derive the covariance function. In addition, the covariance function typically depends on a number of hyperparameters. There exist many different methods to estimate these parameters based on the training data set, e.g. marginal likelihood optimization. However, the involved optimization problems are in general non-convex, such that the marginal likelihood may have multiple local optima [2]. Alternatively, there exists the cross validation approach which deals with a validation and test set to carry out the hyperparameter selection. Still, all of these methods do not guarantee that the covariance function and its hyperparameters fit the data generating process. As consequence, the variance of the GP model may not correctly estimate the real model confidence. A lower bound for the prediction error for GP models with a misspecified covariance is given by [11] whereas an upper bound is still missing. Using GP models in control, the upper bound is highly interesting for stability consideration based on robust control methods.

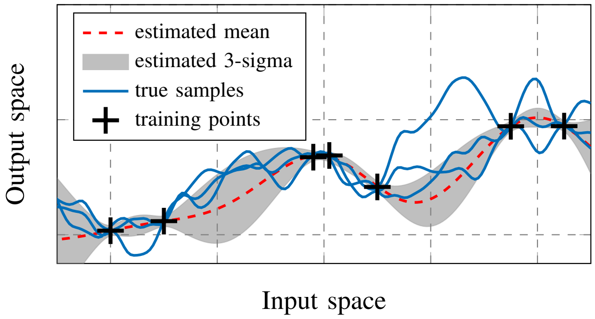

Fig. 1. The variance is misleading in terms of the model confidence.

The contribution of this paper is the derivation of an upper bound for the mean square prediction error (MSPE) between an estimated GP model and a GP model with unknown covariance functions and hyperparameters. For this purpose, a set of possible covariance functions with corresponding hyperparameter sets must be given. We exploit the property that many commonly used covariance functions are pseudo-concave with respect to their hyperparameters. As consequence, the upper bound is the solution of pseudo-concave optimization problems. With additional assumptions, a closed form solution is provided. **Notation:** Vectors are denoted with bold characters. Matrices are described with capital letters. The term $A_{i,:}$ denotes the i-th row of the matrix A. The expression $\mathcal{N}(\mu, \Sigma)$ describes a normal distribution with mean μ and covariance Σ. The notation $\boldsymbol{a} \preccurlyeq \boldsymbol{b}$ describes the componentwise inequality between two vectors $a_i \leq b_i, \forall i$.

The authors are with the Chair of Information-oriented Control (ITR), Department of Electrical and Computer Engineering, Technical University of Munich, 80333 Munich, Germany
{t.beckers, jonas.umlauft, hirche}@tum.de

II. Preliminaries and Problem Setting

A. Gaussian Process Models

Let $(\Omega, \mathcal{F}, P)$ be a probability space with the sample space Ω, the corresponding σ-algebra $\mathcal{F}$ and the probability measure P. The index set is given by $\mathcal{X} \subseteq \mathbb{R}^{n_x}$ with positive integer n_x. Then, a function $f(\boldsymbol{x}, \omega)$, which is a measurable function of $\omega \in \Omega$ with $\boldsymbol{x} \in \mathcal{X}$, is called a stochastic process and is simply denoted by $f(\boldsymbol{x})$. A GP is such a process which is fully described by a mean function $m\colon \mathcal{X} \subseteq \mathbb{R}^{n_x} \to \mathbb{R}$ and a covariance function $k\colon \Phi \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{\varphi}, \boldsymbol{x}, \boldsymbol{x}'))$$

with the hyperparameter vector $\boldsymbol{\varphi} \in \Phi \subseteq \mathbb{R}^l, \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. The mean function is usually defined to be zero, see [2]. The covariance function is a measure for the correlation of two states $(\boldsymbol{x}, \boldsymbol{x}')$ and depends on hyperparameters $\boldsymbol{\varphi}$ whose number $l \in \mathbb{N}$ depends on the function used. A necessary and sufficient condition for the function $k(\cdot,\cdot,\cdot)$ to be a valid covariance function (denoted by the set $\mathcal{K}$) is that the Gram matrix is positive semidefinite for all possible input values [12]. The choice of the covariance function and the determination of the corresponding hyperparameters can be seen as degrees of freedom of the regression. Probably the most widely used covariance function in Gaussian process modeling is the squared exponential (SE) covariance function, see [2]. An overview of the properties of different covariance functions can be found in [13].

In this paper, we use Gaussian process models with the assumption that the mean functions of the GPs are set to zero. Furthermore, a n_x-dimensional input space $\mathcal{X}$ and the output space $\mathbb{R}^{n_y}$ is considered, such that

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) = \begin{cases} f_1(\boldsymbol{x}) \sim \mathcal{GP}(0, k^1(\boldsymbol{\varphi}^1, \boldsymbol{x}, \boldsymbol{x}')) \\ \vdots \qquad \vdots \quad \vdots \\ f_{n_y}(\boldsymbol{x}) \sim \mathcal{GP}(0, k^n(\boldsymbol{\varphi}^n, \boldsymbol{x}, \boldsymbol{x}')) \end{cases} \tag{1}$$

with $\boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathbb{R}^{n_y}$. The Gaussian process for each function f_i depends on the covariance function k^i with the set of hyperparameters $\boldsymbol{\varphi}^i \in \Phi^i \subseteq \mathbb{R}^{l^i}, l^i \in \mathbb{N}$ for all $i \in \{1, \dots, n_y\}$. For the prediction, we concatenate m training inputs $\{\boldsymbol{x}^j\}_{j=1}^m$ and training outputs $\{\boldsymbol{y}^j\}_{j=1}^m$ in an input matrix $X = [\boldsymbol{x}^1, \boldsymbol{x}^2, \dots, \boldsymbol{x}^m]$ and a matrix of outputs $Y^\top = [\boldsymbol{y}^1, \boldsymbol{y}^2, \dots, \boldsymbol{y}^m]$ where $Y_{i,:}$ are corrupted by Gaussian noise with variance σ_i^2. In summary, the training data for the Gaussian processes is described by $\mathcal{D} = \{X, Y\}$. The joint distribution of the i-th component of $\boldsymbol{y}^* \in \mathbb{R}^{n_y}$ for a new test point $\boldsymbol{x}^* \in \mathcal{X}$ and the corresponding vector of the training outputs $Y_{:,i}$ is given by

