**Preprints of the 21st IFAC World Congress (Virtual)**
**Berlin, Germany, July 12-17, 2020**

# Probabilistic $H_2$-norm estimation via Gaussian process system identification ⋆

**Daniel Persson, Anne Koch, Frank Allgöwer**

*Institute for Systems Theory and Automatic Control, University of Stuttgart*

**Abstract:** We present a method for data-based estimation of the $H_2$-norm of a linear time-invariant system from input-output data in a probabilistic setting by employing the recent advances in Gaussian process system identification using stable-spline kernels. Advantages of this starting point include that the norm can be estimated for the continuous-time system and over infinite horizon, even though only a finite number of measurements are available. We approximate the $H_2$-norm distribution as Gaussian, whose expectation can even be obtained analytically, while we use a numerical scheme based on Gaussian process quadrature for the variance. Not only do we utilize the posterior variance of the Gaussian process to derive an error estimate for the $H_2$-norm, but also to tune the estimation by optimizing the input sequence. The performance of the developed scheme is thoroughly evaluated in simulation.

*Keywords:* Nonparametric methods, Bayesian methods, Data-based control, Identification for control, Input and excitation design

## 1. INTRODUCTION

Given the increasing complexity of technical systems, system modeling for control design *ab initio* becomes increasingly difficult and requires extensive experience. Therefore, data-based methods for system analysis and control design grow more important and are facilitated by the increased computing throughput of modern processors. Control design and analysis without system models can instead be based on energy-related system properties such as $H_2$-norm, operator gain (equivalent to $H_\infty$-norm for linear time-invariant (LTI) systems) and passivity. A benefit of energy-based methods for stability analysis is that they generalize to networks of interconnected systems and, in the case of operator gain and passivity, to nonlinear systems.

A straightforward path to data-based estimation of these properties is applying standard system identification to estimate a model from input-output data and then applying a method for computing the desired system property from the identified model. Efficient methods for computing system properties from state-space and transfer function models include e.g., Hara et al. (2010) and Belur and Praagman (2011). However, this path to data-based estimation has the disadvantage of fitting the dynamics to a parametric model, where the accuracy depends on how well the dynamics are reflected by the selected model subspace.

Therefore, there has been an increasing interest in developing non-parametric data-based estimation schemes for system properties. Interesting examples of non-parametric methods include the ones developed by Wahlberg et al. (2010), Müller and Rojas (2019), Romer et al. (2019) and Tu et al. (2018). The disadvantages of these methods are that while the non-parametric methods successfully avoid having to impose naturally uncertain structure on the system dynamics, they provide no way to make use of prior information on the system, estimate properties over finite horizon and Wahlberg et al. (2010); Müller and Rojas (2019) additionally require sequential measurements. Furthermore, they are mostly limited to discrete-time system dynamics.

In this work we provide a method for $H_2$-norm estimation that overcomes these disadvantages while still being non-parametric by making use of the recent advances in Gaussian process (GP) system identification with stable-spline kernels, paving the way for a future extension to $H_\infty$-norm estimation. Additionally we make use of an inherent property of Gaussian process regression (GPR), namely the posterior variance, which has yet been only rarely exploited in GP system identification. The posterior variance does not only provide a measure of (un)certainty, but can also be used for optimal input design. By minimizing over the generalization error, which is defined in terms of the posterior variance, the $H_2$-norm estimation as well as the respective variance can be significantly improved. We verify the developed methods in simulation on randomly generated LTI systems.

## 2. PROBLEM DESCRIPTION

We consider stable causal strictly proper SISO LTI systems, where the $H_2$-norm, which is the main focus of this work, is defined by e.g., Toscano (2013) as

$$||G||_2^2 = \int_0^\infty g^2(t)\, dt = \frac{1}{2\pi}\int_{-\infty}^\infty |G(j\omega)|^2\, d\omega. \qquad (1)$$

⋆ This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 - 390740016. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Anne Koch. For correspondence, `mailto:daniel.o.persson@outlook.com`

The $H_2$-norm captures both the transient response of the system, as it is the $L_2$-norm of the impulse response, and the response to white noise. The $H_2$-norm equivalent to the small gain theorem, called the *mean-square* small gain theorem, is of importance in stochastic control, where a sufficient condition for closed-loop mean-square asymptotic stability is based on the $H_2$-norm of the single agents together with knowledge on the stochastic uncertainty variance on the channels, see e.g., Nandanoori et al. (2018).

