Studies on Mechatronics Thesis

Gaussian Processes: Methods and Applications of Propagating Uncertainty

Jan-Philipp von Bassewitz

Elena Arcari

Prof. Dr. Melanie Zeilinger
Institute for Dynamic Systems and Control
Swiss Federal Institute of Technology Zurich (ETH)

2020

Institute for Dynamic Systems and Control
IDSC
Institut für Dynamische Systeme und Regelungstechnik

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Contents

List of Figures

Acknowledgements

Thanks to Elena Arcari for guiding me through this project and her patience when it came to questions on my side.
Thanks to Professor Zeilinger for giving me the opportunity to do my Studies on Mechatronics project at her lab.

Abstract

Gaussian Processes (GPs) are a powerful data-driven tool for approximating unknown or expensive to evaluate functions. GPs are specifically useful for time-series prediction which is for example used in Control to model dynamical systems [6]. GPs ability to express the uncertainty of the relative prediction renders them especially useful here. GPs come with the downside that their computational effort does not scale well with large amounts of data. They are additionally difficult to evaluate at random inputs as there is no closed-form solution for such evaluations [7]. Unfortunately the conventional ways of approximating the propagation of uncertainty underestimate it. Both of these drawbacks are reducing the applicability of GPs in Control. I will address this issue in the following report by summarizing a range of recent state-of-the-art solutions to these challenges.

1 Introduction

I will start by introducing the mathematics behind Gaussian Processes and the respective function-space and weight-space interpretations. After introducing the mathematics and giving an intuition for the assumptions underlying the formulation, the report will continue by focusing on the evaluation at stochastic inputs, which will result in an expression for a probability distribution of which no closed form solution exists [7, 4].

Solving this equation numerically is computationally infeasible. This issue will then be viewed from different perspectives. Firstly I will introduce the conventional approaches of approximating the stochastic evaluations seen in [5] and highlight their downsides. Using a Gaussian approximation and Taylor expansion [7] will result in relatively good approximations of the mean and variance of predictions. Alternatively approaching the problem of finding the probability distribution by sampling, will open the door for other approximations using both the function-space and weight-space interpretation [5]. Combining these will in the end result in satisfactory results, that significantly improve on the understatement of uncertainty and high computational burden [11]. In this report I will lay special focus on time-series prediction of the scalar autonomous system

$$f(x_{k+1}) = f(x_k) + w_k \tag{1.1}$$

where w is noise with $w \sim \mathcal{N}(0, \sigma_w^2)$. The summarized concepts can be formulated and used for more general cases as well though.

2 Gaussian Process

There are multiple ways of interpreting Gaussian Process models. In the following sections I want to highlight the function-space and weight-space interpretation which are both explained in more detail in [10]. GPs can be applied both in regression and classification problem and stand out because of their inherent ability to associate an uncertainty measurement with every prediction.

2.1 Function-Space Interpretation

In the function-space interpretation one can think of a GP as a distribution over functions, which is entirely defined by a mean function $m(x)$ and a covariance function $k(x, x')$. They are described by

$$m(x) = E[f(x)]$$

$$k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))]$$

and the GP is written as

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

The prior and overall properties of this distribution depend on the covariance function. Examples of using different kernels as the covariance function and drawing function samples from the prior distribution can be seen in Figure 1 and Figure 2.

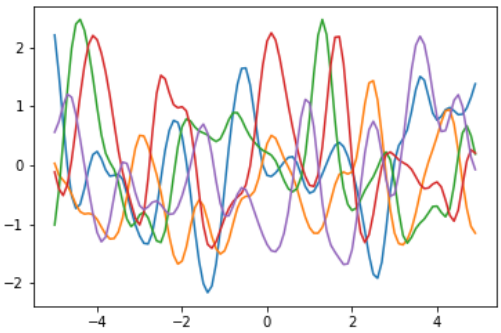


Figure 1: Function samples with squared-exponential kernel

The differences can obviously be large and the choice of $k(x, x')$ has to be made by assuming certain underlying properties of the system one would want to model.