$$\begin{bmatrix} Y_{:,i} \\ y_i^* \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} K^i(\boldsymbol{\varphi}^i, X, X) & \boldsymbol{k}^i(\boldsymbol{\varphi}^i, \boldsymbol{x}^*, X) \\ \boldsymbol{k}^i(\boldsymbol{\varphi}^i, \boldsymbol{x}^*, X)^\top & k^i(\boldsymbol{\varphi}^i, \boldsymbol{x}^*, \boldsymbol{x}^*) \end{bmatrix}\right),$$

where $Y_{:,i}$ is the i-th column of the matrix Y. The function $K^i\colon \Phi^i \times \mathcal{X}^m \times \mathcal{X}^m \to \mathbb{R}^{m \times m}$ is called the Gram matrix whose elements are $K^i_{j',j} = k^i(\boldsymbol{\varphi}^i, X_{:,j'}, X_{:,j}) + \delta(j, j')\sigma_i^2$ for all $j', j \in \{1, \dots, m\}$. The delta function $\delta(j, j') = 1$ for $j = j'$ and zero, otherwise, such that the variance σ_i^2 is added to the diagonal of the Gram matrix. The vector-valued covariance function $\boldsymbol{k}^i\colon \Phi^i \times \mathcal{X} \times \mathcal{X}^m \to \mathbb{R}^m$, with the elements $k^i_j = k^i(\boldsymbol{\varphi}^i, \boldsymbol{x}^*, X_{:,j})$ for all $j \in \{1, \dots, m\}$, expresses the covariance between $\boldsymbol{x}^*$ and the input training data X. A prediction of y_i^* is derived from the joint distribution, see [2] for more details. This conditional probability distribution is Gaussian with the conditional mean

$$\mu_i(\boldsymbol{y}^*|\boldsymbol{x}^*, \mathcal{D}) = \boldsymbol{k}^i(\boldsymbol{\varphi}^i, \boldsymbol{x}^*, X)^\top {K^i}^{-1} Y_{:,i}, \tag{2}$$

and the predicted variance

$$\begin{aligned} \operatorname{var}_i(\boldsymbol{y}^*|\boldsymbol{x}^*, \mathcal{D}) = k^i(\boldsymbol{\varphi}^i, \boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{k}^i(\boldsymbol{\varphi}^i, \boldsymbol{x}^*, X)^\top \\ {K^i}^{-1} \boldsymbol{k}^i(\boldsymbol{\varphi}^i, \boldsymbol{x}^*, X). \end{aligned} \tag{3}$$

Based on (2) and (3), the n_y normal distributed components $y_i^*|\boldsymbol{x}^*, \mathcal{D}$ are combined in a multi-variable Gaussian distribution $\boldsymbol{y}^*|(\boldsymbol{x}^*, \mathcal{D}) \sim \mathcal{N}(\boldsymbol{\mu}(\cdot), \Sigma(\cdot))$

$$\begin{aligned} \boldsymbol{\mu}(\boldsymbol{y}^*|\boldsymbol{x}^*, \mathcal{D}) &= [\mu_1(\cdot), \dots, \mu_{n_y}(\cdot)]^\top \\ \Sigma(\boldsymbol{y}^*|\boldsymbol{x}^*, \mathcal{D}) &= \operatorname{diag}\left[\operatorname{var}_1(\cdot), \dots, \operatorname{var}_{n_y}(\cdot)\right]. \end{aligned}$$

The hyperparameters $\boldsymbol{\varphi}^i$ can be optimized by means of the likelihood function, thus by maximizing the probability of $\boldsymbol{\varphi}^i = \arg\max_{\boldsymbol{\varphi}^i} \log P(Y_{:,i}|X, \boldsymbol{\varphi}^i)$ for all $i \in \{1, \dots, n_y\}$.

B. Problem Setting

We consider two GP models $\mathcal{GP}^1, \mathcal{GP}^2$ following (1) each trained with the same set of data points $\mathcal{D}$. The model $\mathcal{GP}^1$ is based on unknown covariance functions $k^1, \dots, k^{n_y}$ and hyperparameters $\boldsymbol{\varphi}^1, \dots, \boldsymbol{\varphi}^{n_y}$ whereas $\mathcal{GP}^2$ uses the covariance functions $\hat{k}^1, \dots, \hat{k}^{n_y}$ and $\hat{\boldsymbol{\varphi}}^1, \dots, \hat{\boldsymbol{\varphi}}^{n_y}$. The goal is to compute the MSPE between the prediction $\boldsymbol{y} \in \mathbb{R}^{n_y}$ of $\mathcal{GP}^1$ and the mean prediction of $\hat{\boldsymbol{y}} \in \mathbb{R}^{n_y}$ given by $\mathcal{GP}^2$, i.e.

$$\mathrm{E}\left[\|\boldsymbol{y}|(\boldsymbol{x}, \mathcal{D}) - \boldsymbol{\mu}(\hat{\boldsymbol{y}}|\boldsymbol{x}, \mathcal{D})\|^2\right]. \tag{4}$$

Since the covariance functions of $\mathcal{GP}^1$ are unknown, we derive an upper bound for the MSPE.

Remark 1 *The reason for using the predicted mean of $\mathcal{GP}^2$ only is that we compare the MSPE with the predicted variance of $\mathcal{GP}^2$ to show that the variance can be misleading.*

In accordance with the no-free-lunch theorem, it is not possible to give error bounds for the MSPE without any assumptions on $k^1, \dots, k^{n_y}$. Thus, we assume to have knowledge about a possible set of covariance functions $\tilde{\mathcal{K}}$ and a set of ranges for their hyperparameters $\tilde{\Phi}$.