Due to its relevance in control theory, we hence consider the problem of estimating the $H_2$-norm from a finite sequence of $N$ input-output data points given by

$$\begin{aligned}(t_k, y_k), \quad & k \in [0, \ldots, N],\\ u(t), \quad & t \in [0, t_N],\end{aligned}$$

where the complete input history $u(t)$ is assumed to be known and $y_k$ is the measurement of the system output at time $t_k$. The measurement is assumed to be subject to white zero-mean additive Gaussian noise with variance $\sigma_e^2$. We proceed with preliminaries on GP system identification that form the foundation of this work.

## 3. PRELIMINARIES

Let the convolution operator for an input $u(t)$ to the output at time $t_k$ for a causal impulse response be denoted

$$y_k = L_k[g] = \int_0^{t_k} u(t_k - \tau) g(\tau)\, d\tau, \tag{2}$$

where the input is left out of the convolution operator to simplify notation. In the setting of system identification via GPs, the impulse response is considered a GP. To distinguish from the deterministic impulse response, we denote the impulse response Gaussian process (IRGP) as $h(t)$. See e.g., Rasmussen and Williams (2006) for a comprehensive treatment of GPR for general function estimation from data. The GP approach to system identification follows in the same way as in common GPR, with the IRGP considered a priori to be a zero-mean GP with covariance kernel $\kappa(s, t)$. The difference to common GPR is that the impulse response is not measured directly, rather its convolution with the input. As convolution is a linear operator, the output is also Gaussian with covariance as in Pillonetto et al. (2014), namely

$$O(t_i, t_j) = L_i[L_j[\kappa(\cdot, \cdot)]]$$

and the output covariance matrix for measurement time points $\mathbf{t}$ has entries $O_{ij} = O(t_i, t_j)$, for $i, j \in [1, \ldots, N]$ with $N$ being the number of measurements. The expected value and variance of the posterior are then given by

$$\begin{aligned}\mathbb{E}[h(t)|\mathbf{y}] &= O(t, \mathbf{t}) Z^{-1} \mathbf{y},\\ \mathrm{Var}(h(t)|\mathbf{y}) &= \lambda\kappa(t, t) - \lambda O(t, \mathbf{t}) Z^{-1} O(\mathbf{t}, t),\end{aligned} \tag{3}$$

where $O(t, \mathbf{t})$ is the covariance between the IRGP and output, $Z = O(\mathbf{t}, \mathbf{t}) + \frac{\sigma_e^2}{\lambda} I$ is the measurement covariance, $\lambda$ is a scaling factor on the covariance kernel and $\sigma_e^2$ is the measurement noise variance. The problem regarding selection of hypothesis space in traditional system identification is thus replaced by selection of the covariance kernel, which has been a recent research topic. The results are summarized in e.g., Pillonetto et al. (2014), and the main result is a class of kernels for system identification called stable-spline kernels. The first- and second-order stable-spline kernels are given by

$$\begin{aligned}\kappa_1(s, t) &= e^{-\beta \max(s,t)},\\ \kappa_2(s, t) &= \frac{1}{2} e^{-\beta(s+t+\max(s,t))} - \frac{1}{6} e^{-3\beta \max(s,t)},\end{aligned} \tag{4}$$

where $\beta$ is a hyperparameter. Hyperparameters can be selected in numerous ways, the most common of which is marginal likelihood maximization which is also summarized in Pillonetto et al. (2014). In the next section we continue by obtaining the probabilistic $H_2$-norm in terms of the IRGP.

## 4. PROBABILISTIC $H_2$-NORM

We define the *random* squared $H_2$-norm based on the definition of the $H_2$-norm (1) as

$$\Upsilon = \int_0^{\infty} h^2(t)\, dt, \tag{5}$$

where $h(t)$ is the IRGP. Note that this definition is for convenience in notation, we only consider expected value and variance of the norm, who both exist for stable-spline kernels. Care needs to be taken to ensure square integrability when using other kernels. As (5) is a nonlinear transform of $h(t)$, the distribution of the random $H_2$-norm is not Gaussian. Finding the distributions of nonlinearly transformed Gaussian random variables analytically is generally intractable. Therefore the distribution is commonly approximated by moment matching a Gaussian. Moment matching has the property of minimizing the Kullback-Leibler divergence defined in e.g., Rasmussen and Williams (2006). The K-L divergence is zero iff the probability density functions are equal, and can loosely be interpreted as a measure on how similar two distributions are. Hence, approximating the distribution of a nonlinearly transformed random variable as Gaussian by matching the expected value and variance makes the approximation optimal in the sense of Kullback-Leibler divergence. Starting from Equation (5) and using the linearity of the expectation operator, the expected value of $\Upsilon$ is