To restrict the distribution over functions on functions that pass through known points the distribution can be conditioned on given data. This data consists of input values X and function evaluations at these inputs $\mathbf{f} = f(X)$. One defines the data as $\mathcal{D} = \{\mathbf{f}, X\}$. This will allow drawing functions from the GP that comply with this data. An example can be seen in Figure 3 where the known function evaluation are subject to some noise and $f(x)$ is unknown and to be found. To incorporate this knowledge we

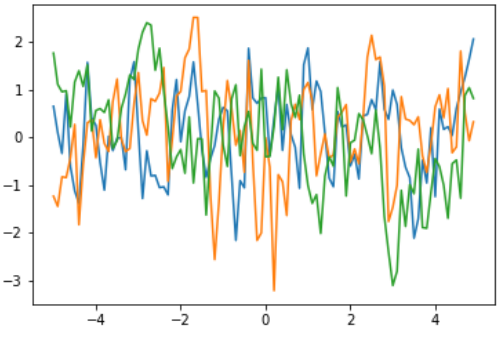


Figure 2: Function samples with Ornstein-Uhlenbeck kernel

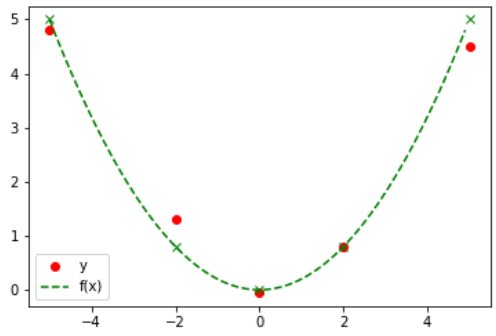


Figure 3: Noisy function evaluations y of $f(x) = x^2$ in red.

can write a joint Gaussian distribution between the given evaluations y and the unknown function values $f_* := f(X_*)$.

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X)+\sigma_n^2 I & K\left(X, X_*\right) \\ K\left(X_*, X\right) & K\left(X_*, X_*\right) \end{bmatrix}\right) \tag{2.1}$$

Here we inherently assume that function evaluations at *close* x-values should result in *close* function values $f(x)$. The distance is described

by the covariance function $k(x, x')$ that describes how *close* some points are. The term $+\sigma_n^2$ in (2.1) results in a covariance larger than zero for all the given function values, which implies that a certain noise is assumed in the data. Which means we assume a relation $y = f(x) + \epsilon$ where ϵ is sampled from some probability density function (pdf). Setting σ to zero will result in function samples that perfectly pass through the given data. By using the rules for finding the conditional distribution of a multivariate Gaussian one can find the mean $\mu(x)$ and variance $\Sigma(x)$ of that conditional at x_* to be

$$\mu(x_*) := E_{f_*}\left[\mathbf{f}_* | D\right] = k_*^T [K + \sigma^2 I]^{-1} \mathbf{y} \tag{2.2}$$

$$\Sigma(x_*) := V_{f_*}\left[\mathbf{f}_* | D\right] = k\left(x_*, x_*\right) - k_*^T \left(K + \sigma^2 I\right)^{-1} k_* \tag{2.3}$$

where $k_* = k(x_*, X)^T$ and $K = k(X, X)$ for notational simplicity. Plotting the resultant mean and variance for the example in Figure 3 with different kernel functions can be seen in Figure 4.

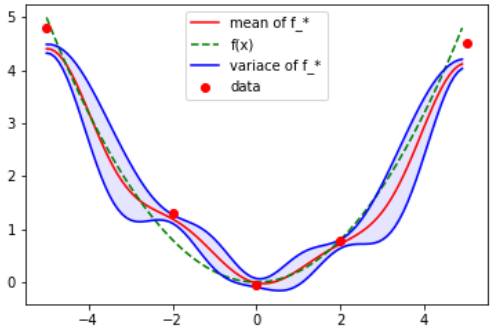
To quantify how good the fit actually is the *marginal likelihood* $p(y|X)$ can be used. It can be computed with the integral

$$p(\mathbf{y} \mid X) = \int p(\mathbf{y} \mid \mathbf{f}, X) p(\mathbf{f} \mid X) d\mathbf{f}$$

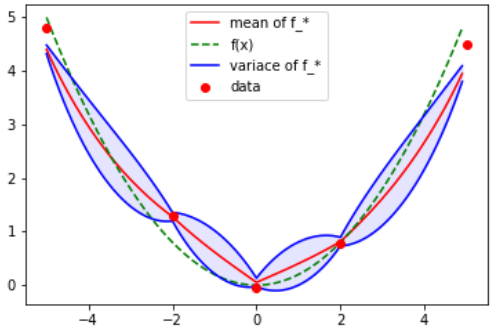
that quantifies how likely the noisy measurements y are as function evaluations at X with the current model. Taking the logarithm allows for easier optimization and results in

$$\log p(\mathbf{y} \mid X) = -\frac{1}{2} \mathbf{y}^\top \left(K + \sigma_n^2 I\right)^{-1} \mathbf{y} - \frac{1}{2} \log \left|K + \sigma_n^2 I\right| - \frac{n}{2} \log 2\pi$$

for the GP formulation.



