2019 18th European Control Conference (ECC)
Napoli, Italy, June 25-28, 2019

Data-driven inference of passivity properties via Gaussian process optimization

Anne Romer, Sebastian Trimpe and Frank Allgöwer

Abstract— **Passivity is an important concept in control design as it pertains to stability properties of the closed loop. We propose a framework to determine to which extent a dynamic system is or is not passive from data. In particular, we develop a probabilistic approach based on Gaussian processes to underestimate the input feedforward passivity index from experiments with measurement noise. We also show how prior knowledge on the input-output behavior can be incorporated in this framework. Besides the offline approach, we present an iterative scheme that in expectation tightens the lower bound on the feedforward passivity index with every additional data sample and gives an upper bound on the conservatism of the resulting passivity measure.**

I. INTRODUCTION

With the growing complexity of engineering systems, the modeling process becomes a more and more time-consuming task, which requires expert knowledge. On the other hand, modern technology allows for gathering large amounts of data about systems and processes. Therefore, there has been a rising interest in learning controllers directly from data [1]. However, while model-based approaches come with an elaborate theory that offers analyzing tools for stability and performance guarantees, such guarantees are still lacking for many data-driven methods. This lack of guarantees often prevents the application of data-driven methods to real-world scenarios and safety-critical systems, which is therefore an active field of research.

One complementary approach to learning controllers directly from data is to determine certain system properties from data first, which are then leveraged to design a controller as depicted in Fig. 1. Knowledge of the $\mathcal{L}_2$-gain, passivity properties or conic relations, for example, allow for the direct application of well-known feedback theorems to design a controller with guarantees for the closed-loop behavior [2], [3]. Hence, learning certain system properties can provide many of the desired advantages of data-driven controllers while still providing guarantees for the closed-loop.

Existing approaches to learn system properties from input-output data samples can be found in [4]–[11]. However, most of these approaches are only applicable for learning passivity properties or the operator gain of linear time-invariant (LTI) systems [4]–[8], and only few consider nonlinear systems [9]–[11]. In [9], the authors address the issue of validating an operator gain from input-output data and extend some insights also to nonlinear operators. In [10] and [11], further ideas for determining system properties such as the operator gain, passivity measures and conic relations of a general nonlinear system from only input-output data are presented. However, these techniques can be quite conservative or require large amounts of input-output data.

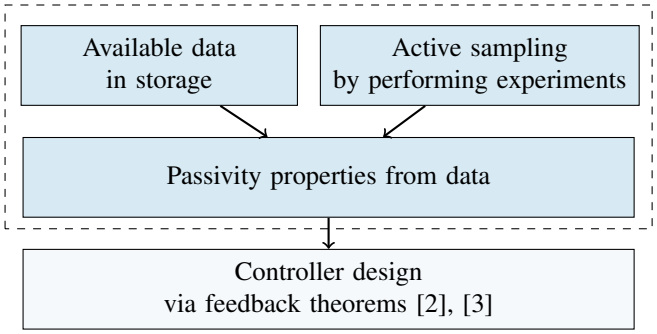

Fig. 1. This work aims to identify passivity properties from data tuples, which are already in storage (offline methods, cf. Sec. III), or from actively sampling additional data (online methods, cf. Sec V). With knowledge on the passivity of the unknown system, well-known feedback theorems can be leveraged for designing controllers with guarantees for the closed-loop input-output behavior [2], [3].

Therefore, we develop a stochastic approach for determining to what extent a general nonlinear system is, or is not, passive from input-output data, when the input-output behavior is unknown. The proposed approach is based on Gaussian processes and does not require identifying the exact input-output relation. In contrast to prior work, the stochastic setting allows for better data efficiency, for incorporating prior system knowledge, and for handling measurement noise. Finally, we use the uncertainty bound quantification of the Gaussian process to introduce a customized active sampling scheme to efficiently perform additional experiments for learning passivity properties.

II. PROBLEM SETUP

We consider a single-input single-output, nonlinear and discrete-time system H, which produces an output y for an input u, as depicted in Fig. 2. H might represent a complex plant, for which the modeling process by first principles becomes a difficult and time-consuming task, or a complex numerical model with no realization as differential equations in closed form at hand. All admissible inputs and corresponding outputs lie in some set $\mathcal{U}, \mathcal{Y} \subset \mathbb{R}^n$. Usually,

A. Romer and F. Allgöwer are with the Institute for Systems Theory and Automatic Control, University of Stuttgart, Germany (email: {anne.romer, frank.allgower}@ist.uni-stuttgart.de). S. Trimpe is with the Max Planck Institute for Intelligent Systems, Stuttgart, Germany (email: trimpe@is.mpg.de).

The authors thank the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology (EXC 310/2) at the University of Stuttgart, along with the Cyber Valley Initiative and the Max Planck Research School (IMPRS) for Intelligent Systems for their support.

$\mathcal{U}$ will contain signals with a certain energy or adhere to certain smoothness conditions corresponding to a frequency bound as both quantities are limited in most physical plants.

Passivity as a system property plays an important role in systems analysis, stability studies and controller design, especially for nonlinear systems [3], [12], [13]. We start with a general input-output definition of passivity for H. The system H is said to be passive if

$$\langle H(u), u\rangle \geq 0 \tag{1}$$

holds for all admissible inputs $u \in \mathcal{U}$, where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product. For input strict passivity, we search for the largest ν such that the inequality

$$\langle H(u), u\rangle \geq \nu\|u\|^2 \tag{2}$$

is true for all admissible inputs $u \in \mathcal{U}$ with $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$ denoting the induced norm. For $\nu > 0$, H is said to be input strictly passive. For $\nu = 0$, we retrieve the passivity condition in (1). If $\nu < 0$, then H is not passive, but the parameter $|\nu|$ corresponds to the feedforward term that renders the system passive. This parameter ν is in general also referred to as the input feedforward passivity index.