$$\mathbb{E}[\Upsilon] = \mathbb{E}\left[\int_0^{\infty} h^2(t)\, dt\right] = \int_0^{\infty} \mathbb{E}[h^2(t)]\, dt \tag{6}$$

and by the definition of covariance, the variance of $\Upsilon$ is

$$\begin{aligned}\mathrm{Var}(\Upsilon) &= \mathbb{E}\left[(\Upsilon - \mathbb{E}[\Upsilon])(\Upsilon - \mathbb{E}[\Upsilon])\right]\\ &= \int_0^{\infty}\int_0^{\infty} \mathrm{Cov}(h^2(s), h^2(t))\, ds\, dt.\end{aligned} \tag{7}$$

### *4.1 Analytic moment evaluation*

We now consider evaluating the moments from the previous section when the GP $h(t)$ is the posterior impulse response from Equation (3). While nonlinear transforms of GPs are in general analytically intractable, here we exploit the fact that in this case the nonlinear transform is a square of the Gaussian process and the structure of the kernel to obtain the expected value of the $H_2$-norm analytically from the posterior IRGP.

*Proposition 1.* The expected $H_2$-norm for a system described by an IRGP is given by

$$\mathbb{E}[\Upsilon|\mathbf{y}] = \int_0^{\infty} O(t, \mathbf{t}) W O(\mathbf{t}, t) + \lambda\kappa(t, t)\, dt \tag{8}$$

where $W = Z^{-1}\mathbf{y}\mathbf{y}^T Z^{-T} - \lambda Z^{-1}$.

**Proof.** To arrive at an expression for (6), the relation between variance and first and second moments as

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

is employed, where $X$ is a random variable. Rearranging and plugging into (6) and inserting the posterior impulse response (3) yields

$$\mathbb{E}[\Upsilon|\mathbf{y}] = \int_0^\infty \mathbb{E}[h(t)|\mathbf{y}]^2 + \mathrm{Var}(h(t)|\mathbf{y})\, dt$$
$$= \int_0^\infty O(t,\mathbf{t})(Z^{-1}\mathbf{y}\mathbf{y}^T Z^{-T} - \lambda Z^{-1})O(\mathbf{t},t) + \lambda\kappa(t,t)\, dt.$$

□

Expanding and changing the orders of integration and summation, the expression becomes a double sum over convolutions as

$$\mathbb{E}[\Upsilon|\mathbf{y}] = \sum_{i,j=1}^{N} W_{ij} L_i \left[ L_j \left[ \int_0^\infty \kappa(t,\cdot)\kappa(t,\cdot)\, dt \right] \right] + \lambda \int_0^\infty \kappa(t,t) dt$$

where $L_i[\cdot]$ is the convolution operator as defined in Equation (2). Hence, if the integral in the above expression can be attained analytically, the mean $H_2$-norm of the posterior can be computed exactly. This is the case for stable-spline kernels as they are linear combinations of decaying exponential functions. As covariance is not a linear operator, to the best of our knowledge there is no similar counterpart for evaluating the $H_2$-norm variance, which motivates the venture into numerical integration presented in the next section.

### *4.2 Gaussian process quadrature moment evaluation*

The goal of this section is to provide a numerical method for computing the $H_2$-norm variance (7) using Gaussian process quadrature (GPQ). GPQ is an application of GPR used to obtain integral approximations of functions from pointwise evaluations, see e.g., the foundational work by O'Hagan (1991). This scheme exploits the fact that integration is a linear operator and linearly transformed GPs are still Gaussian, which makes it possible to obtain a posterior distribution of the integral value similar to (3). In the next subsections we first obtain evaluations of the integrand and then use them to obtain an estimate of the $H_2$-norm variance.

*Obtaining integrand evaluations* Obtaining evaluations of $\mathrm{Cov}(h^2(s), h^2(t))$ from the first and second order moments can be achieved as detailed by Willink (2005) as

$$\mathrm{Cov}(h^2(s), h^2(t)) = \mathbb{E}[h^2(s)h^2(t)] - \mathbb{E}[h^2(s)]\mathbb{E}[h^2(t)]$$
$$= 4\mathrm{Cov}(h(s), h(t))\mathbb{E}[h(s)]\mathbb{E}[h(t)] + 2\mathrm{Cov}(h(s), h(t))^2.$$

Assume that the function $\mathrm{Cov}(h^2(s), h^2(t))$ decays to essentially zero at time $t_f$ such that a time partition $\pi : \mathbf{t}\times\mathbf{t}$ ending at $t_f$ as $0 = t_0 < t_1 < \ldots < t_N = t_f, \mathbf{t} = [t_0, \ldots, t_N]$ can be made without losing information for time larger than $t_f$. Then evaluations of the required moment of the posterior at a grid point $(t_n, t_m)$ in $\pi$ is achieved as $o_{(n-1)N+m} = \mathrm{Cov}(h^2(t_n), h^2(t_m))$ where $o_{(n-1)N+m}$ are entries in the vector $\mathbf{o}$ of integrand evaluations on the grid $\pi$.

