Moment Matching by Kernel-Based Learning

Alessio Moreschini, *Member, IEEE*, Matteo Scandella,
Alessandro Astolfi, *Fellow, IEEE*, and Thomas Parisini, *Fellow, IEEE*

***Abstract*— We introduce a kernel-based moment matching theory which relies upon a novel data-driven model reduction method that employs the estimation of moments within a Reproducing Kernel Hilbert Space. We demonstrate that moment estimation can be enhanced by appropriately tuning the regularization term, regardless of the kernel choice. Additionally, we present conditions to ensure that the Reproducing Kernel Hilbert Space contains only functions which are *bona fide* moments. While exact moment matching with finite data is impractical in this scenario, we introduce the concepts of weak moment matching and moment matching almost everywhere onto the $\mathcal{L}_2$-space. Additionally, we address scenarios in which the dataset contains noisy measurements of outputs that are not yet in a steady-state, which typically biases the estimation due to the effect of the output transients. We further prove that estimating over a Reproducing Kernel Hilbert Space can ensure weak moment matching asymptotically and, with additional assumptions, also moment matching almost everywhere despite these transients. Finally, we provide a probabilistic bound that guarantees weak moment matching for an arbitrarily *finite* amount of data.**

***Index Terms*— Modeling, model reduction, data-driven methods, moment matching, learning for nonlinear systems, statistical model estimation**

I. Introduction

WHILE there is a wealth of measurement and simulation data available, the complexity and high dimensionality of these datasets pose significant challenges across various scientific disciplines [1]. As large-scale dynamic systems generating these datasets continue to grow in complexity, there is an increasing demand for effective model order reduction techniques to interpret the dynamics of these systems and to make simulations, analysis, and controllers more efficient [2]. Large-scale systems are typically described by a large number of equations, which may arise from interconnections of many subsystems [3]–[5], inherent system complexity [6]–[8], or spatial discretization of distributed parameter systems [9], [10].

Main Objective: We aim to enhance the existing paradigm of data-driven model reduction by developing a kernel-based moment matching framework. Given noisy data obtained from measuring the output of an unknown complex dynamic system, we construct a data-driven learning methodology that employs Tikhonov regularization in Reproducing Kernel Hilbert Spaces (RKHS). We employ RKHS theory to estimate the moment function from a Hilbert space based on a data adherence criterion, expressed through a regularized optimization problem. With appropriate tuning, our kernel-based moment matching theory enables the development of models that exhibit low bias, low variance, and strong guarantees on their properties. Finally, we present easily verifiable necessary conditions to ensure that the RKHS contains only functions that can be reliably treated as *bona fide* moments.

State of the Art: Model order reduction involves the development of low-dimensional surrogate models that faithfully represent the essential characteristics of an underlying high-dimensional system. Since no single method universally addresses this task, various techniques have been developed, each tailored to specific properties, contexts, and assumptions, see [11]–[13], and [14] for a recent tutorial. For linear time-invariant (LTI) systems, two major classes of model reduction methods are commonly used: methods based on Singular Value Decomposition (SVD), including Hankel operator techniques and balanced truncation [15]–[23]; and methods based on Krylov subspace projections, such as moment matching and the Loewner framework [24]–[34].

Moment matching is an interpolation method that leverages the concept of *moment* to approximate and interpolate high-order dynamical systems. For LTI systems, the moments are the coefficients of the Laurent series expansion of the associated transfer function at a given point (in which the transfer function is defined) in the complex plane [11, Sec 11.1]. There exists a bijective relation between the moments and the solution of a Sylvester equation as shown in [28], [30]. This connection with the Sylvester equation has enabled [26] to revisit the notion of moment in the time domain and to provide an enhancement for nonlinear systems. This has been achieved by considering an invariant manifold resulting from the cascade interconnection of the system to be reduced with a signal generator system containing the "interpolation signals". Under certain assumptions regarding both the signal generator and the system, it has been demonstrated in [26] that the invariant manifold is a center manifold, thereby linking the moments to the steady-state output response [35] of the cascade interconnection. The center manifold resulting

*This work has been partially supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 739551 (KIOS CoE); by the Italian Ministry for Research in the framework of the 2020 Program for Research Projects of National Interest (PRIN), Grant 2020RTWES4; and by the EPSRC grant EP/X033546.

Alessio Moreschini, Alessandro Astolfi, and Thomas Parisini are with the Department of Electrical and Electronic Engineering, Imperial College London, SW72AZ London, U.K. (e-mails: a.moreschini@imperial.ac.uk, a.astolfi@imperial.ac.uk, t.parisini@imperial.ac.uk). Alessandro Astolfi is also with the Department of Civil Engineering and Computer Science Engineering, University of Rome Tor Vergata, 00133 Rome, Italy. Thomas Parisini is also with the Department of Electronic Systems, Aalborg University, Denmark, and with the Department of Engineering and Architecture, University of Trieste, Italy. 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).

from this cascade interconnection is defined by a nonlinear mapping [36] which is the solution of an *invariance equation*.

The derivation of an analytic solution of this equation is particularly challenging, with the complexity escalating in proportion to the dimensions of both the system and the signal generator. Moreover, in practical applications, the mathematical models that describe the dynamics of high-order systems are frequently either imprecise or unavailable.

With the increasing availability of data and advancements in computational power, data-driven model reduction has gained significant attention in recent years with specific methods, such as Proper Orthogonal Decomposition (POD) and Dynamic Mode Decomposition (DMD), *e.g.*, [37]–[45]. By collecting measurements of the input and output of the system to be reduced, the first data-driven moment matching problem has been posed in [31]. Data-driven moment matching targets datasets generated from measurements only, rather than relying on nonlinear differential equations, and hence the computation of the moment shifts from solving the invariance equation to estimating the moment using the solution of a nonlinear least squares learning problem. An enhancement to moment matching theory, based on noise-free data and agnostic to a priori knowledge of the signal generator, has been presented in [46]. Nevertheless, since noise is an inherent characteristic of sensors, data obtained experimentally by sampling the output of the cascade interconnection are typically corrupted. When measurements are corrupted by stochastic noise, classical least squares estimates for nonlinear models are prone to high variance and systematic bias due to the choice of basis functions. This can result in underfitting when selecting a set of basis functions that is overly conservative due to high systematic bias, or overfitting when using a set of basis functions that is excessively broad due to high variance. Hence, the least squares methods are suitable only in cases with sufficient prior knowledge to allow the selection of the basis functions, *e.g.*, linear or polynomial moment functions. Therefore, in the presence of noisy data, least squares methods may fail to provide accurate or reliable models, hence the need for more flexible alternatives.

On the other hand, it is well-established that Tikhonov regularization in an RKHS setting yields an estimator that minimizes a penalized cost function, where the regularization term constrains the RKHS norm of the estimator. The weight of the penalization term serves as a continuous tuning knob to regulate the complexity of the model, which in turn can be used to better handle the bias-variance trade-off when minimizing the expected mean squared error [47]. An RKHS is an infinite-dimensional Hilbert space in which all the evaluation functionals are linear and bounded [48], [49]. The essence of learning over RKHS lies in the projection of given data, belonging to a finite-dimensional vector space, onto a Hilbert space of functions. The theory of RKHS has become very popular over the past few decades, particularly in learning theory [50]–[57], system identification [58]–[62], and nonlinear systems analysis [63]–[66].

Contributions: As previously mentioned, we employ RKHS theory to estimate the moment function through a regularized optimization problem, which consists of two components: the empirical cost risk and a regularization term. The advantage of using RKHS in moment matching is that the kernel method facilitates the mapping to a higher-dimensional feature space, thereby allowing the complexity to be adjusted through the continuous weight of the regularization term.