In order to determine to what extent the unknown system H is or is not passive, we reformulate the input strict passivity property into an optimization problem given by

$$\nu = \min_{u\in\mathcal{U}, \|u\|^2\neq 0} \rho(u), \quad \rho(u) := \frac{\langle H(u), u\rangle}{\|u\|^2}, \tag{3}$$

with $\rho : \mathcal{U}\backslash\{0\} \to \mathbb{R}$. Since $H : u \to y$ is unknown, we want to find an underestimate of ν in (3) from input-output data obtained from simulations or experiments. While in simulations the numerical errors might usually be very small, measurement noise in experiments oftentimes has a non-negligible impact. Therefore, we consider the case where the output signal is corrupted by additive Gaussian white measurement noise. An extension to more general noise models shall be part of future work. Performing one experiment hence leads to one noisy input-output tuple $(u_i, y_i + e_i)$ (cf. Fig. 2) with $u_i \in \mathcal{U}$, $y_i = H(u_i)$ and $e_i = \begin{pmatrix} e_i(1) & e_i(2) & \dots & e_i(n)\end{pmatrix}^\top$, where we assume $e_i(t)$ to be zero mean Gaussian white noise with variance σ_e^2. Measurement noise on the output signal leads to noisy data-based evaluations of ρ denoted by $\mathcal{D}^N = \{(u_i, f_i)\}_{i=1}^N$ with

$$f_i = \frac{\langle y_i + e_i, u_i\rangle}{\|u_i\|^2} = \rho(u_i) + \frac{\langle e_i, u_i\rangle}{\|u_i\|^2}. \tag{4}$$

The objective of this work is to develop a stochastic approach for learning a lower bound on the input feedforward passivity parameter ν from finitely many noise-corrupted data samples $\mathcal{D}^N$. Within the whole approach, the input-output map H is unknown.

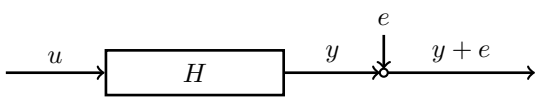

Fig. 2. System H maps any input $u \in \mathcal{U}$ to the corresponding output y. The measured output is corrupted by measurement noise e.

III. Input-strict passivity via Gaussian processes

In this section, we present an approach to learn a lower bound on the input feedforward passivity index ν from available data in storage (offline approach). To find such a lower bound on ρ, we model ρ as a Gaussian process and employ Gaussian process regression to fit the model from data. We start by briefly introducing the necessary background on Gaussian processes before applying it to learn passivity properties.

A. Gaussian process regression

A Gaussian process (GP) is defined as a probability distribution over the space of functions ρ such that every finite subset of function values are jointly Gaussian [14]. A GP, denoted by $\mathcal{GP}(\mu(u), k(u, u'))$, is specified by its mean and covariance functions

$$\begin{aligned}\mu(u) &= \mathbb{E}[\rho(u)]\\ k(u, u') &= \mathbb{E}[(\rho(u) - \mu(u))(\rho(u') - \mu(u'))],\end{aligned}$$

where μ is the expected function value and the covariance function k, also referred to as kernel, captures the covariance between function values and is used to model uncertainty. Probably the most common kernel choice in GP regression is the squared exponential (SE) kernel given by

$$k_{\text{se}}(u, u') = \sigma_f^2 \exp\left(-\frac{1}{2}(u - u')^\top\Lambda^{-1}(u - u')\right) \tag{5}$$

where σ_f^2 is the signal variance and Λ is a positive definite matrix, often chosen to be a diagonal matrix of n squared length scales λ_i^2. These parameters of a kernel are called hyperparameters. Every kernel typically has its own hyperparameters specifying properties of the underlying function such as shape or smoothness, see [14] for details.

In GP regression, we predict the (normal) distribution of function values $\rho(u)$ for $u \in \mathcal{U}$ based on previous data tuples $\mathcal{D}^N$. With (4), we find that the noise on the function evaluation of ρ from experiments has zero mean and variance

$$\mathbb{V}\left(\frac{\langle u_i, e\rangle}{\|u_i\|^2}\right) = \frac{\sigma_e^2}{\|u_i\|^2}.$$

Therefore, the likelihood for observations f_i in (4) is

$$P\left(f_i|\rho(u)\right) = \mathcal{N}\left(\rho(u_i), \frac{\sigma_e^2}{\|u_i\|^2}\right).$$

This leads to the standard closed form predictive posterior mean and variance for noise with input dependent variance [15], which reads

$$\begin{aligned} m_N(u) &= \mathbb{E}[\rho(u)|\mathcal{D}^N] = \mu(u) - k_N^\top(u)(K_N + \Sigma_e^N)^{-1}\hat{\rho}_N\\ \sigma_N^2(u) &= \mathbb{V}[\rho(u)|\mathcal{D}^N]\\ &= k(u, u) - k_N^\top(u)(K_N + \Sigma_e^N)^{-1}k_N(u)\end{aligned} \tag{6}$$

where $k_N(u) = (k(u, u_1), \dots, k(u, u_N))^\top$, $K_N \in \mathbb{R}^{N\times N}$ with $K_{ij} = k(u_i, u_j)$, $\hat{\rho}_N = (\rho_1 - \mu(u_1), \dots, \rho_N - \mu(u_N))^\top$, I_N being the N-dimensional identity matrix and $\Sigma_e^N \in \mathbb{R}^{N\times N}$ with $\Sigma_e^N = \sigma_e^2\text{diag}\left(\|u_1\|^{-2}, \|u_2\|^{-2}, \dots, \|u_N\|^{-2}\right)$. That is, the noise corruption on data tuples (u_i, ρ_i) is

Gaussian depending only on the energy of the input signal and the variance of the measurement noise σ_e^2. Hence, the influence of additive white Gaussian measurement noise can be included in closed form into the standard GP regression.