(a) Squared-exponential kernel



(b) Ornstein-Uhlenbeck kernel

Figure 4: Posterior function distributions for different kernel functions

2.2 Weight-Space Interpretation

As the GP describes a distribution of functions one can write it as a possibly infinite linear combination of functions. This can be written as

$$f(x) = \phi(x)^T \theta \tag{2.4}$$

Sampling from the prior is easy here. It just requires sampling the required weights $\theta \sim \mathcal{N}(0, \Sigma_p)$. This is a property that will later in this report be of great use. The posterior can be drawn from the distribution

$$f_* \mid x_*, X, y \sim \mathcal{N}\left(\frac{1}{\sigma_n^2} \phi\left(x_*\right)^\top A^{-1} \Phi y, \phi\left(x_*\right)^\top A^{-1} \phi\left(x_*\right)\right). \tag{2.5}$$

where $\Phi = \phi(X)$ and $A = \sigma_n^{-2} \Phi \Phi^\top + \Sigma_p^{-1}$

3 GP at Stochastic Inputs

In this section I will introduce GP evaluations at random inputs as described in [7, 4]. This is important in real word applications. Sensors and actuators and therefore signals are mostly subject to noise. Assuming that the input x_* is a random variable its distribution could e.g. be describes by a Gaussian distribution $x_* \sim \mathcal{N}(\mu_{x*}, \sigma_{x*})$. The goal will be to find the distribution f_* given the distribution of x_* and the data $\mathcal{D}$. In the previous section the posterior distribution of f_* was introduced. Here we can write it as $p(f_*|x_*, \mathcal{D})$ where we have to assume x_* to be given. By integrating over the entire input distribution one finds

$$p\left(f_* \mid \mu_{x*}, \sigma_{x*}, \mathcal{D}\right) = \int p\left(f_* \mid x_*, \mathcal{D}\right) p\left(x_* \mid \mu_{x*}, \sigma_{x*}\right) dx_*. \tag{3.1}$$

Unfortunately the resultant distribution can only be evaluated numerically as no closed-form solution exists. In this section I will introduce the common ways of approximating (3.1). They can be split into multiple categories. Firstly I will introduce the ground-truth solution, followed by a procedure that uses the weight-space interpretation. Thirdly the distribution (3.1) can assumed to be Gaussian and the following mean and variance can be estimated with a Taylor expansion. Doing this jointly for the entire trajectory instead of each step iteratively will lead to decent approximation that takes the propagation of uncertainty into account.

3.1 Ground Truth Trajectory Sampling

A numerical scheme proposed in [5] for sampling from this distribution for the scalar autonomous system case (1.1) will be presented here. It will act as a benchmark to all the other sampling based approximations following later in the text. The analogue here to the joint distribution (2.1) is

$$\begin{bmatrix} x_1 \\ \vdots \\ x_{k+1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu\left(x_0\right) \\ \vdots \\ \mu\left(x_k\right) \end{bmatrix}, \begin{bmatrix} k\left(x_0, x_0\right) + \sigma^2 & \dots & k(x_0, x_k) \\ \vdots & \ddots & \vdots \\ k(x_k, x_0) & \dots & k\left(x_k, x_k\right) + \sigma^2 \end{bmatrix} \right). \tag{3.2}$$

For a first given value x_0 the next step can simply be sampled from $x_1 \sim \mathcal{N}(\mu(x_0), k(x_0, x_0) + \sigma^2)$ with $x_1 = \mu\left(x_0\right) + \sqrt{k\left(x_0, x_0\right) + \sigma^2}\ \tilde{w}_0$ where w_0 is drawn from a standard normal distribution. The joint distribution can then be grown one step and x_0 and x_1 sampled from it. This can be generalized to the step $k+1$ with

$$X_{1:k+1} = \mu\left(X_{0:k}\right) + \sqrt{k\left(X_{0:k}, X_{0:k}\right) + I\sigma^2}\tilde{W}_{0:k+1} \tag{3.3}$$

where $\tilde{W}_{0:k+1}$ is drawn from a multi dimensional standard normal distribution and $\sqrt{\cdot}$ is the Cholesky decomposition. This procedure is computationally very expensive as it scales cubically $\mathcal{O}(N^3)$ with the prediction horizon N. But this approach gives accurate samples from the distribution (3.1) as no simplifying assumptions were made.