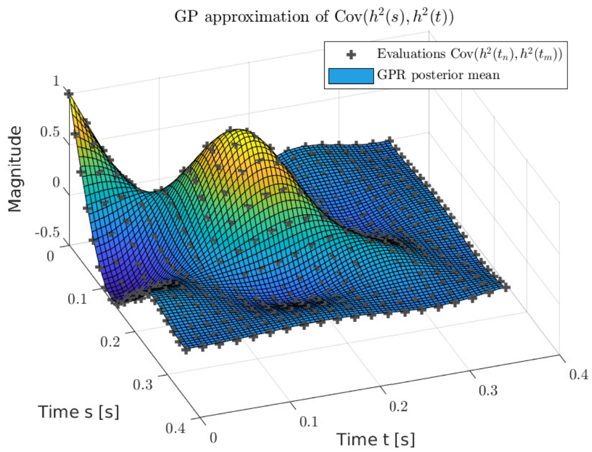


Fig. 1. Evaluations of $\mathrm{Cov}(h^2(s), h^2(t))$ on the grid $\mathbf{t}\times\mathbf{t}$ and the regression for the two-dimensional input domain.

*$H_2$-norm variance* Having obtained evaluations $\mathbf{o}$ at domain points $\pi : \mathbf{t}\times\mathbf{t}$ of $\mathrm{Cov}(h^2(s), h^2(t))$ we proceed to use GPQ to approximate the integral over the domain. To this end we utilize a prior of the form

$$\kappa_c(s,t) = \kappa(s_1,t_1)\kappa(s_2,t_2),$$
$$K_c = K \otimes K,$$

where $s, t \in \mathbb{R}^2_+$, $K$ is the kernel matrix for evaluations along one dimension of the domain, $\otimes$ is the Kronecker tensor product and $K_c$ is the compound kernel formed by combining two kernels, in this case identical ones for each of the two input domain dimensions, since covariance functions are symmetric.

*Remark 1.* The kernel can be freely chosen, however a natural choice is the same stable-spline kernel that was used in the system identification step. This is since the covariance of the squared IRGP also eventually decays to zero, as stable systems are considered.

Suitable prior hyperparameters can be selected in numerous ways, for example by marginal likelihood maximization in a similar fashion as in the system identification step. Even though the evaluations are indeed evaluations and not measurements, setting a small but significant evaluation error variance $\sigma_e^2$ is beneficial as it improves the condition of $K_c$. An example of an estimated covariance function is depicted in Figure 1. Note that the depiction is provided only as an explanatory device. In the $H_2$-norm variance computation, the estimated covariance function is not calculated explicitly but is only used implicitly in the regression.

The value of the posterior integral is computed similarly to e.g., O'Hagan (1991) as

$$k_c(\mathbf{t}) = \int_0^\infty \kappa(s,\mathbf{t})\, ds \otimes \int_0^\infty \kappa(s,\mathbf{t})\, ds,$$
$$\mathrm{Var}(\Upsilon|\mathbf{o},\mathbf{y}) \approx k_c^T(\mathbf{t})(K_c + \frac{\sigma_e^2}{\lambda} I)^{-1}\mathbf{o},$$

where $k_c$ is the covariance between the integrand evaluations and prior integral value. The $H_2$-norm variance is considered to be conditioned on both the integrand evaluations and on the system output, as it is the posterior $H_2$-norm variance and the integrand evaluations introduce error in the variance estimate. An advantage of GPQ is that the variance of the integral resulting from the quadrature can be evaluated, which could be employed as a measure of how well the quadrature estimates the $H_2$-norm variance or for tuning the quadrature scheme. However, these possibilities were not implemented in this work as the quadrature variance was generally small.