Since the noise can become arbitrary large as the input signal approaches zero, we need to require a minimum energy of the input signal. This corresponds to requiring a minimal signal to noise ratio, which is a common assumption wherever measurement noise is taken into account. Therefore, similar to [10], we restrict our attention to the compact input set $\mathcal{U}' = \mathcal{U} \setminus \{u \in \mathbb{R}^n | \|u\| < c\}$ with the assumption that the true minimum of ρ does not lie within the c-punctured set of $\mathcal{U}$.

B. Lower bound on ν

The general idea on how to obtain a lower bound on ν is depicted in Fig. 3. We model ρ as a GP

$$\rho(u) \sim \mathcal{GP}(\mu(u), k(u, u')).$$

Via GP regression, we obtain the posterior distribution of $\rho(u)$, $u \in \mathcal{U}'$, by conditioning on $\mathcal{D}^N$, represented by posterior mean $m_N(u)$ and standard deviation $\sigma_N(u)$ as given in (6). We define a confidence region $m_N(u) \pm \alpha\sigma_N(u)$ as a multiple of the standard deviation with $\alpha > 0$. Assuming that the underlying function lies within these confidence bounds, the minimum of the lower bound on the confidence region yields a lower bound on the passivity parameter ν,

$$\hat{\nu}_N = \min_{u \in \mathcal{U}'} m_N(u) - \alpha\sigma_N(u). \tag{7}$$

The parameter α is hence a design variable that depends on how conservative the estimate shall be, typically chosen to be around $\alpha = 2$.

For the presented problem setup, there are mainly three reasons for choosing such a stochastic approach via GPs:

1) We can account for measurement noise. As shown in Sec. III-A, if we assume additive white Gaussian noise on the output signal, we retrieve additive white Gaussian noise on the cost evaluation which is scaled by the energy of the signal. This insight can be directly incorporated into the GP regression in closed form.
2) Prior knowledge on the unknown system can be incorporated into the construction of the kernel and its hyperparameters as shown in Sec. IV to improve the learning rate and to decrease the required input-output samples.
3) We potentially need fewer data points or receive a tighter bound on the passivity measure since fewer data points are required in regions of low probability of a minimizer. Furthermore, additional samples can be drawn in an efficient manner based on the probabilistic description of ρ (cf. Sec. V).

In the following section, we present how to initialize the GP regression, or more specifically, how to choose the kernel and its hyperparameters from prior knowledge. Finally, Sec. V introduces how one can sample additional data efficiently from the probabilistic description of the cost function ρ.

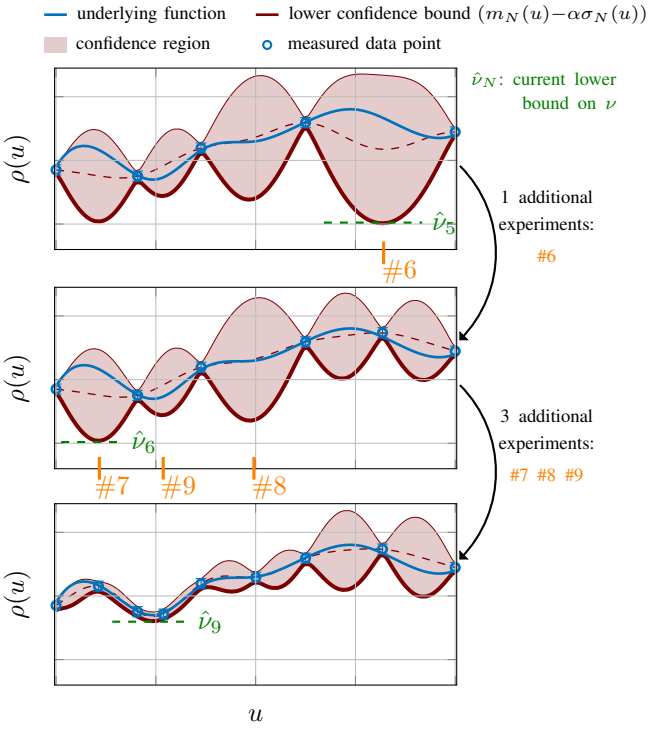

Fig. 3. Via GP regression, we learn a probability distribution of $\rho(u)$ for every $u \in \mathcal{U}'$ from data. From top to bottom, the posterior GP with 5, 6, and 9 data points is shown. We define a confidence region as a factor α of the standard deviation σ. The minimum of the lower bound on the confidence region yields a lower bound estimate on the passivity parameter ν. If additional experiments are performed, these are chosen according to the probabilistic description of ρ for data efficient sampling (details in Sec. V).

IV. Kernel Design for Inference of Input-Strict Passivity

For GP regression, we assume that ρ is a sample of the GP specified by a chosen mean function and kernel. The choice of the kernel is hence an integral step influencing not only the learning speed, but also affecting the inferred values of $\hat{\nu}_N$. By choosing the kernel and its hyperparameters, one should hence include any prior knowledge about the underlying function such as smoothness properties, length scales and expected shape. Therefore, we present in this section approaches on how to choose the kernel according to such insights, incorporating the problem structure as a prior.

