

GP3: A Sampling-based Analysis Framework for Gaussian Processes

Armin Lederer Markus Kessler Sandra Hirche

*Chair of Information-oriented Control (ITR), Department of
Electrical and Computer Engineering, Technical University of Munich,
Germany (e-mail: armin.lederer, markus.kessler, hirche@tum.de).*

Abstract: Although machine learning is increasingly applied in control approaches, only few methods guarantee certifiable safety, which is necessary for real world applications. These approaches typically rely on well-understood learning algorithms, which allow formal theoretical analysis. Gaussian process regression is a prominent example among those methods, which attracts growing attention due to its strong Bayesian foundations. Even though many problems regarding the analysis of Gaussian processes have a similar structure, specific approaches are typically tailored for them individually, without strong focus on computational efficiency. Thereby, the practical applicability and performance of these approaches is limited. In order to overcome this issue, we propose a novel framework called GP3, general purpose computation on graphics processing units for Gaussian processes, which allows to solve many of the existing problems efficiently. By employing interval analysis, local Lipschitz constants are computed in order to extend properties verified on a grid to continuous state spaces. Since the computation is completely parallelizable, the computational benefits of GPU processing are exploited in combination with multi-resolution sampling in order to allow high resolution analysis.

Keywords: Machine learning, Learning for control, Bayesian methods, Gaussian processes, Stability of nonlinear systems, Sampling-based analysis, Learning systems

1. INTRODUCTION

Machine learning is increasingly applied in control approaches, where first principle models are not available or expensive to obtain due to the complexity of systems. In order to apply learning based control approaches to real world applications, it is crucial to certify their safety using theoretically rigorous methods. Although there exists a wide variety of machine learning methods, this theoretical analysis is often difficult, such that safe control approaches typically rely on a few, well-understood learning methods.

Gaussian process (GP) regression is such a machine learning method which bases on solid Bayesian foundations. Due to its inherent bias-variance trade-off it allows efficient generalization from few training data, which makes it an appealing method both for control practitioners and theoreticians, and has lead to its increasing use in control. For the control theoretic analysis of Gaussian processes, several approaches have been developed. In (Beckers and Hirche, 2016a,b) stochastic stability and equilibria of Gaussian processes with certain covariance kernels are investigated analytically. Employing numerical quadrature, an efficient method to analyze stability of the posterior mean function of a Gaussian process state space model with squared exponential kernel is developed in (Vinogradska et al., 2017). A statistical learning error analysis of learned MPC control laws is proposed in (Hertneck et al., 2018) to ensure closed-loop stability. By

extending the Lyapunov stability conditions on a grid to a continuous state space using Lipschitz continuity, the region of attraction of a Gaussian process state space model is learned in (Berkenkamp et al., 2016). A similar approach is employed to determine the region of attraction of a nonlinear system based on the learned infinite horizon cost function in (Lederer and Hirche, 2019). The Lipschitz constants of the posterior mean function used in these approaches are also of interest themselves, e.g., for computing uniform regression error bounds (Lederer et al., 2019) or in batch parallelization of Bayesian optimization in (González et al., 2016).

Despite of the similarity of many of these problems, each one is based on a separate analysis method, which is typically not optimized for computational efficiency. In order to overcome this issue, we propose a novel framework for the analysis of Gaussian process mean functions called GP3: **General Purpose** computation on **Graphics Processing** units for **Gaussian Processes**. By defining a common problem formulation for many problems, we can employ interval analysis to derive local Lipschitz constants on hyperrectangles covering the region of interest. These local Lipschitz constants allow to extend verified properties on a discrete grid to the whole region of interest such that multi-resolution sampling can be used for efficient analysis. As the Lipschitz constants can be computed independently for each hyperrectangle, the method is parallelized using general purpose graphics processing units in order to exploit the full computational power of modern hardware. We demonstrate the flexibility and efficiency of

A. L. gratefully acknowledges financial support from the German Academic Scholarship Foundation.

the GP3 framework by applying it to a region of attraction estimation and a Lipschitz constant bounding problem.

The remaining paper is structured as follows. In Section 2 we define the general property analysis problem. The theoretical background on Gaussian process regression and interval analysis is presented in Section 3. The theoretical foundations of the GP3 framework are explained in Section 4, before it is evaluated in simulations in Section 5.

2. PROBLEM STATEMENT

Consider the posterior mean function¹ $\mu : \mathbb{R}^d \rightarrow \mathbb{R}$ of a Gaussian process with continuous covariance kernel $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, a continuous function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ for comparison and a continuous state transformation $\mathbf{f} : \mathbb{R}^d \rightarrow \mathbb{R}^d$. An abstract problem that finds many practical applications is to find bounds $\epsilon_1, \epsilon_2 \in \mathbb{R}_+$ on the difference

$$-\epsilon_1 \leq g(\mathbf{f}(\mathbf{x})) - \mu(\mathbf{x}) \leq \epsilon_2 \quad \forall \mathbf{x} \in \mathbb{X} \quad (1)$$

