

Kernel-Based Learning of Stable Nonlinear Systems

Matteo Scandella, Michelangelo Bin, *Member, IEEE* and Thomas Parisini, *Fellow, IEEE*

Abstract—Learning models of dynamical systems characterized by specific stability properties is of crucial importance in applications. Existing results mainly focus on linear systems or some limited classes of nonlinear systems and stability notions, and the general problem is still open. This article proposes a kernel-based nonlinear identification procedure to directly and systematically learn stable nonlinear discrete-time systems. The proposed method can be used to enforce, on the learned model, bounded-input-bounded-state stability, the asymptotic gain property, input-to-state stability, and their incremental counterparts. To this aim, we build on the reproducing kernel theory and the regularized least squares method, which are suitably enhanced to handle stability constraints in the kernel properties and in the hyperparameters' selection algorithm. Once the methodology is detailed, and sufficient conditions for stability are singled out, the article reviews some widely used kernels and their applicability within the proposed framework. Finally, numerical results validate the theoretical findings showing, in particular, that stability may have a beneficial impact in long-term simulation with minimal impact on prediction.

Index Terms—Nonlinear system identification, Incremental input-to-state stability, Bounded-input-bounded-state stability, Reproducing kernel Hilbert spaces, Kernel-based regularization

I. INTRODUCTION

A key issue in system identification is to learn models that not only fit the observations, but also possess specific stability properties [1], [2]. Indeed, stability is desirable in many applications as it provides robustness guarantees when the model is used for prediction or simulation, especially with a long time horizon. However, most existing learning procedures do not guarantee that the learned model is stable, even when the observations are generated by a stable system. In this article, we propose a learning approach for nonlinear systems that, instead, systematically guarantees a desired stability property on the learned models.

Matteo Scandella is with the Department of Management, Information and Production Engineering, University of Bergamo, via Marconi 5, 24044, Dalmine (BG), Italy (email: matteo.scandella@unibg.it). Michelangelo Bin is with the Department of Electrical, Electronic and Information Engineering, University of Bologna, Bologna, Italy (email: michelangelo.bin@unibo.it). During the conceptualization and first drafting of this article, M. Bin and M. Scandella were with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK. Thomas Parisini is with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K., He is also with the Department of Electronic Systems, Aalborg University, Denmark, and with the Department of Engineering and Architecture, University of Trieste, Italy (e-mail: t.parisini@imperial.ac.uk).

All correspondence should be sent to Dr. Matteo Scandella (email: matteo.scandella@unibg.it).

While classical identification and machine learning approaches do not deal with the problem of imposing stability on the learned models [3]–[5], in the case of linear time-invariant (LTI) systems such a problem is well-studied and many solutions exist. Among others, it is worth mentioning methods based on ARMAX (AutoRegressive Moving Average eXogenous input) models [6], [7], subspace [8]–[10] and set-membership [11], [12] methods, estimation techniques working in the frequency domain [13], and kernel-based approaches [14], [15]. Instead, in a nonlinear framework the problem is considerably less explored and fewer results are available; moreover, the learning problem is exacerbated by the many available notions of stability and by the wide range of modeling techniques, which have led to more scattered results mainly tailored to particular cases. For instance, there are specific methods guaranteeing stability for linear parameter-varying systems [16], [17], nonlinear finite impulse response systems [18], [19], and linear switching systems [20]. For more generic nonlinear structures, some works rely on neural networks [21]–[24]; more in detail [22], [23] constrain the parameters of a neural network to impose different stability properties, although the stability constraints can easily be violated due to numerical problems arising from the imposition of constraints during the training phase. Other approaches rely on Koopman operators [25], [26], linear matrix inequality constraints [27], [28], or Gaussian process state-space models [29]–[31]; in particular, [29], [30] develop a two-step procedure that first identifies a potentially unstable model fitting the observations, and then learns a virtual control law modifying the previously-estimated model to guarantee stability. Finally, [32], [33] propose a kernel-based procedure to learn contractive maps between function spaces, thereby learning stable input-output models.

In this article, we develop a kernel-based learning approach to directly and systematically learn nonlinear stable discrete-time models. The proposed approach is particularly inspired by [32], [33], although here we focus on predictors and state-space models instead of input-output maps, and we target more general stability notions. Moreover, unlike most of the aforementioned works, the proposed method is systematic, is not tailored on a specific model structure, and the stability constraints are robust to numerical errors in the solution procedure. Furthermore, we target multiple notions of stability of primary interest in control theory, i.e., bounded-input-bounded-state (BIBS) stability [34], asymptotic gain (AG) [35], input-to-state stability (ISS) [36], [37], and their incremental counterparts: δ BIBS stability, δ AG, and δ ISS [38]. The enforced stability guarantees refer both to the exogenous input of the system and to the prediction error. The former

is the main motivation of this article while, as commented in Section II and shown in the numerical validation, stability with respect to the prediction error may lead to a considerable benefit when the learned model is used for simulation over a long (potentially infinite) time horizon.

Kernel-based approaches are popular in system identification and machine learning because they allow systematically enforcing desired properties on the learned models by appropriately shaping the kernel function [39]–[41] or the regularization term [42]–[44]. More specifically, for LTI models, kernel-based techniques enable identifying bounded-input-bounded-output (BIBO) stable models without a priori knowledge of the model's order [14], [15], [45]. The drawback of these methods is that they exploit the properties of the impulse response of an LTI system and its relation to BIBO stability, thereby ruling out possible extensions to the nonlinear case. Existing kernel-based approaches for nonlinear systems are instead based on the estimation of input-output operators [32], [33], [46] or, alternatively, on the direct learning of the predictor function of the state [29], [30] or of the output [19], [41], [46]–[48] instead of the full model. The methodology proposed in this article falls into the latter category. In particular, we equip the technique used in [41], [47] with constraints on the learning algorithm and the hyperparameters' selection method. In this way, the desired stability properties are guaranteed on the learned model while, at the same time, the predictor selection is optimized by fitting the available dataset. Moreover, since the proposed methodology is an extension of the classical kernel approach, it is not tied to a single specific kernel structure nor to a specific hyperparameters' selection method. Instead, it can easily be adapted to different model-assessment rationales, e.g., empirical Bayes [49] or cross-validation techniques [50]–[52], and can be used with different kernel functions.

However, the price to pay for such a generality is that the procedure may lead to a complex bilevel optimization problem that, in some cases, can be challenging to solve numerically. Moreover, the conditions enforced to guarantee stability may be conservative in general, and their enforcement may lead to a loss of prediction performance compared to the unconstrained (but possibly unstable) case. These drawbacks worsen with the predictor's order, so that the proposed method may not scale well with the models' dimension. All these critical aspects, however, concern individual parts of the overall procedure that can be addressed in specific application scenarios without affecting the general methodology.

Paper organization: Section II lays down the problem formulation. Section III briefly recalls the regularized kernel method, and introduces the proposed methodology as an extension of it. Section IV shows how the proposed method can be used to learn BIBS stable and ISS models. Similarly, Section V considers the case of δ BIBS stability and δ ISS. Section VI analyzes some popular kernels within the proposed framework. Finally, Section VII presents a numerical validation of the proposed approach.

Notations: $\mathbb{R}$ and $\mathbb{N}$ denote the set of real and natural numbers, respectively ($0 \in \mathbb{N}$). Given $n, m \in \mathbb{N}$, $\mathbb{R}^{n \times m}$ denotes the set of $n \times m$ matrices and $\mathbb{N}_{\geq m} := \{x \in \mathbb{N} :$

$x \geq m\}$. Tuples of real numbers and column vectors are used interchangeably. Given $n \in \mathbb{N}$, $0_n := (0, \dots, 0) \in \mathbb{R}^n$, and $I_n \in \mathbb{R}^{n \times n}$ denotes the identity matrix. Given two matrices A and B , $\text{col}(A, B)$ and $\text{diag}(A, B)$ denote, respectively, their vertical and diagonal concatenations whenever they are well-defined. $|x|$ denotes the Euclidean norm of $x \in \mathbb{R}^n$. Let $T \subseteq \mathbb{N}$; we denote by $\mathbb{R}^T$ the set of real-valued sequences $x : T \rightarrow \mathbb{R}$ indexed by T . For each $x \in \mathbb{R}^T$ and $i, j \in T$, x_i is the i th element of x and $x_{j:i} := (x_j, \dots, x_i)$. Furthermore, $|x|_\infty := \sup_{t \in T} |x_t| \in [0, \infty]$. A function $\gamma : [0, \infty) \rightarrow [0, \infty)$ is of class $\mathcal{K}$ if it is continuous, strictly increasing, and $\gamma(0) = 0$. A function $\beta : [0, \infty) \times [0, \infty) \rightarrow [0, \infty)$ is of class $\mathcal{KL}$ if, for each $t \in [0, \infty)$, $\beta(\cdot, t)$ is of class $\mathcal{K}$, and for each $s \in [0, \infty)$, $\beta(s, \cdot)$ is continuous, decreasing, and $\lim_{t \rightarrow \infty} \beta(s, t) = 0$.

II. PROBLEM FORMULATION

Given a discrete-time system $\Psi \subseteq \mathbb{R}^{\mathbb{N}} \times \mathbb{R}^{\mathbb{N}}$ relating *input* sequences $u \in \mathbb{R}^{\mathbb{N}}$ to *output* sequences $y \in \mathbb{R}^{\mathbb{N}}$, we consider the general problem of estimating from data a *stable* model of Ψ that can be used for simulation and control purposes. While several *simulation error methods* (SEM) can be found in the literature that optimize the model selection for simulation [53]–[56], these techniques come with considerable issues: (i) the additional complexity of the underlying optimization problem (which is not convex); (ii) the need for a fixed simulation horizon; (iii) the consequent lack of guarantees of error accumulation in infinite-horizon simulation; (iv) the absence of stability guarantees. Instead, we approach the problem at hand as a *prediction-error estimation method with stability constraints*. In this way, we retain a simpler selection method, and we are able to impose stability constraints that also provide guarantees on error accumulation during simulation over infinite time horizons.

More specifically, given $m_u, m_y \in \mathbb{N}$ and an input-output pair $(u, y) \in \Psi$, we define the *prediction sequence* $\hat{y}^{\text{pre}}$ as

$$\forall t \in \mathbb{N}_{\geq \bar{m}}, \quad \hat{y}_t^{\text{pre}} := f(\xi_t, u_t),$$

where $\hat{y}_t^{\text{pre}}$ is the prediction of y_t , $f : \mathbb{R}^{m+1} \rightarrow \mathbb{R}$ is a function to be selected based on the available data, $\bar{m} := \max(m_u, m_y)$, $m := m_u + m_y$ and, for each $t \in \mathbb{N}_{\geq \bar{m}}$, $\xi_t := (y_{t-m_y:t-1}, u_{t-m_u:t-1}) \in \mathbb{R}^m$ collects m_u past values of the input u and m_y past values of the output y . We refer to m as the *model order*, to f as the *predictor*, and we define the *prediction error* sequence as $e := y - \hat{y}^{\text{pre}} \in \mathbb{R}^{\mathbb{N}}$. Once a predictor f is selected, an estimated model $\hat{\Psi}$ of Ψ employable for the purpose of simulation and control can be simply obtained using the predictor f iteratively, as a multiple-step-ahead prediction dynamical model. Namely, given any input u of System Ψ , a sequence $\hat{y} \in \mathbb{R}^{\mathbb{N}}$ is produced by the estimated model $\hat{\Psi}$ by letting, for each $t \in \mathbb{N}_{\geq \bar{m}}$, $\hat{y}_t := f(\hat{\xi}_t, u_t)$ with $\hat{\xi}_t := (\hat{y}_{t-m_y:t-1}, u_{t-m_u:t-1})$.

The aim of the paper is to devise a method to select f in such a way that $\hat{\Psi}$ is a good model and, in addition, as a system with input u and output $\hat{y}$ is stable in some suitable sense. The specific notions of stability that can be enforced

are formally defined later in Sections IV and V by making reference to an *auxiliary dynamical system* defined as follows

$$\Sigma(f) : \quad x_{t+1} = Ax_t + Gf(x_t, v_t) + G\varepsilon_t + Hv_t,$$

in which $x_t \in \mathbb{R}^m$ is the state, $\varepsilon_t \in \mathbb{R}$ and $v_t \in \mathbb{R}$ are two exogenous inputs, and where $A := \text{diag}(S_{m_y}, S_{m_u})$, $G := \text{col}(B_{m_y}, 0_{m_u})$, and $H := \text{col}(0_{m_y}, B_{m_u})$, with,

$$\forall q \in \{m_y, m_u\}, \quad S_q := \begin{bmatrix} 0_{q-1} & I_{q-1} \\ 0 & 0_{q-1}^\top \end{bmatrix}, \quad B_q := \begin{bmatrix} 0_{q-1} \\ 1 \end{bmatrix}.$$

In formalizing the required stability notions, we use $\Sigma(f)$ instead of directly $\hat{\Psi}$ to capture, at the same time, the stability requirements we want to enforce on the estimated model $\hat{\Psi}$ with respect to the input u and also the stability properties that lead to guarantees on the lack of accumulation of prediction errors in simulation. Both points are illustrated hereafter.

First, we notice that, by construction of the matrices A , G and H , the sequence $\hat{\xi}$ is a solution of $\Sigma(f)$ with $\varepsilon = 0$ and $v = u$. Namely, $\Sigma(f)$ with $\varepsilon = 0$ and $v = u$, and with output $\hat{y}_t = f(x_t, u_t)$, is a *state-space model* of $\hat{\Psi}$. As a consequence, requiring stability of $\Sigma(f)$ with respect to the initial conditions and to the input v implies an analogous stability property for the estimated model $\hat{\Psi}$ with respect to u .

Second, we notice that, for every $(u, y) \in \Psi$ and every $t \geq \bar{m}$, the sequence ξ is a solution of $\Sigma(f)$ with $v = u$ and $\varepsilon = e$ because, by the definition of e , $y_t = f(\xi_t, u_t) + e_t$ for every $t \geq \bar{m}$. Hence, $\Sigma(f)$, with inputs $v = u$ and $\varepsilon = e$, and output $y_t = f(x_t, u_t) + e_t$, is a *state-space model* of Ψ . Since the input v is the same in both the previous cases, the difference between y and its estimate $\hat{y}$ can then only be attributed to the different initialization and the different choice of ε . As a consequence, requiring stability of $\Sigma(f)$ with respect to the input ε means providing deterministic guarantees on the extent by which past prediction errors may affect future ones. To better illustrate this latter point, consider the case in which $\Sigma(f)$ is δ ISS (Definition 6 below). Since ξ is a solution of $\Sigma(f)$ with $v = u$ and $\varepsilon = e$, and $\hat{\xi}$ is a solution of $\Sigma(f)$ with $v = u$ and $\varepsilon = 0$, then δ ISS guarantees that $\xi - \hat{\xi}$, and thus $y - \hat{y}$, is bounded by a continuous function of the *maximum* prediction error attained before t . Hence, past prediction errors do not accumulate indefinitely during simulation, as only the maximum prediction error matters in the bound.

To learn a predictor f that fits the available data and satisfies the stability constraints, we assume to have at our disposal a *dataset* $\mathcal{D} := (\bar{u}_i, \bar{y}_i)_{i=1}^n$, which is a collection of $n \in \mathbb{N}_{>\bar{m}}$ data points $(\bar{u}_i, \bar{y}_i)$ taken from an input-output pair of Ψ starting from an unknown time $t_0 \in \mathbb{N}$. In particular, for all $i \in \{1, \dots, n\}$, $(\bar{u}_i, \bar{y}_i) := (u_{t_0+i-1}, y_{t_0+i-1})$, in which $(u, y) \in \Psi$. Then, in qualitative terms, our goal is to devise an algorithm that selects a predictor $f : \mathbb{R}^{m+1} \rightarrow \mathbb{R}$ in such a way that **(P1)** for every input-output pair $(u, y) \in \Psi$, the corresponding error sequence e is “small”, and **(P2)** $\Sigma(f)$ has a given desired stability property with respect to both v and ε . The fitting performance requirement **(P1)** will be approached in Section III as a regularized least-squares problem, as typical in system identification. The stability requirement **(P2)** is instead addressed by the insertion of suitable constraints in the identification procedure, and it will refer to several

specific stability properties. In particular, Section IV focuses on BIBS stability and ISS, while Section V is devoted to their incremental counterparts δ BIBS stability and δ ISS.