Assumption 1 *Let $\tilde{\mathcal{K}}$ be a set of $z \in \mathbb{N}$ covariance functions*

$$\tilde{\mathcal{K}} = \{\tilde{k}^1, \dots, \tilde{k}^z \in \mathcal{K}\} \tag{5}$$

which are positive and pseudo-concave with respect to their hyperparameters. In addition, let $\tilde{\Phi}$ be a set of convex sets

$$\tilde{\Phi} = \{\tilde{\Phi}^1, \dots, \tilde{\Phi}^z | \tilde{\Phi}^j \subseteq \mathbb{R}^{l^j}, l^j \in \mathbb{N}, j \in \{1, \dots, z\}\}, \tag{6}$$

such that all elements of $\tilde{\Phi}^j$ are valid hyperparameters for $\tilde{k}^j$, i.e. $\forall j \in \{1, \dots, z\}, \tilde{k}^j\colon \tilde{\Phi}^j \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. Then, there exists a function $\Psi\colon \{1, \dots, n_y\} \to \{1, \dots, z\}$, such that $k^i = \tilde{k}^{\Psi(i)}, \boldsymbol{\varphi}^i \in \tilde{\Phi}^{\Psi(i)}$ for all $i \in \{1, \dots, n_y\}$.

Following this assumption, it is not necessary to know the exact covariance functions of $\mathcal{GP}^1$ but they must be elements of a set of possible covariance functions given by $\tilde{\mathcal{K}}$. To keep this set as small as possible, statistical hypothesis testing could be used for discarding functions which are too unlikely. Analogously, the exact hyperparameters $\boldsymbol{\varphi}^1, \dots, \boldsymbol{\varphi}^{n_y}$ can be unknown but each of them is in a set of $\tilde{\Phi}^1, \dots, \tilde{\Phi}^z$. In Section III-B, we show that many common covariance functions are pseudo-concave and positive such as the squared exponential, the rational quadratic and the polynomial for specific inputs. A visualization of a possible configuration for the sets $\tilde{\mathcal{K}}$ and $\tilde{\Phi}$ is shown in Fig. 2.

C. Application scenarios

Identification with GP state space models: For learning an unknown dynamics, the GP state space model (GP-SSM) is a common choice in control [14]. Assuming a discrete-time system $\boldsymbol{x}_{\tau+1} = \boldsymbol{f}(\boldsymbol{x}_\tau)$ with $\boldsymbol{x}_\tau \in \mathbb{R}^{n_x}, \boldsymbol{f}\colon R^{n_x} \to R^{n_x}, \tau \in \mathbb{N}$. Based on the dynamics, a set of data $\mathcal{D} = \{\boldsymbol{x}_\tau, \boldsymbol{x}_{\tau+1}\}_{\tau=1}^m$ is generated. For the GP-SSM, the input space $\mathcal{X}$ is the space of current states $\boldsymbol{x}_\tau$ and the output space represents the predicted next step ahead states $\hat{\boldsymbol{x}}_{\tau+1} \in \mathbb{R}^{n_x}$, such that

$$\hat{\boldsymbol{x}}_{\tau+1} = \boldsymbol{\mu}(\hat{\boldsymbol{x}}_{\tau+1} | \boldsymbol{x}_\tau, \mathcal{D}) + \Sigma(\hat{\boldsymbol{x}}_{\tau+1} | \boldsymbol{x}_\tau, \mathcal{D}) \boldsymbol{\zeta}_\tau$$

with $\boldsymbol{\zeta}_\tau \sim \mathcal{N}(0, I)$. The predicted variance correctly represents the model uncertainty if the reproducing kernel Hilbert space norm $\|f_i\|_{k^i}$ is bounded $\forall i \in \{1, \dots, n_x\}$. This is not a strong limitation on the application side since universal covariance functions, e.g. the SE function, approximate any continuous function f_i arbitrarily exactly on a closed set $\mathcal{X}$. However, without knowing the exact covariance function and hyperparameters, the predicted model uncertainty may not be correct. Our result (Theorems 1 and 2) allows to derive an upper bound for the MPSE between the correct but unknown GP-SSM and an estimated GP-SSM. Consequently, the upper bound also captures the error between the estimated GP-SSM and the original discrete-time system.

Reinforcement learning: Following [15], a Gaussian process model is used for the value process $V\colon \mathbb{R}^{n_x} \to \mathbb{R}$ which connects values and rewards in a reinforcement learning scenario. It includes the assumption that the choice of the covariance function reflects the prior concerning the correlation between the values of states and rewards. The presented Theorem 1 can be used to avoid an eventually underestimated MSPE based on the predicted variance with suboptimal hyperparameters. In this scenario, the set $\tilde{\mathcal{K}}$ contains the selected covariance function $\tilde{k}^1$ only. Thus, an upper bound for the MSPE can be computed without knowing the exact hyperparameters.

III. Mean square prediction error

In this section, we present the computation of an upper bound for the MSPE between $\mathcal{GP}^1$ and the mean prediction of $\mathcal{GP}^2$ that is given by[1]

$$\mathrm{E}\left[\|\Delta\|^2\right] = \sum_{i=1}^{n_y} k^i(\boldsymbol{\varphi}^i) - 2\hat{\boldsymbol{k}}^i(\hat{\boldsymbol{\varphi}}^i)^\top \hat{K}^{i^{-1}}(\hat{\boldsymbol{\varphi}}^i)\boldsymbol{k}^i(\boldsymbol{\varphi}^i)$$
$$+ \hat{\boldsymbol{k}}^i(\hat{\boldsymbol{\varphi}}^i)^\top \hat{K}^{i^{-1}}(\hat{\boldsymbol{\varphi}}^i) K^i(\boldsymbol{\varphi}^i) \hat{K}^{i^{-1}}(\hat{\boldsymbol{\varphi}}^i) \hat{\boldsymbol{k}}^i(\hat{\boldsymbol{\varphi}}^i) \qquad (7)$$

with error $\Delta = \boldsymbol{y}|(\boldsymbol{x}, \mathcal{D}) - \boldsymbol{\mu}(\hat{\boldsymbol{y}}|\boldsymbol{x}, \mathcal{D})$. The covariance vector function $\hat{\boldsymbol{k}}$ and the Gram matrix $\hat{K}$ are related to $\mathcal{GP}^2$.