A. Smoothness of the kernel

We start by considering the choice of kernels according to the estimated smoothness of the unknown input-output system H. With application of the chain rule, whenever the input-output operator $u \mapsto H(u)$ is continuously differentiable of order m denoted by $H \in C^m$, we know that $\rho : \mathcal{U}' \to \mathbb{R}$ is continuously differentiable of order m. Two very common choices of kernels are the SE kernel and the Matérn kernels, which could be a good starting point dependent on the knowledge of the smoothness of the unknown input-output operator H. The SE kernel (5) is infinitely differentiable. This means that the GP with

this covariance function is 'very smooth' since it has mean square derivatives of all orders [14]. The Matérn class of covariance functions as presented in [14] is parameterized by a smoothness parameter $\eta > 0$. The process ρ is k-times mean square differentiable if and only if $\eta > k$. For $\eta = \frac{1}{2}$, the Matérn kernel becomes very rough, whereas for $\eta \to \infty$ we retrieve the smooth SE kernel. Depending on m, one can thus decide if the SE kernel is a good choice, or choose a Matérn kernel, e.g., with $\eta \approx m$.

B. Lipschitz continuity

Another parameter that is often available or can be estimated from data [10] is the Lipschitz constant. Let us consider the case where the input-output operator is differentiable and Lipschitz in the sense that $\|H(u_2) - H(u_1)\| \leq L_H \|u_2 - u_1\|$ and hence that

$$\|DH(u)\| \leq L_H \quad \forall u \in \mathcal{U}', \tag{8}$$

where DH denotes the Jacobian matrix of H and $\|DH(u)\|$ denotes the induced matrix norm. With knowledge of L_H, we can infer a Lipschitz constant of the optimization function $u \to \rho(u)$ as follows.

Lemma 1: Assume (8) holds and $H(0) = 0$. Then there is a Lipschitz constant $L_\rho \leq \frac{2\pi L_H}{c}$ such that

$$\|\rho(u_2) - \rho(u_1)\| \leq L_\rho \|u_2 - u_1\|.$$

Proof: We start by estimating the supremum of the gradient of $\rho : \mathcal{U}' \to \mathbb{R}$ by

$$\begin{aligned}
\|\nabla \rho(u)\| &\leq \left\| \frac{H(u) + (DH(u))^\top u}{\|u\|^2} \right\| + \left\| \frac{2u^\top H(u) u}{\|u\|^4} \right\| \\
&\leq \frac{1}{\|u\|} \left(\frac{\|H(u)\|}{\|u\|} + \frac{\|(DH(u))^\top u\|}{\|u\|} + 2\frac{\|u^\top H(u)\|}{\|u\|^2} \right) \\
&\leq \frac{1}{\|u\|} 4L_H \ \leq \frac{1}{c} 4L_H.
\end{aligned}$$

The set $\{u \in \mathbb{R}^n | \|u\| \geq c\}$ is said to be C-quasiconvex, with $C = \frac{\pi}{2}$, since any two points u_1, u_2 can be joined by a curve γ of length at most $\frac{\pi}{2}\|u_1 - u_2\|$ (e.g., an arc of a circle). Integrating the gradient along such a curve can give an upper bound on the Lipschitz constant by

$$\|\rho(u_2) - \rho(u_1)\| \leq \frac{\pi}{2} \|u_1 - u_2\| \sup_{u \in \mathbb{R}^n; \|u\| \geq c} \|\nabla \rho(u)\|.$$

Hence, $L_\rho \leq \frac{2\pi L_H}{c}$. ∎

From [16, Theorem 5], we know that if the covariance kernel is at least four times differentiable and the mean function is at least twice differentiable, then the sample paths are Lipschitz continuous with high probability on a bounded set of $\mathbb{R}^n$. However, since this result does not provide quantitative information on how Lipschitz constant and hyperparameters relate to each other, we will in the following derive a more practical approach.

Generally, the overestimate of the Lipschitz constant from Lemma 1 gives a (potentially conservative) estimate on how the function values of two input signals u_1 and u_2 can differ.

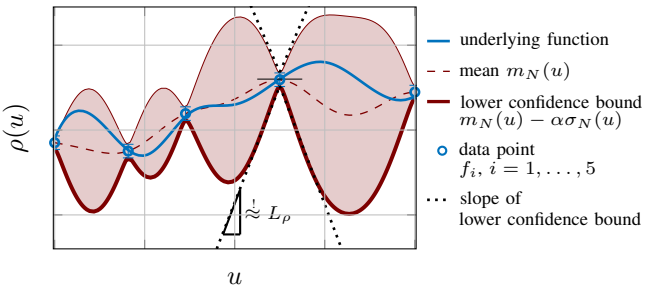

Fig. 4. To find a lower bound the passivity measure, we are mainly interested in the lower bound of the confidence region defined by α times the standard deviation. One possible interpretation of a Lipschitz constant on the underlying function ρ is the slope of the lower confidence bound at a data point. This allows us to incorporate prior knowledge on the Lipschitz constant via the prior on the derivative process.

This can be seen as a parallel concept to isotropic covariance functions, where the covariance between u_1 and u_2 is only dependent on $\|u_1 - u_2\|$. The SE as well as the Matérn covariance class are isotropic covariance functions. Incorporating a large Lipschitz constant would then correspond to choosing the covariance of two close points to be small.

We are now interested in how exactly the prior of a GP can be chosen to incorporate knowledge on Lipschitz continuity. The idea is depicted in Fig. 4. With an upper bound on the Lipschitz constant, we know an upper bound on the slope at each point u. Therefore, in a neighborhood of any data sample u_i, we can restrict the confidence region to $\rho(u_i) - L_\rho\|u - u_i\| \leq \rho(u) \leq \rho(u_i) + L_\rho\|u - u_i\|$. Hence, we want to choose the prior in such a way, that the slope of the lower confidence bound at a data point u_i has the slope of this Lipschitz constant L_ρ. Thus, the Lipschitz constant leads us to a specific condition on the prior of the derivative process.