III. METHODOLOGY

This section introduces the proposed learning method as an extension of the kernel-based regularized regression for the selection of the predictor f enabling the possibility of enforcing desired stability conditions on $\Sigma(f)$. Here, we only present the general methodology, deferring the embedding of specific stability properties to the subsequent sections.

A. Regularized kernel-based learning

The kernel-based regularized regression is a well-established technique in both machine learning [57]–[59] and system identification [14], [41], [60]–[62] selecting the predictor function f within a hypothesis set of candidates according to a predetermined criterion, usually expressed as an optimization problem. The hypothesis set is a Reproducing Kernel Hilbert Space (RKHS) [63], [64] generated by a function $k_\eta : \mathbb{R}^{m+1} \times \mathbb{R}^{m+1} \rightarrow \mathbb{R}$ defined by a parameter $\eta \in \Phi_k \subseteq \mathbb{R}^{n_\eta}$ ($n_\eta \in \mathbb{N}$). We denote by $\mathcal{H}(k_\eta)$ the RKHS associated with k_η , while $\|\cdot\|_{k_\eta}$ and $\langle \cdot, \cdot \rangle_{k_\eta}$ denote the norm and the inner product on $\mathcal{H}(k_\eta)$, respectively. We refer to k as the *kernel structure*, and to k_η as the *actual kernel function* defined using the parameter η and the structure k . Furthermore, we restrict Φ_k to contain only values of η that make k_η a valid kernel function, namely, such that k_η is symmetric and positive semidefinite [63], [65]. Given $z \in \mathbb{R}^{m+1}$, we let $k_\eta^z : \mathbb{R}^{m+1} \rightarrow \mathbb{R}$ be the *representer function* (or *kernel slice*) of z , defined as $k_\eta^z(\cdot) := k_\eta(\cdot, z)$.

Given a kernel structure k and a dataset $\mathcal{D}$, the predictor f is chosen as the solution of the optimization problem

$$\min_{f \in \mathcal{H}(k_\eta)} \sum_{i=\bar{m}+1}^n (f(\bar{z}_i) - \bar{y}_i)^2 + \beta \|f\|_{k_\eta}^2, \quad (1)$$

where $(\beta, \eta) \in (0, \infty) \times \Phi_k$ are parameters to be tuned and, for each $t \in \{\bar{m} + 1, \dots, n\}$, $\bar{z}_t := (\bar{y}_{t-m_y:t-1}, \bar{u}_{t-m_u:t}) \in \mathbb{R}^{m+1}$ is taken from the dataset $\mathcal{D}$. The first term of the cost function weights how well the model fits the dataset. Hence, selecting f according to (1) addresses **(P1)** as is typical in kernel-based methods [41], [66]. Instead, the second term is a regularization term balancing the complexity of f and the adherence to the available data. As a consequence of the *Representer Theorem* [67], it can be shown that (1) has a unique ([68, Prop 8]) solution given by

$$f := \sum_{i=\bar{m}+1}^n c_i k_\eta^{\bar{z}_i} \in \mathcal{H}(k_\eta), \quad (2)$$

in which $c = c_{\bar{m}+1:n} = (K_\eta + \beta I_{n-\bar{m}})^{-1} \bar{y} \in \mathbb{R}^{n-\bar{m}}$, where $K_\eta \in \mathbb{R}^{(n-\bar{m}) \times (n-\bar{m})}$ is the matrix with the (i, j) th element given by $k_\eta(\bar{z}_{\bar{m}+i}, \bar{z}_{\bar{m}+j})$, and $\bar{y} := \bar{y}_{\bar{m}+1:n} \in \mathbb{R}^{n-\bar{m}}$.

The shape of the kernel function k_η determines the properties of all the functions of the RKHS, and, hence, on the estimated predictor f . Under mild assumptions, the literature, e.g. [69], provides probabilistic bounds on the one-step

prediction error obtained with the estimated predictor f . In the case of linear systems, the problem of selecting a good kernel structure is well-studied, and it is possible to select a kernel structure guaranteeing desirable properties including asymptotic stability [39], [62]. The nonlinear case is instead less studied (see, e.g., [29], [41], [47]), and the proposed kernels are typically not as well-motivated as in the linear case. Moreover, it is typically not possible to embed stability constraints directly within k (see Remark 1). Instead, as it will be clarified in the following, the key to impose stability turns out to be the regularization term in (1), which favors functions with smaller norm.

Determining f by solving (1) requires the selection of the hyperparameters β and η . Their value is of primary importance because it determines the trade-off between complexity of the model and fitting performance, thereby causing or avoiding over- and under-fitting. In the literature, this problem is also known as *model selection* (or *assessment*), and it is historically solved using techniques based on cross-validation [65], [70]. In particular, *k-fold cross-validation* [52], [65], *leave-one-out cross-validation* [51], [59], [65], and *generalized cross-validation* [50] are popular methods that can be used in a wide range of applications. More recently, also *Empirical Bayes (EB)* [49] methods have shown to be very efficient [65, Sec 5.4.1]. All the previous techniques have successfully been used in kernel-based identification for both linear [71]–[73] and nonlinear [41], [43], [66] systems when no stability guarantee is contemplated for the identified models, and they all boil down to the minimization of a cost function. Specifically, the general model selection problem can be cast in terms of the following optimization problem

$$\min_{\beta \in [\iota, \infty), \eta \in \Phi_k} J(\beta, \eta), \quad (3)$$

in which the cost function $J : (0, \infty) \times \Phi_k \rightarrow \mathbb{R}$ defines the actual method implemented, and $\iota \in (0, \infty)$ is a parameter introduced to make the feasibility set closed (which is required for the numerical solution of the problem [74, Ch 2]). It is important to notice that J in (3) typically depends on the dataset $\mathcal{D}$, and often explicitly on the solution of (1). Therefore, (1) and (3) may form a *bilevel optimization problem* [75] whose numerical solution may require a complex iterative procedure solving (1) and (3) multiple times.

B. Modifying the kernel method for stability guarantees

The method explained so far in Section III-A only concerns Point (P1), but it does not address Point (P2), since no specific stability property is a priori guaranteed on the learned model $\Sigma(f)$. To address Point (P2), we modify the optimization problems (1) and (3) to enforce a specific stability property on the learned model. In particular, we first modify Problem (1) by adding an upper bound on the norm of f . The modified problem reads as follows

$$\begin{cases} \min_{f \in \mathcal{H}(k_\eta)} & \sum_{i=\bar{m}+1}^n (f(\bar{z}_i) - \bar{y}_i)^2 + \alpha \|f\|_{k_\eta}^2 \\ \text{subject to} & m_y \|f\|_{k_\eta}^2 \leq \chi \end{cases} \quad (4)$$

in which $\alpha \in [0, \infty)$ is a regularization parameter and $\chi \in (0, 1)$ is a parameter less than, but close to, 1 needed to ensure that the feasibility set is closed. Second, we restrict the kernel hyperparameter η to a set $\Omega_k \subseteq \Phi_k$, called *viability set*, whose definition determines the actual desired stability property on $\Sigma(f)$. The construction of the viability set Ω_k is discussed in detail in Sections IV and V. Once Ω_k is given, Problem (3) is modified as follows

$$\min_{\alpha \in [\iota, \infty), \eta \in \Omega_k} J(\alpha, \eta). \quad (5)$$

In Sections IV and V, we show that, by defining a suitable viability set Ω_k , the model obtained by solving the modified problems (4)–(5) is guaranteed to be BIBS stable, ISS, δ BIBS stable and/or δ ISS. To this aim, we follow a rationale similar to [32], [33], also used in the preliminary conference version of this article [76]. We first derive a sufficient condition on the predictor f guaranteeing that $\Sigma(f)$ has the desired stability property; then, we define the viability set Ω_k to ensure that, for each $\eta \in \Omega_k$, the predictor f , obtained by solving (5) and (4), satisfies such a sufficient condition. It is worth noticing that the proposed approach does not rely on a specific cost function J used in (3). Hence, it is possible to use all the various model assessment methods discussed in Section III-A that can be written in the form of (3).

Regarding solvability, we notice that Problem (4) with $\alpha = 0$ is an instance of *Ivanov regularization* [77], and it is equivalent to (1) for a specific value of β [78]. Lemma 1 below (proved in Appendix A) extends this result to the case of $\alpha > 0$, which is needed to cover (4). The lemma uses the function $\gamma : (0, \infty) \rightarrow \mathbb{R}$ defined as

$$\gamma(a) := m_y \bar{y}^\top (K_\eta + aI_{n-\bar{m}})^{-1} K_\eta (K_\eta + aI_{n-\bar{m}})^{-1} \bar{y} - \chi. \quad (6)$$

Lemma 1. *The solution of (4) can be obtained by solving (1) with $\beta = \max(\bar{\alpha}, \alpha)$, where $\bar{\alpha} \in [0, \infty)$ is such that $\gamma(\bar{\alpha}) = 0$, if it exists, or $\bar{\alpha} = 0$ otherwise.*

Lemma 1 implies that the constraint on the RKHS norm in (4) is potentially translated to a lower bound on β in (1). This is in line with the typical effect of increasing the regularization parameter, which forces the optimal predictor to have a lower norm. The predictor f that solves (4) may have a larger one-step prediction error with respect to the unconstrained optimization problem (1) as the added constraint limits the feasible set. Nevertheless, in view of Lemma 1, the bounds on the one-step prediction error known for unconstrained problems (e.g. [69]) can still be applied, albeit with a potentially larger regularization parameter.

Finally, we stress that, as in the classic kernel method described in Section III-A, also Problems (5) and (4) may give rise to a bilevel optimization problem with the consequent complications, since J presents the same dependency on the dataset and/or on the solution of (4) as before. Nevertheless, we underline that the viability set Ω_k does not depend on the dataset. Therefore, if (5) is only solved for feasibility and not optimality, Problems (5) and (4) can be solved once and sequentially.

IV. LEARNING BIBS STABLE AND ISS MODELS

This section shows how the methodology proposed in Section III-B can be used to guarantee that the learned model is BIBS stable, ISS, or has a *weak Asymptotic Gain* (wAG) property. We focus only on the case where $m_y > 0$ because, if $m_y = 0$ (which corresponds to a *finite impulse response* (FIR) dynamics), then $\Sigma(f)$ does not depend on f and reduces to a contractive LTI system driven by v . Hence, in this case the stability properties of the one-step ahead predictor and the simulator can be simply inferred from the continuity and boundedness properties of f , as done e.g. in [19, Prop 11].

A. Stability notions and results

BIBS stability, ISS, and wAG are formalized hereafter.

Definition 1 ([34]). *System $\Sigma(f)$ is said to be BIBS stable if for every $q \in [0, \infty)$, there exists $w \in [0, \infty)$, such that every solution (x, v, ε) of $\Sigma(f)$ satisfies*

$$|x_0| \leq q, |v|_\infty \leq q, |\varepsilon|_\infty \leq q \Rightarrow |x|_\infty \leq w.$$

Definition 2 ([36], [37]). *System $\Sigma(f)$ is said to be ISS if there exist a class- $\mathcal{KL}$ function α and two class- $\mathcal{K}$ functions κ_u and κ_e such that every solution (x, v, ε) of $\Sigma(f)$ satisfies*

$$\forall t \in \mathbb{N}, |x_t| \leq \alpha(|x_0|, t) + \kappa_u(|v_{0:t}|_\infty) + \kappa_e(|\varepsilon_{0:t}|_\infty).$$

Definition 3. *System $\Sigma(f)$ is said to have the wAG property if there exist $b \in [0, \infty)$ and a class- $\mathcal{K}$ function κ such that every solution (x, v, ε) of $\Sigma(f)$ satisfies*

$$\limsup_{t \rightarrow \infty} |x_t| \leq b + \kappa \left(\limsup_{t \rightarrow \infty} \max(|v_t|, |\varepsilon_t|) \right).$$

In particular, b and κ are called the asymptotic bias and asymptotic gain, respectively. If $b = 0$, we say that $\Sigma(f)$ has the asymptotic gain (AG) property [35].

The following proposition establishes some sufficient conditions on the predictor f guaranteeing that $\Sigma(f)$ has one or more of the considered stability properties.

Proposition 1. *Suppose that there exist $\mu \in [0, 1)$, $\rho \in [0, \infty)$, $d \in [0, \infty)$, and a class- $\mathcal{K}$ function ω such that, for every $x \in \mathbb{R}^m$ and $v \in \mathbb{R}$, the following conditions hold*

$$|x|^2 + |v|^2 \geq \rho \Rightarrow |f(x, v)|^2 \leq \frac{\mu}{m_y} |x|^2 + \omega(|v|), \quad (7a)$$

$$|x|^2 + |v|^2 < \rho \Rightarrow |f(x, v)| \leq d. \quad (7b)$$

Then:

- A. $\Sigma(f)$ is BIBS stable.
- B. If f is continuous, $\Sigma(f)$ has the wAG property with asymptotic bias $b(\rho)$, for some class- $\mathcal{K}$ function b .
- C. If (7) holds with $\rho = 0$, $\Sigma(f)$ is ISS.

The proof of Proposition 1 is omitted, since it follows similar arguments of the proof of Proposition 3 stated in the next section.

B. Learning with stability constraints

To show that the procedure explained in Section III-B can be used to guarantee that the learned models are BIBS stable and/or ISS, we proceed in two steps:

- S1.** We derive a condition on the kernel function ensuring that all the elements of the associated RKHS satisfy Conditions (7) with $\rho, \mu, d \in [0, \infty)$ and ω of class- $\mathcal{K}$.
- S2.** We design the viability set Ω_k in order to guarantee that the predictor f selected by solving (4) satisfies Conditions (7) with $\mu \in [0, 1)$.

Steps **S1** and **S2** guarantee that every feasible predictor satisfies the assumptions of Proposition 1 thus yielding a stable $\Sigma(f)$.

Remark 1. We stress that satisfaction of the assumptions of Proposition 1 cannot be achieved in one step by just shaping the kernel structure k . In particular, the only RKHS whose functions all satisfy (7a) with $m_y > 0$ is the trivial space, which only contains the 0 function. Indeed, if for some $\eta \in \Phi_k$, $f \in \mathcal{H}(k_\eta) \setminus \{0\}$ satisfies (7a) with $\mu \in (0, 1)$, then $\mu^{-1}f \in \mathcal{H}(k_\eta)$ may not satisfy (7a).

Proposition 2 (proved in Appendix B) fulfills Step **S1** by providing sufficient conditions on a kernel function k_η for which (7) holds for all functions $f \in \mathcal{H}(k_\eta)$, although with μ possibly larger than 1.

Proposition 2. *Let k be a kernel structure and $\eta \in \Phi_k$. Assume that there exist $\nu, s \in [0, \infty)$ such that*

$$\forall a \in \mathbb{R}^{m+1}, \quad |a|^2 \geq \nu \Rightarrow k_\eta(a, a) \leq |a|^2, \quad (8a)$$

$$\forall a \in \mathbb{R}^{m+1}, \quad |a|^2 < \nu \Rightarrow k_\eta(a, a) \leq s. \quad (8b)$$

Then, every function $f \in \mathcal{H}(k_\eta)$ satisfies conditions (7) with $\mu = m_y |f|_{k_\eta}^2$, $\rho = \nu$, $d = \sqrt{s} |f|_{k_\eta}$ and $\omega(p) = ap^2$ with any $a > |f|_{k_\eta}^2$.

Let k_η be a kernel function satisfying the assumptions of Proposition 2. Then, each $f \in \mathcal{H}(k_\eta)$ satisfies (7) with $\mu = m_y |f|_{k_\eta}^2$. Since the constraint $m_y |f|_{k_\eta}^2 \leq \chi < 1$ is already present in (4), then Step **S2** is fulfilled if we can properly define the viability set Ω_k in such a way that all the solutions of (5) satisfy the assumptions of Proposition 2. This leads to the following definition.

Definition 4. *Let k be a kernel structure. Given $\rho \in [0, \infty]$, we define the ρ -viability set of k as*

$$\Theta_k^\rho := \{ \eta \in \Phi_k : \exists \nu \in [0, \rho] \cap \mathbb{R}, \exists s \in [0, \infty), (8) \text{ holds} \}$$

If $\Theta_k^\rho \neq \emptyset$, we say that the kernel structure k is ρ -viable.

Joining Steps **S1** and **S2** and using Proposition 1 directly leads to Theorem 1 below, which is the main result of the present section.