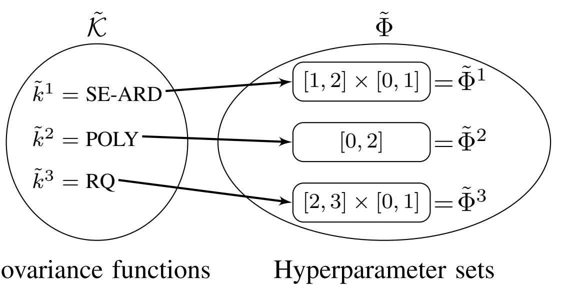

Fig. 2. Example configuration for Assumption 1

Remark 2 *If the estimated covariance function and its hyperparameters are correct, i.e.* $k^i = \hat{k}^i, \boldsymbol{\varphi}^i = \hat{\boldsymbol{\varphi}}^i$ *for all* i*, the mean square error is simplified to*

$$\mathrm{E}\left[\|\Delta\|^2\right] = \mathrm{Tr}\left(\Sigma(\hat{\boldsymbol{y}}|\boldsymbol{x}, \mathcal{D})\right), \qquad (8)$$

which is the trace of the posterior variance matrix.

It is obvious, that the true covariance functions k^i are needed to compute this error. To overcome this issue, we derive an upper bound based on a set of covariance functions and hyperparameters. For determining this bound, the maximum of (7) has to be computed without knowing the covariance function k^i and the corresponding hyperparameters $\boldsymbol{\varphi}^i$. With Assumption 1, this problem is a non-convex, mixed-integer optimization problem. For simplicity in notation in the following derivations, parts of (7) are renamed as

$$\alpha^i(\boldsymbol{x}) = k^i(\boldsymbol{\varphi}^i) \qquad (9)$$
$$\beta^i(\boldsymbol{x}) = \hat{\boldsymbol{k}}^i(\hat{\boldsymbol{\varphi}}^i)^\top \hat{K}^{i^{-1}} \boldsymbol{k}^i(\boldsymbol{\varphi}^i) \qquad (10)$$
$$\gamma^i(\boldsymbol{x}) = \hat{\boldsymbol{k}}^i(\hat{\boldsymbol{\varphi}}^i)^\top \hat{K}^{i^{-1}} K^i(\boldsymbol{\varphi}^i) \hat{K}^{i^{-1}} \hat{\boldsymbol{k}}^i(\hat{\boldsymbol{\varphi}}^i) \qquad (11)$$

with $\alpha^i, \beta^i, \gamma^i\colon \mathcal{X} \to \mathbb{R}$.

Lemma 1 *For any* $k^i \in \tilde{\mathcal{K}}$*, the inequality*

$$k^i(\boldsymbol{\varphi}^i, \boldsymbol{x}, \boldsymbol{x}') \leq \max_{j \in \{1,\dots,z\}} \max_{\tilde{\boldsymbol{\varphi}}^j \in \tilde{\Phi}^j} \tilde{k}^j(\tilde{\boldsymbol{\varphi}}^j, \boldsymbol{x}, \boldsymbol{x}')$$

holds for $\boldsymbol{\varphi}^i \in \tilde{\Phi}^{\Psi(i)}, \forall i \in \{1, \dots, n_y\}$ *and* $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$.

Proof: Since k^i is an element of $\tilde{\mathcal{K}}$, the maximization over all covariance functions $\tilde{k}^j$ with their hyperparameter sets $\tilde{\Phi}^j$ must be an upper bound for k^i. The optimization problem can be separated in an outer maximization over the finite number of covariance functions $\tilde{k}^j$ and an inner maximization over the convex hyperparameter sets. ■

Lemma 2 *Under Assumption 1, there exists a lower bound* $\underline{\beta}^i(\boldsymbol{x})\colon \mathcal{X} \to \mathbb{R}$ *for* (10) *given by*

$$\underline{\beta}^i(\boldsymbol{x}) = \sum_{p=1}^m \min\left\{h_p^i, 0\right\} \max_j \max_{\tilde{\boldsymbol{\varphi}}^j} \tilde{k}^j(\tilde{\boldsymbol{\varphi}}^j, \boldsymbol{x}, X_{:,p}) \qquad (12)$$
$$\boldsymbol{h}^i = \hat{\boldsymbol{k}}^i(\hat{\boldsymbol{\varphi}}^i)^\top \hat{K}^{i^{-1}}, \ \boldsymbol{h}^i \in \mathbb{R}^m$$

with $j \in \{1, \dots, z\}$ *and* $\tilde{\boldsymbol{\varphi}}^j \in \tilde{\Phi}^j, \forall \boldsymbol{x} \in \mathcal{X}, \forall i \in \{1, \dots, n_y\}$.

[1] For notational convenience we do not write the arguments $\boldsymbol{x}$ and X

Proof: The term (10) can be lower bounded by

$$\beta^i(\boldsymbol{x}) \geq \sum_{p=1}^{m} \min\left\{h_p^i, 0\right\} \max \tilde{k}^j(\tilde{\boldsymbol{\varphi}}^j, \boldsymbol{x}, X_{:,p}) \\ + \max\left\{h_p^i, 0\right\} \min \tilde{k}^j(\tilde{\boldsymbol{\varphi}}^j, \boldsymbol{x}, X_{:,p}) \qquad (13) \\ \text{s.t. } j \in \{1, \dots, z\}, \tilde{\boldsymbol{\varphi}}^j \in \tilde{\Phi}^j,$$

because the negative elements of $\boldsymbol{h}$ are multiplied with the maximum value of all covariance functions in $\tilde{\mathcal{K}}$ and vice versa. The minimum of $\tilde{k}^j(\tilde{\boldsymbol{\varphi}}^j, \boldsymbol{x}, X_{:,p})$ is always positive following Assumption 1, so that

$$\beta^i(\boldsymbol{x}) \geq \sum_{p=1}^{m} \min\left\{h_p^i, 0\right\} \max \tilde{k}^j(\tilde{\boldsymbol{\varphi}}^j, \boldsymbol{x}, X_{:,p})$$

s.t. $j \in \{1, \dots, z\}, \tilde{\boldsymbol{\varphi}}^j \in \tilde{\Phi}^j$ holds. With Lemma 1, we obtain the lower bound (12). ■