on a compact set $\mathbb{X} \subset \mathbb{R}^d$. Various choices for the functions $\mathbf{f}(\cdot), g(\cdot)$ immediately come to mind: if we choose $g(\cdot)$ as the function generating the training data of the Gaussian process and set $\mathbf{f}(\mathbf{x}) = \mathbf{x}$ we can determine the maximum learning error. Furthermore, $g(\cdot)$ can be the Gaussian process mean $\mu(\cdot)$ itself and $\mathbf{f}(\cdot)$ can be defined such that it returns the closest point in a discrete set. Thereby the variation of the mean function with respect to the discrete set is analyzed. Finally, we can consider $\mathbf{f}(\cdot)$ as autonomous, discrete-time dynamics and choose $g(\cdot) = \mu(\cdot)$, such that we can immediately investigate if the mean function $\mu(\cdot)$ satisfies the second condition of Lyapunov's theorem. Although this problem formulation offers such a high flexibility, it allows a straight forward, uniform treatment through interval analysis, which we exploit in the following sections.

3. THEORETICAL BACKGROUND

3.1 Gaussian Process Regression

Gaussian process regression is a supervised machine learning method, which is frequently applied in control and system identification due to its Bayesian foundations. The Gaussian process distribution assigns to any finite subset $\{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subset \mathbb{X}$ from a continuous input domain $\mathbb{X} \subset \mathbb{R}^d$ a joint Gaussian distribution (Rasmussen and Williams, 2006). It is completely defined through the mean function $m : \mathbb{R}^d \rightarrow \mathbb{R}$ and the covariance function $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. Although prior information like approximate models can be incorporated as mean function $m(\cdot)$ into GP regression, such knowledge is often not available, such that the mean function $m(\cdot)$ is usually set to zero. We also assume this in the following. In contrast, a wide variety of different covariance functions $k(\cdot, \cdot)$ is applied in GP regression to encode prior information such as smoothness, periodicity and stationarity. Frequently used covariance functions are the squared exponential kernel

$$k_{\text{SE}}(r) = \sigma_f^2 \exp\left(-\frac{1}{2}r^2\right) \quad (2)$$

¹ Lower/upper case bold symbols denote vectors/matrices, \mathbb{R}_+ denotes all real positive numbers, \mathbf{I}_n the $n \times n$ identity matrix and $\|\cdot\|$ the Euclidean norm.

and the Matérn class kernels

$$k_{m+\frac{1}{2}}(r) = \sigma_f^2 p_m(r) \exp\left(-\sqrt{(2m+1)r}\right) \quad (3)$$

with signal variance $\sigma_f^2 \in \mathbb{R}_+$, $p_m(\cdot)$ a polynomial of order $m \in \mathbb{N}$ and automatic relevance determination distance (Neal, 1996)

$$r = \sqrt{\sum_{i=1}^d \frac{(x_i - x'_i)^2}{l_i^2}} \quad (4)$$

with length scales $l_i \in \mathbb{R}_+$. Matérn class kernels are commonly used with $m = 1$ or $m = 2$ which results in the polynomials

$$p_1(r) = 1 + \sqrt{3}r \quad (5)$$

$$p_2(r) = \left(1 + \sqrt{5}r + \frac{5}{3}r^2\right). \quad (6)$$

A major reason for their popularity is the fact that a GP with either of these covariance functions is a universal approximator (Steinwart, 2001), i.e., any continuous function can be approximated with arbitrary precision.

GP regression is based on the assumption that the training data set $\mathbb{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^N$ is generated through noisy observations of a function $h : \mathbb{R}^d \rightarrow \mathbb{R}$, i.e.,

$$y^{(i)} = h(\mathbf{x}^{(i)}) + \omega^{(i)}, \quad (7)$$

where $\omega^{(i)} \sim \mathcal{N}(0, \sigma_n^2)$ are i.i.d. random variables. By conditioning the prior joint Gaussian distribution of a prediction $h(\mathbf{x})$ and the training outputs $y^{(i)}$ on the training data \mathbb{D} , we obtain the predictive mean

$$\mu(\mathbf{x}) = \mathbb{E}[h(\mathbf{x})|\mathbb{D}, \mathbf{x}] = \mathbf{k}^T(\mathbf{x})\boldsymbol{\lambda}, \quad (8)$$

where

$$\boldsymbol{\lambda} = (\mathbf{K} + \sigma_n^2 \mathbf{I}_N)^{-1} \mathbf{y} \quad (9)$$

and the data covariance matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$ and the covariance vector $\mathbf{k}(\mathbf{x}) \in \mathbb{R}^N$ are defined through $K_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ and $k_i(\mathbf{x}) = k(\mathbf{x}^{(i)}, \mathbf{x})$. The observation noise variance σ_n^2 , the signal variance σ_f^2 and the length scales l_i are considered hyperparameters of the GP regression and can be determined by maximization of the log-likelihood (Rasmussen and Williams, 2006).

3.2 Interval Analysis for Property Analysis

Interval analysis is a method to approach the problem of calculating bounds of functions. Instead of operating on exact values, interval analysis uses real compact intervals $[a] = [\underline{a}, \bar{a}] = \{x \in \mathbb{R} | \underline{a} \leq x \leq \bar{a}\}$. Basic mathematical operations for intervals are defined as

$$[a] + [b] = [\underline{a} + \underline{b}, \bar{a} + \bar{b}], \quad (10)$$

$$[a] - [b] = [\underline{a} - \bar{b}, \bar{a} - \underline{b}], \quad (11)$$

$$[a] \cdot [b] = [\min\{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\}, \max\{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\}]. \quad (12)$$

A more thorough introduction into operations on intervals can be found, e.g., in (Alefeld and Mayer, 2000).