The derivative of a GP is yet another GP since differentiation is a linear operator [14]. In fact, since both the expectation operator and the differentiation operator are commutative, the derivative of the posterior mean is equal to the mean of the distribution of derivatives [17]. The mean and covariance function of the derivative process $\rho_d \sim \mathcal{GP}(\mu_d, k_d)$ are thus defined by $\mu_d(u) = \nabla_u \mu(u)$ and $k_d(u, u') = \nabla_u \nabla_{u'} k(u, u')$. This insight allows us to incorporate prior information in the derivative process.

Let us further consider the SE kernel (5) and a constant prior mean function. The covariance function of a GP describing the derivative of a GP with SE kernel reads

$$\nabla_u \nabla_{u'} k(u, u') = \Lambda^{-1} \left(I - (u - u')(u - u')^\top \Lambda^{-1} \right) k(u, u')$$

and hence

$$\nabla_u \nabla_{u'} k(u, u')|_{u=u'} = \Lambda^{-1} \sigma_f^2.$$

Since the Lipschitz constant should mark the border of the confidence region, the hyperparameters Λ, σ_f should approximately be chosen such that

$$\frac{\sigma_f}{\lambda_i} = \left(\frac{1}{\alpha} L_\rho \right)^{\frac{1}{2}} \quad i = 1, \ldots, n$$

holds. In fact, for the SE kernel, a relation between the Lipschitz constants and the ratio between prior variance and lengthscales $\sigma_f \lambda_i^{-1}$ has also been mentioned in [18]. While this specific relation of hyperparameters and Lipschitz constant considers the SE kernels, this approach can be transferred to other sufficiently differentiable isotropic kernels.

In some application scenarios, the estimate of the Lipschitz constant of an input-output operator H might be an estimate with non-negligible uncertainty, e.g. when obtained from data. This uncertainty of the Lipschitz constant, modeled for example as $p(L_\rho) = \mathcal{N}\left(L_\rho, \sigma_L^2\right)$, can again be incorporated as a prior on the hyperparameters by choosing, e.g.,

$$\lambda_i^{-2} \sim \mathcal{N}\left(\frac{L_\rho}{\alpha\sigma_f^2}, \frac{\sigma_L^2}{\alpha^2\sigma_f^4}\right),$$

where σ_L^2 expresses the uncertainty in the estimate of the Lipschitz constant L_ρ. Such hyper-priors $p(\Theta)$ can then be used to find the posterior over the hyperparameters by optimizing $\max_\Theta p(f_i|u_i, \Theta)p(\Theta)$ [14].

C. Parametric kernel design

In many applications, the unknown input-output system can be decomposed into a part with only parametric uncertainty and a general nonlinear part. For example, the unknown input-output system can be composed of a linear LTI system and nonlinear residuals leading to

$$\rho(u) = \rho_{\text{LTI}}(u) + \rho_{\text{NL}}(u). \tag{9}$$

This provides us with additional structural information on $u \to \rho(u)$, which can be incorporated into the kernel design to improve the performance of the GP regression. For scalar discrete-time LTI systems and a given input sequence $u(t), t = 1, ..., N$, the input to output map reads

$$\begin{pmatrix} y(1) \\ \cdot \\ \cdot \\ \vdots \\ y(n) \end{pmatrix} = \begin{pmatrix} g_0 & 0 & 0 & ... & 0 \\ g_1 & g_0 & 0 & ... & 0 \\ g_2 & g_1 & g_0 & ... & 0 \\ \vdots & & & \ddots & \vdots \\ g_{n-1} & g_{n-2} & ... & g_1 & g_0 \end{pmatrix} \begin{pmatrix} u(1) \\ \cdot \\ \cdot \\ \vdots \\ u(n) \end{pmatrix}$$

where g_k, $k = 1, 2, \ldots$, denotes the impulse response sequence, which we consider unknown. This input-output relation is in the following denoted by $y = Gu$. For such LTI systems, the cost function ρ_{LTI} hence reads

$$\frac{u^\top G u}{\|u\|^2} = \underbrace{\frac{1}{\|u\|^2}\begin{pmatrix} u_1^2 + \ldots + u_n^2 \\ u_1u_2 + \ldots + u_{N-1}u_N \\ \vdots \\ u_1u_n \end{pmatrix}}_{h(u)}^\top \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{n-1} \end{pmatrix}. \tag{10}$$

The vector $h(u)$ can be interpreted as a set of fixed basis functions in u_i, and $g = (g_0, \ldots, g_{n-1})^\top$ as a set of a-priori unknown parameters. We model g as a Gaussian random variable, $g \sim \mathcal{N}(\mu_g, \Sigma_g)$, with μ_g, for example, according to some measured impulse response of the system, or zero if no other information is available. The variance models uncertainty about the impulse response and can also be used to encode smoothness properties and other prior knowledge.

From (9), with (10) and $\rho_{\text{NL}} \sim \mathcal{GP}(0, k(u, u'))$ a general GP, we then obtain another GP reading $\rho(u) \sim \mathcal{GP}\left(h(u)^\top \mu_g, k(u, u') + h(u)^\top \Sigma_g h(u')\right)$ for which the predictions can be found in [14, p. 28]. While we presented here some problem specific approaches on the kernel choice for determining the input feedforward passivity index, there exists a whole line of research on kernel methods for system identification and function estimation more generally, see e.g. [19] and the references therein.

V. Active sampling scheme

In many applications, additional simulations or experiments can be performed. This corresponds to actively sampling additional data points (online method). In this section, sampling schemes for choosing these additional data points customized to learning a lower bound on the input feedforward passivity parameter are introduced and discussed.