3.2 Approximating Samples with Basis Functions

The weight-space interpretation of a GP described in [11] serves as an interesting alternative to the function-space interpretation and can in this case be used to approximate function samples from a posterior. Let $\phi_i(x)$ be a l-dimensional finite set of basis functions. The

GP approximation then takes the form

$$f(x) = \sum_{i=1}^{l} \theta_i \phi_i(x). \tag{3.4}$$

Inference is here done on the weight distribution of θ. Choosing a prior distribution as $\theta \sim \mathcal{N}(0, \Sigma_p)$ the posterior distribution $\theta|y \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$ of the weights is described with

$$\begin{aligned} \mu_\theta &= \left(\Phi^\top \Phi + \sigma^2 I\right)^{-1} \Phi^\top y \\ \Sigma_\theta &= \left(\Phi^\top \Phi + \sigma^2 I\right)^{-1} \sigma^2. \end{aligned} \tag{3.5}$$

where $\Phi = \phi(X)$. Drawing from the posterior is cheap as the computation of $\sqrt{\Sigma_\theta}$ scales with $\mathcal{O}(l^3)$. Unfortunately this approach also underestimates the propagation of uncertainty.

4 Gaussian Approximation with Taylor expansion

To simplify expression (3.3) the authors of paper [7] assume the distribution to be Gaussian and compute its mean and variance. By using the law of iterated expectation values and the law of iterative variances one gets

$$E_{\mathbf{f}_*}[\mathbf{f}_*] = E_{x_*}\left[E_{\mathbf{f}_*}[\mathbf{f}_*|x_*]\right] = E_{x_*}\left[\mu(x_*)\right] \tag{4.1}$$

$$\begin{aligned} V_{\mathbf{f}_*}[\mathbf{f}_*] = E_{x_*}\left[V_{\mathbf{f}_*}[\mathbf{f}_*|x_*]\right] + V_{x_*}\left[E_{\mathbf{f}_*}[\mathbf{f}_*|x_*]\right] = \\ E_{x_*}\left[\Sigma(x_*)\right] + V_{x_*}\left[\mu(x_*)\right]. \end{aligned} \tag{4.2}$$

4.1 Independence Assumption

Evaluating these expressions is still not trivial. Preceding to derive a simpler expression for the scalar autonomous case (1.1) with a first order Taylor expansion of $E[x_{k+1}]$ and a zeroth order Taylor expansion of $V[x_{k+1}]$ will yield

$$\begin{aligned} E[x_{k+1}] = E[\mu(x_k)] + E[w_k] \approx \\ E[\mu(\mu_{x_k}) + \nabla\mu(\mu_{x_k})(x_k - \mu_{x_k})] = \mu(\mu_{x_k}). \end{aligned} \tag{4.3}$$

$$\begin{aligned} V[x_{k+1}] = E[\Sigma(x_k)] + V[\mu(x_k)] + \sigma_w^2 \approx \\ \Sigma(\mu_{x_k}) + \sigma_w^2 \end{aligned} \tag{4.4}$$

These equations describe the mean and variance of each of the random variables x_i. This can then be jointly written to lead to a naive approximation of (3.3)

$$\begin{bmatrix} x_1 \\ \vdots \\ x_{k+1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(x_0) \\ \vdots \\ \mu(x_k) \end{bmatrix}, \begin{bmatrix} k(x_0,x_0)+\sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & k(x_k,x_k)+\sigma^2 \end{bmatrix} \right). \tag{4.5}$$

Higher order Taylor expansions can be used for more accuracy but will increase the computational cost. Sampling from this distribution can again be done with a Cholesky similarly to (3.3). The sampling is cheap, does not require the iterative approach and the predictive mean is accurate. The major downside is that it underestimates the uncertainty of times-steps past the first because the propagation of uncertainty is not accounted for as can be seen from the off-diagonal parts being all zero.