## 5. EMPLOYING THE INHERENT ERROR ESTIMATE

One of the benefits of GPR is that it inherently provides an assessment of the expected error through the posterior variance. In this section we proceed to make use of the posterior variance in GP system identification in general and specifically when using the IRGP to obtain the $H_2$-norm as in the previous section. Due to the nonlinear relationship between the IRGP and probabilistic $H_2$-norm ($\Upsilon$), it is not trivial to directly make use of the $H_2$-norm variance to optimize or tune the estimation. Therefore, in the following sections, we first define a proximate goal, the generalization error, that we use to achieve the ultimate goal of reducing the $H_2$-norm estimate error. We define the generalization error for an IRGP in a similar way as for common GPR.

Experiment design in the system identification setting concerns designing experiments such that the maximal amount of information about the system is extracted given limited resources such as input energy and measurements. Here we focus specifically on optimal input design. Classical schemes are reviewed by e.g., Mehra (1974). Examples of optimal input design for discrete-time GP system identification include Fujimoto and Sugie (2018) and Mu and Chen (2018).

### *5.1 Generalization error*

The generalization error is defined in e.g., Rasmussen and Williams (2006) as the expected squared error between a function drawn from a prior and the predicted mean averaged over some prediction domain. In the case of an IRGP, assuming $h(t)$ is drawn from a zero mean GP with covariance kernel $\kappa(s,t)$, the expected squared error of the impulse response at time $t$ is

$$\mathbb{E}[(h(t) - O^T(t)Z^{-1}\mathbf{y})^2] = \lambda\kappa(t,t) - \lambda O^T(t)Z^{-1}O(t),$$

if the regression is correctly specified, meaning the prior is the actual GP from which the impulse response is drawn. We recognize this as the posterior variance at $t$. A common definition of the generalization error is an average of the error over some interval of $t$. In our case, a natural choice seems to be integrating from 0 to $\infty$ as stable systems are considered, whose impulse responses are integrable. The generalization error for correctly specified regression is then

$$\varepsilon(\mathbf{t}) = \int_0^\infty \underbrace{\lambda\kappa(t,t) - \lambda O^T(t,\mathbf{t})Z^{-1}O(\mathbf{t},t)}_{\mathrm{Var}(h(t))}\, dt \tag{9}$$

where $\mathbf{t}$ is the vector of time points for the measurements. This expression looks familiar, and is indeed part of the mean $H_2$-norm computation where $\mathbb{E}[h(t)]^2 + \mathrm{Var}(h(t))$ is integrated, cf. Equation (8). The machinery for computing the generalization error for a fixed set of measurement time points is hence already in place.

### *5.2 Relation to $H_2$-norm variance*

Having defined the generalization error for an IRGP, we proceed to derive the relation between the generalization error and the $H_2$-norm variance (7), i.e. how uncertainty in the IRGP impacts the uncertainty of the $H_2$-norm estimate. Intuitively, if the assumptions made are fulfilled, the only way for uncertainty to enter the $H_2$-norm estimate is through the IRGP posterior variance, hence there should be some link between them. The relation is stated in the following proposition.

*Proposition 2.* For an IRGP with stable-spline kernel, the $H_2$-norm variance is related to the impulse response generalization error defined in Equation (9) as

$$\mathrm{Var}(\Upsilon) \le C \int_0^\infty \mathrm{Var}(h(t))\, dt$$

where $C$ is a positive constant.

**Proof.** The squared $H_2$-norm variance (7) can be bounded using the Cauchy-Schwarz inequality for expectation as

$$\int_0^\infty \int_0^\infty \mathrm{Cov}(h^2(s), h^2(t))\, ds\, dt \le \left(\int_0^\infty \sqrt{\mathrm{Var}(h^2(t))}\, dt\right)^2. \tag{10}$$

As $h(t)$ is Gaussian at every time $t$, its higher order moments are available at each $t$ from the mean and variance as detailed by e.g., Willink (2005). In this case, we have

$$\begin{aligned}\mathrm{Var}(h^2(t)) &= \mathbb{E}[h^4(t)] - \mathbb{E}[h^2(t)]^2 \\ &= \mathrm{Var}(h(t))\underbrace{\left(4\mathbb{E}[h(t)]^2 + 2\mathrm{Var}(h(t))\right)}_{\le 4\mathbb{E}[h^2(t)]}.\end{aligned}$$