Lemma 3 *Under Assumption 1, there exists an upper bound* $\bar{\gamma}^i(\boldsymbol{x})\colon \mathcal{X} \to \mathbb{R}$ *for* (11) *given by*

$$\bar{\gamma}^i(\boldsymbol{x}) = \sum_{p,q=1,\dots,m} \max\left\{h_p^i h_q^i, 0\right\} \max_j \max_{\tilde{\boldsymbol{\varphi}}^j} \tilde{k}^j(\tilde{\boldsymbol{\varphi}}^j, X_{:,q}, X_{:,p}) \qquad (14) \\ \boldsymbol{h}^i = \hat{\boldsymbol{k}}^i(\hat{\boldsymbol{\varphi}}^i)^\top \hat{K}^{i^{-1}}, \; \boldsymbol{h}^i \in \mathbb{R}^m$$

with $j \in \{1, \dots, z\}$ *and* $\tilde{\boldsymbol{\varphi}}^j \in \tilde{\Phi}^j, \forall \boldsymbol{x} \in \mathcal{X}, \forall i \in \{1, \dots, n_y\}$.

Proof: It is analogous to the proof of Lemma 2. ■

Theorem 1 *Consider the MSPE between the output* $\boldsymbol{y}$ *of* $\mathcal{GP}^1$ *and the mean* $\hat{\boldsymbol{y}}$ *of* $\mathcal{GP}^2$ (4). *With Assumption 1, there exists an upper bound for the MSPE given by*

$$\mathrm{E}\left[\|\Delta\|^2\right] \leq n_y \bar{\alpha}(\boldsymbol{x}) + \sum_{i=1}^{n_y} \bar{\gamma}^i(\boldsymbol{x}) - 2\underline{\beta}^i(\boldsymbol{x}) \qquad (15)$$

$$\bar{\alpha}(\boldsymbol{x}) = \max_{j \in \{1,\dots,z\}} \max_{\tilde{\boldsymbol{\varphi}}^j \in \tilde{\Phi}^j} \tilde{k}^j(\tilde{\boldsymbol{\varphi}}^j, \boldsymbol{x}, \boldsymbol{x}) \qquad (16)$$

with $\underline{\beta}^i$ *of* (12) *and* $\bar{\gamma}^i$ *of* (14).

Proof: The mean square error is upper bounded by the sum of the upper bounds for each term of (7). An upper bound of (9) with Assumption 1 can be computed by (16) following Lemma 1. The bound $\bar{\alpha}$ is independent of the training data $\mathcal{D}$ and thus, independent of i, so that it is summed up by $n_y\bar{\alpha}$. With Lemmas 2 and 3, the second and third term is bounded which results in (16). ■

Remark 3 *The minimum of* (13) *is set to zero because the numerical computation would be hard to obtain since* $\tilde{k}$ *is only pseudo-concave. In this form, the solution of (15) can be computed by standard optimization algorithms [16].*

A. Closed form solution

With additional assumptions, it is possible to provide a closed form solution for (15) of Theorem 1.

Assumption 2 *Each convex set of hyperparameters* $\tilde{\Phi}^j \in \tilde{\Phi}$ *of* (6) *can be described by two vectors* $\underline{\boldsymbol{\varphi}}^j, \bar{\boldsymbol{\varphi}}^j \in \mathbb{R}^{l^j}$

$$\tilde{\Phi}^j = \left\{\tilde{\boldsymbol{\varphi}}^j \in \mathbb{R}^{l^j} | \underline{\boldsymbol{\varphi}}^j \preceq \tilde{\boldsymbol{\varphi}}^j \preceq \bar{\boldsymbol{\varphi}}^j\right\}, \forall j \in \{1, \dots, z\}.$$

Assumption 3 *Each covariance function* $\tilde{k}^j, j \in \{1, \dots, z\}$ *of* (5) *is componentwise strictly increasing with respect to its hyperparameters* $\tilde{\boldsymbol{\varphi}}^j$*, i.e.* $\forall \tilde{\varphi}_i^j, \tilde{v}_i^j$ *such that* $\tilde{\varphi}_i^j < \tilde{v}_i^j$ *one has* $\tilde{k}^j(\tilde{\boldsymbol{\varphi}}^j, \boldsymbol{x}, \boldsymbol{x}') < \tilde{k}^j(\tilde{\boldsymbol{v}}^j, \boldsymbol{x}, \boldsymbol{x}') \; \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}, \tilde{\boldsymbol{\varphi}}^j, \tilde{\boldsymbol{v}}^j \in \tilde{\Phi}^j$ *for all* $i \in \{1, \dots, l^j\}$.

Assumption 2 requires that each of the convex hyperparameter sets $\Phi^j \subseteq \mathbb{R}^{l^j}$ is a l^j-dimensional hyperrectangle which is a weak restriction in practice. In Section III-B, we show that Assumption 3 holds for some commonly used covariance functions. Based on these assumptions, there exists a closed form solution of Theorem 1 because the maximum of the covariance function $\tilde{k}^j$ is now always at $\bar{\boldsymbol{\varphi}}^j$, see Fig. 3.

Theorem 2 *Consider the MSPE between the output* $\boldsymbol{y}$ *of* $\mathcal{GP}^1$ *and the mean* $\hat{\boldsymbol{y}}$ *of* $\mathcal{GP}^2$ (4). *With Assumptions 1 to 3, there exists an upper bound for the MSPE given by*

$$\mathrm{E}\left[\|\Delta\|^2\right] \leq \sum_{i=1}^{n_y} \max_j \left\{\tilde{k}^j(\bar{\boldsymbol{\varphi}}^j, \boldsymbol{x}, \boldsymbol{x}) + \kappa^i(\boldsymbol{x}) - \eta^i(\boldsymbol{x})\right\}$$

$$\eta^i(\boldsymbol{x}) = 2\sum_{p=1}^{m} \min\left\{h_p^i, 0\right\} \tilde{k}^j(\bar{\boldsymbol{\varphi}}^j, \boldsymbol{x}, X_{:,p}) \\ + \max\left\{h_p^i, 0\right\} \tilde{k}^j(\underline{\boldsymbol{\varphi}}^j, \boldsymbol{x}, X_{:,p}) \\ \kappa^i(\boldsymbol{x}) = \sum_{p,q=1,\dots,m} \max\left\{h_p^i h_q^i, 0\right\} \tilde{k}^j(\bar{\boldsymbol{\varphi}}^j, X_{:,q}, X_{:,p}) \\ + \min\left\{h_p^i h_q^i, 0\right\} \tilde{k}^j(\underline{\boldsymbol{\varphi}}^j, X_{:,q}, X_{:,p}) \qquad (21)$$

with $\boldsymbol{h}^i = \hat{\boldsymbol{k}}^i(\hat{\boldsymbol{\varphi}}^i)^\top \hat{K}^{i^{-1}}$.