4.2 Dropping the Independence Assumption

The same idea of linearization applied to the sampling based approach (3.3) is presented in [5]. As mentioned the first sample will be accurate with $x_1 = \mu(x_0) + \sqrt{k(x_0,x_0)+\sigma^2}\,\tilde{w}_0$. The following iterations will be costly due to the recalculation of the Cholesky decomposition for every time step

though. So by jointly linearing $\mu(x)$ in

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \mu(x_0) \\ \mu(x_1) \end{bmatrix} + \sqrt{\begin{bmatrix} k(x_0,x_0)+\sigma^2 & k(x_0,x_1) \\ k(x_1,x_0) & k(x_1,x_1)+\sigma^2 \end{bmatrix}} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \tag{4.6}$$

around $\mu_1 = \mu(x_0)$ so that $\mu(x_1) = \mu(\mu_1) + \nabla\mu(\mu_1)(x_1 - \mu_1)$ one gets

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \approx \begin{bmatrix} \mu_1 \\ \mu(\mu_1) \end{bmatrix} + \begin{bmatrix} I & 0 \\ \nabla\mu(\mu_1) I \end{bmatrix} \sqrt{\begin{bmatrix} k(x_0,x_0)+\sigma^2 k(x_0,x_1) \\ k(x_1,x_0) & k(x_1,x_1)+\sigma^2 \end{bmatrix}} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \tag{4.7}$$

The samples x_1 and x_2 value can now be used to iteratively expand the procedure to further time-steps. Generalizing the notation for N time-steps leads to

$$X_{1:N} \approx M_{1:N} + A\sqrt{k(M_{0:N-1}, M_{0:N-1}) + I\sigma^2}\tilde{W}. \tag{4.8}$$

with $\tilde{W}$ being sampled from a N dimensional standard normal distribution, $M = [\mu_0, ..., \mu_N]^T$ and $A_{i,j} = \prod_{l=j}^{i-1} \nabla\mu(\mu_l)$. This approach does account for the propagating of uncertainty as can be seen in the structure of the covariance matrix which is not diagonal. It furthermore scales quadratically $\mathcal{O}(N^2)$ with the prediction horizon N compared to the ground truth sampling which scales cubically.

5 State-of-the-Art Approximations

In this section I will summarize the current state-of-the-art approximation of (3.1). A split of the posterior into the prior and an update as a function of the prior will allow for a combination of both the weight-space interpretation with its cheap evaluations of the prior and the accurate representation of the uncertainty with the function space interpretation. This will again be a sampling approach to the problem.

5.1 Splitting Posterior

To combine these advantages of the weight-space interpretation and the function-space interpretation paper [11] introduces a split of the posterior. According to Matheron's rule one can sample from a conditional distribution $a|b$ by first sampling the two random variables a and b from the joint distribution and then use the update for

$$(a \mid b = \beta) = a + \mathrm{Cov}(a,b)\,\mathrm{Cov}(b,b)^{-1}(\beta - b).$$

This can analogously be done for the posterior distribution of a GP with

$$\underbrace{f_*|y}_{\text{posterior}} = \underbrace{f_*}_{\text{prior}} + \underbrace{k_*^T(K+\sigma^2 I)^{-1}(y - f - \epsilon)}_{\text{update}}. \tag{5.1}$$

Substituting the prior f_* in (5.1) with the prior from the weight-space interpretation $\phi(x)^T\theta$ where θ is sampled from a normal distribution will reduce the computational burden and not result in less accuracy. Computing the Cholesky for the prior which scales with $\mathcal{O}(N^3)$ with prediction horizon N is not necessary anymore.

$$\underbrace{f_*|y}_{\text{posterior}} \approx \underbrace{\sum_{j=1}^{l} \phi_j(x_*)\theta_j}_{\text{prior}} + \underbrace{\sum_{i=1}^{N} v_i k(x_*, x_i)}_{\text{update}} \tag{5.2}$$

where v_i are the components of $V = (K(X,X) + \sigma^2 I)^{-1}(y - \Phi\theta - \epsilon)$. This presents a very potent approximation because it scales with $\mathcal{O}(N)$ and results in very accurate representations of the uncertainty which are almost identical to the ground truth. An additional benefit is that this expression can be differentiated which is not possible for the standard GP. This is due to the fact that the prior is a linear combination of functions.

6 Applications

Motivating the applicability of the described methods, I would like to shortly introduce a selection of research papers that build on this research.