Based on this interval arithmetic, it is possible to use intervals as inputs to functions $f : \mathbb{R} \rightarrow \mathbb{R}$ and calculate output intervals $[y] = f([x])$, where y serves as a lower bound and \bar{y} as an upper bound of the function values on the input interval. It is straight forward to adapt this approach to higher dimensional functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ by considering so called *hyperrectangles* instead of intervals. A hyperrectangle is completely defined by its center \mathbf{c}

and length parameter \mathbf{b} , such that it defines the multi-dimensional interval $[\mathbf{c} - \mathbf{b}, \mathbf{c} + \mathbf{b}]$ with edge widths $2\mathbf{b}$. Using a grid of hyperrectangles, interval analysis allows to efficiently expand the validity of (1) on the hyperrectangle centers to the area covered by the hyperrectangles. Using multi-resolution grids as proposed, e.g., in (Bobiti and Lazar, 2018), this enables efficient determination of valid constants ϵ_1 and ϵ_2 in (1).

4. SAMPLING-BASED ANALYSIS OF GAUSSIAN PROCESSES

Due to the strong nonlinearity of typical Gaussian process mean functions, standard interval operations are not directly applicable to their analysis. Therefore, we develop an efficient multi-resolution sampling algorithm for the analysis of Gaussian processes in Section 4.1. As this algorithm requires upper and lower bounds for the derivative of covariance kernels depending on the training data, we investigate such bounds for squared exponential and Matérn class kernels in Section 4.2.

4.1 Multi-resolution Analysis of Gaussian Processes

Gaussian processes exhibit a strongly nonlinear mean function in general. In order to efficiently analyze their mean functions we develop a multi-resolution sampling algorithm in this section. Exploiting Lipschitz continuity we derive a theorem to calculate the bounds ϵ_1 and ϵ_2 in (1) inside hyperrectangles. In order to evaluate the bound, we determine a local Lipschitz constant of the GP on a hyperrectangle. Due to the sampling structure, the analysis of hyperrectangles can be efficiently parallelized using GPUs in each grid refinement iteration.

The basis of our approach lies in the independent analysis of the bounds ϵ_1 and ϵ_2 on hyperrectangles. This analysis is founded on the following theorem, which relies on Lipschitz continuity of all involved functions.

Theorem 1. Consider a function $\mathbf{f}(\cdot)$ with local Lipschitz constant L_f and a posterior mean function $\mu(\cdot)$ of a Gaussian process with local Lipschitz constant L_μ on a hyperrectangle with center \mathbf{c} and edge widths $2\mathbf{b}$. Furthermore, assume the function $g(\cdot)$ has a local Lipschitz constant L_g on the hyperrectangle with center $\mathbf{f}(\mathbf{c})$ and edge widths $2L_f\mathbf{b}$. Then, (1) holds with

$$\epsilon_1 = g(\mathbf{f}(\mathbf{c})) - \mu(\mathbf{c}) - (L_f L_g + L_\mu)\|\mathbf{b}\| \quad (13)$$

$$\epsilon_2 = g(\mathbf{f}(\mathbf{c})) - \mu(\mathbf{c}) + (L_f L_g + L_\mu)\|\mathbf{b}\| \quad (14)$$

for all $\|\mathbf{x} - \mathbf{c}\| \leq \mathbf{b}$, where the absolute values and the comparison are performed element-wise.

Proof. Due to Lipschitz continuity of $\mathbf{f}(\cdot)$ and $g(\cdot)$ it follows that

$$|g(\mathbf{f}(\mathbf{x})) - g(\mathbf{f}(\mathbf{c}))| \leq L_f L_g \|\mathbf{b}\| \quad \forall \mathbf{x} : \|\mathbf{x} - \mathbf{c}\| \leq \|\mathbf{b}\|.$$

Furthermore, Lipschitz continuity of the GP yields

$$|\mu(\mathbf{x}) - \mu(\mathbf{c})| \leq L_\mu \|\mathbf{b}\|.$$

Applying the triangle inequality we finally obtain the bounds (13) and (14).

The application of this theorem crucially relies on the Lipschitz constant of the GP mean function. The computation of this Lipschitz constant on a hyperrectangle can be performed independently for each hyperrectangle

and depends merely on the parameter vector $\boldsymbol{\lambda}$, the kernel function $k(\cdot, \cdot)$ and the data points $\mathbf{x}^{(i)}$, $i = 1, \dots, N$, as shown in the following theorem.