Inserting the last expression into (10) and applying the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\mathrm{Var}(\Upsilon) &\le \left(\int_0^\infty \sqrt{4\mathbb{E}[h^2(t)]\mathrm{Var}(h(t))}\, dt\right)^2 \\ &\le \int_0^\infty 4\mathbb{E}[h^2(t)]\, dt \int_0^\infty \mathrm{Var}(h(t))\, dt \le C\int_0^\infty \mathrm{Var}(h(t))\, dt\end{aligned}$$

where the last inequality is due to the fact that $\mathbb{E}[h^2(t)]$ is integrable for stable-spline kernels, which is a direct consequence of integrability as shown in Pillonetto et al. (2014) and boundedness by arguments in Carmeli et al. (2006). □

### *5.3 Optimal input design algorithm*

This input design scheme is largely based on the one developed by Fujimoto and Sugie (2018), but is extended to continuous-time system identification by using a different objective. Where Fujimoto and Sugie use the mutual information between prior and posterior, we use the generalization error for IRGPs as defined in (9). As the algorithm depends on the prior of the IRGP, either prior knowledge or a preliminary experiment is a prerequisite.

We limit the scope to considering the special case of obtaining a sequence of piecewise constant inputs for continuous-time GP system identification. This limitation is motivated by the fact that in practical applications the hardware is often limited to piecewise constant inputs and by making the optimization problem finite dimensional.

Additionally we assume that the measurement sequence is complete and with the same sample rate as the input sequence. Before stating the input design optimization problem, we define the lower-triangular Toeplitz matrix of the input sequence and the integrated kernel matrix as

$$U \in \mathbb{R}^{N\times N} \text{ with } U = \begin{bmatrix} u_1 & 0 & \dots & & \\ u_2 & u_1 & 0 & \dots & \\ u_3 & u_2 & u_1 & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & \end{bmatrix},$$
$$H \in \mathbb{R}^{N\times N} \text{ with } H_{i,j} = \int_{t_i}^{t_{i+1}} \int_{t_j}^{t_{j+1}} \kappa(s,t)\, ds\, dt, \tag{11}$$
$$H(t) \in \mathbb{R}^{N} \text{ with } H(t)_i = \int_{t_i}^{t_{i+1}} \kappa(s,t)\, ds,$$

where $N$ is the length of the input and measurement sequences.

*Proposition 3.* The input sequence $\mathbf{u}_*$ that solves the optimization problem

$$\min_{\mathbf{u}\in\Gamma} \tilde{\varepsilon}(\mathbf{u})$$

is the best input sequence for GP system identification in the sense of minimizing the aggregated posterior variance with respect to the constraint $\mathbf{u} \in \Gamma$. The objective is defined as

$$\tilde{\varepsilon}(\mathbf{u}) = -\sum_{i,j=1}^{N} (U^T(UHU^T + \frac{\sigma_e^2}{\lambda} I)^{-1} U)_{ij} W_{ij} \tag{12}$$

with $W = \int_0^\infty H(t)H^T(t)\, dt$. The optimal input sequence also minimizes the generalization error as defined in (9).

**Proof.** For the special case of a piecewise constant input sequence, the generalization error (9) can be rewritten using the Toeplitz matrix of the input sequence and integrated kernel matrix (11) as

$$\tilde{\varepsilon}(\mathbf{u}) = -\int_0^\infty H^T(t) U^T (UHU^T + \frac{\sigma_e^2}{\lambda} I)^{-1} U H(t)\, dt$$
$$= -\sum_{i,j=1}^{N} (U^T(UHU^T + \frac{\sigma_e^2}{\lambda} I)^{-1} U)_{ij} W_{ij},$$

with $W$ as in (12) and where the prior part of the generalization error has been dropped as it does not depend on the optimization variable and the notation $\tilde{\varepsilon}$ is used to distinguish between the generalization error and the objective. □

The objective $\tilde{\varepsilon}(\mathbf{u})$ is non-convex with multiple local optima, but even local optima are good compared to white noise input sequences as we show in the simulation section. The gradient of the objective is available analytically as

$$\frac{\partial \tilde{\varepsilon}}{\partial u_k} = -\sum_{i,j=1}^{N} W_{ij} (U_k^T M^{-1} U + U^T M^{-1} U_k$$
$$+ U^T M^{-1} (U_k H U^T + U H U_k^T) M^{-1} U)_{ij},$$

with $M = UHU^T + \frac{\sigma_e^2}{\lambda} I$, where $U_k$ is the derivative of the Toeplitz matrix $U$ with respect to each input piece $u_k$. The gradient can hence be utilized in optimizing the objective using for example *fmincon* in matlab or a projected gradient method as done in Fujimoto and Sugie (2018). An example of the method applied is depicted in Figure 2. In particular, note that the optimal input is an exponentially decaying oscillation, similar to what was obtained by Fujimoto and Sugie (2018) for discrete-time, and that the spectrum of the input sequence is concentrated to the lower end of the spectrum, quickly dropping off above the system bandwidth of $\sim 3.8$ rad/s. The result is a reduced posterior variance along the impulse response.