Remark 4 *The solution of* (21) *is a closed form expression in the sense that it can be evaluated in a finite number of operations because the maximization is over a finite set.*

Proof: Assume that we choose $j \in \{1, \dots, z\}$ of each maximization such that $\tilde{k}^j$ of (21) is equal to the covariance function k^i. With Assumption 3, the covariance function k^i with the hyperparameters $\boldsymbol{\varphi}^i$ is always equal or less then with $\bar{\boldsymbol{\varphi}}^i$ and vice versa, i.e. $k^i(\boldsymbol{\varphi}^i, \boldsymbol{x}, \boldsymbol{x}') \geq k^i(\underline{\boldsymbol{\varphi}}^i, \boldsymbol{x}, \boldsymbol{x}')$. Thus, we have to prove

$$\mathrm{E}\left[\|\Delta\|^2\right] \leq \sum_{i=1}^{n_y} \Big\{k^i(\bar{\boldsymbol{\varphi}}^i, \boldsymbol{x}, \boldsymbol{x}) \\ + \sum_{p,q=1,\dots,m} \max\left\{h_p^i h_q^i, 0\right\} k^i(\bar{\boldsymbol{\varphi}}^i, X_{:,q}, X_{:,p}) \\ + \min\left\{h_p^i h_q^i, 0\right\} k^i(\underline{\boldsymbol{\varphi}}^i, X_{:,q}, X_{:,p}) \\ - 2\sum_{p=1}^{m} \min\left\{h_p^i, 0\right\} k^i(\bar{\boldsymbol{\varphi}}^i, \boldsymbol{x}, X_{:,p}) \\ + \max\left\{h_p^i, 0\right\} k^i(\underline{\boldsymbol{\varphi}}^i, \boldsymbol{x}, X_{:,p})\Big\}, \qquad (22)$$

where each term of (22) upper bounds the corresponding term of $\mathrm{E}\left[\|\Delta\|^2\right]$ in (7) analogous to the idea in the proof of Lemma 2. Since (21) maximizes over all $\tilde{k}^j$ and, considering Assumption 1, the covariance function k^i is element of $\tilde{\mathcal{K}}$, there exists a j such that the assumption at the beginning of the proof is fulfilled. (17) ■

Corollary 1 *If $k^1 = \dots = k^{n_y}$ and $\boldsymbol{\varphi}^1 = \ldots = \boldsymbol{\varphi}^{n_y}$, the closed form solution* (21) *of Theorem 2 is equivalent to the posterior variance given by* (8) *for $\tilde{\mathcal{K}} = \{k^1\}$ and $\tilde{\Phi} = \{\boldsymbol{\varphi}^1\}$*

Proof: This is a result of (22) if the set $\tilde{\mathcal{K}}$ only contains the covariance functions $k^1 = \dots = k^{n_y}$ and the set $\tilde{\Phi}$ only the corresponding hyperparameters $\boldsymbol{\varphi}^1 = \ldots = \boldsymbol{\varphi}^{n_y}$. ■

Remark 5 *Corollary 1 shows the convergence of the upper bound to the true MSPE* (4) *between $\mathcal{GP}^1$ and $\mathcal{GP}^2$ for the minimum-size sets $\tilde{\mathcal{K}}, \tilde{\Phi}$.*

B. Pseudo-concave covariance functions

In the following, we show that many common covariance functions fulfill Assumptions 1 and 3.

Proposition 1 *The covariance functions* (17) *to* (20) *with the corresponding parameters are pseudo-concave and componentwise monotonically increasing with respect to their hyperparameters on the designated domain.*

Proof: The following proof considers each covariance function separately.
Polynomial: The polynomial function k is strictly increasing on $\varphi \in \mathbb{R}_{\geq 0}$ for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{n_x}_{\geq 0}$ and hence, pseudo-concave [17] and componentwise monotonically increasing.
Rational quadratic: The covariance function is quasi-concave if $\det H_3(\boldsymbol{\varphi}) > 0$ and $\det H_2(\boldsymbol{\varphi}) < 0$, where the matrix H_r is the r-th order leading principal submatrix of the bordered Hessian of k in respect to $\boldsymbol{\varphi}$, see [17]. The principal submatrices are given by

$$H_2 = \frac{-4d^2p^2\varphi_2^4}{\varphi_1^2(2p\varphi_1^2+d)^2\left(\frac{2p\varphi_1^2+d}{2p\varphi_1^2}\right)^{2p}} < 0,$$

$$H_3 = \frac{8d\varphi_2^4 p(dp+d+6p\varphi_1^2)}{\varphi_1^2(2p\varphi_1^2+d)^2\left(\frac{2p\varphi_1^2+d}{2p\varphi_1^2}\right)^{3p}} > 0,$$

with $d = \|\boldsymbol{x} - \boldsymbol{x}'\|^2 > 0, \forall p \in \mathbb{N}_{>0}, \boldsymbol{\varphi} \in \mathbb{R}^2_{>0}$ so that the function is quasi-concave. Since $k \in \mathcal{C}^1$ and $\partial k/\partial \boldsymbol{\varphi} \neq \boldsymbol{0}$ on its domain, the function is also pseudo-concave [17]. It is obviously also componentwise monotonically increasing.
Squared exponential: The covariance function can be rewritten as

$$k(\boldsymbol{\varphi}, \boldsymbol{x}, \boldsymbol{x}') = \exp\left(\log(\varphi_{n_x+1}^2) + \sum_{i=1}^{n_x} -\frac{|x_i - x_i'|^2}{2\varphi_i^2}\right),$$

where the argument of the exponential functions is quasi-concave, since this sum of concave functions is concave on all $\boldsymbol{\varphi} \in \mathbb{R}^{n_x+1}_{>0}$ for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{n_x}$. The composition with the strictly increasing exponential function results in an overall quasi-concave function [18, Theorem 8.5]. Since k is continuous and $\partial k/\partial \boldsymbol{\varphi} \neq \boldsymbol{0}$ on its domain, the function is also pseudo-concave. Since the exponential and the logarithm function are monotonically increasing, the covariance function is componentwise monotonically increasing.