Theorem 1. *Let $\rho \in [0, \infty]$, k be a ρ -viable kernel structure, and (α, η) be an optimal solution of (5) obtained with $\Omega_k = \Theta_k^\rho$. Let f be the optimal solution of (4). Then, $\Sigma(f)$ is BIBS stable. If, in addition, f is continuous, then $\Sigma(f)$ has the wAG property with asymptotic bias $b(\rho)$, where b is a class- $\mathcal{K}$ function. Finally, if $\rho = 0$, then $\Sigma(f)$ is ISS.*

It is important to observe that a solution of (5) exists only if $\Theta_k^\rho \neq \emptyset$, namely, if k is ρ -viable for the chosen $\rho \in [0, \infty]$. ρ -viability of some of the most commonly used kernels is discussed in Section VI, where we also characterize their ρ -viability set.

We observe that ∞ -viability is sufficient for BIBS stability, while 0-viability is sufficient for ISS. Since $\rho \mapsto \Theta_k^\rho$ is increasing, achieving BIBS stability is easier than ISS, which matches with the fact that ISS is a stricter stability condition. Moreover, ρ -viability for $\rho \in (0, \infty)$ can be interpreted as an “intermediate” stability property between ISS and BIBS stability. Finally, we underline that Theorem 1 also states that $\Sigma(f)$ yields the wAG property if k is ρ -viable for some ρ and f is continuous. In this respect, we point out that, since f is a linear combination of representer functions, f is continuous when the functions $k(z, \cdot)$ are continuous for each $z \in \mathbb{R}^{m+1}$.

Remark 2. Regarding the BIBS stability requirement, we stress that, in principle, BIBS stability may be obtained whenever the predictor is a bounded function. Hence, every bounded kernel produces a BIBS stable system. However, the bound achievable in this way corresponds to a limit case of Proposition 1 in which $\rho = \infty$. As clear from the proof of the proposition, this gives no information on how the amplitude of the input affects that of the output. The result of Theorem 1, instead, provides a bound composed of two terms, the first depending on ρ and independent of the input, and the second given by a class- $\mathcal{K}$ function of the norm of the input. Therefore, for $\rho = 0$, Theorem 1 also provides a characterization of *continuity* of the input-output operator associated with the predictor. For $\rho > 0$, a bias is introduced scaling with ρ , thereby providing a combined effect of continuity and uniform boundedness.

Remark 3. 0-viability, needed in Theorem 1 to prove ISS, implies $f(0_{m+1}) = 0$. Conversely, ISS implies that (x, v, ε) , with $x_t = 0_m$ and $\varepsilon_t = v_t = 0$ for all $t \in \mathbb{N}$, is a solution of $\Sigma(f)$, which in turn implies $f(0_{m+1}) = 0$. Therefore, Theorem 1 only characterizes ISS with respect to the origin. This property is also reflected in Condition (7a) that implies $k(0_{m+1}, 0_{m+1}) = 0$ and, therefore, $f(0_{m+1}) = 0$ for each $f \in \mathcal{H}(k_\eta)$ [64, Prop 2.3]. Imposing ISS with respect to a different known equilibrium pair (x^*, v^*) is possible by applying the preliminary data transformation $(x, v) \mapsto (x - x^*, v - v^*)$ before using Theorem 1.

V. LEARNING δ BIBS STABLE AND δ ISS MODELS

This section shows how the methodology proposed in Section III-B can be used to guarantee that the learned model is δ BIBS stable, δ ISS, and/or has a *weak incremental asymptotic gain* (w δ AG) property. We follow the same line of reasoning of Section IV and provide the “incremental version” of the results therein. As in Section IV, also this section focuses on the case $m_y > 0$ for similar reasons.

A. Incremental stability notions and results

Consider two solutions $(x^a, v^a, \varepsilon^a)$ and $(x^b, v^b, \varepsilon^b)$ of $\Sigma(f)$ and let $\tilde{x} := x^a - x^b$, $\tilde{v} := v^a - v^b$, and $\tilde{\varepsilon} := \varepsilon^a - \varepsilon^b$. The

notions of δ BIBS stability, δ ISS, and w δ AG are formalized in the following definitions.

Definition 5. System $\Sigma(f)$ is said to be δ BIBS stable if, for every $q \in [0, \infty)$, there exists $w \in [0, \infty)$, such that every two solutions $(x^a, v^a, \varepsilon^a)$ and $(x^b, v^b, \varepsilon^b)$ of $\Sigma(f)$ satisfy

$$|\tilde{x}_0| \leq q, |\tilde{v}|_\infty \leq q, |\tilde{\varepsilon}|_\infty \leq q \Rightarrow |\tilde{x}|_\infty \leq w.$$

Definition 6 ([38]). System $\Sigma(f)$ is said to be δ ISS if there exist a class- $\mathcal{KL}$ function α and two class- $\mathcal{K}$ functions κ_u and κ_e such that every two solutions $(x^a, v^a, \varepsilon^a)$ and $(x^b, v^b, \varepsilon^b)$ of $\Sigma(f)$ satisfy

$$\forall t \in \mathbb{N}, |\tilde{x}_t| \leq \alpha(|\tilde{x}_0|, t) + \kappa_u(|\tilde{v}_{0:t}|_\infty) + \kappa_e(|\tilde{\varepsilon}_{0:t}|_\infty).$$

Definition 7. System $\Sigma(f)$ is said to have w δ AG property if there exist $b \in [0, \infty)$ and a class- $\mathcal{K}$ function κ such that every two solutions $(x^a, v^a, \varepsilon^a)$ and $(x^b, v^b, \varepsilon^b)$ of $\Sigma(f)$ satisfy

$$\limsup_{t \rightarrow \infty} |\tilde{x}_t| \leq b + \kappa \left(\limsup_{t \rightarrow \infty} \max(|\tilde{v}_t|, |\tilde{\varepsilon}_t|) \right).$$

In particular, b and κ are called incremental asymptotic bias and incremental asymptotic gain, respectively. If $b = 0$, we say that $\Sigma(f)$ has the incremental asymptotic gain (δ AG) property.

Proposition 3 (proved in Appendix C) stated below parallels Proposition 1 by giving sufficient conditions on the predictor f guaranteeing that $\Sigma(f)$ is δ BIBS stable, δ ISS, and has the w δ AG property.

Proposition 3. Suppose that there exist $\mu \in [0, 1)$, $\rho \in [0, \infty)$, $d \in [0, \infty)$, and a class- $\mathcal{K}$ function ω such that, for all $x^a, x^b \in \mathbb{R}^m$ and $v^a, v^b \in \mathbb{R}$, the following hold

$$|\tilde{x}|^2 + |\tilde{v}|^2 \geq \rho \Rightarrow |\tilde{f}|^2 \leq \frac{\mu}{m_y} |\tilde{x}|^2 + \omega(|\tilde{v}|), \quad (9a)$$

$$|\tilde{x}|^2 + |\tilde{v}|^2 < \rho \Rightarrow |\tilde{f}| \leq d, \quad (9b)$$

where $\tilde{x} = x^a - x^b$, $\tilde{v} = v^a - v^b$, and $\tilde{f} = f(x^a, v^a) - f(x^b, v^b)$. Then:

- A. $\Sigma(f)$ is δ BIBS stable.
- B. If f is uniformly continuous, $\Sigma(f)$ has the w δ AG property with incremental asymptotic bias $b(\rho)$, for some class- $\mathcal{K}$ function b .
- C. If (9) holds with $\rho = 0$, $\Sigma(f)$ is δ ISS.

B. Learning with incremental stability constraints

To show that the procedure explained in Section III-B can be used to guarantee that the learned models are δ BIBS stable and/or δ ISS, we follow the same line of reasoning of Section IV-B, and we proceed in two steps (cf. **S1-S2**):

- δ S1.** We derive a condition on the kernel function ensuring that all the elements of the associated RKHS satisfy Conditions (9) with $\rho, \mu, d \in [0, \infty)$ and ω of class- $\mathcal{K}$.
- δ S2.** We design the viability set Ω_k in order to guarantee the predictor f selected by solving (4) satisfies Conditions (9) with $\mu \in [0, 1)$.

Step δ S1 is fulfilled by the following proposition, which is analogous to Proposition 2.

Proposition 4. Let k be a kernel structure and $\eta \in \Phi_k$. Assume that there exist $\nu, s \in [0, \infty)$ such that

$$\forall a, b \in \mathbb{R}^{m+1}, \quad |a - b|^2 \geq \nu \Rightarrow h_\eta(a, b) \leq |a - b|^2, \quad (10a)$$

$$\forall a, b \in \mathbb{R}^{m+1}, \quad |a - b|^2 < \nu \Rightarrow h_\eta(a, b) \leq s, \quad (10b)$$

with $h_\eta(a, b) := k_\eta(a, a) - 2k_\eta(a, b) + k_\eta(b, b)$. Then, every function $f \in \mathcal{H}(k_\eta)$ satisfies Conditions (9) with $\mu = m_y |f|_{k_\eta}^2$, $\rho = \nu$, $d = \sqrt{s} |f|_{k_\eta}$, and $\omega(p) = ap^2$ with any $a > |f|_{k_\eta}^2$.

The function h_η defined in Proposition 4 is the squared kernel metric of k_η [79, Sec 4.2]. Hence, Condition (10) is equivalent to asking that the canonical feature map of k_η is non-expansive for points whose distance is large, and incrementally bounded for points that are closer.

As before, Step $\delta S2$ is fulfilled if we can properly define the viability set Ω_k in such a way that all the solutions of (5) satisfy the assumptions of Proposition 4. This leads to the following definition (cf. Definition 4).

Definition 8. Let k be a kernel structure. Given $\rho \in [0, \infty]$, we define the ρ - δ viability set of k as

$$\Delta_k^\rho := \{\eta \in \Phi_k : \exists \nu \in [0, \rho] \cap \mathbb{R}, \exists s \in [0, \infty), (10) \text{ holds}\}.$$

If $\Delta_k^\rho \neq \emptyset$, we say that the kernel structure k is ρ - δ viable.

The ρ - δ viability of some of the most commonly used kernels is discussed in Section VI along with their viability sets. Joining $\delta S1$ and $\delta S2$ and using Proposition 4 directly leads to Theorem 2 below (analogous to Theorem 1).

Theorem 2. Let $\rho \in [0, \infty]$, k be a ρ - δ viable kernel structure, and (α, η) be an optimal solution of (5) obtained with $\Omega_k = \Delta_k^\rho$. Let f be the optimal solution of (4). Then, $\Sigma(f)$ is $\delta BIBS$ stable. If, in addition, f is uniformly continuous, then $\Sigma(f)$ has the $w\delta AG$ property with incremental asymptotic bias $b(\rho)$, where b is a class- $\mathcal{K}$ function. Finally, if $\rho = 0$, then $\Sigma(f)$ is δISS .

Since $\rho \mapsto \Delta_k^\rho$ is increasing, the same relationship among the conditions for BIBS stability and ISS discussed in Section IV-B holds for their incremental version as well. Moreover, since f is a linear combination of kernel functions, it is uniformly continuous when so are the functions k_η^z for each $z \in \mathbb{R}^{m+1}$. Finally, we underline that similar conclusions of those given in Remark 2 apply to $\delta BIBS$ stability as well.

VI. KERNEL VIABILITY

In this section, we consider some of the most commonly used kernels, and we derive their ρ -viability and ρ - δ viability sets, thereby connecting them with the stability results presented in the previous sections. Before delving into specific kernels, it is possible to state two general results and draw some preliminary considerations. Specifically, Proposition 5 below states that, if the kernel is continuous (resp. uniformly continuous), Condition (8b) (resp. Condition (10b)) is always satisfied and ρ -viability only depends on Condition (8a) (resp. the ρ - δ viability only depends on (10a)). The proof of Proposition 5 directly follows by the fact that continuous functions are locally bounded, and a similar property holds for uniformly

continuous functions (see Lemma 2 in Appendix C.3). Proposition 6, instead, states that all bounded kernels are ∞ -viable and ∞ - δ viable.

Proposition 5 (Continuous kernels). Let k be a kernel structure such that, for all $\eta \in \Phi_k$, the kernel function k_η is continuous (resp. uniformly continuous). Then, for each $\rho \in [0, \infty]$, it holds that $\Theta_k^\rho = \{\eta \in \Phi_k : \exists \nu \in [0, \rho] \cap \mathbb{R}, (8a) \text{ holds}\}$ (resp. $\Delta_k^\rho = \{\eta \in \Phi_k : \exists \nu \in [0, \rho] \cap \mathbb{R}, (10a) \text{ holds}\}$).

Proposition 6 (Bounded kernels). Let k be a kernel structure such that, for each $\eta \in \Phi_k$, k_η is bounded. Then, k is ∞ -viable and ∞ - δ viable with $\Theta_k^\infty = \Delta_k^\infty = \Phi_k$.

Proof. Let $b_\eta \in (0, \infty)$ be such that $|k_\eta(a, b)| \leq b_\eta$ for all $a, b \in \mathbb{R}^{m+1}$. Condition (8b) trivially holds with $\nu = s = b_\eta$. Similarly, also (8a) holds with $\nu = b_\eta$ since $|a|^2 \geq b_\eta \Rightarrow |a|^2 \geq k_\eta(a, a)$ for all $a \in \mathbb{R}^{m+1}$. Hence, $\Theta_k^\infty = \Phi_k$. Finally, $\Delta_k^\infty = \Phi_k$ can be proved similarly by noticing that, for every $a, b \in \mathbb{R}^{m+1}$, $h_\eta(a, b) \leq 4b_\eta$. $\square$

A. Degenerate kernels

This section considers a *degenerate* kernel structure k [65, Def 4.1], namely, such that, for each $\eta \in \Phi_k$, there exist $p_\eta \in \mathbb{N}$ and $\Gamma_\eta : \mathbb{R}^{m+1} \rightarrow \mathbb{R}^{p_\eta}$, such that $k_\eta(a, b) = \Gamma_\eta(a)^\top \Gamma_\eta(b)$ for every $a, b \in \mathbb{R}^{m+1}$. Since $k_\eta(a, a) = |\Gamma_\eta(a)|^2$ and $h_\eta(a, b) = |\Gamma_\eta(a) - \Gamma_\eta(b)|^2$, Conditions (8) and (10) can be easily rewritten in terms of Γ_η . Moreover, k_η is continuous or bounded if Γ_η enjoys the same property. Therefore, Proposition 6 and the first statement of Proposition 5 can easily be rewritten in terms of Γ_η instead of k_η .

A relevant degenerate kernel is the linear affine kernel analyzed in Proposition 7 below (proved in Appendix E). It defines the space of all linear affine functions. Hence, it produces linear models. Although there are more specialized methods for kernel-based identification of linear systems [14], Proposition 7 shows that, for 0-viability, the hyperparameter that regulates the affine part of the kernel needs to be 0. In this way, we guarantee that the space only contains linear functions. Furthermore, the 0- δ viability set, the ∞ - δ viability set, and the ∞ -viability are the same. This is consistent with the fact that δISS , $\delta BIBS$ stability, and BIBS stability are all equivalent notions for linear systems.

Proposition 7 (Linear affine kernel). Let k be a kernel structure such that $\Phi_k = [0, \infty)^2$ and $k_\eta(a, b) = \tau a^\top b + \sigma$ for all $a, b \in \mathbb{R}^{m+1}$ and $\eta = (\tau, \sigma) \in \Phi_k$. Then, for each $\rho \in [0, \infty]$, it follows that $\Theta_k^\rho = \{(\tau, \sigma) \in \Phi_k : \tau \in [0, 1] \wedge \sigma \in [0, \rho(1 - \tau)]\} \cup (1, 0)$ and $\Delta_k^\rho = [0, 1] \times [0, \infty)$.

Another common degenerate kernel is the polynomial kernel analyzed in Proposition 8 below (proved in Appendix F). This kernel is widely used in the literature [80]–[82] because it defines the space of polynomials with an arbitrary degree [57, Prop 2.1]. Unfortunately, this kernel is only viable when it is reduced to the linear kernel previously analyzed.

Proposition 8 (Polynomial kernel). Let k be a kernel structure such that $\Phi_k = \{2, \dots\} \subseteq \mathbb{N}$ and $k_\eta(a, b) = (a^\top b)^\eta$ for all $a, b \in \mathbb{R}^{m+1}$ and $\eta \in \Phi_k$. Then, for every $\rho \in [0, \infty]$, $\Theta_k^\rho = \Delta_k^\rho = \emptyset$.

Although Proposition 8 gives a somewhat negative result, nonlinear polynomial kernels (as well as other degenerate kernels) can always be modified to enable their usage to learn stable systems. This is briefly discussed in Section VI-C.

B. Stationary Kernels

Stationary kernels enjoy many relevant properties (see [65, Sec 4.2.1] or [57, Sec 4.4] for more details), and they include many widely-used kernels such as the *Gaussian Kernel* [83] and all *isotropic kernels*.