In the present problem setup, the goal is to quickly receive a tight estimate on the minimum of ρ. While $\hat{\nu}_N$ from (7) yields the current lower bound estimate of the passivity parameter ν, it concurrently provides information where an additional data point increases this lower bound. This corresponds to choosing the next data point u_{N+1} such that $u_{N+1} = \text{argmin}_{u \in \mathcal{U}'} U_{\text{LCB}}(u)$ with

$$U_{\text{LCB}}(u) = m_N(u) - \alpha\sigma_N(u), \tag{11}$$

where U_{LCB} is called the Lower Confidence Bound (LCB) acquisition function determining where the next query point should be (analogous to Upper Confidence Bound (UCB) [20] for maximization). This active sampling scheme is illustrated in Fig. 3, where the next data point is chosen at the current minimum of the lower confidence bound. For specific choices of varying α, particular regret bounds are proven for the UCB approach [21]. Since we are interested in a lower bound with high probability, and hence the minimum of the lower confidence bound, we choose a fixed α.

Another approach to receive a tight bound on ν with high data efficiency is to reduce the confidence region below the minimum of ρ as quickly as possible. The corresponding acquisition function according to the defined goal reads

$$\begin{aligned} U_{\text{LVM}}(u) = \int_{\mathcal{U}'} \max\big(&\bar{\nu}_N - m_N(v) \\ &+ \alpha\mathbb{E}[\sigma_{N+1}(v)|\mathcal{D}^N, u_{N+1}{=}u], 0\big)\,\mathrm{d}v. \end{aligned} \tag{12}$$

minimizing the expected confidence region below the expected minimum of ρ denoted by $\bar{\nu}_N = \min_{u \in \mathcal{U}'} m_N(u)$, as depicted in Fig. 5. We will call this acquisition function the Lower Variance Minimum (LVM) approach. Since the variance only depends on u_{N+1} independently of f_{N+1}, $\mathbb{E}[\sigma_{N+1}(v)|\mathcal{D}^N, u_{N+1}{=}u] = \sigma_{N+1}(v, u)$, and hence (12) simplifies to

$$U_{\text{LVM}}(u) = \int_{\mathcal{U}'} \max\left(\bar{\nu}_N - m_N(v) + \alpha\sigma_{N+1}(v, u), 0\right)\mathrm{d}v.$$

Both acquisition functions (11) and (12) are depicted in Fig. 5. Both sampling approaches demonstrated comparable

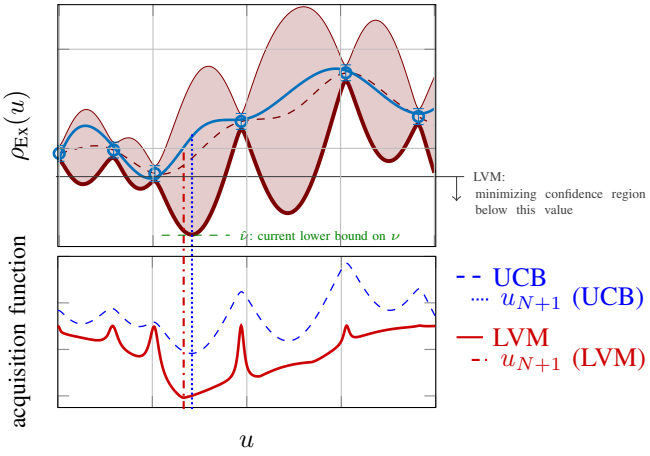

Fig. 5. Comparison of the two acquisition functions UCB and LVM applied to the illustrative example ρ_{Ex}. While LCB from (11) samples at the current lower bound on ν, LVM from (12) tries to minimize the confidence region below the current expected minimum.

Algorithm 1: Active sampling for learning passivity

```
Specify GP prior (cf. Sec. IV);
Specify confidence region by choosing α;
Specify stopping criteria ϵ (cf. Prop. 2);
Initialize D^N with set of prior data samples;
repeat
    Update GP posterior (m_N, σ_N) in (6);
    Compute next input signal u_{N+1} via (11);
    Perform experiment with u_{N+1} and calculate f_{N+1};
    Set D^{N+1} = {D^N, (u_{N+1}, f_{N+1})};
    Set N = N + 1 ;
until 2ασ_{N-1}(u_N) > ϵ;
return ν̂_{N-1} in (7)
```

performance in terms of data efficiency in an initial comparison. However, the acquisition function (11) is significantly less expensive to compute. While a more comprehensive comparison of both acquisition functions is left to future work, (11) will be our preferential choice in the following.

Algorithm 1 summarizes the active sampling scheme introduced above. The core approach that is used herein for efficient sampling of the passivity function is essentially Bayesian optimization. Bayesian optimization generally is a global (black-box) optimization technique that builds a probabilistic description (typically a GP) of the unknown cost function and uses this to select next samples in an efficient way [22]. Bayesian optimization has been successfully applied in other learning-based control contexts, such as for direct data-driven controller tuning, e.g. in [18], [23], [24].

We next prove some properties about the proposed approach. The following proposition shows that Algorithm 1 is expected to improve the lower bound in each step.

Proposition 1: The lower bound on ν from Algorithm 1 is in expectation monotonically increasing in the sense that, given $\mathcal{D}^{N-1}$ at step $N-1$,

$$\mathbb{E}\left[\hat{\nu}_N - \hat{\nu}_{N-1} | \mathcal{D}^{N-1}\right] \geq 0.$$