This article provides the theoretical foundation for the preliminary conference paper [67], in which we introduced the use of RKHS for moment matching on the basis of the solution of a regularized optimization problem. We present a formal theoretical analysis of data-driven moment matching and moment estimation, and we establish finite-sample performance guarantees for the estimator. In general, moment estimation cannot be exact when only a finite amount of noisy data is available. This makes exact moment matching in data-driven settings challenging, if not infeasible. To address this issue and ensure consistency with the problem at hand, we introduce two key concepts in this article: *moment matching almost everywhere* (*i.e.*, the matching for all points in an $\mathcal{L}_2$ space, except possibly on a set of measure zero), and *weak moment matching* (*i.e.* the difference in the $\mathcal{L}_2$ norm of the two moments is finite and bounded by a constant). If the dataset contains noisy measurements of an output that is not yet in steady-state, the estimation of the moment can be biased by the output transient. The theory developed in this article demonstrates that—despite the transient's effect on the dataset—estimating over an RKHS can ensure moment matching. More specifically, under different assumptions regarding noise and the regularization term, we prove the following.

- *Weak moment matching asymptotically in expectation.* As the number of data increases, the difference in the $\mathcal{L}_2$ norm of the estimated moment and the true moment converges in expectation.
- *Moment matching almost everywhere asymptotically in expectation.* The estimated moment converges to the true moment, except in a subset of its domain with measure zero, in expectation.
- *Weak moment matching asymptotically almost surely.* As the number of data increases, the difference in the $\mathcal{L}_2$ norm of the estimated moment and the true moment converges with probability one.
- *Moment matching almost everywhere asymptotically almost surely.* The estimated moment converges to the true moment, except in a subset of its domain with measure zero, with probability one.
- *Weak moment matching with an a priori probabilistic bound.* For a given finite dataset, we establish an a priori probability that the difference in the $\mathcal{L}_2$ norm between the estimated moments and the true moments is bounded by a certain constant. This result significantly strengthens finite sampling performance by providing an a priori bound, thereby enhancing confidence in the quality of the estimated moments.

Furthermore, we provide an additional condition that renders the notions of moment matching almost everywhere and exact moment matching equivalent.

Organization: Section II recalls the preliminary notions used throughout the article such as vector-valued RKHS (Section II-B) and the notions of moment and moment matching

for continuous-time and discrete-time systems (Section II-C). Section III contains the formal statement of the problem under study, as well as the two notions of matching almost everywhere and weak moment matching. Section IV describes the proposed approach for learning a parameterized model that achieves moment matching directly from input-output data. Then, Section V presents an in-depth analysis of the performance of the proposed kernel-based method both in the asymptotic and in the finite sample regimes. Section VI numerically validates the proposed methodology on the RC ladder benchmark of [68] with two signal generators. Finally, Section VII offers some concluding remarks.

II. Preliminaries

A. Main Notations

The sets of real, natural, and integer numbers are denoted by $\mathbb{R}$, $\mathbb{N}$, and $\mathbb{Z}$, respectively. The sets of nonnegative and positive real numbers are denoted by $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$, respectively. The set of $n \times m$ real matrices is denoted by $\mathbb{R}^{n\times m}$. Tuples of real numbers and column vectors are used interchangeably. The identity matrix is denoted by $\mathbf{I}_n \in \mathbb{R}^{n\times n}$. The matrices with all entries equal to one (zero) are denoted by $\mathbf{1}_{n\times m} \in \mathbb{R}^{n\times m}$ ($\mathbf{0}_{n\times m} \in \mathbb{R}^{n\times m}$). For compactness, when clear from the context, we denote with 0 the zero matrix of appropriate dimension. The Kronecker product is denoted by $\otimes$. Given a matrix $A \in \mathbb{R}^{n\times n}$, $A^\top$ is the transpose of A, A^{-1} is the inverse of A (provided it exists), and $\mathrm{Tr}(A)$ is the trace of A. Given a Euclidean n-space, $|\cdot| : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ denotes the Euclidean norm. The space of square-integrable functions that maps $\mathcal{A} \subseteq \mathbb{R}^n$ to $\mathcal{B} \subseteq \mathbb{R}^m$ is denoted by $\mathcal{L}_2(\mathcal{A}, \mathcal{B})$, *i.e.*, $\mathcal{L}_2(\mathcal{A}, \mathcal{B})$ contains all the measurable functions $f : \mathcal{A} \to \mathcal{B}$ such that $\int_{\mathcal{A}} |f(x)|^2 dx < \infty$. When it is clear from the context, the space $\mathcal{L}_2(\mathcal{A}, \mathcal{B})$ is simply denoted by $\mathcal{L}_2$. The inner product of functions $f \in \mathcal{L}_2(\mathcal{A}, \mathcal{B})$ and $g \in \mathcal{L}_2(\mathcal{A}, \mathcal{B})$, and the induced norm are defined as

$$\langle f, g\rangle_{\mathcal{L}_2} \coloneqq \int_{\mathcal{A}} f(x)^\top g(x) dx\,, \qquad |f|_{\mathcal{L}_2} \coloneqq \sqrt{\langle f, f\rangle_{\mathcal{L}_2}}\,,$$

respectively. All mappings are assumed to be sufficiently smooth, if not otherwise stated. Given a real Hilbert space $\mathcal{H}$ equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, the norm induced by the inner product is given by $|\cdot|_{\mathcal{H}} \coloneqq \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$. Given two real Hilbert spaces $\mathcal{H}_a$ and $\mathcal{H}_b$ and a bounded linear operator $A : \mathcal{H}_a \to \mathcal{H}_b$, the adjoint operator of A is denoted by A^*, *i.e.*, $A^* : \mathcal{H}_b \to \mathcal{H}_a$ is the unique bounded linear operator for which $\langle w, Av\rangle_{\mathcal{H}_b} = \langle A^* w, v\rangle_{\mathcal{H}_a}$, for all $v \in \mathcal{H}_a$ and $w \in \mathcal{H}_b$. Let $(\mathcal{S}, \mathcal{F}, \mathcal{P})$ be a probability space in which $\mathcal{S}$ is the sample space, $\mathcal{F}$ is the sigma algebra over $\mathcal{S}$, and $\mathcal{P}$ the probability measure. All random variables considered hereafter are defined as measurable functions on the same probability space $(\mathcal{S}, \mathcal{F}, \mathcal{P})$. The expected value of a vector-valued random variable $v : \mathcal{S} \to \mathbb{R}^n$ is defined as $\mathbb{E}[v] \coloneqq \int_{\mathcal{S}} v(s)\mathcal{P}(ds)$, provided the integral exists.

B. Vector-Valued RKHS

In this section, we revisit the RKHSs in as much detail as is necessary for the scope of this article (for a comprehensive discussion, see, e.g., [49], [52], [53], [56]). We consider vector-valued RKHS containing functions that map $\mathbb{R}^n$ to $\mathbb{R}^m$, with $n \in \mathbb{N}$ and $m \in \mathbb{N}$, formally defined as follows.

Definition 1 (RKHS)**.** *A Hilbert space $\mathcal{H}$ of functions that map $\mathbb{R}^n$ to $\mathbb{R}^m$ is an RKHS if and only if the linear functional that maps $v \in \mathcal{H}$ to $y^\top v(x) \in \mathbb{R}$ is continuous for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$.*

Definition 2 (Reproducing Kernel)**.** *A function $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{m\times m}$ is a vector-valued reproducing kernel*[1] *if **(i)** for each pair $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$, $k(a, b) = k(b, a)^\top$, and **(ii)** for each finite set of points $\{(x_i, y_i)\}_{i=1}^p \subset \mathbb{R}^n \times \mathbb{R}^m$,*

$$\sum_{i=1}^{p}\sum_{j=1}^{p} y_i^\top k(x_i, x_j) y_j \geq 0\,.$$

For every kernel $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{m\times m}$ there exists a unique vector-valued RKHS, $\mathcal{H}$, of functions that map $\mathbb{R}^n$ to $\mathbb{R}^m$ such that for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ the function $F_{x,y} : \mathbb{R}^n \to \mathbb{R}^m$ such that $F_{x,y}(z) \coloneqq k(z, x)y$, for all $z \in \mathbb{R}^n$, belongs to $\mathcal{H}$ (see [52, Th 1] and [52, Prop 1]). Hence, for each $x \in \mathbb{R}^n$, we can define the linear function $k_x : \mathbb{R}^m \to \mathcal{H}$ such that $k_x y = F_{x,y}$. Furthermore, the linear span $\bar{\mathcal{H}}$ of the set $\{F_{x,y} : x \in \mathbb{R}^n, y \in \mathbb{R}^m\}$ is dense in $\mathcal{H}$ [69, Sec 2]. The relationship between a kernel, k, and its RKHS, $\mathcal{H}$, provides a means of evaluating the functions that belong to the Hilbert space $\mathcal{H}$. In particular, for each $v \in \mathcal{H}$, $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, it can be shown [53, Sec 2.2] that $y^\top v(x) = \langle v, F_{x,y}\rangle_{\mathcal{H}}$. Then, given $k_x^* : \mathcal{H} \to \mathbb{R}^m$, *i.e.* the adjoint operator of k_x, we have

$$y^\top v(x) = \langle v, F_{x,y}\rangle_{\mathcal{H}} = \langle v, k_x y\rangle_{\mathcal{H}} = y^\top k_x^* v\,.$$