Definition 9. A kernel structure k is said to be stationary if, for every $\eta \in \Phi_k$, there exists a function $\bar{k}_\eta : \mathbb{R}^{m+1} \rightarrow \mathbb{R}$, such that $k_\eta(a, b) = \bar{k}_\eta(a - b)$ for all $a, b \in \mathbb{R}^{m+1}$.

The following result characterizes viability of stationary kernels.

Theorem 3 (Stationary kernels). *Let k be a stationary kernel structure. Then, for each $\rho \in [0, \infty]$,*

$$\begin{aligned}\Theta_k^\rho &= \{\eta \in \Phi_k : \bar{k}_\eta(0_{m+1}) \leq \rho\}, \\ \Delta_k^\rho &= \{\eta \in \Phi_k : \forall z \in \mathbb{R}^{m+1}, \\ &\quad |z|^2 \geq \rho \Rightarrow 2\bar{k}_\eta(0_{m+1}) - 2\bar{k}_\eta(z) \leq |z|^2\}.\end{aligned}$$

Moreover, $\Delta_k^\rho \supseteq \{\eta \in \Phi_k : 4\bar{k}_\eta(0_{m+1}) \leq \rho\}$.

Theorem 3 is proved in Appendix G. A first immediate consequence of Theorem 3 is that all stationary kernels are always ∞ -viable and ∞ - δ viable and, hence, they can be used to learn BIBS stable and δ BIBS stable models without additional restrictions on the hyperparameters. This is formalized by the following corollary.

Corollary 1. *Let k be a stationary kernel structure. Then, $\Theta_k^\infty = \Delta_k^\infty = \Phi_k$.*

A second relevant consequence of Theorem 3 is that no non-trivial stationary kernel is 0-viable. Hence, stationary kernels are generally not suitable to learn ISS systems.

Corollary 2. *Let k be a stationary kernel structure. Then, $\Theta_k^0 = \{\eta \in \Phi_k : \forall a, b \in \mathbb{R}^{m+1}, k_\eta(a, b) = 0\}$, that is, no non-trivial stationary kernel is 0-viable.*

Proof. The claim directly follows from Theorem 3 since $\bar{k}_\eta(0_{m+1}) \leq 0$ if and only if $k_\eta(a, a) = 0$ for all $a \in \mathbb{R}^{m+1}$, which is true if and only if $k_\eta(a, b) = 0$ for all $a, b \in \mathbb{R}^{m+1}$ (see (22)). $\square$

Two important stationary kernels are the Gaussian Kernel and the Matérn Kernel [84]. They are analyzed in Propositions 9 and 10 below (proved in Appendix H and Appendix I, respectively). Proposition 10 only considers the first degree Matérn Kernel for simplicity, but analogous results can be obtained with larger degrees. Both kernels are 0- δ viable, provided the hyperparameters satisfy certain conditions for δ ISS.

Proposition 9 (Gaussian Kernel). *Let k be a stationary kernel structure such that $\Phi_k = (0, \infty) \times [0, \infty)^2$ and, for every $z \in$*

$\mathbb{R}^{m+1}$ and $\eta = (\gamma, \tau, \sigma) \in \Phi_k$, $\bar{k}_\eta(z) = \tau \exp(-\gamma|z|^2) + \sigma$. Then, for each $\rho \in [0, \infty]$, we have

$$\begin{aligned}\Theta_k^\rho &= \{(\tau, \gamma, \sigma) \in \Phi_k : \tau + \sigma \leq \rho\}, \\ \Delta_k^\rho &= \begin{cases} \{(\tau, \gamma, \sigma) \in \Phi_k : 2\tau\gamma \leq 1\}, & \text{if } \rho = 0, \\ \{(\tau, \gamma, \sigma) \in \Phi_k : v(\tau, \gamma) \leq \rho\} & \text{if } \rho \in (0, \infty), \\ \Phi_k, & \text{if } \rho = \infty, \end{cases}\end{aligned}$$

in which $v(\tau, \gamma) = 2\tau + \gamma^{-1}W(-2\gamma\tau e^{-2\gamma\tau})$ and W is the principal branch of the Lambert W function [85]. Additionally, $\Delta_k^\rho \supseteq [0, \frac{\rho}{2}] \times [0, \infty)^2$.

Proposition 10 (Matérn Kernel). *Let k be a stationary kernel structure such that $\Phi_k = [0, \infty)^3$ and $\bar{k}_\eta(z) = \tau(1 + \gamma\sqrt{3}|z|) \exp(-\gamma\sqrt{3}|z|) + \sigma$, for all $z \in \mathbb{R}^{m+1}$ and $\eta = (\tau, \gamma, \sigma) \in \Phi_k$. Then, for each $\rho \in [0, \infty]$, we have*

$$\begin{aligned}\Theta_k^\rho &= \{(\tau, \gamma, \sigma) \in \Phi_k : \tau + \sigma \leq \rho\}, \\ \Delta_k^0 &= \{(\tau, \gamma, \sigma) \in \Phi_k : 3\tau\gamma^2 \leq 1\}.\end{aligned}$$

Finally, Proposition 11 below (proved in Appendix J) analyzes the kernel proposed in [41], which is specifically designed for the identification of predictors for NARX systems. Loosely speaking, this kernel is designed so that older measurements weight less than recent ones in the computation of the predicted value (with the parameter ξ playing the role of a forgetting factor). Since this is a stationary kernel, according to Corollary 2 it is not suitable for learning ISS systems, but it guarantees that the learned model is δ BIBS and BIBS stable without constraining the hyperparameters' selection optimization problem, i.e. (3) and (5) are equivalent.

Proposition 11. *Let k be a stationary kernel structure such that $\Phi_k = [0, \infty)^3 \times \{1, \dots, \bar{m}\}$ and, for all $z = (y_{1:m_y}, u_{1:m_u+1}) \in \mathbb{R}^{m+1}$ and $\eta = (\tau, \gamma, \xi, p) \in \Phi_k$,*

$$\bar{k}_\eta(z) = \tau \sum_{t=0}^{\bar{m}-p} \exp(-\xi t - \gamma|z_t|^2),$$

in which $z_t = (y_{t+1:p_y}, u_{t+1:p_u})$ where $p_y := \min(t + p, m_y)$ and $p_u := \min(t + p, m_u + 1)$. Then

$$\begin{aligned}\Theta_k^0 &= \{(\tau, \gamma, \xi, p) \in \Phi_k : \tau = 0\}, \\ \Delta_k^0 &= \{(\tau, \gamma, \xi, p) \in \Phi_k : 2\gamma\tau\pi(\xi, p) \leq 1\},\end{aligned}$$

where $\pi(\xi, p) := \bar{m} - p + 1$ if $\xi = 0$ and $\pi(\xi, p) := \frac{1 - e^{-(\bar{m}-p+1)\xi}}{1 - e^{-\xi}}$ if $\xi > 0$.

We emphasize that Proposition 11 validates the rationale used in [41] behind the construction of the kernel. Firstly, notice that $\pi(\cdot, p)$ is a continuous non-increasing function in $[0, \infty)$ whose image is $(1, \bar{m} - p + 1]$. Therefore, a necessary condition for 0- δ viability is that $2\gamma\tau \leq 1$. If this is the case, we need to select ξ large enough to guarantee that $2\gamma\tau\pi(\xi, p) \leq 1$. Thus, in qualitative terms, the predictor must lead to a system forgetting past initial conditions sufficiently fast.

C. Composition of viable kernels

It is well known that the sum of symmetric positive semidefinite kernels is a positive semidefinite kernel [57,

Prop 13.1]. Proposition 12 below (proved in Appendix K) gives sufficient conditions under which the linear combination of viable kernels is again viable.

Proposition 12 (Sum of kernels). *Let $(w_1)_{i=1}^q$ be $q \in \mathbb{N}$ kernel structures and k be the kernel structure such that $\Phi_k = (0, \infty)^q \times \Phi_{w_1} \times \cdots \times \Phi_{w_q}$ and*

$$\forall a, b \in \mathbb{R}^{m+1}, \quad k_\eta(a, b) := \sum_{i=1}^q \tau_i w_{i, \eta_i}(a, b),$$

for all $\eta = (\tau_{1:q}, \eta_{1:q}) \in \Phi_k$. Then, for each $\rho \in [0, \infty]$, the following inclusions hold

$$\Theta_k^\rho \supseteq \left\{ ((\tau_i)_{i=1}^q, (\eta_i)_{i=1}^q) \in (0, \infty)^q \times \Theta_w^\rho : \sum_{i=1}^q \tau_i \leq 1 \right\},$$

$$\Delta_k^\rho \supseteq \left\{ ((\tau_i)_{i=1}^q, (\eta_i)_{i=1}^q) \in (0, \infty)^q \times \Delta_w^\rho : \sum_{i=1}^q \tau_i \leq 1 \right\},$$

where $\Theta_w^\rho = \Theta_{w_1}^\rho \times \cdots \times \Theta_{w_q}^\rho$ and $\Delta_w^\rho = \Delta_{w_1}^\rho \times \cdots \times \Delta_{w_q}^\rho$.

As implied by Corollary 2, stationary kernels are not suitable for learning ISS systems. To address this problem, Proposition 13 below (proved in Appendix L) provides a way to modify a stationary kernel in order to obtain a non-trivial 0-viable kernel structure, which is therefore compatible with the conditions for ISS.

Proposition 13. *Let ℓ and w be kernel structures such that w is stationary. Let k be a kernel structure such that $\Phi_k = \Phi_\ell \times \Phi_w$ and $k_\eta(a, b) = \ell_{\eta_1}(a, b) w_{\eta_2}(a, b)$ for all $a, b \in \mathbb{R}^{m+1}$ and all $\eta = (\eta_1, \eta_2) \in \Phi_k$. Then, for each $\rho \in [0, \infty]$, $\Theta_k^\rho \supseteq \{(\eta_1, \eta_2) \in \Theta_\ell^\rho \times \Phi_w : \bar{w}_{\eta_2}(0_{m+1}) \leq 1\}$.*

An example of application of Proposition 13 can be easily constructed by taking ℓ as a degenerate kernel, and w as one of the stationary kernels analyzed in Section VI-B. In particular, let $p \in \mathbb{N}$ and $\Gamma : \mathbb{R}^{m+1} \rightarrow \mathbb{R}^p$ be such that $|\Gamma(a)|^2 \leq |a|^2$, for all $a \in \mathbb{R}^{m+1}$. Then, the kernel

$$k_\eta(a, b) := \Gamma(a)^\top \Gamma(b) (\tau \exp(-\gamma|a - b|^2) + \sigma) \quad (11)$$

is 0-viable with $\Theta_k^\rho \supseteq \{(\tau, \gamma, \sigma) \in \Phi_k : \tau + \sigma \leq 1\}$, where we have used the properties of the Gaussian kernel analyzed in Proposition 9.

VII. NUMERICAL EXPERIMENTS

This section provides some numerical experiments validating the theoretical findings. Section VII-A presents a Monte Carlo analysis on two models that satisfy the assumptions of Proposition 1 and 3, respectively. This experiment allows us to validate the proposed method in the case where the dataset is generated by a system having the same structure as the identified one. Section VII-B, instead, reports a more realistic application concerning the identification of the potassium ion gate's dynamics in the *Hodgkin-Huxley's* neuron model [86].

A. Learning stable systems

We consider two discrete-time models A and B with input u and output y that obey the two difference equations

$$\text{A : } y_t = 0.2|p_t|\sqrt{\sin(|p_t|) + 1}, \quad \text{B : } y_t = 0.2\sin(|p_t|)^2$$

where $p_t := (y_{t-2}, y_{t-1}, u_{t-2}, u_{t-1})$. Using Propositions 1 and 3, we can show that System A is ISS and B is δ ISS. For the identification of model $i \in \{A, B\}$ we employ a dataset $\mathcal{D}_i$ containing 200 samples of the input and output measurements. In particular, for every $t \in \{1, \dots, 200\}$, the pairs $(\bar{u}_t, \bar{y}_t) \in \mathcal{D}_i$ are obtained by sampling the random variables

$$\bar{u}_t = u_{t+2}, \quad \bar{y}_t = y_{t+2} + w_{t+2}^i,$$

where $w_t^A \sim \mathcal{N}_0(0.05)$, $w_t^B \sim \mathcal{N}_0(0.02)$ ($\mathcal{N}_0(S)$ denotes the 0-mean Normal distribution with covariance matrix S); the signal-to-noise ratio is close to 10 in both cases. Furthermore, $y_0 \sim \mathcal{N}_0(1)$, $y_1 \sim \mathcal{N}_0(1)$, and the input satisfies $u_t \sim \mathcal{N}_0(1)$ for every $t \in \mathbb{N}$. All the introduced random variables are mutually independent. For every model $i \in \{A, B\}$, the performances are evaluated on a validation dataset $\mathcal{D}_i^v$ sampled from the same distribution as the dataset $\mathcal{D}_i$.

To better validate the model, we consider both the performance in 1-step-ahead prediction and in simulation. In particular, the 1-step-ahead prediction sequence $\hat{y}^{\text{pre}}$ and the simulated output sequence $\hat{y}$ are as defined in Section II using the inputs of the validation dataset $\mathcal{D}_i^v$. Both sequences are equal to the output of the validation datasets for the first $\bar{m}$ samples to guarantee a sensible initial condition for the simulation. Prediction and simulation are evaluated in terms of their adherence to the output sequence $\bar{y}_t$. In particular, we define the performance indexes $q_{\text{pre}} := \frac{1}{n-\bar{m}} \sum_{t=\bar{m}+1}^n |\bar{y}_t - \hat{y}_t^{\text{pre}}|$ and $q_{\text{sim}} := \frac{1}{n-\bar{m}} \sum_{t=\bar{m}+1}^n |\bar{y}_t - \hat{y}_t|$. The analysis is carried out using Monte Carlo simulations with 501 runs. For the learning procedure, we select $m_y = m_u = 2$, $\iota = 10^{-10}$, $\chi = 0.99$, and the hyperparameters are computed using the EB method, suitably modified to account for the stability constraint for the proposed method. After selecting the hyperparameters, we employ the procedure presented in [87] to reduce the computational complexity of the estimated model by enforcing sparsity of the solution.

Since, in the context of this validation, A is known to be ISS, the developed theory suggests using kernel (11) with $\Gamma(a) = a$ for all $a \in \mathbb{R}^{m+1}$. We thus consider the two predictors:

Aa. f obtained by solving (1)-(3),

Ab. f obtained by solving (4)-(5) with $\Omega_k = \Theta_k^0$.

Therefore, **Ab** guarantees that the estimated model is ISS while **Aa** does not. As Figure 1 shows, imposing the right stability notion, ISS, yields similar performance to the standard predictor in both prediction and simulation. However, the models estimated with **Ab** are ISS as proven in Theorem 1, while **Aa** gives no guarantees of ISS. Similar conclusions can also be obtained for model B. Since B is δ ISS, the learning procedure is carried out using the Gaussian Kernel analyzed in Proposition 9, and we consider two predictors:

Ba. f obtained by solving (1)-(3),

Bb. f obtained by solving (4)-(5) with $\Omega_k = \Delta_k^0$.

In particular, **Bb** guarantees that the estimated model is δ ISS while **Ba** does not. As shown in Figure 2, we can draw the same observations as in the previously analyzed case.

These two examples show that, when the underlying system can be modeled by $\Sigma(f)$ for some predictor f that satisfies the assumptions of Propositions 1 and 3, the proposed method

is able to guarantee that the estimated models have a desired stability property without significantly deteriorating the prediction and simulation capabilities.