Proof: We start by applying the linearity of the expectation operator, which directly yields $\mathbb{E}\left[\hat{\nu}_N - \hat{\nu}_{N-1}|\mathcal{D}^{N-1}\right] = \mathbb{E}\left[m_N(u) - m_{N-1}(u)|\mathcal{D}^{N-1}\right] - \alpha\,\mathbb{E}\left[\sigma_N(u) - \sigma_{N-1}(u)|\mathcal{D}^{N-1}\right]$. From [25], we know that $\sigma_N(u) \leq \sigma_{N-1}(u)$ holds independently of the next drawn data sample (u_N, f_N). Furthermore, the expected value for the next data sample ρ_N on the basis of $\mathcal{D}^{N-1}$ is expressed by $m_{N-1}(u)$. Therefore, $\mathbb{E}\left[m_N(u)|\mathcal{D}^{N-1}\right] = m_{N-1}(u)$ holds, and hence

$$\begin{aligned}\mathbb{E}\left[m_N(u) - \alpha\sigma_N(u) - m_{N-1}(u) - \alpha\sigma_{N-1}(u)|\mathcal{D}^{N-1}\right] \\ = -\alpha\,\mathbb{E}\left[\sigma_N(u) - \sigma_{N-1}(u)|\mathcal{D}^{N-1}\right] \geq 0.\end{aligned}$$

Since this must hold for all $u \in \mathcal{U}'$, it also holds for the minimum and thus, $\mathbb{E}\left[\nu_N - \nu_{N-1}|\mathcal{D}^{N-1}\right] \geq 0$. ∎

Furthermore, by means of the stopping criteria ϵ one can choose how tight the lower bound on ν shall be and bound the conservatism of the resulting $\hat{\nu}_N$ of Algorithm 1.

Proposition 2: When $2\alpha\sigma_N(u_{N+1}) < \epsilon$ (stopping criteria) is met, then the input feedfoward parameter ν is maximally ϵ larger than the lower bound resulting from Algorithm 1, i.e. $\nu - \hat{\nu}_N \leq \epsilon$, with a probability of at least $\Phi(\alpha)$, where

$$\Phi(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-\frac{t^2}{2}} \mathrm{d}t.$$

Proof: This follows directly from the marginalization property of the GP with

$$P(\rho(u_{N+1}) < m_N(u_{N+1}) + \alpha\sigma_N(u_{N+1})) = \Phi(\alpha),$$

where Φ is the cumulative distribution function of the normal distribution. ∎

To calculate the actual probability mass below the resulting estimate $\hat{\nu}_N$, one can use Monte Carlo Methods via sampling from the GP or, for example, use the approach in [26] to calculate the probability mass via discretization.

VI. Example

We present a numerical example to illustrate the proposed approach. To facilitate comparison, we adopt the example in [10], the Van der Pol oscillator. The underlying system is governed by $\dot{x}_1(t) = x_2(t)$, $\dot{x}_2(t) = 4(1 - x_1(t)^2)x_2(t) - x_1(t) - u(t)$, and $y(t) = x_1(t)$, with $x_1(0) = x_2(0) = 0$, $H : u \mapsto y$ considering the (closed) unit interval $t \in [0, 1]$. As in [10], we restrict our input to a four dimensional subspace $\mathcal{V} = \text{span}(\{P_1, P_2, P_3, P_4\})$ of $\mathcal{L}_2$, P_i being the first four Legendre polynomials

$$\begin{aligned} P_1(t) &= 1, & P_2(t) &= 2t - 1, \\ P_3(t) &= 6t^2 - 6t + 1, & P_4(t) &= 20t^3 - 30t^2 + 12t - 1, \end{aligned}$$

which constitute a four dimensional orthonormal basis. All coefficients of the admissible input are bounded by $b_i = 5$, $i = 1, \ldots, 4$, similar to [10] and we choose $c = 0.8$ to retain a meaningful signal to noise ratio.

We discretize the system with $\Delta t = 0.001$. The Gaussian measurement noise on the discretized output signal has zero mean and standard deviation $\sigma_n = 3$. We define the confidence region to be $\alpha = 2$ times the standard deviation. Furthermore, we choose the SE kernel (5) due to the smoothness of the example system. According to an

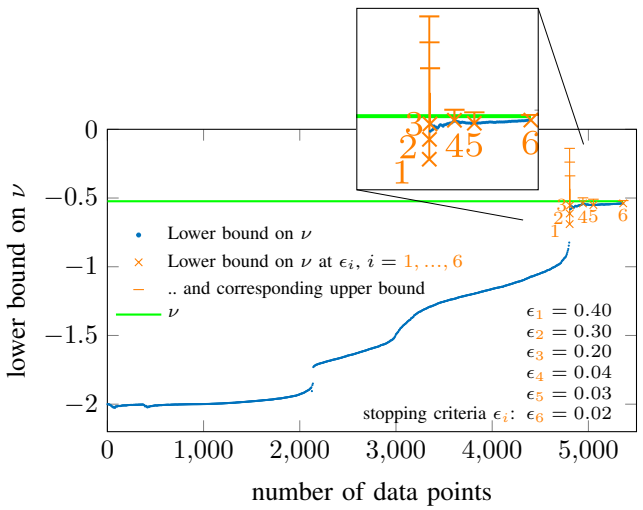

Fig. 6. Applying the iterative approach introduced in Sec. V, we computed lower bounds on the input feedforward passivity parameter of the example problem in [10], a Van der Pol oscillator.

estimated upper bound on the Lipschitz constant from data of approximately $L_\rho = 4.5$, we choose the hyperparameters to be $\sigma_f = 1$ and $\lambda_i = 2/3$ for all $i = 1, \ldots, 4$.

The resulting lower bound on ν is shown in Fig. 6, where the values for different stopping criteria ϵ_i, $i = 1, \ldots, 6$ are highlighted with the corresponding upper confidence bound obtained from Prop. 2. Starting at around 4810 data samples, we receive a quite tight bound on the passivity parameter, which never fails to underestimate the input feedforward passivity parameter $\nu = \min_{u \in \mathcal{U}'} \rho(u)$. Moreover, all upper bounds on the conservatism are valid upper bounds on ν. Hence, we receive a tighter bound with fewer data points than [10], while also including measurement noise. However, the computational cost is higher.