Therefore, the evaluation of v in x is given by $v(x) = k_x^* v$. This relationship is called the reproducing property of the RKHS. Hence, we can relate the kernel to the norm $|\cdot|_{\mathcal{H}}$ over $\mathcal{H}$ of functions belonging to $\bar{\mathcal{H}} \subseteq \mathcal{H}$. Note that, for each $x_1 \in \mathbb{R}^n$, $x_2 \in \mathbb{R}^n$, $y_1 \in \mathbb{R}^m$, and $y_2 \in \mathbb{R}^m$, we have

$$\langle F_{x_1,y_1}, F_{x_2,y_2}\rangle_{\mathcal{H}} = \langle F_{x_1,y_1}, k_{x_2} y_2\rangle_{\mathcal{H}} = y_2^\top k(x_2, x_1) y_1\,.$$

Therefore, the squared norm of $v = \sum_{i=1}^p F_{x_i,y_i} \in \bar{\mathcal{H}}$ with $\{(x_i, y_i)\}_{i=1}^p \subset \mathbb{R}^n \times \mathbb{R}^m$ yields

$$|v|_{\mathcal{H}}^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} \big\langle F_{x_i,y_i}, F_{x_j,y_j}\big\rangle_{\mathcal{H}} = \sum_{i=1}^{p}\sum_{j=1}^{p} y_j^\top k(x_j, x_i) y_i \geq 0\,. \tag{1}$$

C. The Notion of Moment for Nonlinear Systems

Consider a multi-input, multi-output, nonlinear dynamical system described by the equations

$$\sigma x(t) = f(x(t), u(t))\,, \qquad x(0) = x_0\,, \tag{2a}$$
$$y(t) = h(x(t), u(t))\,, \tag{2b}$$

with state $x(t) \in \mathbb{R}^{d_{\mathrm{x}}}$ ($d_{\mathrm{x}} \in \mathbb{N}$), input $u(t) \in \mathbb{R}^{d_{\mathrm{u}}}$ ($d_{\mathrm{u}} \in \mathbb{N}$), output $y(t) \in \mathbb{R}^{d_{\mathrm{y}}}$ ($d_{\mathrm{y}} \in \mathbb{N}$), and time $t \in \mathbb{T}$. Note that, for $\mathbb{T} = \mathbb{Z}$ the operator σ describes the shift forward operator, *i.e.*, $\sigma x(t) = x(t+1)$, whereas for $\mathbb{T} = \mathbb{R}$ the operator σ describes the differential operator, *i.e.*, $\sigma x(t) = \dot{x}(t)$. The mappings

[1] For brevity, vector-valued reproducing kernels are referred to as kernels.

$f : \mathbb{R}^{d_\mathrm{x}} \times \mathbb{R}^{d_\mathrm{u}} \to \mathbb{R}^{d_\mathrm{x}}$ and $h : \mathbb{R}^{d_\mathrm{x}} \times \mathbb{R}^{d_\mathrm{u}} \to \mathbb{R}^{d_\mathrm{y}}$ are locally defined around the origin and assumed sufficiently smooth. Without loss of generality $f(0,0) = 0$ and $h(0,0) = 0$. The notions of local observability [70] and local accessibility [71] are revisited below for systems of the form (2).

Definition 3 (Observability)**.** *Let $\mathcal{X} \subset \mathbb{R}^{d_\mathrm{x}}$ be a set containing some open ball centered in* 0*. The system* (2) *is (locally) observable if for any $x_a(0) \in \mathcal{X}$ and $x_b(0) \in \mathcal{X}$ such that $x_a(0) \neq x_b(0)$, and any $u : \mathbb{T} \to \mathbb{R}^{d_\mathrm{u}}$, the output trajectories are not identical,* i.e. *$h(x_a(t), u(t)) \neq h(x_b(t), u(t))$ for some $t \in \mathbb{T}$ with $x_a(t) \in \mathcal{X}$ and $x_b(t) \in \mathcal{X}$.*

Definition 4 (Accessibility)**.** *Let $\mathcal{X} \subset \mathbb{R}^{d_\mathrm{x}}$ be a set containing some open ball centered in* 0*. Let $\mathcal{R}(x(0), T)$ be the set (reachable set) containing all $\bar{x}$ for which there exists a u such that the evolution of* (2a) *from $x(0) \in \mathcal{X}$ satisfies $x(t) \in \mathcal{X}$, for $0 \leq t \leq T$, and $x(T) = \bar{x}$. The system* (2) *is (locally) accessible if for all $x(0) \in \mathcal{X}$ the set $\bigcup_{t \leq T} \mathcal{R}(x(0), t)$ contains a non-empty open subset of $\mathcal{X}$ for all $T > 0$.*

Assumption 1. *The system* (2) *is minimal,* i.e.*, locally observable and locally accessible. The origin of* (2a)*, for $u = 0$, is locally exponentially stable.*

To define moments for systems of the form (2), we introduce a nonlinear dynamical system called the *signal generator*, described by the equations

$$\sigma\omega(t) = s(\omega(t))\,, \qquad \omega(0) = \omega_0\,, \tag{3a}$$
$$u(t) = \ell(\omega(t))\,, \tag{3b}$$

with state $\omega(t) \in \Omega$, $\Omega \subset \mathbb{R}^{d_\omega}$ a sufficiently small compact invariant neighborhood with $0 \in \Omega$, and output $u(t) \in \mathbb{R}^{d_\mathrm{u}}$. The mappings $s : \Omega \to \mathbb{R}^{d_\omega}$ and $\ell : \Omega \to \mathbb{R}^{d_\mathrm{u}}$ are assumed sufficiently smooth and such that $s(0) = 0$ and $\ell(0) = 0$. The signal generator enjoys particular stability properties as established in the following definition, see [35, Sec 8.1].

Definition 5 (Neutral stability)**.** *The system* (3) *is said to be neutrally stable if the origin $\omega = 0$ is a stable equilibrium, and there is an open neighborhood of the origin in which every point is Poisson stable*[2]*.*

Assumption 2. *The signal generator* (3) *is locally observable and neutrally stable.*

It is clear that a neutrally stable signal generator implies that for any $\omega_0 \in \Omega$ the corresponding trajectory ω is persistent in time on the domain Ω, and does not decay to zero as time tends to infinity. Building on Assumptions 1 and 2, in [26] the notion of moment has been further associated with the steady-state output response [35] of the interconnection of the system (2) with the signal generator (3), that is, the system

$$\sigma\omega(t) = s(\omega(t))\,, \tag{4a}$$
$$\sigma x(t) = f(x(t), \ell(\omega(t)))\,, \tag{4b}$$
$$y(t) = h(x(t), \ell(\omega(t)))\,. \tag{4c}$$

[2]A point ω_0 is *Poisson stable* if, for each time $T > 0$ and each neighborhood U^0 of ω_0, the trajectory ω with initial condition ω_0 passes through U^0 for some $t_1 > T$ and $t_2 < -T$, see [35, Sec 8.1].