B. Potassium channel of an excitable cell

We consider the K^+ channel dynamics of the Hodgkin-Huxley's neuron model, which is given by the following differential equation [86, Eq (7,26)]

$$H : \begin{cases} \dot{\kappa} = \frac{(V + 10)(1 - \kappa)}{100 \left(e^{\frac{V}{10} + 1} - 1 \right)} - \frac{e^{\frac{V}{80}}}{8} \kappa \\ I = 36(V - 12)\kappa^4 \end{cases}$$

where V is the input (the neuron's membrane potential) and I is the output (the K^+ current flowing across the membrane). The objective is to learn a discrete-time predictor of the output I . For the learning procedure, we consider a dataset $\mathcal{D}_H$ generated by sampling (with sampling time 0.1 s), for each $t \in \{1, \dots, n\}$, the random signals

$$\bar{u}_t = V(49.9 + 0.1t), \quad \bar{y}_t = r(49.9 + 0.1t, u, \kappa_0),$$

where $n = 201$, $\kappa_0 \sim \mathcal{N}_0(1)$, $r(h, u, \kappa_0)$ is the forced output of system H at time h given the initial condition κ_0 , and $V(t)$ is chosen as the random variable

$$V(t) = \sum_{i=1}^{50} A_i \sin(2\pi\nu_i t + \varphi_i),$$

where, for every $i \in \{1, \dots, 50\}$, $A_i \sim \mathcal{U}(0.1, 0.5)$, $\nu_i \sim \mathcal{U}(0, 1)$, $\varphi_i \sim \mathcal{U}(0, 2\pi)$ ($\mathcal{U}(a, b)$ denotes the uniform distribution on $[a, b] \subseteq \mathbb{R}$). Hence, $\bar{u}$ and $\bar{y}$ are measurements of V and I , respectively. The random variables $\{\kappa_0\} \cup \{A_i, \nu_i, \varphi_i\}_{i=1}^{50}$ are all mutually independent. For this experiment, we select $m_u = m_y = 2$, $\iota = 10^{-10}$, $\chi = 0.99$, and the hyperparameters are computed using the EB method modified to account for the stability constraint for the proposed method. After selecting the hyperparameters, we employ the procedure presented in [87] to reduce the computational complexity of the estimated model by enforcing sparsity of the solution. The identification algorithm is carried out using the Gaussian kernel (described in Proposition 9), and we consider three predictors:

Ha. f obtained by solving (1)-(3),

Hb. f obtained by solving (4)-(5) with $\Omega_k = \Delta_k^\infty$,

Hc. f obtained by solving (4)-(5) with $\Omega_k = \Delta_k^0$.

The obtained predictors are evaluated using a validation dataset sampled from the same distribution as the dataset $\mathcal{D}_H$ with $n = 5001$. The analysis is carried out using Monte Carlo simulations with 501 runs. The performance in prediction and simulation of the estimated models are reported in Figures 3 and 4, respectively. We notice that **Ha** and **Hb** have similar prediction performance. As explained in Remark 2, since the Gaussian kernel is bounded, the two methods generate BIBS stable models. However, **Hb** also guarantees that the system is δ BIBS and that it has the wAG and $w\delta$ AG properties without performance deterioration. Instead, **Hc** causes a small decrease of predictive performance. However, the estimated models are guaranteed to be δ ISS. Thus, as explained in Section II, the simulation error remains bounded by a class $\mathcal{K}$ function of the

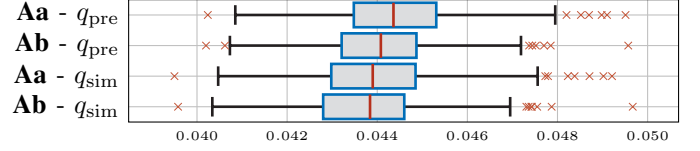


Fig. 1. Box plots of the performance indexes for system **A**.

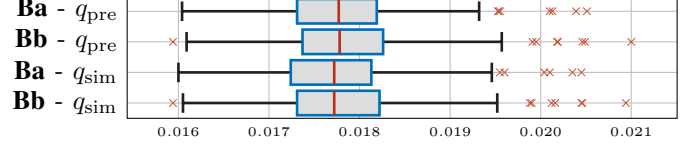


Fig. 2. Box plots of the performance indexes for system **B**.

infinity norm of the past prediction error (see Definition 6). The beneficial effect of this property can be seen in the simulation performance shown in Figure 4 where we can notice that the simulation error of **Hc** remains of the same order of magnitude of the prediction error. Instead, **Ha** presents a significantly larger simulation error that renders the models unsuitable for simulation purposes. Similar conclusions can also be drawn from Figure 5 where it is possible to see that the prediction error is small for both models, but the simulation error of **Ha** is significantly larger in magnitude.

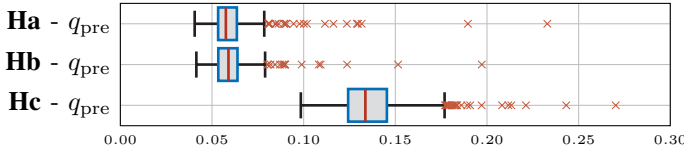
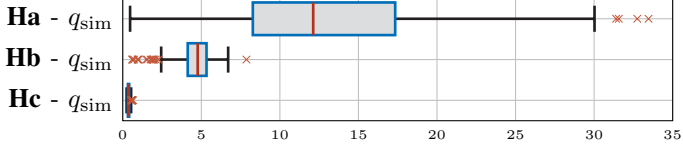
Finally, case **Hb** shows that only guaranteeing the weaker stability property of δ BIBS with $w\delta$ AG allows for better simulation performance without significantly impacting the prediction capability of the estimated model. Thus, it provides a middle ground between the lack of guarantees of **Ha** and the strong ones of **Hc**.

VIII. CONCLUDING REMARKS

Identifying nonlinear models possessing certain desired stability properties is of crucial importance in applications. Yet, existing results are limited in scope, and a systematic procedure to learn stable nonlinear systems is still a challenging problem. In this article, we make a step towards the solution of this problem showing how stability properties can be effectively included in a kernel-based learning procedure. In particular, we embedded these properties in the hyperparameters' selection algorithm, and stability is guaranteed by upper-bounding the RKHS norm of the predictor.

The proposed approach has the merit of offering a systematic non-parametric procedure to guarantee stability of the learned models. It is also a flexible solution allowing different kernel functions to be used. Moreover, the stability guarantees that can be enforced refer to both the input to the system and to the prediction error. In turn, stability with respect to the prediction error is shown to have considerable benefits when the model is used for simulation.

As a drawback, however, the stability conditions enforced (formally given in Propositions 1 and 3) may be quite conservative in general, and their restrictiveness increases with the model dimension. In practice, this issue can have a detrimental impact on the data-fitting performances in certain cases. Future research efforts will focus on (i) quantifying this detriment;

Fig. 3. Box plots of the performance indexes in prediction of system **H**.Fig. 4. Box plots of the performance indexes in simulation of system **H**.

(ii) addressing this problem, which may require continuity conditions weaker than Lipschitz-continuity; (iii) providing formal guarantees on the out-of-sample prediction performances of the estimated predictor.

APPENDIX

A. Proof of Lemma 1

For every $c \in \mathbb{R}^{n-\bar{m}}$, define $\phi_c := \sum_{i=\bar{m}+1}^n c_i k_{\eta}^{\bar{z}_i} \in \mathcal{H}(k_{\eta})$ as in (2). Then, by invoking [67, Th 2] with

$$\Omega(\omega) := \begin{cases} \alpha |\omega|_{\mathcal{H}(k_{\eta})}^2 & \text{If } m_y |\omega|_{\mathcal{H}(k_{\eta})}^2 \leq \chi, \\ \infty & \text{Otherwise,} \end{cases}$$

we can notice that there exists $c \in \mathbb{R}^{n-\bar{m}}$ such that ϕ_c is the solution of (4). Hence, Problem (4) boils down to

$$\min_{c \in \mathbb{R}^{n-\bar{m}}} B(c) \quad \text{s.t.} \quad C(c) \leq 0, \quad (12)$$

where

$$B(c) = \sum_{i=\bar{m}+1}^n (\phi_c(\bar{z}_i) - \bar{y}_i)^2 + \alpha |\phi_c|_{k_{\eta}}^2$$

$$C(c) = m_y |\phi_c|_{k_{\eta}}^2 - \chi.$$

By definition of K_{η} , and since $k_{\eta}(a, b) = \langle k_{\eta}^a, k_{\eta}^b \rangle_{k_{\eta}}$ [79, Lem. 4.19], for every $c \in \mathbb{R}^{n-\bar{m}}$ we have

$$|\phi_c|_{k_{\eta}}^2 = \sum_{i=\bar{m}+1}^n \sum_{j=\bar{m}+1}^n c_i c_j \langle k_{\eta}^{\bar{z}_i}, k_{\eta}^{\bar{z}_j} \rangle_{k_{\eta}} = c^{\top} K_{\eta} c.$$

Then, we obtain $B(c) = c^{\top} (K_{\eta} + \alpha I_{n-\bar{m}}) K_{\eta} c - 2\bar{y}^{\top} K_{\eta} c + \bar{y}^{\top} \bar{y}$ and $C(c) = m_y c^{\top} K_{\eta} c - \chi$. Additionally, $\nabla B(c) = 2c^{\top} (K_{\eta} + \alpha I_{n-\bar{m}}) K_{\eta} - 2\bar{y}^{\top} K_{\eta}$ and $\nabla C(c) = 2m_y c^{\top} K_{\eta}$. As a consequence, B and C are smooth, convex functions and, hence, the *KKT conditions* are sufficient for optimality [88, Sec 5.5.3]. Namely, if $(c, \lambda) \in \mathbb{R}^{n-\bar{m}} \times [0, \infty)$ satisfies

$$\nabla B(c) + \lambda \nabla C(c) = 0_{n-\bar{m}}^{\top}, \quad \lambda C(c) = 0, \quad C(c) \leq 0, \quad (13)$$

then c is an optimal solution of (12). The first of (13) is satisfied by $c = (K_{\eta} + (\alpha + \lambda m_y) I_{n-\bar{m}})^{-1} \bar{y}$. Thus, the solutions of (1) and (4) are equivalent if $\beta = \alpha + \lambda m_y$ with λ satisfying (13). Hence, if there exists a $\beta \in [0, \infty)$ such that

$$C2 : \frac{\beta - \alpha}{m_y} \gamma(\beta) = 0, \quad C3 : \gamma(\beta) \leq 0,$$

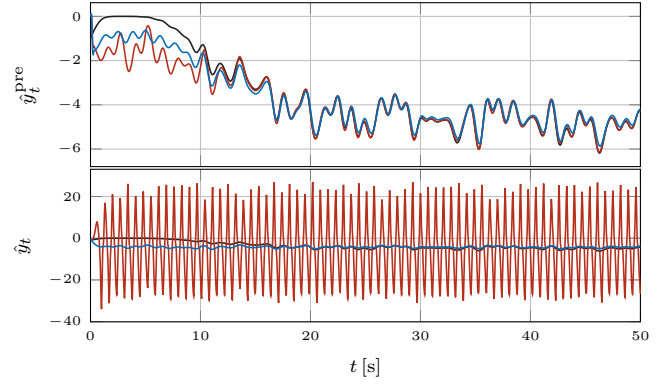


Fig. 5. The top (bottom) plot shows the comparison between the prediction (simulation) sequence of **Ha** (solid red line), the prediction (simulation) sequence of **Hc** (solid blue line), and the output of System **H** (solid black line). All the sequences are computed by taking the estimated model from the Monte Carlo run with median performance index using a newly generated validation dataset.

where γ is defined in (6), then (13) are satisfied by $c = (K_{\eta} + \beta I_{n-\bar{m}})^{-1} \bar{y}$ and $\lambda = \frac{\beta - \alpha}{m_y}$.

We now find β satisfying C2 and C3, thereby completing the proof. First, notice that γ is non-increasing and $\lim_{a \rightarrow \infty} \gamma(a) = -\chi$. Hence, if γ is not always negative, there exists a $\bar{\alpha} \in [0, \infty)$ such that $\gamma(\bar{\alpha}) = 0$. Otherwise, we let $\bar{\alpha} = 0$. If $\alpha \leq \bar{\alpha}$, then C2, C3, and $\beta \geq \alpha$ are all satisfied by the choice $\beta = \bar{\alpha}$ since $\gamma(\bar{\alpha}) \leq 0$ by definition of $\bar{\alpha}$. Otherwise, if $\alpha > \bar{\alpha}$, we notice that $\beta = \alpha$ trivially fulfills C2. Moreover, C3 is also satisfied because γ is a non-increasing function and $\gamma(\bar{\alpha}) \leq 0$. Hence, we conclude the proof by noticing that, in all cases, $\beta = \max(\bar{\alpha}, \alpha)$.

B. Proof of Proposition 2

Pick $f \in \mathcal{H}(k_{\eta})$. Using the Cauchy-Schwarz inequality, for every $a \in \mathbb{R}^{m+1}$, we obtain

$$|f(a)|^2 = |\langle f, k_{\eta}^a \rangle_{k_{\eta}}|^2 = \langle f, k_{\eta}^a \rangle_{k_{\eta}}^2 \leq |f|_{k_{\eta}}^2 \langle k_{\eta}^a, k_{\eta}^a \rangle_{k_{\eta}}.$$

Since $\langle k_{\eta}^a, k_{\eta}^a \rangle_{k_{\eta}} = k_{\eta}(a, a)$, then $|f(a)|^2 \leq |f|_{k_{\eta}}^2 k_{\eta}(a, a)$. Therefore, Conditions (8) lead to the implications

$$|a|^2 \geq \nu \Rightarrow |f(a)|^2 \leq |f|_{k_{\eta}}^2 |a|^2,$$

$$|a|^2 < \nu \Rightarrow |f(a)| \leq |f|_{k_{\eta}} \sqrt{k_{\eta}(a, a)} \leq |f|_{k_{\eta}} \sqrt{s},$$

which coincide with (7) for $a = \text{col}(x, v)$ ($x \in \mathbb{R}^m$ and $v \in \mathbb{R}$), $\mu = m_y |f|_{k_{\eta}}^2$, $\rho = \nu$, $d = |f|_{k_{\eta}} \sqrt{s}$ and $\omega(b) = |f|_{k_{\eta}}^2 b^2$.

C. Proof of Proposition 3

We divide the proof in several subsections for clarity of exposition.

1) δ ISS Lyapunov Candidate: Let $P := \text{diag}(1, \dots, m_y, 1, \dots, m_u)$ and define the function $V : \mathbb{R}^m \rightarrow [0, \infty]$ as $V(x) := x^{\top} P x$. Let A , G and H be the matrices defined in Section II. Then, P satisfies

$$A^{\top} P A - P = -I_m, \quad G^{\top} P G = m_y, \quad H^{\top} P H = m_u,$$

$$G^{\top} P H = 0, \quad A^{\top} P G = 0_m, \quad A^{\top} P H = 0_m, \quad (14)$$

and

$$\forall x \in \mathbb{R}^m, \quad |x|^2 \leq V(x) \leq \bar{m}|x|^2. \quad (15)$$

Fix $x^a, x^b \in \mathbb{R}^m$, $v^a, v^b, \varepsilon^a, \varepsilon^b \in \mathbb{R}$, and let

$$\begin{aligned} \tilde{x} &:= x^a - x^b, & \tilde{v} &:= v^a - v^b, \\ \tilde{\varepsilon} &:= \varepsilon^a - \varepsilon^b, & \tilde{f} &:= f(x^a, v^a) - f(x^b, v^b). \end{aligned} \quad (16)$$

By using the Young's inequality $2\tilde{f}\tilde{\varepsilon} \leq (1-\mu)\tilde{f}^2 + (1-\mu)^{-1}\tilde{\varepsilon}^2$ [89] and (14) we obtain

$$\begin{aligned} V(A\tilde{x} + G\tilde{f} + G\tilde{\varepsilon} + H\tilde{v}) - V(\tilde{x}) \\ = -|\tilde{x}|^2 + m_y\tilde{f}^2 + 2m_y\tilde{f}\tilde{\varepsilon} + m_y\tilde{\varepsilon}^2 + m_u\tilde{v}^2 \\ \leq -|\tilde{x}|^2 + m_y(2-\mu)\tilde{f}^2 + m_y\frac{2-\mu}{1-\mu}\tilde{\varepsilon}^2 + m_u\tilde{v}^2. \end{aligned} \quad (17)$$

Using (9a) and (15), we thus obtain

$$\begin{aligned} |\tilde{x}|^2 + |\tilde{v}|^2 \geq \rho \Rightarrow V(A\tilde{x} + G\tilde{f} + G\tilde{\varepsilon} + H\tilde{v}) \\ \leq \lambda V(\tilde{x}) + \varrho_u(|\tilde{v}|) + \varrho_e(|\tilde{\varepsilon}|), \end{aligned} \quad (18)$$

with $\lambda := 1 - \frac{(1-\mu)^2}{\bar{m}_2} \in [0, 1)$, $\varrho_u(s) := m_y(2-\mu)\omega(s) + m_us^2$, and $\varrho_e(s) := m_y\frac{2-\mu}{1-\mu}s^2$. Notice that ϱ_u and ϱ_e are of class- $\mathcal{K}$.