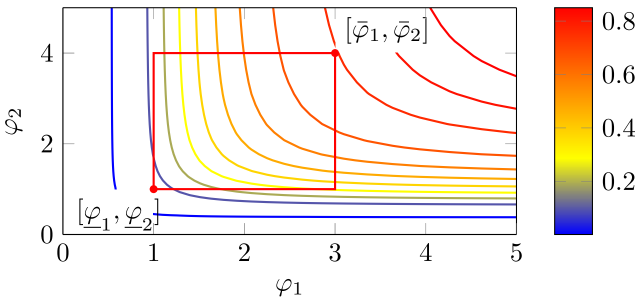

Fig. 3. The SE function $k(\boldsymbol{\varphi}, \boldsymbol{x}, \boldsymbol{x})$ over its hyperparameters $\boldsymbol{\varphi}$. With Assumptions 2 and 3, the maximum is at the corner of the hyperrectangle $\bar{\boldsymbol{\varphi}}$.

Matérn: For $p \in \mathbb{N}_{>0}, \nu = p + 1/2$, the function can be simplified to

$$k_{\boldsymbol{\varphi}}(d) = \varphi_2^2 \exp\left(-\frac{\sqrt{2\nu}d}{\varphi_1}\right)\frac{p!}{(2p)!}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\left(\frac{\sqrt{8\nu}d}{\varphi_1}\right)^{p-i}.$$

Analogous to the rational quadratic covariance, for the principal submatrices, it holds $\det \bar{H}_2 < 0$ and $\det \bar{H}_2 > 0$. With $k \in \mathcal{C}^1$ and $\partial k/\partial \boldsymbol{\varphi} \neq \boldsymbol{0}$ on its domain, the function is pseudo-concave. Since the exponential function grows faster than the polynomial, the covariance function is also componentwise monotonically increasing. ■

IV. Simulation

In this section, we present a numerical example for the result of Theorem 2 with GP-SSMs. For this purpose, we assume that a discrete-time, one-dimensional system can be correctly modeled by $\mathcal{GP}^1$ with Matérn covariance function where $p = 1$ and the hyperparameters $\boldsymbol{\varphi}^1 = [5.2, 1.6]^\top$. The training set contains 10 uniformly distributed measurements. Since the correct covariance function is usually unknown in real-world applications, the squared exponential (SE) covariance function is often used to learn the system dynamics. Following that approach, $\mathcal{GP}^2$ with SE covariance function is trained with the measurements of the system. The hyperparameters are optimized according to the likelihood

TABLE I

Pseudo-concave and componentwise monotonically increasing covariance functions

Covariance function	Expression $k(\boldsymbol{\varphi}, \boldsymbol{x}, \boldsymbol{x}') =$		Parameters	Domain
Polynomial	$(\boldsymbol{x}\boldsymbol{x}' + \varphi^2)^p$	(17)	$p \in \mathbb{N}, \varphi \in \mathbb{R}_{\geq 0}$	$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{n_x}_{\geq 0}$
Rational quadratic	$\varphi_2^2\left(1 + \frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2p\varphi_1^2}\right)^{-p}$	(18)	$p \in \mathbb{N}_{>0}, \boldsymbol{\varphi} \in \mathbb{R}^2_{>0}$	$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{n_x}$
Squared exponential	$\varphi_{n_x+1}^2 \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{x}')^\top P^{-1}(\boldsymbol{x}-\boldsymbol{x}')}{2}\right)$	(19)	$P = \operatorname{diag}(\varphi_1^2, \dots, \varphi_{n_x}^2), \boldsymbol{\varphi} \in \mathbb{R}^{n_x+1}_{>0}$	$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{n_x}$
Matérn	$\varphi_2^2 \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|\boldsymbol{x}-\boldsymbol{x}'\|}{\varphi_1}\right)^\nu \mathfrak{K}_\nu\left(\frac{\sqrt{2\nu}\|\boldsymbol{x}-\boldsymbol{x}'\|}{\varphi_1}\right)$	(20)	$\nu = p + 1/2, p \in \{0, 1, 2\}, \boldsymbol{\varphi} \in \mathbb{R}^2_{\geq 0}$	$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{n_x}$

TABLE II

Covariance functions in $\tilde{\mathcal{K}}$ with hyperparameter sets in $\tilde{\Phi}$.

Covariance functions	Hyperparameter sets
$\tilde{k}^1$: Matérn $p = 1$	$\{0.9\boldsymbol{\varphi}^1 \preceq \tilde{\boldsymbol{\varphi}}^1 \preceq 1.1\boldsymbol{\varphi}^1\}$ $\{[0;0] \prec \tilde{\boldsymbol{\varphi}}^1 \preceq 2\boldsymbol{\varphi}^1\}$ $\{[0;0] \prec \tilde{\boldsymbol{\varphi}}^1 \preceq 3\boldsymbol{\varphi}^1\}$
$\tilde{k}^2$: Matérn $p = 0, 2$	$\{[1;1.5] \preceq \tilde{\boldsymbol{\varphi}}^2 \preceq [10;2]\}$
$\tilde{k}^3$: Rational quadratic $p = 1$	$\{[1;0.1] \preceq \tilde{\boldsymbol{\varphi}}^3 \preceq [20;1]\}$
$\tilde{k}^4$: Squared exponential	$\{[0.1;0.01] \preceq \tilde{\boldsymbol{\varphi}}^4 \preceq [10;1]\}$

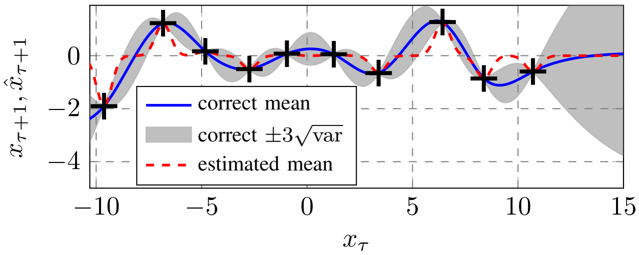

Fig. 4. Based on the training data, the estimated mean generates a misleading impression of the underlying process.

function with a conjugate gradient method which results in the hyperparameters $\hat{\varphi}^1 = [0.36, 0.32]^\top$.