Theorem 2. Consider the posterior mean function $\mu(\cdot) = \mathbf{k}^T(\cdot)\boldsymbol{\lambda}$ of a GP with covariance kernel $k(\cdot, \cdot)$ on a hyperrectangle with center \mathbf{x} and length $2\mathbf{b}$. Let $\mathbf{L}_{k,i}^{\partial j}$ denote a vector of partial derivative bounds of the i -th element of $\mathbf{k}(\cdot)$ with respect to x_j . Then, a local Lipschitz constant of the mean on the hyperrectangle is given by

$$L_\mu(\mathbf{x}, \mathbf{b}) = \sqrt{\sum_{j=1}^d \max \left\{ \left(\sum_{i=1}^N \mathbf{R}(\lambda_i) \lambda_i \mathbf{L}_{k,i}^{\partial j} \right)^2 \right\}}, \quad (15)$$

where

$$\mathbf{R}(\lambda_i) = \begin{cases} \mathbf{I}_2 & \lambda_i > 0 \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & \lambda_i \leq 0. \end{cases} \quad (16)$$

Proof. Due to the scalar product of parameter vector $\boldsymbol{\lambda}$ and the kernel vector $\mathbf{k}(\mathbf{x})$ in (8), the partial derivatives of the mean function are given by

$$\frac{\partial}{\partial x_j} \mu(\mathbf{x}) = \left[\frac{\partial}{\partial x_j} k_1(\mathbf{x}) \dots \frac{\partial}{\partial x_j} k_N(\mathbf{x}) \right] \boldsymbol{\lambda}.$$

The vectors $\mathbf{L}_{k,i}^{\partial j}$ contain upper and lower bound on the partial derivatives $\frac{\partial}{\partial x_j} k_i(\mathbf{x})$ in the first and second row, respectively. By multiplying these vectors with the matrix $\mathbf{R}(\lambda_i)$, the order of elements is changed if λ_i is negative, such that upper derivative bounds are multiplied with positive λ_i s and lower bounds with negative ones in the first row of the resulting vector. The multiplication is performed in the inverse combination in the second row. By summing up each row, upper and lower bounds on the partial derivatives are obtained. We take the maximum squared value of these two rows as squared Lipschitz constant in the j -th direction and finally, calculate the overall Lipschitz $L_\mu(\mathbf{x}, \mathbf{b})$ constant by taking the square root of the sum of squared Lipschitz constants in all directions.

Theorem 1 and Theorem 2 allow for a straightforward implementation and integration in a multi-resolution, GPU parallelized sampling algorithm, which is depicted in Algorithm 1. Initially, a sampling grid is created by dividing the analyzed region \mathbb{X} into M hyperrectangles in line 2, where the centers \mathbf{c}_i are concatenated in \mathbf{C} and edge-lengths \mathbf{b}_i are concatenated in \mathbf{B} . Generally, any initial sampled grid is possible for this step, since our approach does not crucially depend on it. The grid is refined in line 6 by using an indication value s_i , which divides hyperrectangles that must be further refined for $s_i = \mathbf{false}$ from those that require no further refinement. For the refinement procedure itself, several ways are possible in general. A simple method is to resample the grid with smaller hyperrectangle sizes \mathbf{B} and skip the calculation where $s_i = \mathbf{true}$ already. Another example for a refinement method is the 2-refinement used in (Bobiti and Lazar, 2018), where hyperrectangles with $s_i = \mathbf{false}$ are divided into 2^d smaller hyperrectangles with new edge widths $\frac{1}{2}\mathbf{b}_i$.

The computation of the bounds ϵ_1, ϵ_2 based on Theorems 1 and 2 takes place inside the parallelized For (**ParFor**) loop of line 7. By shifting the execution to a highly parallel GPU, the computation time of the algorithm can

Algorithm 1 Multi-resolution bound calculation

```

1: function BOUNDS( $g(\cdot), \mu(\cdot), \mathbb{X}, M, \bar{\epsilon}_1(\cdot), \bar{\epsilon}_2(\cdot), b_{\min}$ )
2:   [ $\mathbf{C}; \mathbf{B}$ ]  $\leftarrow$  SAMPLINGGRID( $\mathbb{X}, M$ )
3:    $\epsilon_1, \epsilon_2 \leftarrow$  zeros[ $\text{len}(\mathbf{C})$ ]
4:    $\mathbf{s} \leftarrow$  false[ $\text{len}(\mathbf{C})$ ]
5:   while any( $s_i$ ) = false do
6:     [ $\mathbf{C}; \mathbf{B}$ ]  $\leftarrow$  REFINEGRID( $\mathbf{C}; \mathbf{B}; s_i$ )
7:     ParFor  $i = 1$  to  $\text{len}(\mathbf{C})$  do
8:       Calculate  $L_\mu(\mathbf{c}_i, \mathbf{b}_i)$  using Theorem 2
9:       Calculate  $\epsilon_{1,i}, \epsilon_{2,i}$  based on Theorem 1
10:       $s_i \leftarrow (\epsilon_{1,i} \leq \bar{\epsilon}_1(\mathbf{c}_i) \wedge \epsilon_{2,i} \leq \bar{\epsilon}_2(\mathbf{c}_i)) \vee$ 
11:        ( $\|\mathbf{b}_i\| \leq b_{\min}$ )
12:     end
13:   return  $\mathbf{C}, \epsilon_1, \epsilon_2$ 

```

be significantly reduced, which is an important feature of this approach. The termination of the refinement process of the algorithm is ensured by setting $s_i = \text{true}$, once the bounds ϵ_1, ϵ_2 inside the hyperrectangle satisfy pre-defined desired bounding functions $\bar{\epsilon}_1(\mathbf{c}_i)$ and $\bar{\epsilon}_2(\mathbf{c}_i)$ or the size $\|\mathbf{b}_i\|$ of the hyperrectangle falls below a specified minimum size b_{\min} .