2) Proof of A: Fix $q \in [0, \infty)$ and pick arbitrarily two solutions $(x^a, v^a, \varepsilon^a)$ and $(x^b, v^b, \varepsilon^b)$ of $\Sigma(f)$ such that $|x_0^a - x_0^b| \leq q$, $|v^a - v^b|_\infty \leq q$, and $|\varepsilon^a - \varepsilon^b|_\infty \leq q$. For simplicity, we use the error variables (16). Moreover, define

$$\begin{aligned} c_1 &:= \sqrt{\frac{2(\varrho_u(q) + \varrho_e(q))}{(1-\lambda)\bar{m}}}, & c_2 &:= \max\{q, c_1, \sqrt{\rho}\}, \\ c_3 &:= \bar{m}c_2^2, & c_4 &:= \sup_{|\tilde{x}| \leq c_2, |\tilde{v}| \leq q, |\tilde{\varepsilon}| \leq q} |\tilde{f}^2 + 2\tilde{f}\tilde{\varepsilon}| + 2q^2, \\ c_5 &:= c_3 + 2\bar{m}c_4. \end{aligned}$$

As $\lambda \in [0, 1)$, then $c_1 > 0$. Moreover, c_4 exists since, in view of (9), f is bounded whenever so are $\tilde{x}$ and $\tilde{v}$. Also, in view of the first equality in (17), we have

$$\forall t \in \mathbb{N}, \quad V(\tilde{x}_t) \leq c_3 \Rightarrow V(\tilde{x}_{t+1}) \leq c_3 + \bar{m}c_4 \leq c_5. \quad (19)$$

The proof is carried out by contradiction. Suppose that there exists no $w \in [0, \infty)$ such that $|\tilde{x}|_\infty \leq w$. Then, in view of (15), there exists $\bar{t}$ such that $V(\tilde{x}_{\bar{t}}) \geq |\tilde{x}_{\bar{t}}|^2 > c_5$ (otherwise $|\tilde{x}|_\infty \leq w$ with $w = \sqrt{c_5}$). In view of (15), for every $t \in \mathbb{N}$, $|\tilde{x}_t| \leq c_2$ implies $V(\tilde{x}_t) \leq \bar{m}|\tilde{x}_t|^2 \leq \bar{m}c_2^2 = c_3$ and thus, in view of (19), $V(\tilde{x}_{t+1}) \leq c_5$. Therefore, since $|\tilde{x}_0| \leq q \leq c_2$, we have $V(\tilde{x}_0) \leq c_3$ and $V(\tilde{x}_1) \leq c_5$.

Thus, since $V(\tilde{x}_{\bar{t}}) > c_5$, there exist $t_0, t_1 \in \mathbb{N}$ satisfying $1 \leq t_0 < t_1 \leq \bar{t}$ and such that $V(\tilde{x}_t) \in [c_3, c_5]$ for every $t \in \{t_0, \dots, t_1-1\}$ and $V(\tilde{x}_{t_1}) > c_5$. In view of (19), $V(\tilde{x}_{t_1-1}) \geq c_3$. Hence,

$$|\tilde{x}_{t_1-1}|^2 + |\tilde{v}_{t_1-1}|^2 \geq |\tilde{x}_{t_1-1}|^2 \geq \frac{V(\tilde{x}_{t_1-1})}{\bar{m}} \geq \frac{c_3}{\bar{m}} = c_2^2 \geq \rho.$$

Therefore, from (18), we obtain $V(\tilde{x}_{t_1}) \leq \lambda V(\tilde{x}_{t_1-1}) + \varrho_u(q) + \varrho_e(q)$. Since $c_3 \leq V(\tilde{x}_{t_1-1}) \leq c_5$ and $c_3 = \bar{m}c_2^2 \geq \bar{m}c_1^2$, we can write

$$\begin{aligned} V(\tilde{x}_{t_1}) &\leq V(\tilde{x}_{t_1-1}) - (1-\lambda)V(\tilde{x}_{t_1-1}) + \varrho_u(q) + \varrho_e(q) \\ &\leq c_5 - (1-\lambda)c_3 + \varrho_u(q) + \varrho_e(q) \\ &\leq c_5 - \bar{m}(1-\lambda)c_1^2 + \bar{m}\frac{1-\lambda}{2}c_1^2 \leq c_5 - \frac{\bar{m}(1-\lambda)}{2}c_1^2. \end{aligned}$$

Since $c_1 \geq 0$, we obtain $V(\tilde{x}_{t_1}) \leq c_5$, a contradiction.

3) Preliminaries to the proof of B: Before proving **B**, we state and prove two technical lemmas dealing with uniformly continuous functions.

Lemma 2. Let $p \in \mathbb{N}$, $r \in [0, \infty)$, and $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}$ be a uniformly continuous function. Then $(x, y) \mapsto \varphi(x) - \varphi(y)$ is bounded on $\{(x, y) \in \mathbb{R}^p \times \mathbb{R}^p : |x - y| \leq r\}$.

Proof. Since φ is uniformly continuous, there exists $\delta > 0$ such that, for every $a, b \in \mathbb{R}^p$, $|a - b| \leq \delta \Rightarrow |\varphi(a) - \varphi(b)| \leq 1$. Let $N \in \mathbb{N}_{>0}$ be such that $N \geq \frac{r}{\delta}$. Pick $x, y \in \mathbb{R}^p$ satisfying $|x - y| \leq r$. For each $k \in \{0, \dots, N\}$, let $z_k := \frac{N-k}{N}x + \frac{k}{N}y$. Then, $|z_{k+1} - z_k| = \frac{1}{N}|x - y| \leq \frac{r}{N} \leq \delta$. Since $x = z_0$ and $y = z_N$, we then obtain $|\varphi(x) - \varphi(y)| = |\varphi(z_0) - \varphi(z_N)| \leq \sum_{k=0}^{N-1} |\varphi(z_k) - \varphi(z_{k+1})| \leq N$. $\square$

Lemma 3. Let $p \in \mathbb{N}$, $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}$ be a uniformly continuous function, and $g : \mathbb{R} \rightarrow [0, \infty)$ be a continuous function such that $g(0) = 0$. Then, the function¹ $k : [0, \infty) \rightarrow [0, \infty)$ defined as $k(s) := \sup_{|x-y| \leq s} g(\varphi(x) - \varphi(y))$ is continuous, non-decreasing, and $k(0) = 0$.

Proof. The only non-trivial claim to prove is continuity of k . In the following, for $h \in [0, \infty)$, we let $\Omega_h := \{(x, y) \in \mathbb{R}^p \times \mathbb{R}^p : |x - y| \leq h\}$. Pick $s \in [0, \infty)$ and $\iota \in (0, \infty)$, and let $M_{s+1} := \sup_{(x,y) \in \Omega_{s+1}} |\varphi(x) - \varphi(y)| \geq 0$, which exists in view of Lemma 2. Since g is continuous, it is uniformly continuous on $[-M_{s+1}, M_{s+1}]$. Hence, there exists $\delta^g(\iota) \in (0, \infty)$ such that, for every $a, b \in [-M_{s+1}, M_{s+1}]$, $|a - b| \leq \delta^g(\iota) \Rightarrow |g(a) - g(b)| \leq \frac{\iota}{2}$. Since φ is uniformly continuous, there exists $\delta^\varphi(\iota) \in (0, \infty)$ such that, for every $(x, y), (x', y') \in \Omega_{s+1}$, $|x - x'| + |y - y'| \leq \delta^\varphi(\iota) \Rightarrow |(\varphi(x) - \varphi(y)) - (\varphi(x') - \varphi(y'))| \leq |\varphi(x) - \varphi(x')| + |\varphi(y) - \varphi(y')| \leq \delta^g(\iota)$, which in turn implies $|g(\varphi(x) - \varphi(y)) - g(\varphi(x') - \varphi(y'))| \leq \frac{\iota}{2}$.

With $\delta(\iota) := \min\{1, \delta^\varphi(\iota)\}$, let $(\bar{x}, \bar{y}) \in \Omega_{s+\delta(\iota)}$ be such that $k(s + \delta(\iota)) \leq g(\varphi(\bar{x}) - \varphi(\bar{y})) + \frac{\iota}{2}$ (which exists by definition of k), and define $\bar{x}' := \bar{x} - \frac{\delta(\iota)}{2(s+\delta(\iota))}(\bar{x} - \bar{y})$ and $\bar{y}' := \bar{y} + \frac{\delta(\iota)}{2(s+\delta(\iota))}(\bar{x} - \bar{y})$. Then, $|\bar{x}' - \bar{y}'| = \frac{s}{s+\delta(\iota)}|\bar{x} - \bar{y}| \leq s$. Hence, $(\bar{x}', \bar{y}') \in \Omega_s$. Moreover, $|\bar{x} - \bar{x}'| + |\bar{y} - \bar{y}'| = \frac{\delta(\iota)}{s+\delta(\iota)}|\bar{x} - \bar{y}| \leq \delta(\iota)$. As $\delta(\iota) \leq 1$ and $\delta(\iota) \leq \delta^\varphi(\iota)$, then

$$\begin{aligned} |\varphi(\bar{x}) - \varphi(\bar{y})| &\leq M_{s+1}, \quad |\varphi(\bar{x}') - \varphi(\bar{y}')| \leq M_{s+1}, \\ |g(\varphi(\bar{x}) - \varphi(\bar{y})) - g(\varphi(\bar{x}') - \varphi(\bar{y}'))| &\leq \frac{\iota}{2}. \end{aligned}$$

Therefore, since k is non-decreasing, and by definition of $(\bar{x}, \bar{y})$, for every $\eta \in (0, \delta(\iota))$, we obtain

$$\begin{aligned} |k(s + \eta) - k(s)| &= k(s + \eta) - k(s) \leq k(s + \delta(\iota)) - k(s) \\ &\leq k(s + \delta(\iota)) - g(\varphi(\bar{x}') - \varphi(\bar{y}')) \\ &\leq g(\varphi(\bar{x}) - \varphi(\bar{y})) + \frac{\iota}{2} - g(\varphi(\bar{x}') - \varphi(\bar{y}')) \\ &\leq |g(\varphi(\bar{x}) - \varphi(\bar{y})) - g(\varphi(\bar{x}') - \varphi(\bar{y}'))| + \frac{\iota}{2} \leq \iota. \end{aligned}$$

Right-continuity thus follows by the arbitrariness of s and ι . Left-continuity can be proven with a similar argument. $\square$

¹Notice that k is defined on $[0, \infty)$ in view of Lemma 2.

4) **Proof of B:** Assume that f is uniformly continuous. Pick arbitrarily two solutions $(x^a, \tilde{v}^a, \tilde{\varepsilon}^a)$ and $(x^b, \tilde{v}^b, \tilde{\varepsilon}^b)$ of $\Sigma(f)$. For simplicity, we use the error variables (16). Let $q' := \max\{|\tilde{v}|_\infty, |\tilde{\varepsilon}|_\infty\}$ and define the quantities

$$c'_1 := \sqrt{\frac{2(\varrho_u(q') + \varrho_e(q'))}{(1-\lambda)\bar{m}}}, \quad c'_2 := \max\{q', c'_1, \sqrt{\rho}\},$$

$$c'_3 := \bar{m}(c'_2)^2, \quad c'_4 := \sup_{|\tilde{x}| \leq c'_2, |\tilde{v}| \leq q', |\tilde{\varepsilon}| \leq q'} \left| \tilde{f}^2 + 2\tilde{f}\tilde{\varepsilon} \right| + 2(q')^2.$$

As in Section C.2, (9) guarantee that $c'_4 \geq 0$ exists. Moreover, proceeding as in Section C.2, we find that, for all $t \in \mathbb{N}$, $V(\tilde{x}_t) \geq c'_3$ implies $|\tilde{x}_t|^2 \geq (c'_2)^2 \geq \rho$ and $V(\tilde{x}_{t+1}) \leq V(\tilde{x}_t) - \frac{(1-\lambda)\bar{m}}{2}(c'_1)^2$. Furthermore, the first equality of (17) yields $V(\tilde{x}_t) \leq c'_3 \Rightarrow V(\tilde{x}_{t+1}) \leq c'_3 + \bar{m}c'_4$. Assuming without loss of generality $q' > 0$, these two inequalities and (15) suffice to claim the existence of a $\bar{t} \in \mathbb{N}$ such that

$$\forall t \geq \bar{t}, \quad |\tilde{x}_t| \leq \sqrt{c'_3 + \bar{m}c'_4} =: \beta(q', \rho), \quad (20)$$

where we highlighted the dependency of β from q' and ρ .

Next, we can write

$$\sup_{\substack{|\tilde{x}| \leq \sqrt{c'_3}, \\ |\tilde{v}| \leq q'}} \tilde{f}^2 \leq \sup_{|\tilde{x}, \tilde{v}| \leq 2 \max\{\sqrt{c'_3}, q'\}} \tilde{f}^2 =: \omega_1(\max\{\sqrt{c'_3}, q'\}).$$

Lemma 3 applied with $p = m + 1$, $\varphi = f$, and $g(s) = s^2$ implies that ω_1 is continuous, non-decreasing, and $\omega_1(0) = 0$. We also have $\omega_1(\max\{\sqrt{c'_3}, q'\}) \leq \omega_1(\sqrt{c'_3}) + \omega_1(q')$. As a consequence, we can bound c'_4 as

$$c'_4 \leq 2 \cdot \sup_{|\tilde{x}| \leq \sqrt{c'_3}, |\tilde{v}| \leq q'} \tilde{f}^2 + 3(q')^2 = \frac{\omega_2(c'_3) - c'_3 + \omega_3(q')}{\bar{m}}$$

in which $\omega_2(c'_3) := c'_3 + 2\bar{m}\omega_1(\sqrt{c'_3})$ and $\omega_3(q') := 2\bar{m}\omega_1(q') + 3\bar{m}(q')^2$ are both of class- $\mathcal{K}$ since they are the sum of a class- $\mathcal{K}$ function (the map $c'_3 \mapsto c'_3$ and $q' \mapsto 3\bar{m}(q')^2$, respectively) and the functions $2\bar{m}\omega_1(\sqrt{\cdot})$ and $2\bar{m}\omega_1$, with ω_1 that has the previously-claimed properties. As a consequence, we have $\beta(q', \rho)^2 \leq \omega_2(c'_3) + \omega_3(q')$. Since $c'_3 \leq 2\bar{m}((\max\{c'_1, q'\})^2 + \rho)$, and $c'_1 = c'_1(q')$, seen as a function of q' , is of class- $\mathcal{K}$, then by using the triangle inequality $\omega_2(s + z) \leq \omega_2(2s) + \omega_2(2z)$ [36, Eq (12)], we can finally write $\beta(q', \rho) \leq b(\rho) + \kappa(q')$, where $\kappa(q') := \sqrt{\omega_2(4\bar{m} \max\{c'_1(q'), q'\}^2) + \omega_3(q')}$ and $b(\rho) := \sqrt{\omega_2(4\bar{m}\rho)}$. Using the definition of q' , from (20) we get

$$\limsup_{t \rightarrow \infty} |\tilde{x}_t| \leq b(\rho) + \kappa(\max\{|\tilde{v}|_\infty, |\tilde{\varepsilon}|_\infty\}),$$

and the result follows by the same arguments of [35, Lem. II.1] and since b is of class- $\mathcal{K}$.

5) **Proof of C:** If $\rho = 0$, then δ ISS directly follows from (18), because the function $(x, y) \mapsto V(x - y)$ is a δ ISS Lyapunov function. See, [90, Thm. 8] and [76] for more details.

D. Proof of Proposition 4

Pick $f \in \mathcal{H}(k_\eta)$. For every $a, b \in \mathbb{R}^{m+1}$, $|f(a) - f(b)|^2 = |\langle f, k_\eta^a \rangle_{k_\eta} - \langle f, k_\eta^b \rangle_{k_\eta}|^2 = \langle f, k_\eta^a - k_\eta^b \rangle_{k_\eta}^2 \leq |f|_{k_\eta}^2 \langle k_\eta^a - k_\eta^b, k_\eta^a - k_\eta^b \rangle_{k_\eta}$, where we used the Cauchy-Schwarz inequality. Since, for every $a, b \in \mathbb{R}^{m+1}$, $k_\eta(a, b) =$

$\langle k_\eta^a, k_\eta^b \rangle_{k_\eta}$, we obtain $|f(a) - f(b)|^2 \leq |f|_{k_\eta}^2 h_\eta(a, b)$, in which, we recall, $h_\eta(a, b) := k_\eta(a, a) - 2k_\eta(a, b) + k_\eta(b, b)$. Then, in view of (10), the proof follows by the same arguments of that of Proposition 2 (see Appendix B).