In Fig. 4, the estimated mean $\boldsymbol{\mu}(\hat{x}_{\tau+1}|x_\tau, \mathcal{D})$ together with mean and variance of the true generating process are shown. It is obvious that the mean does not correspond to the true process, although it represents the training data effectively. As consequence, the mean square error between the estimated mean and the correct model is radically underestimated in the state space and in the time domain as presented in Fig. 5. To overcome this issue, we use Theorem 2 to compute an upper bound of the MSPE without exact knowledge of the correct covariance function. For this purpose, we consider a set of covariance functions with their corresponding hyperparameter sets shown in Table II. For comparison of different interval ranges, we use three different interval sizes around the true hyperparameters. Figure 5 shows the estimated and true mean square prediction error which is normally unknown. The estimated error obviously underestimates the true MPSE. In contrast, the derived upper bound given by Theorem 2 based on the functions of Table II successfully confines the true MSPE. With a wider range of the interval the bound becomes loser.

Conclusion

We derive an upper bound for the mean square prediction error between an estimated GP model and a GP model with unknown covariance function. For the proposed upper bound, no exact knowledge about the underlying covariance function is required. Instead, only a set of possible covariance functions with their hyperparameter sets are necessary. With additional weak assumptions, a closed form solution is provided. A numerical example demonstrates that this bound confines the usually unknown mean square prediction error.

Acknowledgments

The research leading to these results has received funding from the ERC Starting Grant "Control based on Human Models (con-humo)" agreement n°337654.

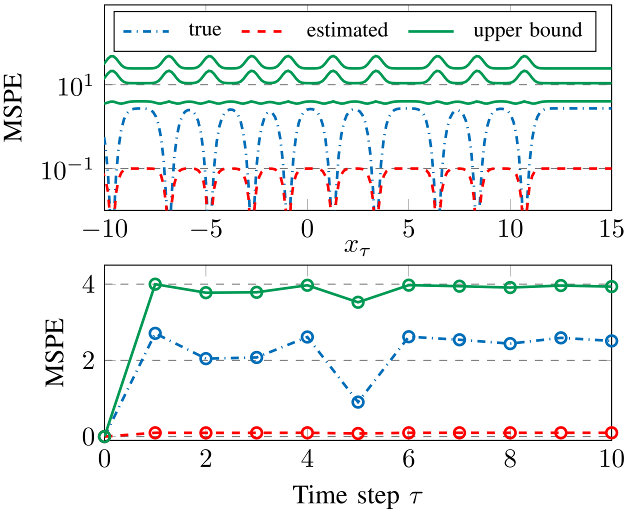

Fig. 5. Top: The estimated, the true and the upper bound of the MSPE for a 10%, 100%, and 200% error interval (from bottom to top) around the correct hyperparameter values. Bottom: The comparison in time domain with the 10% bound.

References

[1] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.

[2] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, vol. 1. MIT press Cambridge, 2006.

[3] G. Chowdhary, J. How, and H. Kingravi, "Model reference adaptive control using nonparametric adaptive elements," in *Proc. of Conference on Guidance Navigation and Control*, 2012.

[4] J. Umlauft, T. Beckers, M. Kimmel, and S. Hirche, "Feedback linearization using Gaussian processes," in *Proc. of the Conference on Decision and Control*, 2017.

[5] K. Zhou and J. C. Doyle, *Essentials of robust control*, vol. 104. Prentice hall Upper Saddle River, NJ, 1998.

[6] J. R. Medina, T. Lorenz, and S. Hirche, "Synthesizing anticipatory haptic assistance considering human behavior uncertainty," *IEEE Transactions on Robotics*, vol. 31, no. 1, pp. 180–190, 2015.

[7] T. Beckers, J. Umlauft, D. Kuli, and S. Hirche, "Stable Gaussian process based tracking control of lagrangian systems," in *Proc. of the Conference on Decision and Control*, 2017.

[8] J. Kocijan, R. Murray-Smith, C. E. Rasmussen, and A. Girard, "Gaussian process model based predictive control," in *Proc. of the American Control Conference*, 2004.

[9] M. Seeger, "Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers," in *Advances in neural information processing systems*, pp. 603–609, 2000.

[10] G. Pillonetto and G. De Nicolao, "Kernel selection in linear system identification part I: A Gaussian process perspective," in *Proc. of the Decision and Control and European Control Conference*, 2011.

[11] J. Wågberg, D. Zachariah, T. B. Schön, and P. Stoica, "Prediction performance after learning in Gaussian process regression," in *Proc. of the Int. Conference on Artificial Intelligence and Statistics*, 4 2017.

[12] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[13] C. M. Bishop *et al.*, *Pattern recognition and machine learning*, vol. 4. Springer New York, 2006.

[14] R. Frigola, Y. Chen, and C. Rasmussen, "Variational Gaussian process state-space models," in *Advances in Neural Information Processing Systems*, pp. 3680–3688, 2014.

[15] Y. Engel, S. Mannor, and R. Meir, "Reinforcement learning with Gaussian processes," in *Proc. of the 22nd International Conference on Machine learning*, 2005.

[16] J. E. Higgins and E. Polak, "Minimizing pseudoconvex functions on convex compact sets," *Journal of Optimization Theory and Applications*, vol. 65, no. 1, pp. 1–27, 1990.

[17] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear programming: theory and algorithms*. John Wiley & Sons, 2013.

[18] R. K. Sundaram, *A first course in optimization theory*. Cambridge university press, 1996.