4.2 Derivative Intervals of Kernels

Typical covariance kernels are highly nonlinear functions themselves, which complicates the interval analysis of their derivatives. However, many kernels exhibit a structure such that derivatives are monotonous on large intervals. We exploit this behavior for deriving upper and lower bounds for squared exponential and Matérn class kernels in the following theorem.

Theorem 3. Consider a multivariate squared exponential or Matérn class kernel $k(\cdot)$ with length scale \mathbf{l} on a hyperrectangle with center \mathbf{x} and edge lengths $2\mathbf{b}$. Then, the derivative bounds for a training point $\mathbf{x}^{(i)}$ with respect to the j -th component is given by

$$\mathbf{L}_{k,i}^{\partial j} = \mathbf{T}(x_j^{(i)} - x_j) \boldsymbol{\kappa}(\mathbf{x}^{(i)} - \mathbf{x}) \quad (17)$$

where

$$\boldsymbol{\kappa}(\boldsymbol{\Delta}) = \begin{cases} \begin{bmatrix} -\frac{\partial}{\partial \Delta_j} k(\rho_j(\Delta - \bar{\tau}, |\Delta_j| - b_j)) \\ -\frac{\partial}{\partial \Delta_j} k(\rho_j(\Delta - \underline{\tau}, |\Delta_j| + b_j)) \end{bmatrix} & |\Delta_j| > b_j + \tilde{l}_j \\ \begin{bmatrix} -\frac{\partial}{\partial \Delta_j} k(\rho_j(\Delta - \bar{\tau}, |\Delta_j| + b_j)) \\ -\frac{\partial}{\partial \Delta_j} k(\rho_j(\Delta - \underline{\tau}, |\Delta_j| - b_j)) \end{bmatrix} & |\Delta_j| < \tilde{l}_j - b_j \\ \begin{bmatrix} -\frac{\partial}{\partial \Delta_j} k(\rho_j(\Delta - \bar{\tau}, \tilde{l}_j)) \\ \min \left\{ -\frac{\partial}{\partial \Delta_j} k(\rho_j(\Delta - \underline{\tau}, |\Delta_j| \pm b_j)) \right\} \end{bmatrix} & \|\Delta_j - \tilde{l}_j\| \leq b_j \end{cases} \quad (18)$$

$$\mathbf{T}(\Delta_j) = \begin{cases} \mathbf{I}_2 & \Delta_j > 0 \\ \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} & \Delta_j \leq 0 \end{cases}, \quad (19)$$

with modified weighted distance $\rho_j(\boldsymbol{\Delta}, \tilde{\Delta})$, maximum point \tilde{l}_j , maximum distance $\bar{\tau}$, minimum distance $\underline{\tau}$ and index set \mathbb{I}_j defined as follows

$$\rho_j(\boldsymbol{\Delta}, \tilde{\Delta}) = \sqrt{\frac{\tilde{\Delta}^2}{l_j^2} + \sum_{i \in \mathbb{I}_j} \frac{\Delta_i^2}{l_i^2}} \quad (20)$$

$$\tilde{l}_j = \begin{cases} l_j & \text{squared exponential kernel} \\ \frac{l_j}{\sqrt{3}} & \text{Matérn kernel with } m = 1 \\ \frac{5 + \sqrt{5}}{10} l_j & \text{Matérn kernel with } m = 2 \end{cases} \quad (21)$$

$$\bar{\tau} = \min \left\{ \|\mathbf{b}_{\neq j}\|, \|\mathbf{x}_{\neq j}^{(i)} - \mathbf{x}_{\neq j}\| \right\} \frac{\mathbf{x}^{(i)} - \mathbf{x}}{\|\mathbf{x}_{\neq j}^{(i)} - \mathbf{x}_{\neq j}\|} \quad (22)$$

$$\underline{\tau} = \|\mathbf{b}_{\neq j}\| \frac{\mathbf{x} - \mathbf{x}^{(i)}}{\|\mathbf{x}_{\neq j}^{(i)} - \mathbf{x}_{\neq j}\|} \quad (23)$$

$$\mathbb{I}_j = \{1, \dots, j-1, j+1, \dots, d\}. \quad (24)$$

Proof. We start this proof by first showing that (17) holds for the squared exponential before we highlight the differences in the proof for Matérn kernels. Due to the exponent product rule and the automatic relevance determination (4), the derivative of the multivariate squared exponential kernel can be split into

$$\frac{\partial}{\partial x_j} k(\mathbf{x}^{(i)}, \mathbf{x}) = -\frac{\partial}{\partial \Delta_j} k\left(\frac{|\Delta_j|}{l_j}\right) k\left(\sqrt{\sum_{i \in \mathbb{I}_j} \frac{\Delta_j^2}{l_i^2}}\right),$$