E. Proof of Proposition 7

Since, for every $\eta = (\tau, \sigma) \in \Phi_k$, k_η is continuous, then, by Proposition 5, k is ρ -viable if Condition (8a) holds. Moreover, $k_\eta(a, a) = \tau|a|^2 + \sigma$. Hence, Condition (8a) is equivalent to the existence of $\nu \in [0, \rho] \cap \mathbb{R}$ such that, for every $z \in [\nu, \infty)$, $g(z) := (\tau - 1)z + \sigma \leq 0$. If $\tau \in (1, \infty)$, g is increasing and strictly positive in $(0, \infty)$. Therefore, the condition is not satisfied for any $\rho \in [0, \infty]$. Instead, if $\tau \in [0, 1]$, $g(z)$ is non-increasing. Therefore, its maximum in the interval $[\nu, \infty)$ is $g(\nu) = (\tau - 1)\nu + \sigma$. As $g(\nu) \geq g(\rho)$ for all $\nu \leq \rho$, then we conclude that k is ρ -viable if and only if $\tau \in [0, 1]$ and $\sigma \in [0, \rho(1 - \tau)]$.

As for what concerns ρ - δ viability, notice that, for every $(\tau, \sigma) \in \Phi_k$, $h_\eta(a, b) = \tau|a - b|^2$. Hence, (10b) trivially holds for every $(\tau, \sigma) \in \Phi_k$ with any $s > \tau\nu$, whereas (10a) holds if and only if $\tau \leq 1$.

F. Proof of Proposition 8

We recall that $k_\eta(a, b) = \Gamma_\eta(a)^\top \Gamma_\eta(b)$ for every $a, b \in \mathbb{R}^{m+1}$ and $\eta \in \Phi_k$, where Γ_η is the vector-valued function whose entries are all the possible η th degree ordered products of its argument [57, Prop 2.1]. By definition, k is ρ -viable only if (8a) holds, namely, only if there exists $\nu \in [0, \rho]$ such that, for all $|a| \geq \nu$, $|\Gamma_\eta(a)|^2 \leq |a|^2$, which is never satisfied because Γ_η is a polynomial function with degree $\eta > 1$. Therefore, $\Theta_k^\rho = \emptyset$. The equality $\Delta_k^\rho = \emptyset$ can be obtained by similar arguments.

G. Proof of Theorem 3

First, we focus on Θ_k^ρ . Since $k_\eta(a, a) = \bar{k}_\eta(0_{m+1})$, Conditions (8) read

$$\forall a \in \mathbb{R}^{m+1}, \quad |a|^2 \geq \nu \Rightarrow \bar{k}_\eta(0_{m+1}) \leq |a|^2,$$

$$\forall a \in \mathbb{R}^{m+1}, \quad |a|^2 < \nu \Rightarrow \bar{k}_\eta(0_{m+1}) \leq s.$$

The second condition is fulfilled with any $\nu \in [0, \infty)$ and with $s := \bar{k}_\eta(0_{m+1})$. Instead, the first condition holds if and only if $\bar{k}_\eta(0_{m+1}) \leq \nu$. Therefore, the claim of the theorem holds since $\nu \in [0, \rho] \cap \mathbb{R}$ such that $\bar{k}_\eta(0_{m+1}) \leq \nu$ exists if and only if $\rho \geq \bar{k}_\eta(0_{m+1})$.

Regarding Δ_k^ρ , let $g(z) := 2\bar{k}_\eta(0_{m+1}) - 2\bar{k}_\eta(z)$ for all $z \in \mathbb{R}^{m+1}$. Then, for all $a, b \in \mathbb{R}^{m+1}$, $h_\eta(a, b) = g(a - b)$. Therefore, Conditions (10) are equivalent to

$$\forall z \in \mathbb{R}^{m+1}, \quad |z|^2 \geq \nu \Rightarrow g(z) \leq |z|^2, \quad (21a)$$

$$\forall z \in \mathbb{R}^{m+1}, \quad |z|^2 < \nu \Rightarrow g(z) \leq s. \quad (21b)$$

Now, notice that, for every $z \in \mathbb{R}^{m+1}$,

$$|\bar{k}_\eta(z)| = |k_\eta(z, 0_{m+1})| = |\langle k_\eta^z, k_\eta^{0_{m+1}} \rangle_{k_\eta}|$$

$$\leq \sqrt{k_\eta(z, z)k_\eta(0_{m+1}, 0_{m+1})} = \bar{k}_\eta(0_{m+1}). \quad (22)$$

Therefore, $g(z) \in [0, 4\bar{k}_\eta(0_{m+1})]$ for all $z \in \mathbb{R}^{m+1}$. Hence, (21b) always holds with any $\nu \in [0, \infty)$ and with $s = 4\bar{k}_\eta(0_{m+1})$. Thus, since (21a) holds for some ν only if it holds for all $\nu' \geq \nu$, then $\Delta_k^\rho = \{\eta \in \Phi_k : \forall z \in \mathbb{R}^{m+1}, |z|^2 \geq \rho \Rightarrow 2\bar{k}_\eta(0_{m+1}) - 2\bar{k}_\eta(z) \leq |z|^2\}$. Finally, the fact that, for every $\rho \in [0, \infty]$, $\Delta_k^\rho \supseteq \{\eta \in \Phi_k : 4\bar{k}_\eta(0_{m+1}) \leq \rho\}$ directly follows from (21a) and $g(z) \leq 4\bar{k}_\eta(0_{m+1})$.

H. Proof of Proposition 9

Since k is a stationary kernel structure with $\bar{k}_\eta(0_{m+1}) = \tau + \sigma$ for every $\eta \in \Phi_k$, the claimed expression of Θ_k^ρ directly follows from Theorem 3. In view of Theorem 3, Δ_k^ρ is the set of $\eta \in \Phi_k$ such that, for every $z \in \mathbb{R}^{m+1}$, $|z|^2 \geq \rho$ implies $2\tau - 2\tau \exp(-\gamma|z|^2) \leq |z|^2$ or, equivalently,

$$\forall \zeta \in [\rho, \infty), g(\zeta) := 2\tau - 2\tau \exp(-\gamma\zeta) - \zeta \leq 0. \quad (23)$$

Notice that g is smooth, $g(0) = 0$, and its derivative is $g'(\zeta) = 2\tau\gamma \exp(-\gamma\zeta) - 1$. First, pick $\rho = 0$. We have $g'(\zeta) \leq 0$ for every $\zeta \geq 0$ if and only if $2\tau\gamma \leq 1$ and, in this case, $g(\zeta) \leq 0$ for all $\zeta \geq 0$. Otherwise, g is increasing in 0. Hence, $\Delta_k^0 = \{(\tau, \gamma, \sigma) \in \Phi_k : 2\tau\gamma \leq 1\}$. Next, pick $\rho \in (0, \infty]$. If $2\tau\gamma \leq 1$, then $(\tau, \gamma, \sigma) \in \Delta_k^0 \subseteq \Delta_k^\rho$. Instead, if $2\tau\gamma > 1$, g is increasing in the interval $[0, \gamma^{-1} \log(2\gamma\tau))$ and decreasing in $[\gamma^{-1} \log(2\gamma\tau), +\infty)$. Therefore, $g(\zeta) \leq 0$ for all $\zeta \geq \bar{\zeta}$, in which $\bar{\zeta}$ is the unique strictly positive solution of $g(\bar{\zeta}) = 0$, which is given by $v(\gamma, \tau)$. Then, (23) holds if and only if $\rho \geq v(\gamma, \tau)$, implying $\Delta_k^\rho = \{(\tau, \gamma, \sigma) \in \Phi_k : v(\gamma, \tau) \leq \rho\}$. For $\rho = \infty$, $\Delta_k^\infty = \Phi_k$ is a direct consequence of Theorem 3. Finally, since $W(z) \leq 0$, for all $z \leq 0$, we obtain $\Delta_k^\rho \supseteq \{(\tau, \gamma, \sigma) \in \Phi_k : 2\tau \leq \rho\} = [0, \frac{\rho}{2}] \times [0, \infty)^2$.

I. Proof of Proposition 10

Since k is a stationary kernel structure with $\bar{k}_\eta(0_{m+1}) = \tau + \sigma$ for every $\eta \in \Phi_k$, the claimed expression of Θ_k^ρ directly follows from Theorem 3. In view of Theorem 3, Δ_k^0 is the set of $\eta \in \Phi_k$ such that, for every $z \in \mathbb{R}^{m+1}$, $2\tau - 2\tau(1 + \gamma\sqrt{3}|z|) \exp(-\gamma\sqrt{3}|z|) \leq |z|^2$ or, equivalently, for every $\zeta \in [0, \infty)$, $g(\zeta) := 2\tau - 2\tau(1 + \gamma\sqrt{3}\zeta) \exp(-\gamma\sqrt{3}\zeta) - \zeta^2 \leq 0$. Notice that g is smooth, $g(0) = 0$, and its derivative is $g'(\zeta) = 2\zeta(3\gamma^2\tau \exp(-\gamma\sqrt{3}\zeta) - 1)$. We have $g'(\zeta) \leq 0$ for every $\zeta \geq 0$ if and only if $3\gamma^2\tau \leq 1$ and, in this case, $g(\zeta) \leq 0$ for all $\zeta \geq 0$. Instead, if $3\gamma^2\tau > 1$, g is increasing in 0. Hence, $\Delta_k^0 = \{(\tau, \gamma, \sigma) \in \Phi_k : 3\gamma^2\tau \leq 1\}$.

J. Proof of Proposition 11

Since k is a stationary kernel structure with $\bar{k}_\eta(0_{m+1}) = \tau \sum_{t=0}^{\bar{m}-p} e^{-\xi t}$ for every $\eta \in \Phi_k$, the claimed expression of Θ_k^ρ directly follows from Theorem 3. Regarding Δ_k^ρ , notice that we can write $\bar{k}_\eta(z) = 2\gamma\tau \sum_{t=0}^{\bar{m}-p} e^{-\xi t} w_{\tilde{\eta}}(z_t)$ where $w_{\tilde{\eta}}$ is the Gaussian Kernel (described in Proposition 9) with parameter $\tilde{\eta} = ((2\gamma)^{-1}, \gamma, 0) \in \Phi_w$. Since $2(2\gamma)^{-1}\gamma = 1$, then $\tilde{\eta} \in \Delta_w^0$. Therefore, according to Theorem 3, we have $2w_{\tilde{\eta}}(0_{m+1}) -$

$2w_{\tilde{\eta}}(z_t) \leq |z_t|^2 \leq |z|^2$. Then, we obtain

$$\begin{aligned} 2\bar{k}_\eta(0_{m+1}) - 2\bar{k}_\eta(z) &= 2\gamma\tau \sum_{t=0}^{\bar{m}-p} e^{-\xi t} (2w_{\tilde{\eta}}(0_{m+1}) - 2w_{\tilde{\eta}}(z_t)) \\ &\leq |z|^2 2\gamma\tau \sum_{t=0}^{\bar{m}-p} e^{-\xi t}, \end{aligned}$$

from which the claim follows by Theorem 3, since $\sum_{t=0}^{\bar{m}-p} e^{-\xi t} = \pi(\xi, p)$.

K. Proof of Proposition 12

First, let us focus on Θ_k^ρ . Conditions (8) read

$$\forall a \in \mathbb{R}^{m+1}, \quad |a|^2 \geq \nu \Rightarrow \sum_{i=1}^q \tau_i w_{i, \eta_i}(a, a) \leq |a|^2, \quad (24)$$

$$\forall a \in \mathbb{R}^{m+1}, \quad |a|^2 < \nu \Rightarrow \sum_{i=1}^q \tau_i w_{i, \eta_i}(a, a) \leq s. \quad (25)$$

Pick $\rho \in [0, \infty]$. For every $i \in \{1, \dots, q\}$, $\eta_i \in \Theta_{w_i}^\rho$ implies that there exist $\nu_i \in [0, \rho] \cap \mathbb{R}$ and $s_i \in [0, \infty)$ such that

$$\forall a \in \mathbb{R}^{m+1}, \quad |a|^2 \geq \nu_i \Rightarrow w_{i, \eta_i}(a, a) \leq |a|^2, \quad (26)$$

$$\forall a \in \mathbb{R}^{m+1}, \quad |a|^2 < \nu_i \Rightarrow w_{i, \eta_i}(a, a) \leq s_i. \quad (27)$$

Let $\nu := \max_{i=1, \dots, q} \nu_i$ and $s := \max_{i=1, \dots, q} \max\{\nu_i, s_i\}$. If $a \in \mathbb{R}^{m+1}$ satisfies $|a|^2 \geq \nu$, then $|a|^2 \geq \nu_i$ for all $i = 1, \dots, q$. Hence, by (26), if $\sum_{i=1}^q \tau_i \leq 1$, we obtain $\sum_{i=1}^q \tau_i w_{i, \eta_i}(a, a) \leq |a|^2$. Thus, (24) holds. Consider now the case in which $|a|^2 < \nu$. For each i such that $|a|^2 < \nu_i$, (27) yields $w_{i, \eta_i}(a, a) \leq s_i \leq s$; otherwise, (26) yields $w_{i, \eta_i}(a, a) \leq |a|^2 \leq \nu_i \leq \nu \leq s$. In both cases, $w_{i, \eta_i}(a, a) \leq s$. Then, (25) holds.

The claim regarding Δ_k^ρ follows by similar arguments.

L. Proof of Proposition 13

Let $\eta_1 \in \Theta_\ell^\rho$. Then, there exist $s_1 \in (0, \infty)$ and $\nu_1 \in [0, \rho]$ such that

$$\forall a \in \mathbb{R}^{m+1}, \quad |a|^2 \geq \nu_1 \Rightarrow \ell_{\eta_1}(a, a) \leq |a|^2,$$

$$\forall a \in \mathbb{R}^{m+1}, \quad |a|^2 < \nu_1 \Rightarrow \ell_{\eta_1}(a, a) \leq s_1.$$

Since $w_{\eta_2}(a, a) = \bar{w}_{\eta_2}(0_{m+1})$, Conditions (8) read as

$$\forall a \in \mathbb{R}^{m+1}, \quad |a|^2 \geq \nu \Rightarrow \ell_{\eta_1}(a, a) \bar{w}_{\eta_2}(0_{m+1}) \leq |a|^2,$$

$$\forall a \in \mathbb{R}^{m+1}, \quad |a|^2 < \nu \Rightarrow \ell_{\eta_1}(a, a) \bar{w}_{\eta_2}(0_{m+1}) \leq s.$$

Therefore, the second condition is satisfied with $s = s_1 \bar{w}_{\eta_2}(0_{m+1})$ and $\nu = \nu_1$. Instead, with $\nu = \nu_1$, the first condition is satisfied if $\bar{w}_{\eta_2}(0_{m+1}) \leq 1$.

REFERENCES

- [1] L. Ljung, "Perspectives on system identification," *Annu Rev Control*, vol. 34, no. 1, pp. 1–12, 2010.
- [2] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.-Y. Glorennec, H. Hjalmarrsson, and A. Juditsky, "Nonlinear black-box modeling in system identification: A unified overview," *Automatica*, vol. 31, no. 12, pp. 1691–1724, 1995.