Definition 6 (Center Manifold)**.** *A smooth function $\pi : \Omega \to \mathbb{R}^{d_\mathrm{x}}$, locally defined in Ω and satisfying $\pi(0) = 0$, is a center manifold of* (4) *if there exists an open neighborhood $\mathcal{B}$ of $(\omega, x) = (0, 0)$ such that*

$$x(0) = \pi(\omega(0)) \implies x(t) = \pi(\omega(t))\,, \tag{5}$$

for all $t \in \mathbb{T}$ such that $(\omega(t), x(t)) \in \mathcal{B}$.

The implication (5) is an *invariance property* and yields that the graph of π, *i.e.* the subset $\mathcal{M} \coloneqq \{(x, \omega) : x = \pi(\omega)\} \cap \mathcal{B}$, is invariant with respect to the solutions of (4). Leveraging Assumptions 1 and 2, and invoking the Center Manifold Theorem [36, Th 1], it can be shown that there exists a mapping $\pi : \Omega \to \mathbb{R}^{d_\mathrm{x}}$ which defines a locally attractive center manifold for the interconnected system (4), see [35, Prop 8.1.1]. Therefore, depending on the time domain, for all $t \in \mathbb{T}$ such that $(\omega(t), x(t)) \in \mathcal{B}$, the mapping π is defined as the unique solution of the so-called *invariance equation*, *i.e.*

$$\frac{\partial \pi}{\partial \omega}(\omega) s(\omega) = f(\pi(\omega), \ell(\omega))\,, \qquad \text{if } \mathbb{T} = \mathbb{R}, \tag{6a}$$
$$\pi(s(\omega)) = f(\pi(\omega), \ell(\omega)), \qquad \text{if } \mathbb{T} = \mathbb{Z}\,. \tag{6b}$$

Regardless of the time domain, for each trajectory ω with $\omega_0 \in \Omega$, the mapping π guarantees the existence of a (local) steady-state output response $y_{\mathrm{ss}}(t) \coloneqq h(\pi(\omega(t)), \ell(\omega(t)))$, see [35, Sec 8.2], which is such that

$$\lim_{t \to \infty} y(t) - y_{\mathrm{ss}}(t) = 0\,. \tag{7}$$

Definition 7 (Moment)**.** *The (local) moment of the system* (2) *at (s, ℓ) is defined as the mapping $W : \mathbb{R}^{d_\omega} \to \mathbb{R}^{d_\mathrm{y}}$ such that*

$$\forall \omega \in \Omega, \qquad W(\omega) \coloneqq h(\pi(\omega), \ell(\omega))\,. \tag{8}$$

D. Moment Matching for Nonlinear Systems

For a given signal generator of the form (3) defining the moment W of the system (2) at (s, ℓ), the moment matching problem is to find another system such that its moment $\overline{W}$ at (s, ℓ), satisfies the matching condition

$$W = \overline{W}. \tag{9}$$

Definition 8 (Moment Matching Model)**.** *A system of dimension $d_\xi \in \mathbb{N}$ described by the equations*

$$\sigma\xi(t) = \bar{f}(\xi(t), u(t))\,, \qquad \xi(0) = \xi_0\,, \tag{10a}$$
$$\bar{y}(t) = \bar{h}(\xi(t), u(t))\,, \tag{10b}$$

with state $\xi(t) \in \mathbb{R}^{d_\xi}$, input $u(t) \in \mathbb{R}^{d_\mathrm{u}}$, output $\bar{y}(t) \in \mathbb{R}^{d_\mathrm{y}}$, and moment $\overline{W}$ at (s, ℓ) is said to achieve moment matching at (s, ℓ) if (9) *holds. Furthermore, if $d_\xi < d_\mathrm{x}$ then* (10) *is a reduced order model of* (2)*.*

Remark 1. A model of dimension $d_\xi = d_\omega$ achieving moment matching at (s, ℓ) was originally proposed in [26]. Recently in [72], a family of models of dimension $d_\xi \geq d_\omega$ achieving moment matching has been described by the equations

$$\sigma\xi_a = \bar{f}_a(\xi_a, \xi_b, u)\,, \quad \sigma\xi_b = \bar{f}_b(\xi_a, \xi_b, u)\,, \quad \bar{y} = \bar{h}(\xi_a, \xi_b, u)\,,$$

with $\xi_a(t) \in \mathbb{R}^{d_\omega}$ and $\xi_b(t) \in \mathbb{R}^{d_\xi - d_\omega}$, $u(t) \in \mathbb{R}^{d_\mathrm{u}}$, $\bar{y}(t) \in \mathbb{R}^{d_\mathrm{y}}$, and mappings $\bar{f}_a$, $\bar{f}_b$, and $\bar{h}$ of appropriate

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3618165

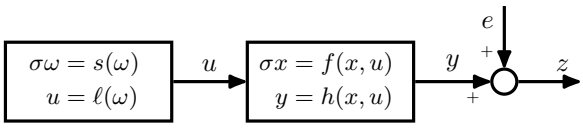

Fig. 1. Block diagram of the moment matching setup in which the measurement of the output $\boldsymbol{y}$ is affected by the additive noise $\boldsymbol{e}$.

dimensions such that $\bar{f}_a(0,0,0) = 0$, $\bar{f}_b(0,0,0) = 0$, $\bar{h}(0,0,0) = 0$, $\bar{f}_a(\xi_a,0,\ell(\xi_a)) = s(\xi_a)$, $\bar{f}_b(\xi_a,0,\ell(\xi_a)) = 0$, and $\bar{h}(\xi_a,0,\ell(\xi_a)) = W(\xi_a)$.

III. Problem Statement

As outlined in Section II-C, it is possible to identify a model that achieves moment matching using the model (10) once the moment W is derived. In this regard, the first data-driven moment matching problem for nonlinear system has been posed in [31], in which the problem has been cast in terms of a least mean squares estimation problem and the (approximate) moment of (2) at (s,ℓ) has been constructed by collecting samples from the signal generator (4a) and the output (4c). Yet, in a practical experimental setup, the measured output is affected by noise. In particular, assuming that the input signal u, generated by the known signal generator, is perfectly known and unaffected by any noise, we consider that the output measurements y are subject to additive noise, *i.e.*

$$\forall t \in \mathbb{T}, \qquad z(t) \coloneqq y(t) + e(t), \tag{11}$$

where $y(t) \in \mathbb{R}^{d_\mathrm{y}}$ is the output in (4c) and $e(t) \in \mathbb{R}^{d_\mathrm{y}}$ is a random variable that represent the measurement noise. The statistical properties of the stochastic process e are formally introduced in Section V (see Assumptions 4 and 5), and the interconnected model with noisy measurements and noise-free input is shown in Figure 1.

In this work, we aim to devise a kernel-based algorithm that, given the signal generator (3), provides a kernel-based model that approximately achieves moment matching from data using measurements obtained from the interconnected system (4) without solving the invariance equation (6). In turn, the mapping f, h, π, and W and the initial condition $x_0 \in \mathbb{R}^{d_\mathrm{x}}$ are assumed unknown.