The paper [2] on Thompson sampling uses GPs as a tool to approximate costly to evaluate or black-box functions for optimization purposes and makes use of some iterative methods described. [9] introduces the idea of splitting the posterior seen in (5.2) to facilitate a more scalable version of Thompson sampling that makes use of the cheap and accurate sampling.

On another note the papers [3, 8] further improve the methods of using GPs in the context of model predictive control in similar fashion to the discussed methods for the scalar autonomous case with success.

7 Discussion

To compare the discussed methods Figure 5 classifies them qualitatively with respect to the attributes *accuracy of uncertainty propagation* and *computational efficiency*. The ground-truth trajectory sampling scales cubicly but is highly accurate. The Gaussian approximation approach with Taylor expansion accounts for the propagation of uncertainty while only scaling quadratically. The computational complexity is further decreased with the independence assumption but at the cost of accuracy. The same is true for the basis function approximation that scales linearly but is again incaccurate. The posterior split method on the other hand combines an accurate representation of the uncertainty plus it only scales linearly with the prediction horizon.

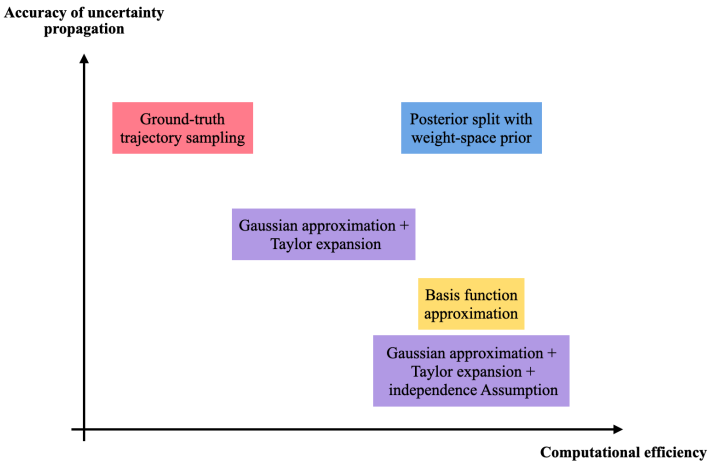


Figure 5: Qualitative comparison of methods

8 Conclusion

In conclusion there are many methods proposed in the literature to deal with the drawbacks of GPs, specifically lack of scalability and inaccurate approximations of uncertainty propagation. While they all improve on the naive trajectory sampling (4.5) they differ in computational cost and accuracy. The method that stands out is the combination of function-space and weight-space interpretation (5.2), as it both preserves the uncertainty propagation and scales only linearly with the prediction horizon.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[2] E. Bradford, A. M. Schweidtmann, and A. Lapkin. Efficient multiobjective optimization employing gaussian processes, spectral sampling and a genetic algorithm, 2020.

[3] L. I. Eric Bradford and E. A. del Rio-Chanona. Nonlinear model predictive control with explicit back-offs for gaussian process state space models, 2017.

[4] A. Girard, C. Rasmussen, J. Q. n. Candela, and R. Murray-Smith. Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, pages 545–552. MIT Press, 2003.

[5] L. Hewing, E. Arcari, L. P. Fröhlich, and M. N. Zeilinger. On simulation and trajectory prediction with gaussian process dynamics, 2020.

[6] L. Hewing, J. Kabzan, and M. N. Zeilinger. Cautious model predictive control using gaussian process regression. *IEEE Transactions on Control Systems Technology*, 28(6):2736–2743, Nov 2020.

[7] J. Quiñonero-Candela, A. Girard, and C. Rasmussen. Prediction at an uncertain input for gaussian processes and relevance vector machines - application to multiple-step ahead time-series forecasting. Technical Report IMM-2003-18, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2003.

[8] J. Umlauft, T. Beckers, and S. Hirche. A scenario-based optimal control approach for gaussian process state space models. 06 2018.

[9] S. Vakili, V. Picheny, and A. Artemev. Scalable thompson sampling using sparse gaussian process models, 2020.

[10] C. E. R. . C. K. I. Williams. *Gaussian Processes for Machine Learning*. the MIT Press, 2006.

[11] J. T. Wilson, V. Borovitskiy, A. Terenin, P. Mostowsky, and M. P. Deisenroth. Efficiently sampling functions from gaussian process posteriors, 2020.