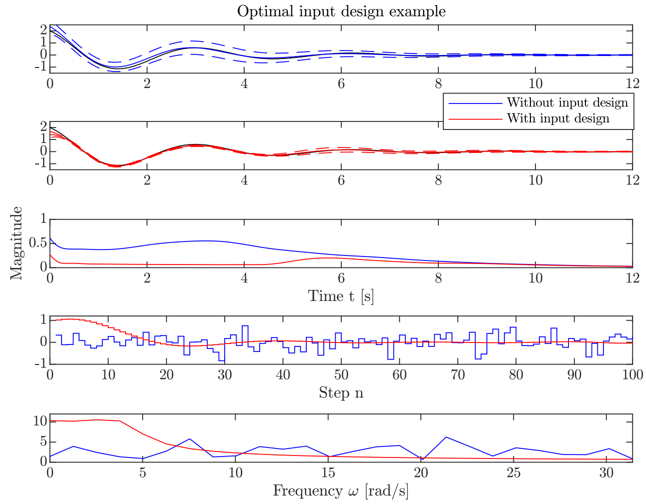


Fig. 2. Experiment design example on the system $G(s) = \frac{2s}{s^2+0.8s+4}$ using a second-order stable-spline kernel (4) with an oscillating component, $\kappa(s,t) = \kappa_2(s,t)(\cos(\beta_2|s-t|)+1)$, as suggested in Chen (2018). 1: GP system identification for white noise input scaled to energy limit. 2: GP system identification for optimal input. 3. Comparison of $3\sigma$ confidence bounds. 4: Input sequences in time domain. 5: Input sequences in frequency domain.

## 6. SIMULATION

The developed methods were evaluated in numerical simulation on randomly generated LTI systems in a similar fashion to Pillonetto et al. (2014), where *rmodel* was used to generate random systems of specified order, with an additional constraint $NT_s > t_f$ with $T_s = 1/(4|\lambda_{\max}|)$ where $\lambda_{\max}$ is the pole with largest absolute value, $N$ is the number of samples and $t_f$ is the time where the impulse response has decayed to essentially zero. This constraint is imposed in order to be able to select a fast enough sample rate while still having enough measurements to capture the whole impulse response. Note that this is only imposed to ensure comparability by using a fixed number of measurements for all generated systems and constitutes no general restriction. A first measurement sequence is generated using *lsim* with a zero-order hold white noise input sequence scaled to have a certain energy, i.e. $\mathbf{u}^T\mathbf{u} = E_l$. Then GP system identification is performed for each kernel and the mean and variance of the $H_2$-norm are computed from the posterior IRGP. The optimal input sequence, limited to the same energy as the white noise input sequence, is obtained for each kernel and used to simulate another sequence of measurements, and the norm estimation process is repeated for the new sequences. Two metrics are used to compare performance between different kernels and schemes, relative error (RE) and relative error overestimation (REO), who we define as

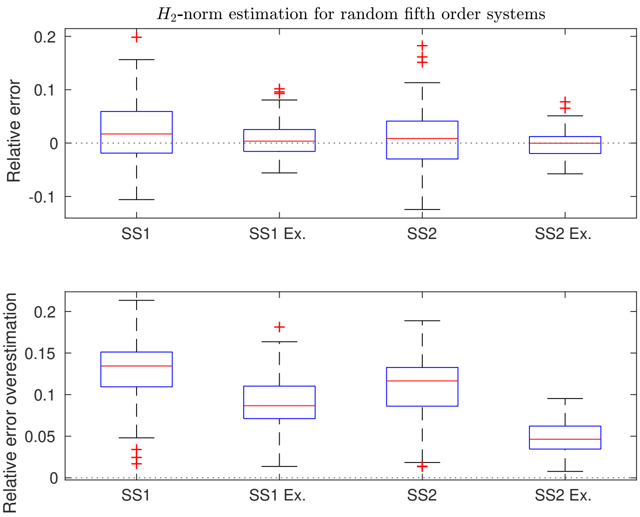


Fig. 3. RE and REO boxplots using first- and second-order stable-spline kernels for random fifth-order systems, 100 random systems, 100 measurements and additive white measurement noise with SNR 20. Ex. denotes that optimal input design was used to select the input sequence. The input sequences are limited as $\mathbf{u}^T\mathbf{u} = 100$. The boxes indicate the median and the 25th and 75th percentiles, the whiskers indicate the extreme values and + denotes outliers.