IV. A Kernel-Based Moment Matching Framework

In this section, we illustrate our methodology to obtain a model that approximately achieves moment matching. First, we estimate the moment $W : \mathbb{R}^{d_\omega} \to \mathbb{R}^{d_\mathrm{y}}$ of the system, and, then, we use it to define a parametric model using (10). Since the method relies on data, we assume that we have collected $n \in \mathbb{N}$ data samples from the output measurements z in (11) and the trajectory ω generated by (3a). In particular, we define the dataset

$$\mathcal{D} \coloneqq \left\{(\bar{t}_i, \bar{\omega}_i, \bar{z}_i)\right\}_{i=1}^{n} \subseteq \mathbb{T} \times \Omega \times \mathbb{R}^{d_\mathrm{y}}, \tag{12}$$

where, for all $i \in \{1,\dots,n\}$, $\bar{t}_i \in \mathbb{T}$ denotes the i-th sampling time, $\bar{\omega}_i = \omega(\bar{t}_i)$, and $\bar{z}_i = z(\bar{t}_i)$. Furthermore, we assume that $\bar{t}_i$ is a non-decreasing divergent sequence. Since we assume to know the signal generator, the samples of ω are generated solving (3a) without using sensors. Thus, the samples of ω are not corrupted by noise. The measurements (11) can be rewritten in terms of the moment and of the output transient, *i.e.*

$$\forall t \in \mathbb{T}, \qquad z(t) = W(\omega(t)) + \tau(t) + e(t),$$

where $\tau : \mathbb{T} \to \mathbb{R}^{d_\mathrm{y}}$ is the output transient given by the difference

$$\forall t \in \mathbb{T}, \qquad \tau(t) \coloneqq y(t) - y_\mathrm{ss}(t). \tag{13}$$

For all $i \in \{1,\dots,n\}$ and sampling time $\bar{t}_i \in \mathbb{T}$, we denote $\bar{\tau}_i = \tau(t_i)$ and $\bar{e}_i = e(t_i)$ as the output transient and the noise affecting the i-th measurement, respectively. We also define the sequence

$$T_n \coloneqq \sum_{i=1}^{n}\sum_{j=1}^{n} \bar{\tau}_i^\top \bar{\tau}_j. \tag{14}$$

If the system has a feedthrough term, the moment W is defined using the functions h, π, and ℓ. Therefore, in this case, since the function ℓ is assumed to be known, instead of learning directly the moment W, we learn the function $\mu^\star : \mathbb{R}^{d_\omega} \times \mathbb{R}^{d_\mathrm{u}} \to \mathbb{R}^{d_\mathrm{y}}$ such that $\mu^\star(\omega, \ell(\omega)) = W(\omega)$. Instead, if the system is strictly proper, the function ℓ does not enter directly into the model, and we directly identify the moment W. To consider simultaneously the aforementioned cases, we define $d_\mathrm{q} \coloneqq d_\omega$ if it is known that the system is strictly proper or $d_\mathrm{q} \coloneqq d_\omega + d_\mathrm{u}$ otherwise. We also introduce the function $q : \mathbb{R}^{d_\omega} \to \mathbb{R}^{d_\mathrm{q}}$ such that

$$q(\omega) = \begin{cases} \omega & \text{if } d_\mathrm{q} = d_\omega, \\ (\omega, \ell(\omega)) & \text{if } d_\mathrm{q} = d_\omega + d_\mathrm{u}. \end{cases}$$

In summary, the objective is to estimate the function $\mu^\star : \mathbb{R}^{d_\mathrm{q}} \to \mathbb{R}^{d_\mathrm{y}}$ such that $W(\omega) = \mu^\star(q(\omega))$. For compactness, for all $i \in \{1,\dots,n\}$, we also define $\bar{q}_i = q(\bar{\omega}_i) \in \mathbb{R}^{d_\mathrm{q}}$.

A. Kernel-Based Estimation of Moments

Data-driven methods for moment matching rely on the fact that (7) implies that $z(t) - W(\omega(t)) \approx e(t)$ for large $t \in \mathbb{T}$. Therefore, we define an estimator that minimizes the squared deviation between the output z and the function $W(\omega)$ at the same time instant. We thus define the estimator as the function that solves the infinite-dimensional optimization problem

$$\min_{\mu \in \mathcal{H}} \quad \sum_{i=1}^{n} |\bar{z}_i - \mu(\bar{q}_i)|^2 + \rho_n |\mu|_{\mathcal{H}}^2, \tag{15}$$

where $\mathcal{H}$ is a vector-valued RKHS, as in Definition 1, containing functions that map $\mathbb{R}^{d_\mathrm{q}}$ to $\mathbb{R}^{d_\mathrm{y}}$ with kernel $k : \mathbb{R}^{d_\mathrm{q}} \times \mathbb{R}^{d_\mathrm{q}} \to \mathbb{R}^{d_\mathrm{y} \times d_\mathrm{y}}$, and $\rho_n \in \mathbb{R}_{\geq 0}$ is a parameter to be tuned that can vary with the number of data n. The cost function describing the optimization problem (15) is the sum of two terms. The first one is the squared deviation that inspires the method, while the second one is a regularization term that is used to penalize functional complexity to avoid overfitting the available data. Here, the complexity of the function is defined via the RKHS norm used as a hypothesis space for the learning problem. Thus, the kernel of the RKHS influences the meaning of the regularization term as explained in (1). The parameter ρ_n weights the relative contribution of both terms.

To solve the possibly infinite-dimensional problem (15), the Representer theorem [52, Th 4] can be applied.

Theorem 1 (Representer Theorem)**.** *If $\hat{\mu}_n$ minimizes* (15)*, then there exist $\{c_i\}_{i=1}^n \subset \mathbb{R}^{d_y}$ such that*

$$\hat{\mu}_n = \sum_{i=1}^{n} F_{\bar{q}_i, c_i} \tag{16}$$

where $F_{s,c} : \mathbb{R}^{d_y} \to \mathcal{H}$ is as defined in Section II-B for every $s \in \mathbb{R}^{d_q}$ and $c \in \mathbb{R}^{d_y}$.

Theorem 1 transforms (15) from an infinite-dimensional problem on a Hilbert Space $\mathcal{H}$ to a finite-dimensional problem on $\mathbb{R}^{nd_y}$. Note that the resulting estimator $\hat{\mu}_n$ in (16) depends on the finite amount of data n that establishes the dimension of the optimization problem. In particular, using (1) and the definition of $F_{q,i}$, the estimator is given by (16) with $\{c_i\}_{i=1}^n$ that solve the finite-dimensional optimization problem

$$\min_{c \in \mathbb{R}^{nd_y}} \quad |\bar{z} - Kc|^2 + \rho_n \, c^\top K c,$$

where $K \in \mathbb{R}^{nd_y \times nd_y}$ is a block matrix of which the (i,j)-th block is $k(\bar{q}_i, \bar{q}_j)$, $c \coloneqq (c_1, \dots, c_n) \in \mathbb{R}^{nd_y}$, and $\bar{z} \coloneqq (\bar{z}_1, \dots, \bar{z}_n) \in \mathbb{R}^{nd_y}$. Since this is a quadratic problem, its solution can be found by solving the linear equation $K\big(K + \rho_n \mathbf{I}_{nd_y}\big)c = K\bar{z}$. Although the matrix K can be singular, one can readily verify that $c = \big(K + \rho_n \mathbf{I}_{nd_y}\big)^{-1}\bar{z}$ is always a valid solution of the quadratic optimization problem. As a result, we have

$$\forall s \in \mathbb{R}^{d_q}, \qquad \hat{\mu}_n(s) = \sum_{i=1}^{n} F_{\bar{q}_i, c_i}(s) = \sum_{i=1}^{n} k(s, \bar{q}_i) c_i \,,$$

and, more compactly, $\hat{\mu}_n(q) = k^\star(s)c$ with $k^\star(s) \coloneqq [k(\bar{q}_1, s), \cdots, k(\bar{q}_n, s)] \in \mathbb{R}^{d_y \times nd_y}$. Finally, the kernel-based model that approximately achieves moment matching takes the form (10) provided that the output mapping $\bar{h}$ in (10b) satisfies

$$\bar{h}(\xi, \ell(\xi)) = \overline{W}_n \coloneqq \hat{\mu}_n(q(\xi)), \tag{17}$$

which highlights the dependency of the moment of the reduced-order model on the amount of data n.

B. Kernel Function Selection

The methodology in Section IV-A is general and applies to any RKHS $\mathcal{H}$. However, in the context of model reduction by moment matching, the choice of $\mathcal{H}$ is crucial, as not all RKHS contain functions that are *bona fide* moments. Indeed, the resulting estimator obtained from a general RKHS may not correspond to a valid moment about the system equilibrium (e.g., the origin for system (2)), leading to reduced-order models that misrepresent behavior near that point. This dependence is formalized in the following proposition.