where $\boldsymbol{\Delta} = \mathbf{x}^{(i)} - \mathbf{x}$. This separation allows to determine interval bounds for both factors independently and combine the maximizers or minimizers, respectively, with the help of $\rho_j(\boldsymbol{\Delta}, \Delta_j)$ based on (12). Due to odd symmetry of the univariate derivative, it is sufficient to derive only bounds for the positive real line and obtain the interval bounds for the negative real line by multiplication with $\mathbf{T}(\Delta_j)$. The kernel derivative is monotonous on the intervals $[0, l_j]$ and $[l_j, \infty]$ with a maximum at l_j . Therefore, we obtain three intervals with maximizers ν

$$\begin{aligned} |\Delta_j| > b_j + l_j &: \nu = |\Delta_j| - b_j \\ |\Delta_j| < l_j - b_j &: \nu = |\Delta_j| + b_j \\ \|\Delta_j - l_j\| \leq b_j &: \nu = l_j \end{aligned}$$

which can be analogously obtained for the minimizers. The multivariate kernel is trivially maximized by considering the minimal distance to the training point, which is given by $\boldsymbol{\Delta} - \bar{\tau}$, while it is minimized by $\boldsymbol{\Delta} - \underline{\tau}$. Therefore, (17) provides an upper and lower bound for the derivative of the squared exponential kernel on the hyperrectangle and we proceed with Matérn class kernels. Although the derivatives of Matérn class kernels cannot be separated as the squared exponential kernel, they exhibit a similar behavior in that they are monotonous in the non-derived directions $i \neq j$. Therefore, we can choose the same maximizer and minimizer in these directions as for the squared exponential kernel. Furthermore, the derivatives of Matérn class kernels exhibit the same odd symmetry in the derived direction j and have two monotonous intervals on the positive real line with maximum at $\tilde{l}_j = l_j/\sqrt{3}$ for $m = 1$ and $\tilde{l}_j = (5 + \sqrt{5})l_j/10$. Therefore, we can upper and lower bound the derivative for Matérn class kernels on the positive real line using (19), which concludes the proof.

Although this theorem might appear rather complicated, it allows a straightforward implementation with few conditional operators, which is beneficial for GPU paralleliza-

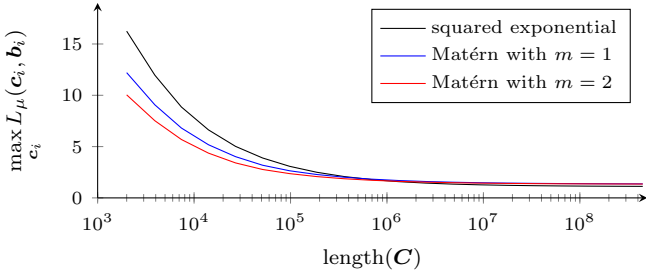


Fig. 1. Decrease rate of Lipschitz constants for increasing number of analyzed hyperrectangles
 tion (Owens et al., 2008). Therefore, it allows efficient, parallelized analysis of Gaussian process mean functions in combination with Theorems 1 and 2.

5. NUMERICAL EVALUATION

5.1 Efficiency of the GP3 Framework

In order to demonstrate the advantages of the GP3 framework² over a global Lipschitz constant computation and a CPU parallelized implementation, we compare these approaches on the example proposed in (Lederer et al., 2019). A crucial step therein is the derivation of a Lipschitz constant for a posterior mean representing the function

$$f(\mathbf{x}) = 1 - \sin(x_1) + \frac{1}{1 + \exp(-x_2)}. \quad (26)$$

The Gaussian process is trained with 100 samples which are uniformly spaced over the analyzed region $\mathbb{X} = [-6, 4] \times [-4, 4]$. We train three GPs with a squared exponential and Matérn kernels with $m = 1$ and $m = 2$. The hyperparameters obtained via log-likelihood maximization for each of these GPs are depicted in Table 1.

We compute the Lipschitz constants of the posterior mean functions using the GP3 framework for different numbers of hyperrectangles and the naive, global Lipschitz constant proposed in (Lederer et al., 2019). The Lipschitz constants obtained from the GP3 approach are depicted in Figure 1. Since the naive approach yields Lipschitz constants of 343.62, 245.66 and 268.20 for the squared exponential and Matérn kernel with $m = 1$ and $m = 2$, respectively, the corresponding constant curves are not displayed in the figure. In contrast, the Lipschitz constants obtained by the GP3 approach can provide Lipschitz constants smaller than 20 with less than 2000 hyperrectangles. Furthermore, with approximately 10^6 hyperrectangles the Lipschitz constant has almost converged to a constant value.