- [3] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Wiley, 2015.
- [4] R. Pintelon and J. Schoukens, *System Identification*. Wiley, 2012.
- [5] P. van Overschee and B. L. de Moor, *Subspace Identification for Linear Systems*. Springer, 1996.
- [6] T. Söderström and P. Stoica, "On the stability of dynamic models obtained by least-squares identification," *IEEE Trans Autom Control*, vol. 26, no. 2, pp. 575–577, 1981.
- [7] U. Nallasivam, B. Srinivasan, V. Kuppuraj, M. N. Karim, and R. Rengaswamy, "Computationally efficient identification of global ARX parameters with guaranteed stability," *IEEE Trans Autom Control*, vol. 56, no. 6, pp. 1406–1411, 2011.
- [8] S. L. Lacy and D. S. Bernstein, "Subspace identification with guaranteed stability using constrained optimization," *IEEE Trans Autom Control*, vol. 48, no. 7, pp. 1259–1263, 2003.
- [9] J. Umenberger and I. R. Manchester, "Scalable identification of stable positive systems," in *IEEE CDC*, 2016, pp. 4630–4635.
- [10] J. Umenberger, J. Wågberg, I. R. Manchester, and T. B. Schön, "Maximum likelihood identification of stable linear dynamical systems," *Automatica*, vol. 96, pp. 280–292, 2018.
- [11] V. Cerone, D. Piga, and D. Regruto, "Enforcing stability constraints in set-membership identification of linear dynamic systems," *Automatica*, vol. 47, no. 11, pp. 2488–2494, 2011.
- [12] M. Lauricella and L. Fagiano, "Set membership identification of linear systems with guaranteed simulation accuracy," *IEEE Trans Autom Control*, vol. 65, no. 12, pp. 5189–5204, 2020.
- [13] J. Lataire and T. Chen, "Transfer function and transient estimation by Gaussian process regression in the frequency domain," *Automatica*, vol. 72, pp. 217–229, 2016.
- [14] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.
- [15] M. Scandella, M. Mazzoleni, S. Formentin, and F. Previdi, "Kernel-based identification of asymptotically stable continuous-time linear dynamical systems," *Int J Control*, vol. 95, no. 6, pp. 1668–1681, 2022.
- [16] M. A. H. Darwish, P. B. Cox, I. Proimadis, G. Pillonetto, and R. Tóth, "Prediction-error identification of LPV systems: A nonparametric Gaussian regression approach," *Automatica*, vol. 97, pp. 92–103, 2018.
- [17] V. Cerone, D. Piga, D. Regruto, and R. Tóth, "Input-output LPV model identification with guaranteed quadratic stability," in *IFAC SYSID*, vol. 45, no. 16, 2012, pp. 1767–1772.
- [18] S. J. Qin and T. J. McAvoy, "Nonlinear FIR modeling via a neural net PLS approach," *Comput Chem Eng*, vol. 20, no. 2, pp. 147–159, 1996.
- [19] G. Pillonetto, "System identification using kernel-based regularization: New insights on stability and consistency issues," *Automatica*, vol. 93, pp. 321–332, 2018.
- [20] V. De Iuliis, F. Smarra, C. Manes, and A. D'Innocenzo, "Stability analysis of switched ARX models and application to learning with guarantees," *Nonlinear Anal Hybrid Syst*, vol. 46, p. 101250, 2022.
- [21] J. de Jesus Rubio and W. Yu, "Stability analysis of nonlinear system identification via delayed neural networks," *IEEE Trans Circuits Syst II: Express Br*, vol. 54, no. 2, pp. 161–165, 2007.
- [22] F. Bonassi, M. Farina, and R. Scattolini, "On the stability properties of gated recurrent units neural networks," *Syst Control Lett*, vol. 157, p. 105049, 2021.
- [23] F. Bonassi and R. Scattolini, "Recurrent neural network-based internal model control design for stable nonlinear systems," *Eur J Control*, vol. 65, p. 100632, 2022.
- [24] S. De Carli, D. Previtalli, L. Pitturelli, M. Mazzoleni, A. Ferramosca, and F. Previdi, "Infinity-norm-based input-to-state-stable long short-term memory networks: a thermal systems perspective," in *ECC*, 2025, pp. 911–916.
- [25] P. Bevanda, M. Beier, S. Kerz, A. Lederer, S. Sosnowski, and S. Hirche, "Diffeomorphically learning stable Koopman operators," *IEEE Control Syst Lett*, vol. 6, pp. 3427–3432, 2022.
- [26] M. Khosravi, "Representer theorem for learning Koopman operators," *IEEE Trans Autom Control*, vol. 68, no. 5, pp. 2995–3010, 2023.
- [27] J. Umenberger and I. R. Manchester, "Specialized interior-point algorithm for stable nonlinear system identification," *IEEE Trans Autom Control*, vol. 64, no. 6, pp. 2442–2456, 2019.
- [28] M. F. Shakib, R. Tóth, A. Y. Pogromsky, A. Pavlov, and N. van de Wouw, "Kernel-based learning of stable nonlinear state-space models," in *IEEE CDC*, 2023, pp. 2897–2902.
- [29] J. Umlauf and S. Hirche, "Learning stochastically stable Gaussian process state-space models," *IFAC J Syst Control*, vol. 12, p. 100079, 2020.
- [30] W. Xiao, A. Lederer, and S. Hirche, "Learning stable nonparametric dynamical systems with Gaussian process regression," in *IFAC World Congr*, vol. 53, 2020, pp. 1194–1199.
- [31] T. Beckers and S. Hirche, "Stability of Gaussian process state space models," in *ECC*, 2016, pp. 2275–2281.
- [32] H. J. van Waarde and R. Sepulchre, "Training Lipschitz continuous operators using reproducing kernels," in *LADC*, ser. Proc Mach Learn Res, vol. 168, 2022, pp. 221–233.
- [33] —, "Kernel-based models for system analysis," *IEEE Trans Autom Control*, vol. 68, no. 9, pp. 5317–5332, 2023.
- [34] V. Andriano, A. Bacciotti, and G. Beccari, "Global stability and external stability of dynamical systems," *Nonlinear Anal Theory Methods Appl*, vol. 28, no. 7, pp. 1167–1185, 1997.
- [35] E. D. Sontag and Y. Wang, "New characterizations of input-to-state stability," *IEEE Trans Autom Control*, vol. 41, no. 9, pp. 1283–1294, 1996.
- [36] E. D. Sontag, "Smooth stabilization implies coprime factorization," *IEEE Trans Autom Control*, vol. 34, no. 4, pp. 435–443, 1989.
- [37] Z.-P. Jiang and Y. Wang, "Input-to-state stability for discrete-time nonlinear systems," *Automatica*, vol. 37, no. 6, pp. 857–869, 2001.
- [38] D. Angeli, "A Lyapunov approach to incremental stability properties," *IEEE Trans Autom Control*, vol. 47, no. 3, pp. 410–421, 2002.
- [39] F. Dinuzzo, "Kernels for linear time invariant system identification," *SIAM J Control Optim*, vol. 53, no. 5, pp. 3299–3317, 2015.
- [40] C. A. Micchelli, Y. Xu, and H. Zhang, "Universal kernels," *J Mach Learn Res*, vol. 7, no. 95, pp. 2651–2667, 2006.
- [41] G. Pillonetto, M. H. Quang, and A. Chiuso, "A new kernel-based approach for nonlinear system identification," *IEEE Trans Autom Control*, vol. 56, no. 12, pp. 2825–2840, 2011.
- [42] M. Mazzoleni, M. Scandella, and F. Previdi, "A comparison of manifold regularization approaches for kernel-based system identification," in *IFAC ALCOS*, vol. 52, no. 29, 2019, pp. 180–185.
- [43] S. Formentin, M. Mazzoleni, M. Scandella, and F. Previdi, "Nonlinear system identification via data augmentation," *Syst Control Lett*, vol. 128, pp. 56–63, 2019.
- [44] S. Formentin and A. Chiuso, "Control-oriented regularization for linear system identification," *Automatica*, vol. 127, p. 109539, 2021.
- [45] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.
- [46] A. Dalla Libera, R. Carli, and G. Pillonetto, "Kernel-based methods for Volterra series identification," *Automatica*, vol. 129, p. 109686, 2021.
- [47] M. Mazzoleni, M. Scandella, S. Formentin, and F. Previdi, "Enhanced kernels for nonparametric identification of a class of nonlinear systems," in *ECC*, 2020, pp. 540–545.
- [48] Y. Bhujwalla, V. Laurain, and M. Gilson, "An RKHS approach to systematic kernel selection in nonlinear system identification," in *IEEE CDC*, 2016, pp. 3898–3903.
- [49] D. J. C. MacKay, "Bayesian interpolation," *Neural Comput*, vol. 4, no. 3, pp. 415–447, 1992.
- [50] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.
- [51] D. M. Allen, "The relationship between variable selection and data augmentation and a method for prediction," *Technometrics*, vol. 16, no. 1, pp. 125–127, 1974.
- [52] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J R Stat Soc B*, vol. 36, no. 2, pp. 111–147, 1974.
- [53] G. Pillonetto, A. Aravkin, D. Gedon, L. Ljung, A. H. Ribeiro, and T. B. Schön, "Deep networks for system identification: A survey," *Automatica*, vol. 171, p. 111907, 2025.
- [54] M. Forgiogno and D. Piga, "Continuous-time system identification with neural networks: Model structures and fitting criteria," *Eur J Control*, vol. 59, pp. 69–81, 2021.
- [55] L. Piroddi, "Simulation error minimisation methods for NARX model identification," *Int J Model Identif Control*, vol. 3, no. 4, p. 392, 2008.
- [56] K. Krikelis, J.-S. Pei, K. van Berkel, and M. Schoukens, "Identification of structured nonlinear state-space models for hysteretic systems using neural network hysteresis operators," *Measurement*, vol. 224, p. 113966, 2024.
- [57] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2018.
- [58] V. N. Vapnik, *Statistical learning theory*. Wiley, 1998.
- [59] G. Wahba, *Spline Models for Observational Data*. SIAM, 1990.
- [60] J. Lataire, R. Pintelon, D. Piga, and R. Tóth, "Continuous-time linear time-varying system identification with a frequency-domain kernel-based estimator," *IET Control Theory Appl*, vol. 11, no. 4, pp. 457–465, 2017.

- [61] G. Pillonetto, T. Chen, A. Chiuso, G. D. Nicolao, and L. Ljung, *Regularized System Identification*. Springer, 2022.
- [62] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of transfer functions, regularizations and Gaussian processes — revisited," *Automatica*, vol. 48, no. 8, pp. 1525–1535, 2012.
- [63] N. Aronszajn, "Theory of reproducing kernels," *Trans Am Math Soc*, vol. 68, no. 3, pp. 337–404, 1950.
- [64] S. Saitoh and Y. Sawano, *Theory of Reproducing Kernels and Applications*. Springer, 2016.
- [65] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- [66] M. Mazzoleni, A. Chiuso, M. Scandella, S. Formentin, and F. Previdi, "Kernel-based system identification with manifold regularization: A Bayesian perspective," *Automatica*, vol. 142, p. 110419, 2022.
- [67] F. Dinuzzo and B. Schölkopf, "The representer theorem for Hilbert spaces: a necessary and sufficient condition," in *Adv Neural Inf Process Syst*, vol. 25, 2012.
- [68] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri, "Some properties of regularized kernel methods," *J Mach Learn Res*, vol. 5, pp. 1363–1390, 2004.
- [69] A. Rastogi, G. Blanchard, and P. Mathé, "Convergence analysis of Tikhonov regularization for non-linear statistical inverse problems," *Electron J Stat*, vol. 14, no. 2, pp. 2798–2841, 2020.
- [70] P. A. Herzberg, "The parameters of cross-validation," *Monogr Suppl Psychometrika*, vol. 34, pp. 1–70, 1969.
- [71] G. Pillonetto and A. Chiuso, "Tuning complexity in regularized kernel-based regression and linear system identification: The robustness of the marginal likelihood estimator," *Automatica*, vol. 58, pp. 106–117, 2015.
- [72] B. Mu, T. Chen, and L. Ljung, "Asymptotic properties of generalized cross validation estimators for regularized system identification," in *IFAC SYSID*, vol. 51, no. 15, 2018, pp. 203–208.
- [73] —, "On asymptotic properties of hyperparameter estimators for kernel-based regularization methods," *Automatica*, vol. 94, pp. 381–395, 2018.
- [74] D. P. Bertsekas, *Nonlinear programming*, 3rd ed. Athena Sci, 2016.
- [75] J. F. Bard, *Practical Bilevel Optimization*. Springer, 1998.
- [76] M. Scandella, M. Bin, and T. Parisini, "Kernel-based identification of incrementally input-to-state stable nonlinear systems," in *IFAC World Congress*, vol. 56, 2023, pp. 5127–5132.
- [77] S. Page and S. Grünewälder, "Ivanov-regularised least-squares estimators over large RKHSs and their interpolation spaces," *J Mach Learn Res*, vol. 20, no. 120, pp. 1–49, 2019.
- [78] L. Oneto, S. Ridella, and D. Anguita, "Tikhonov, Ivanov and Morozov regularization for support vector machine learning," *Mach Learn*, vol. 103, no. 1, pp. 103–136, 2015.
- [79] I. Steinwart and A. Christmann, *Support vector machines*. Springer, 2008.
- [80] L. A. Aguirre, M. V. Corêa, and C. C. S. Cassini, "Nonlinearities in NARX polynomial models: representation and estimation," *IEE Proc Control Theory Appl*, vol. 149, no. 4, pp. 343–348, 2002.
- [81] K. Karami, D. Westwick, and J. Schoukens, "Applying polynomial decoupling methods to the polynomial NARX model," *Mech Syst Signal Process*, vol. 148, p. 107134, 2021.
- [82] M. Farina and L. Piroddi, "Identification of polynomial input/output recursive models with simulation error minimisation methods," *Int J Syst Sci*, vol. 43, no. 2, pp. 319–333, 2012.
- [83] I. Steinwart, D. Hush, and C. Scovel, "An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels," *IEEE Trans Inf Theory*, vol. 52, no. 10, pp. 4635–4643, 2006.
- [84] M. L. Stein, *Interpolation of spatial data: Some theory for kriging*. Springer, 1999.
- [85] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert-W function," *Adv Comput Math*, vol. 5, no. 1, pp. 329–359, 1996.
- [86] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *J Physiol*, vol. 117, no. 4, pp. 500–544, 1952.
- [87] M. Scandella, M. Mazzoleni, S. Formentin, and F. Previdi, "A note on the numerical solutions of kernel-based learning problems," *IEEE Trans Autom Control*, vol. 66, no. 2, pp. 940–947, 2021.
- [88] S. Boyd and L. Vandenberghe, *Convex Optimization*. Camb Univ Press, 2019.
- [89] W. H. Young, "On classes of summable functions and their Fourier series," *Proc R Soc Lond, Contain Pap Math Phys Character*, vol. 87, no. 594, pp. 225–229, 1912.
- [90] D. N. Tran, B. S. Rüffer, and C. M. Kellett, "Incremental stability properties for discrete-time systems," in *IEEE CDC*, 2016, pp. 477–482.



based methods.



ternal model theory, adaptive systems, excitable systems and neuromorphic control.



the University of Trieste, Italy, where from 2009 to 2012, he was the Deputy Rector. In 2023, he held a "Scholar-in-Residence" visiting position with Digital Futures-KTH, Sweden. He has coauthored a research monograph in the Communication and Control Series (Springer Nature) and over 400 publications, including journal articles, book chapters, and conference papers. In 2023 he was the recipient of the Knighthood of the Order of Merit of the Italian Republic for scientific achievements abroad awarded by the Italian President of the Republic. In 2018 he received the Honorary Doctorate from the University of Aalborg, Denmark and in 2024, the IEEE CSS Transition to Practice Award. In 2026, he has been elected as member of the Accademia Europaea. Moreover, he was awarded the 2007 IEEE Distinguished Member Award, and was co-recipient of the IFAC Best Application Paper Prize of the Journal of Process Control for the period 2011–2013 and of the 2004 Outstanding Paper Award of IEEE Transactions on Neural Networks. In 2016, he was awarded as Principal Investigator with Imperial of the H2020 European Union flagship Teaming Project KIOS Research and Innovation Centre of Excellence led by the University of Cyprus with an overall budget of over 40 million Euros. He was the 2021–2022 President of the IEEE Control Systems Society, the Editor-in-Chief of IEEE Transactions on Control Systems Technology (2009–2016). He is currently an Editor of Automatica and the Editor-in-Chief of the European Journal of Control. He is a Fellow of IFAC, a Member of IEEE TAB Periodicals Review and Advisory Committee and chairs the IEEE CSS Awards Committee.

Matteo Scandella received the Ph.D. degree in engineering and applied science in 2020 from the University of Bergamo, Italy. Since February 2024, he has been with the Department of Management, Information and Production Engineering, University of Bergamo, Italy. From 2020 to 2024, he was with the Department of Electrical and Electronic Engineering, Imperial College London, UK. He serves as Associate Editor of European Journal of Control. His research focuses on system identification and learning-

Michelangelo Bin (Member, IEEE) received the Ph.D. degree in Control Theory in 2019 from the University of Bologna, Italy. Since October 2022, he has been with the Department of Electrical, Electronic and Information Engineering, University of Bologna, Italy. From 2019 to 2023, he was with the Department of Electrical and Electronic Engineering, Imperial College London, UK. He serves as Associate Editor of Systems & Control Letters. His research interests include systems theory and nonlinear control, regulation and in-

Thomas Parisini (Fellow, IEEE) received the Ph.D. degree in electronic engineering and computer science from the University of Genoa, Italy, in 1993. He was an Associate Professor with Politecnico di Milano, Italy. He currently holds the Chair of industrial control and is the Head of the Control and Power Research Group, Imperial College London, U.K. He also holds a Distinguished Professorship at Aalborg University, Denmark. Since 2001, he has been the Danieli Endowed Chair of automation engineering with