Proposition 1. *The RKHS $\mathcal{H}$ contains exclusively valid moment functions only if $k(0,0) = \mathbf{0}_{d_y \times d_y}$.*

Proof. First, recall that $\pi(0) = 0$, $\ell(0) = 0$ and $h(0,0) = 0$ as assumed in Section II-C. Therefore, from Definition 7, we have that $W(0) = h(\pi(0), \ell(0)) = \mu^\star(0) = 0$. Hence, $\mathcal{H}$ contains only valid moment functions only if $g(0) = 0$, for all $g \in \mathcal{H}$. Then, for the reproducing property, we also have $k_0^* g = 0$. Due to the linearity of the operator, k_0^* and its adjoint k_0 are the zero operators of their respective operator spaces. Thus, for each $y \in \mathbb{R}^{d_y}$, we obtain $0 = (k_0 y)(0) = F_{0,y}(0) = k(0,0)y$, which holds true only if $k(0,0) = \mathbf{0}_{d_y \times d_y}$. □

If $\mathcal{H}$ does not satisfy Proposition 1, the estimator exists but does not represent a moment of the system (2) around the origin. Owing to Proposition 1, the widely used stationary kernels (see [73, Sec 4.2.1] or [50, Sec 4.4] for more detail) are not suitable for this application. For instance, popular kernels like the Gaussian Kernel and all the various types of radial basis functions [74] do not define an RKHS that contains only valid moment functions. This is formalized hereafter in Definition 9 and in Corollary 1.

Definition 9. *A kernel $k : \mathbb{R}^{d_q} \times \mathbb{R}^{d_q} \to \mathbb{R}^{d_y \times d_y}$ is stationary if there exist a function $\bar{k} : \mathbb{R}^{d_q} \to \mathbb{R}^{d_y \times d_y}$ such that $k(a,b) = \bar{k}(a-b)$, for all $a, b \in \mathbb{R}^{d_q}$.*

Corollary 1. *Let $k : \mathbb{R}^{d_q} \times \mathbb{R}^{d_q} \to \mathbb{R}^{d_y \times d_y}$ be a stationary kernel that satisfies the condition of Proposition 1. Then, $k(a,b) = 0$ for all $a, b \in \mathbb{R}^{d_q}$.*

Proof. Since k is stationary, there exists a function $\bar{k} : \mathbb{R}^{d_q} \to \mathbb{R}^{d_y \times d_y}$ such that $k(a,b) = \bar{k}(a-b)$ for all $a, b \in \mathbb{R}^{d_q}$. Directly from the definition of kernel (see Definition 2), $y^\top \bar{k}(s) y \geq 0$, for all $y \in \mathbb{R}^{d_y}$ and $s \in \mathbb{R}^{d_q}$. Also,

$$\begin{aligned} y^\top \bar{k}(s) y = y^\top k(s,0) y &= \langle F_{s,y}, F_{0,y} \rangle \\ &\leq \sqrt{\langle F_{s,y}, F_{s,y} \rangle \langle F_{0,y}, F_{0,y} \rangle} \\ &= \sqrt{y^\top k(s,s) y \, y^\top k(0,0) y} = y^\top \bar{k}(0) y. \end{aligned}$$

Thus, from Proposition 1, we have $0 \leq y^\top \bar{k}(s) y \leq y^\top \bar{k}(0) y = 0$. We complete the proof by noting that the only function $\bar{k}$ that satisfies this property is $\bar{k}(s) = 0$. □

In the scalar case, *i.e.*, $d_y = 1$, a suitable kernel is the polynomial kernel with degree $p \in \mathbb{N}$ defined as $k(a,b) = (a^\top b)^p$, for all $a, b \in \mathbb{R}^{d_q}$. Other examples can be found by exploiting the property that the multiplication of two kernels is also a kernel [50, Prop 13.2]. Hence, a kernel suitable for moment matching can be constructed by multiplying a generic kernel with a kernel satisfying the condition in Proposition 1.

For the non-scalar case, *i.e.*, $d_y > 1$, a commonly used strategy to design the kernel is to use a separable kernel [55, Sec 4]. In particular, if $k_1 : \mathbb{R}^{d_q} \times \mathbb{R}^{d_q} \to \mathbb{R}$ is a scalar kernel that satisfies the condition in Proposition 1, we define the vector-valued kernel as

$$\forall a, b \in \mathbb{R}^{d_q}, \qquad k(a,b) \coloneqq k_1(a,b) B_\alpha,$$

where $\alpha \in [0,1]$ is a parameter to be tuned, and $B_\alpha \coloneqq \alpha \mathbf{1}_{d_y \times d_y} + (1-\alpha)\mathbf{I}_{d_y}$. More detail on this type of separable kernel can be found in [55, Sec 4.1].

We conclude this section by describing the kernel used for the numerical results illustrated in Section VI as an example of a kernel which is well-defined for our purposes. In particular, we consider the kernel $k : \mathbb{R}^{d_q} \times \mathbb{R}^{d_q} \to \mathbb{R}^{d_y \times d_y}$ such that, for all $q_a \in \mathbb{R}^{d_q}$ and $q_b \in \mathbb{R}^{d_q}$,

$$k(q_a, q_b) \coloneqq k_{\mathsf{p}}(q_a, q_b) k_{\mathsf{g}}(q_a, q_b) B_\alpha, \tag{18}$$

where $\alpha \in [0,1]$, k_{p} is a polynomial kernel [50, Eq (1.61)], and k_{g} is a Gaussian kernel [50, Eq (1.62)]. In particular, the kernel k_{p} and k_{g} are such that, for all $q_a \in \mathbb{R}^{d_{\mathrm{q}}}$ and $q_b \in \mathbb{R}^{d_{\mathrm{q}}}$,

$$k_{\mathrm{p}}(q_a, q_b) := \begin{cases} \left(\tau_\omega q_b^\top q_a\right)^p, & \text{if } d_{\mathrm{q}} = d_\omega, \\ \left(\tau_\omega \omega_b^\top \omega_a + \tau_u u_b^\top u_a\right)^p, & \text{if } d_{\mathrm{q}} = d_\omega + d_{\mathrm{u}}, \end{cases}$$

$$k_{\mathrm{g}}(q_a, q_b) := \begin{cases} e^{-\gamma_\omega |q_a - q_b|^2}, & \text{if } d_{\mathrm{q}} = d_\omega, \\ e^{-\gamma_\omega |\omega_a - \omega_b|^2 - \gamma_u |u_a - u_b|^2}, & \text{if } d_{\mathrm{q}} = d_\omega + d_{\mathrm{u}}, \end{cases}$$

where $\omega_a \in \mathbb{R}^{d_\omega}$, $\omega_b \in \mathbb{R}^{d_\omega}$, $u_a \in \mathbb{R}^{d_{\mathrm{u}}}$ and $u_b \in \mathbb{R}^{d_{\mathrm{u}}}$ are such that $q_a = (\omega_a, u_a)$, and $q_b = (\omega_b, u_b)$. Finally, note that $\alpha \in [0,1]$, $\tau_\omega \in \mathbb{R}_{\geq 0}$, $\tau_u \in \mathbb{R}_{\geq 0}$, $p \in \mathbb{N}$, $\gamma_\omega \in \mathbb{R}_{\geq 0}$, and $\gamma_u \in \mathbb{R}_{\geq 0}$ are parameters to tune.

C. Hyperparameters Selection

The method illustrated in Section IV-A requires the tuning of several parameters: the regularization term $\rho_n \in [0, \infty)$ and the parameters $\psi \in \mathbb{R}^{d_\psi}$ that define the kernel (for example, $\psi := (\tau_w, \tau_u, \gamma_w, \gamma_u, \alpha, p)$ for the kernel (18)). These parameters are usually called hyperparameters in learning problems and their value determines the trade-off between the complexity of the model and the fitting performance. A good selection helps to avoid problems related to over- and under-fitting the available dataset. In the literature, this problem is known as *model assessment*, see [47, Sec 7], and it involves dealing with the bias-variance trade-off [47, Sec 7.2]. In particular, it is possible to show that, for each $\omega \in \Omega$, the mean square error (MSE) of the estimator can be equivalently rewritten as the sum of two terms

$$\begin{aligned} &\mathrm{MSE}(\omega) := \mathbb{E}\left[\left|\hat{\mu}_n(q(\omega)) - W(\omega)\right|^2\right] \\ &= \underbrace{\left|\mathbb{E}[\hat{\mu}_n(q(\omega))] - W(\omega)\right|^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[\left|\hat{\mu}_n(q(\omega)) - \mathbb{E}[\hat{\mu}_n(q(\omega))]\right|^2\right]}_{\text{Variance}}. \end{aligned}$$

Hyperparameters that induce more complex models tend to result in high variance and low bias, while those that lead to less complex models exhibit the opposite phenomenon. Hence, it is of utmost importance to select hyperparameters in a way that achieves an optimal balance between the two components of the error. In our context, the most important hyperparameter is ρ_n, as a large ρ_n leads to high bias and low variance.