Additionally, we compare the time required to compute the Lipschitz constants on the hyperrectangles with a GPU and CPU parallelization. We compute the Lipschitz constants on a system with a NVIDIA TITAN V GPU, which has 2:1 single to double precision unit ratio, two

² Code is available at <https://gitlab.lrz.de/alederer/gp3>

kernel	σ_n^2	σ_f^2	\mathbf{l}
squared exponential	0.1	0.956	$[1.762 \ 5.537]^T$
Matérn with $m = 1$	0.1	1.274	$[3.755 \ 15.052]^T$
Matérn with $m = 2$	0.1	1.012	$[2.333 \ 8.496]^T$

Table 1. Hyperparameters of the different covariance kernels

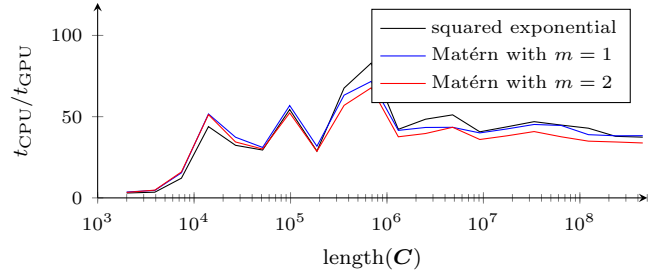


Fig. 2. Average speedup of GPU over CPU parallelization over the number of analyzed hyperrectangles

AMD EPYC 16-Core CPUs and 1TB RAM. The speedup of the GPU compared to the CPU parallelization averaged over 10 runs of both implementations are displayed for the three different covariance kernels in Figure 2. Although we parallelize the Lipschitz constant computation with 64 threads on the CPU, the GPU parallelization achieves a speedup of at least 30 for large numbers of hyperrectangles. Merely at low numbers of hyperrectangles the speedup can be in the single digit region due to the computational overhead of GPU computation. Therefore, GP3 allows to exploit the advantages of modern general purpose GPU computing to achieve low computation times, while requiring a reasonable amount of hyperrectangles for convergence of the obtained Lipschitz constant.

5.2 Region of Attraction for Power Systems

As an example for a problem of the form (1), we analyze the region of attraction of a nonlinear autonomous system. We consider the single machine infinite bus system (Münz and Romeres, 2013)

$$m_1 \ddot{\phi}_1 + d_1 \dot{\phi}_1 = -a_{12}(\sin(\theta_1 + \phi_1) - \sin(\theta_1)), \quad (27)$$

which models a synchronous machine with inertia m_1 , damping $d_1 = 20$ and steady state phase $\theta_1 = \arcsin(0.05)$ as generator bus connected to an infinite bus with $\theta_2 = \phi_2 = 0$, such that $a_{12} = 10$ is the product of the susceptance between both buses and the root mean square voltages u_1 and u_2 at bus 1 and 2, respectively. For analyzing the regions, where the system is discrete-time stable, we consider the cost of finite time trajectories proposed in (Bobiti and Lazar, 2018) as Lyapunov function, i.e.,

$$V(\mathbf{x}_0) = \sum_{k=0}^K \mathbf{x}^2(k\Delta t, \mathbf{x}_0) \quad (28)$$

where $\mathbf{x}(t, \mathbf{x}_0) = [\dot{\phi}(t) \ \phi(t)]$ denotes the solution of the differential equation (27) for initial state \mathbf{x}_0 and $\Delta t \in \mathbb{R}_+$ is the sampling time. For determining the states at the sampling times $k\Delta t$, we numerically integrate the system using the Bogacki-Shampine method (Bogacki and Shampine, 1989). In order to avoid the computational complexity of performing this numerical integration for many test points, we compute the value of (28) with $K = 1000$ and $\Delta t = 0.01$ only for $N = 1024$ initial states \mathbf{x}_0 uniformly spread over the rectangle $\mathbb{X} = [-5, 5]^2$ and train a Gaussian process with the data to obtain a learned Lyapunov function $\tilde{V}(\cdot) = \mu(\cdot)$.

In order to analyze the region of attraction of the system, we follow the approach proposed in (Lederer and Hirche, 2019). First, we determine the regions \mathbb{W} of the state space satisfying the inequality

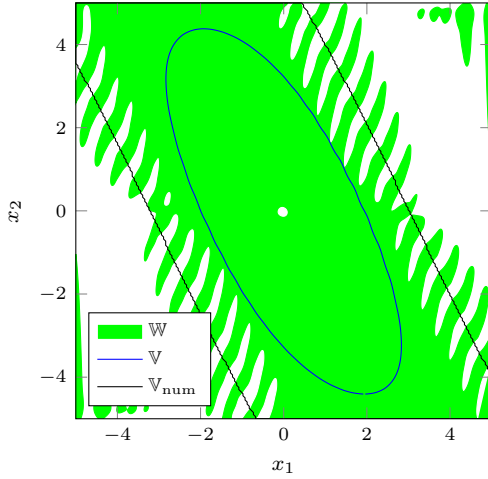


Fig. 3. Lyapunov decrease region \mathbb{W} and region of attraction \mathbb{V} obtained with Alg. 1 as well as numerically approximated region of attraction \mathbb{V}_{num} for the single machine infinite bus system

$$\Delta V(\mathbf{x}_0) = \tilde{V}(\mathbf{x}(\Delta t, \mathbf{x}_0)) - \tilde{V}(\mathbf{x}_0) \leq 0, \quad (29)$$

which corresponds to verifying (1) with $\epsilon_1 = \infty$, $\epsilon_2 = 0$, $g(\cdot) = \mu(\cdot)$ and $\mathbf{f}(\cdot) = \mathbf{x}(\Delta t, \cdot)$. Based on the decrease region \mathbb{W} , the region of attraction is a level set of the learned Lyapunov given by

$$\mathbb{V} = \{\mathbf{x} \in \mathbb{X} : \tilde{V}(\mathbf{x}) \leq \min_{\mathbf{x} \in \mathbb{X} : \Delta \tilde{V}(\mathbf{x}) > 0} \tilde{V}(\mathbf{x})\}, \quad (30)$$

which can also be determined using the GP3 framework. Since the Lipschitz based analysis of our approach does not allow the verification of the decrease condition close to the origin, we assume stability in a ball around $\mathbf{0}$ with radius $r = 0.1$ similarly as in (Lederer and Hirche, 2019).