Varying the hyperparameters, similar ratios $\sigma_f \lambda_i^{-1}$ led to similar results, while larger ratios generally required larger amounts of data to meet the stopping criteria and smaller ratios started to fail to underestimate ν at some iterations.

The computational costs and the curse of dimensionality are in general still open problems and a field of active research for Bayesian optimization techniques. Some approaches to Bayesian optimization with high-dimensional problems are found in [22], including some results on identifying relevant dimensions or exploiting sparsity. A tailored approach to decrease the computational complexity of Bayesian optimization for passivity properties is left to future work.

VII. Conclusion

The presented approach based on GP regression and Bayesian optimization is an attractive option for learning passivity properties from data (offline or online) since it allows for handling measurement noise, it admits incorporating prior knowledge, and, above all, it can be significantly less conservative than a deterministic approach. The lower number of necessary data points, however, comes at the cost of higher computational expenses.

References

[1] Z.-S. Hou and Z. Wang, "From model-based control to data-driven control: Survey, classification and perspective," *Information Sciences*, vol. 235, pp. 3–35, 2013.

[2] G. Zames, "On the input-output stability of time-varying nonlinear feedback systems part I: Conditions derived using concepts of loop gain, conicity, and positivity," *IEEE Trans. Automat. Control*, vol. 11, no. 2, pp. 228–238, 1966.

[3] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*. SIAM, 1975.

[4] K. van Heusden, A. Karimi, and D. Bonvin, "Data-driven estimation of the infinity norm of a dynamical system," in *Proc. 46th IEEE Conf. on Decision and Control*, pp. 4889–4894, 2007.

[5] B. Wahlberg, M. B. Syberg, and H. Hjalmarsson, "Non-parametric methods for $\mathcal{L}_2$-gain estimation using iterative experiments," *Automatica*, vol. 46, no. 8, pp. 1376–1381, 2010.

[6] C. R. Rojas, T. Oomen, H. Hjalmarsson, and B. Wahlberg, "Analyzing iterations in identification with application to nonparametric H_∞-norm estimation," *Automatica*, vol. 48, no. 11, pp. 2776–2790, 2012.

[7] A. Romer, J. M. Montenbruck, and F. Allgöwer, "Sampling strategies for data-driven inference of passivity properties," in *Proc. 56th IEEE Conf. on Decision and Control*, pp. 6389–6394, 2017.

[8] M. I. Müller, P. E. Valenzuela, A. Proutiere, and C. R. Rojas, "A stochastic multi-armed bandit approach to nonparametric H_∞ estimation," in *Proc. 56th IEEE Conf. on Decision and Control*, 2017.

[9] K. Poolla, P. Khargonekar, A. Tikku, J. Krause, and K. Nagpal, "A time-domain approach to model validation," *IEEE Trans. Automat. Control*, vol. 39, no. 5, pp. 951–959, 1994.

[10] J. M. Montenbruck and F. Allgöwer, "Some problems arising in controller design from big data via input-output methods," in *Proc. 55th IEEE Conf. on Decision and Control*, pp. 6525–6530, 2016.

[11] A. Romer, J. M. Montenbruck, and F. Allgöwer, "Determining dissipation inequalities from input-output samples," in *Proc. 20th IFAC World Congress*, pp. 7789–7794, 2017.

[12] A. Van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*. Springer, 2000.

[13] J. Bao and P. L. Lee, *Process Control: The Passive Systems Approach*. Springer, 2007.

[14] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[15] P. Goldberg, C. K. I. Williams, and C. M. Bishop, "Regression with input-dependent noise: A Gaussian process treatment," in *Advances in Neural Information Processing Systems*, pp. 493–499, 1998.

[16] S. Ghosal and A. Roy, "Posterior consistency of Gaussian process prior for nonparametric binary regression," *The Annals of Statistics*, vol. 34, no. 5, pp. 2413–2429, 2006.

[17] C. E. Rasmussen, "Gaussian processes to speed up hybrid Monte Carlo for expensive Bayesian integrals," *Bayesian Statistics*, vol. 7, pp. 651–659, 2003.

[18] F. Berkenkamp, A. Krause, and A. P. Schoellig, "Bayesian optimization with safety constraints: Safe and automatic parameter tuning in robotics," *arXiv:1602.04450*, 2016.

[19] G. Pillonetto, F. Dinuzzo, T. Chen, G. D. Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, pp. 657–682, 2014.

[20] P. Auer, "Using confidence bounds for exploitation-exploration tradeoffs," *J. Machine Learning Research*, vol. 3, pp. 397–422, 2002.

[21] N. Srinivas, A. Krause, S. Kakad, and M. Seeger, "Gaussian process optimization in the bandit setting: no regret and experimental design," in *Proc. 27th International Conference on Machine Learning, Haifa, Israel*, pp. 1015–1022, 2010.

[22] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proceedings of the IEEE*, vol. 104, pp. 148–175, 2016.

[23] A. Marco, P. Hennig, J. Bohg, S. Schaal, and S. Trimpe, "Automatic LQR tuning based on Gaussian process global optimization," in *Proc. IEEE Intl. Conf. on Robotics and Automation*, pp. 270–277, 2016.

[24] A. Marco, P. Hennig, S. Schaal, and S. Trimpe, "On the design of LQR kernels for efficient controller learning," in *Proc. 56th IEEE Conf. on Decision and Control*, pp. 5193–5200, 2017.

[25] C. K. Williams and F. Vivarelli, "Upper and lower bounds on the learning curve for Gaussian processes," *Machine Learning*, vol. 40, pp. 77–102, 2000.

[26] J. P. Cunningham, P. Hennig, and S. Lacoste-Julien, "Gaussian probabilities and expectation propagation," *arXiv:1111.6832*, 2011.