The proposed method does not rely on a specific set of hyperparameters, making it possible to employ most of the methods available in the literature, each with its own advantages and disadvantages. Although optimizing hyperparameter selection is beyond the scope of this article, it is worth noting that the most common model assessment methods rely on cross-validation techniques, which aim to approximate $MSE(\omega)$ from available data for various values of $\omega \in \Omega$. In particular, *Monte Carlo cross-validation* [75], *k-fold cross-validation* [73], [76], *leave-one-out cross-validation* [51], [73], [77], and *generalized cross-validation* [78] are methods in this family. Alternative methods, such as *Empirical Bayes* [79], rely on Bayesian theory and Gaussian processes, and they are very computationally efficient [73, Sec 5.4.1]. All these methods select the optimal parameters by solving an optimization problem that can also be equipped with constraints on the hyperparameters.

D. The Linear Time Invariant case

The presented methodology is designed for nonlinear models, either in continuous or discrete time. In this section, we analyze how to use the proposed approach for LTI models with an LTI signal generator, *i.e.*, the maps s, ℓ, f, and h are linear. As shown in [26], in this case, the maps π and W are also linear and, thus, the map $\mu^\star$ that we aim to learn from the data is linear. Since the space of linear function that maps $\mathbb{R}^{d_{\mathrm{q}}}$ to $\mathbb{R}^{d_{\mathrm{y}}}$ is an RKHS [50, Prop 2.1] with kernel $k(a,b) = (a^\top b) \otimes \mathbf{I}_{d_{\mathrm{y}}}$, for each $a, b \in \mathbb{R}^{d_{\mathrm{q}}}$, the proposed method can also be used for LTI models. In particular, defining $Q \in \mathbb{R}^{d_{\mathrm{q}} \times n}$ as the matrix in which the i-th column is $\bar{q}_i$, we have $K = (Q^\top Q) \otimes \mathbf{I}_{d_{\mathrm{y}}}$ and $k^\star(q) = (q^\top Q) \otimes \mathbf{I}_{d_{\mathrm{y}}}$, for all $q \in \mathbb{R}^{d_{\mathrm{q}}}$. Then, the estimator becomes

$$\hat{\mu}_n(q) = k^\star(q)c = \left[q^\top Q\left(Q^\top Q + \rho_n \mathbf{I}_n\right)^{-1}\right] \otimes \mathbf{I}_{d_{\mathrm{y}}} \bar{z}.$$

V. Analysis of the Kernel-Based Moment

In this section, we analyze the relationship between the estimator $\overline{W}_n$ described in Section IV and the moment W of the underlying system. To begin with, we first recall that by smoothness of h, ℓ, and π the moment W defined in (8) is smooth and thus continuous. Moreover, since Ω is compact, W is also bounded, *i.e.* $\int_\Omega |W(\omega)|^2 d\omega < \infty$. Hence, the moment W is a square-integrable function, *i.e.* $W \in \mathcal{L}_2(\Omega, \mathbb{R}^{d_{\mathrm{y}}})$.

To proceed with the analysis, we introduce the following assumption on the kernel and we prove a technical lemma.

Assumption 3. *Let $\mathcal{Q} := \{q(\omega) : \omega \in \Omega\} \subseteq \mathbb{R}^{d_{\mathrm{q}}}$. The kernel $k : \mathbb{R}^{d_{\mathrm{q}}} \times \mathbb{R}^{d_{\mathrm{q}}} \to \mathbb{R}^{d_{\mathrm{y}} \times d_{\mathrm{y}}}$ is such that $\zeta^2 := \sup_{s \in \mathcal{Q}} \mathrm{Tr}(k(s,s))$ is finite.*

Lemma 1. *Suppose that Assumption 3 holds and let $\widehat{\mathcal{H}}$ be the RKHS with the restriction of k on $\mathcal{Q} \times \mathcal{Q}$ as kernel. Then, $\widehat{\mathcal{H}} \subseteq \mathcal{L}_2(\mathcal{Q}, \mathbb{R}^{d_{\mathrm{y}}})$ and $\overline{W}_n$ is a square-integrable function,* i.e. *$\overline{W}_n \in \mathcal{L}_2(\Omega, \mathbb{R}^{d_{\mathrm{y}}})$.*

Proof. The first claim follows from [53, Cor 4.6], by Assumption 3. Then, the restriction of $\hat{\mu}_n \in \mathcal{H}$ belongs to $\widehat{\mathcal{H}} \subseteq \mathcal{L}_2(\mathcal{Q}, \mathbb{R}^{d_{\mathrm{y}}})$. Therefore,

$$\int_\Omega \left|\overline{W}_n(\omega)\right|^2 d\omega = \int_\Omega \left|\hat{\mu}_n(q(\omega))\right|^2 d\omega = \int_{\mathcal{Q}} \left|\hat{\mu}_n(s)\right|^2 ds < \infty$$

and thus $\overline{W}_n \in \mathcal{L}_2(\Omega, \mathbb{R}^{d_{\mathrm{y}}})$. □

In summary, both the moment W and its estimator $\overline{W}_n$ are square-integrable functions, and it is possible to define a distance between W and $\overline{W}_n$ in the $\mathcal{L}_2$ space. Bearing this in mind, we introduce two useful weaker notions of moment matching in the $\mathcal{L}_2$ space.

Definition 10 (Moment matching almost everywhere)**.** *The model* (10) *is said to achieve moment matching at (s, ℓ) almost everywhere if its moment $\overline{W}_n \in \mathcal{L}_2$ is such that*

$$\left|\overline{W}_n - W\right|^2_{\mathcal{L}_2(\Omega, \mathbb{R}^{d_{\mathrm{y}}})} = 0.$$

Definition 11 (Weak moment matching)**.** *Let $\varrho \in \mathbb{R}_{\geq 0}$. The model* (10) *is said to achieve ϱ-weak moment matching at (s, ℓ) if its moment $\overline{W}_n \in \mathcal{L}_2$ is such that*

$$\left|\overline{W}_n - W\right|^2_{\mathcal{L}_2(\Omega, \mathbb{R}^{d_{\mathrm{y}}})} \leq \varrho.$$

Note that the notion of moment matching almost everywhere is weaker than the notion of moment matching in the sense of (9) because it implies that the moments W and $\overline{W}_n$ can differ in a subset of their domain as long as it has measure zero. Nonetheless, by the principle that two continuous functions equal almost everywhere on a domain are equal everywhere [80, Ch 3], moment matching almost everywhere implies moment matching. Specifically, if $\overline{W}_n$ is continuous and (10) is a model achieving moment matching at (s, ℓ) almost everywhere with moment $\overline{W}_n \in \mathcal{L}_2$ then (10) achieves moment matching, *i.e*, $\overline{W}_n = W$.

Moreover, as established in [53, Prop 5.1], all functions within the space $\mathcal{H}$ are continuous if the kernel associated with $\mathcal{H}$ is continuous. Consequently, for continuous kernels (*e.g.*, (18)) defining a continuous estimator function, we can infer exact moment matching in the sense of (9).

Before presenting the main results of our analysis, we first introduce the statistical properties of the stochastic process e defined in (11). Specifically, we consider two distinct scenarios, each characterized by the following assumptions.