We apply Alg. 1 with minimal size $b_{\min} = 10^{-9}$ of the hyperrectangles and a conservative Lipschitz constant $L_f = 20$ to both problems. The resulting decrease region \mathbb{W} as well as region of attraction \mathbb{V} of the single machine infinite bus system are illustrated in Fig. 3. Additionally, an estimate of the region of attraction \mathbb{V}_{num} obtained through examining the convergence of trajectories after 10^4 simulation steps for 4000 initial states \mathbf{x}_0 is depicted. The verified decrease region is different from the approximated true region of attraction \mathbb{V}_{num} at its boundary. Due to the saw tooth behavior at the boundary of the verified decrease region \mathbb{W} , the resulting region of attraction \mathbb{V} is smaller than the numerical approximation \mathbb{V}_{num} . However, this underestimation results mainly from imprecision of the learning, while the GP3 approach allows to analyze the Gaussian process very accurately, since the boundary of \mathbb{V} visually touches the boundary of \mathbb{W} at $\mathbf{x} = [1.2 \ 1.7]^T$. Therefore, better estimates of the region of attraction can easily be obtained by training the Gaussian process with more data of the discrete-time Lyapunov function (28).

6. CONCLUSION

This paper introduces a novel framework for the analysis of Gaussian process mean functions called GP3: general purpose computation on graphics processing units for Gaussian processes. Based on interval analysis to compute local Lipschitz constants, the posterior mean function is analyzed using multi-resolution sampling. Due indepen-

dence of the computations for each sample, the method can be parallelized on a GPU for computational efficiency. In order to demonstrate the computational benefits of the GP3 framework, it is applied to a Lipschitz constant bounding and a region of attraction estimation problem.

REFERENCES

- Alefeld, G. and Mayer, G. (2000). Interval analysis: theory and applications. *Journal of Computational and Applied Mathematics*, 121, 421–464.
- Beckers, T. and Hirche, S. (2016a). Equilibrium distributions and stability analysis of Gaussian Process State Space Models. In *Proceedings of the IEEE Conference on Decision and Control*, 6355–6361.
- Beckers, T. and Hirche, S. (2016b). Stability of Gaussian Process State Space Models. In *Proceedings of the European Control Conference*, 2275–2281.
- Berkenkamp, F., Moriconi, R., Schoellig, A.P., and Krause, A. (2016). Safe Learning of Regions of Attraction for Uncertain, Nonlinear Systems with Gaussian Processes. In *Proceedings of the IEEE Conference on Decision and Control*, 4661–4666.
- Bobiti, R. and Lazar, M. (2018). Automated-Sampling-Based Stability Verification and DOA Estimation for Nonlinear Systems. *IEEE Transactions on Automatic Control*, 63(11), 3659–3674.
- Bogacki, P. and Shampine, L.F. (1989). A 3(2) pair of Runge - Kutta formulas. *Applied Mathematics Letters*, 2(4), 321–325.
- González, J., Dai, Z., Hennig, P., and Lawrence, N.D. (2016). Batch Bayesian Optimization via Local Penalization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 648–657.
- Hertneck, M., Köhler, J., Trimpe, S., and Allgöwer, F. (2018). Learning an Approximate Model Predictive Controller with Guarantees. *IEEE Control Systems Letters*, 2(3), 543–548.
- Lederer, A. and Hirche, S. (2019). Local Asymptotic Stability Analysis and Region of Attraction Estimation with Gaussian Processes. In *Proceedings of the IEEE Conference on Decision and Control*.
- Lederer, A., Umlauf, J., and Hirche, S. (2019). Uniform Error Bounds for Gaussian Process Regression with Application to Safe Control. In *Advances in Neural Information Processing Systems*.
- Münz, U. and Romeres, D. (2013). Region of attraction of power systems. *IFAC Proceedings Volumes*, 46(27), 49–54.
- Neal, R.M. (1996). *Lecture Notes in Statistics: Bayesian Learning for Neural Networks*.
- Owens, J.D., Houston, M., Luebke, D., Green, S., Stone, J.E., and Phillips, J.C. (2008). GPU Computing. *Proceedings of the IEEE*, 96(5), 879–899.
- Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA.
- Steinwart, I. (2001). On the Influence of the Kernel on the Consistency of Support Vector Machines. *Journal of Machine Learning Research*, 2, 67–93.
- Vinogradskaya, J., Bischoff, B., Nguyen-Tuong, D., and Peters, J. (2017). Stability of Controllers for Gaussian Process Dynamics. *Journal of Machine Learning Research*, 18, 1–37.