$$\text{RE} = \frac{\mathbb{E}[\Upsilon] - ||G||_2^2}{||G||_2^2}, \quad \text{REO} = \frac{3\sqrt{\text{Var}(\Upsilon)}}{||G||_2^2} - |\text{RE}|.$$

where $||G||_2^2$ is the actual squared $H_2$-norm, $\mathbb{E}[\Upsilon]$ is the estimated squared $H_2$-norm and $3\sqrt{\text{Var}(\Upsilon)}$ is the $3\sigma$ confidence bound on the norm estimate. RE is hence a measure of how well the $H_2$-norm is estimated and REO is a measure of how well the uncertainty is captured by the norm variance. REO larger than zero indicates that the confidence bound holds, and small REO indicates that the confidence bound is less conservative. A boxplot of the RE and REO for random fifth-order systems is depicted in Figure 3 for 100 systems using 100 sequential measurements. It is notable that the performance is good, even for fairly high order systems, and that the optimal input design produces significantly better estimates.

## 7. CONCLUSION

We introduced a framework to estimate the $H_2$-norm via GP system identification. The proposed method does not only provide a way to impose prior knowledge while still being non-parametric, it also enables estimating system properties from the continuous-time dynamics with infinite horizon using a finite number of measurements. Furthermore, it inherently handles missing measurements, which just corresponds to conditioning the prior on fewer data points, who do not need to be evenly spaced in time. In the proposed probabilistic framework, confidence bounds based on the posterior variance can be obtained. Furthermore, the variance can be utilized for optimal input design. To conclude, estimating system properties using GP system identification is a very promising lane, not only considering the results that were obtained in this work, but also due to possible future research directions, such as the extension to multiple-input multiple-output systems and other system properties such as the $H_\infty$-norm.

## REFERENCES

Belur, M.N. and Praagman, C. (2011). An efficient algorithm for computing the $\mathcal{H}_\infty$-norm. *IEEE Transactions on Automatic Control*, 56(7), 1656–1660.

Carmeli, C., De Vito, E., and Toigo, A. (2006). Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4(4), 377–408.

Chen, T. (2018). On kernel design for regularized LTI system identification. *Automatica*, 90, 109–122.

Fujimoto, Y. and Sugie, T. (2018). Informative input design for kernel-based system identification. *Automatica*, 89, 37–43.

Hara, S., Iwasaki, T., and Tanaka, H. (2010). $\mathcal{H}_2$- and $\mathcal{H}_\infty$-norm computations for LTI systems with generalized frequency variables. In *Proc. 2010 American Control Conference*, 1862–1867.

Mehra, R. (1974). Optimal input signals for parameter estimation in dynamic systems–Survey and new results. *IEEE Transactions on Automatic Control*, 19(6), 753–768.

Mu, B. and Chen, T. (2018). On input design for regularized LTI system identification: Power-constrained input. *Automatica*, 97, 327–338.

Müller, M.I. and Rojas, C.R. (2019). Gain estimation of linear dynamical systems using Thompson Sampling. In *Proc. 22nd International Conference on Artificial Intelligence and Statistics*, 1535–1543.

Nandanoori, S.P., Diwadkar, A., and Vaidya, U. (2018). Mean square stability analysis of stochastic continuous-time linear networked systems. *IEEE Transactions on Automatic Control*, 63(12), 4323–4330.

O'Hagan, A. (1991). Bayes–Hermite quadrature. *J. of Statistical Planning and Inference*, 29(3), 245–260.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., and Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50, 657–682.

Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Romer, A., Berberich, J., Köhler, J., and Allgöwer, F. (2019). One-shot verification of dissipativity properties from input-output data. *IEEE Control Systems Letters*, 3(3), 709–714.

Toscano, R. (2013). *Structured Controllers for Uncertain Systems*. Advances in Industrial Control. Springer London, London.

Tu, S., Boczar, R., and Recht, B. (2018). On the approximation of Toeplitz operators for nonparametric $\mathcal{H}_\infty$-norm estimation. In *Proc. 2018 American Control Conference*, 1867–1872.

Wahlberg, B., Barenthin Syberg, M., and Hjalmarsson, H. (2010). Non-parametric methods for L2-gain estimation using iterative experiments. *Automatica*, 46(8), 1376–1381.

Willink, R. (2005). Normal moments and Hermite polynomials. *Statistics & Probability Letters*, 73, 271–275.