Assumption 4. *The noise e is a stochastic process defined on $\mathbb{T}$ such that there exists $\Sigma \in \mathbb{R}_{\geq 0}$ with the following properties: for all $t_1 \in \mathbb{T}$ and $t_2 \in \mathbb{T} \setminus \{t_1\}$, we have $\mathbb{E}[e(t_1)] = 0$, $\mathbb{E}[e(t_1)^\top e(t_2)] = 0$ and $\mathbb{E}[|e(t_1)|^2] \leq \Sigma^2$.*

Assumption 5. *The noise e is a stochastic process defined on $\mathbb{T}$ such that there exist $\Sigma \in \mathbb{R}_{\geq 0}$ and $\Psi \in \mathbb{R}_{\geq 0}$ with the following properties: for all $t_1 \in \mathbb{T}$ and $t_2 \in \mathbb{T} \setminus \{t_1\}$, we have $\mathbb{E}[e(t_1)] = 0$, $\mathbb{E}[e(t_1)^\top e(t_2)] = 0$ and*

$$\forall m \in \mathbb{N}_{\geq 2}, \quad \mathbb{E}[|e(t_1)|^m] \leq \frac{m!}{2}\Sigma^2\Psi^{m-2}.$$

Note that Assumption 5 implies Assumption 4, thereby making the latter a strictly weaker condition.

Building upon the preceding discussion, we now present the main results of our analysis, which constitute the foundation of the proposed theory. To enhance readability, all proofs have been moved to Section V-A. The first theorem illustrates the property of the method asymptotically in expectation.

Theorem 2 (Matching in Expectation)**.** *Suppose that Assumptions 1, 2, 3, and 4 hold and that $\mu^\star \in \mathcal{H}$. Then, the following hold.*

R21. *If there exists $\bar{\rho} \in \mathbb{R}_{>0}$ such that*

$$\limsup_{n\to\infty} \rho_n = \bar{\rho}, \tag{19}$$

then the model (10) *satisfying* (17) *achieves $\bar{\rho}|\mu^\star|_{\mathcal{H}}$-weak moment matching asymptotically in expectation,* i.e.

$$(19) \implies \limsup_{n\to\infty} \mathbb{E}\Big[\big|W - \overline{W}_n\big|_{\mathcal{L}_2}^2\Big] \leq \bar{\rho}|\mu^\star|_{\mathcal{H}}^2.$$

R22. *If*

$$\lim_{n\to\infty} \rho_n = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \frac{1}{n\rho_n} = 0, \tag{20}$$

then the model (10) *satisfying* (17) *achieves moment matching almost everywhere asymptotically in expectation,* i.e.,

$$(20) \implies \lim_{n\to\infty} \mathbb{E}\Big[\big|W - \overline{W}_n\big|_{\mathcal{L}_2}^2\Big] = 0.$$

Theorem 2 contains two results: the first addresses the property of the estimator when ρ_n converge to a positive quantity, while the second provides a condition on ρ_n to guarantee moment matching almost everywhere in expectation asymptotically provided (20) holds. The weak moment matching condition **R21** aligns with general learning theory, which states that the regularization introduces a bias in the estimator [58, Sec 6.7]. This bias does not vanish asymptotically, making it impossible to achieve moment matching almost everywhere in expectation with a not-vanishing ρ_n. However, the second result in Theorem 2 indicates that if ρ_n decreases at a certain rate with the amount of data then the bias vanishes, and it is possible to achieve moment matching almost everywhere asymptotically in expectation.

Although Theorem 2 provides information about the asymptotic statistical properties of the method, it does not offer any insight into the probability of achieving these properties. In this regard, the following theorem establishes that moment matching can be achieved almost surely, provided that stronger convergence conditions on ρ_n and the noise term e are satisfied.

Theorem 3 (Matching Almost Surely)**.** *Suppose that Assumptions 1, 2, 3, and 5 hold and that $\mu^\star \in \mathcal{H}$. Then, the following hold.*

R31. *If there exists $\bar{\rho} \in \mathbb{R}_{>0}$ and* (19) *holds, then the model* (10) *satisfying* (17) *achieves $\bar{\rho}|\mu^\star|_{\mathcal{H}}$-weak moment matching asymptotically almost surely,* i.e.

$$(19) \implies \limsup_{n\to\infty} \big|W - \overline{W}_n\big|_{\mathcal{L}_2}^2 \leq \bar{\rho}|\mu^\star|_{\mathcal{H}}^2, \text{ almost surely.}$$

R32. *If*

$$\lim_{n\to\infty} \rho_n = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \frac{\log(2n)}{\sqrt{n}\rho_n} = 0, \tag{21}$$

then the model (10) *satisfying* (17) *achieves moment matching almost everywhere asymptotically almost surely,* i.e.,

$$(21) \implies \lim_{n\to\infty} \big|W - \overline{W}_n\big|_{\mathcal{L}_2}^2 = 0, \text{ almost surely.}$$

Theorem 3 provides results analogous to Theorem 2, but stated in terms of probability rather than expectation. Therefore, it is possible to achieve weak moment matching asymptotically almost surely or moment matching almost everywhere asymptotically almost surely with different conditions. However, a stricter convergence rate, *i.e.* (21) is required in comparison with (20) of Theorem 2.

Note that both Theorems 2 and 3 provide information applicable only with an infinite amount of data (*i.e.* $n \to \infty$) that can also be used as an approximation for large n. In this connection, the following theorem analyzes the property of the method in the presence of a fixed finite amount of data.

Theorem 4 (Matching with Probabilistic Bounds)**.** *Suppose that Assumptions 1, 2, 3, and 5 hold and that $\mu^\star \in \mathcal{H}$. Moreover, suppose that there exists a function $\tilde{\mu}^\star \in \mathcal{L}_2$, such that $\mu^\star = \iota^* \tilde{\mu}^\star$, and that*[3] *$\rho_n \geq \delta_n$. Then, the model* (10)

[3]Note that δ_n is a quantity which decreases with n. It describes the norm of the operator which is formally defined in (26).

satisfying (17) *is a kernel-based model which achieves* ϑ*-weak moment matching with probability at least* $1-\varepsilon$*, where* $\varepsilon \in (0,1]$ *and*

$$\vartheta \coloneqq 2\rho_n|\tilde{\mu}^\star|_{\mathcal{L}_2} + \frac{\zeta\sqrt{T_n}}{n\sqrt{\rho_n-\delta_n}} + \frac{2\zeta\left(\frac{\Psi}{n}+\frac{\Sigma}{\sqrt{n}}\right)}{\sqrt{\rho_n-\delta_n}}\log\left(\frac{2}{\varepsilon}\right). \tag{22}$$

Theorem 4 establishes the probability of achieving ϑ-weak-moment matching under certain additional assumptions. It is important to note that ϑ comprises three components: the first is associated with the regularization terms, and it decreases as ρ_n decreases; the second component represents the uncertainty introduced by the presence of the transient term (14), which introduces a bias to the estimator that decreases as either n or ρ_n increase; and the third component accounts for the uncertainty due to the stochastic nature of the noise, which also decreases as either n or ρ_n increase.

A. Proofs of the Main Results

Before presenting the proofs of the main results, we need to introduce some instrumental notions. First, following Lemma 1, we define the canonical embedding operator $\iota : \widehat{\mathcal{H}} \to \mathcal{L}_2(\mathcal{Q},\mathbb{R}^{d_{\mathrm{y}}})$, *i.e.*, $\iota(v) = v$, for all $v \in \mathcal{H}$. The operator ι is linear and bounded [53, Prop 4.4], thus its adjoint $\iota^* : \mathcal{L}_2(\mathcal{Q},\mathbb{R}^{d_{\mathrm{y}}}) \to \widehat{\mathcal{H}}$ exists, and it is also linear and bounded.

Given $Y \coloneqq \mathbb{R}^{d_{\mathrm{y}}}$, consider the product Euclidean space Y^n endowed with the inner product $\langle a,b\rangle_{Y^n} = n^{-1}\sum_{i=1}^n a_i^\top b_i$, and the induced norm $|b|_{Y^n}^2 = n^{-1}\sum_{i=1}^n |b_i|^2$, where $a \coloneqq (a_1,\dots,a_n) \in Y^n$ and $b \coloneqq (b_1,\dots,b_n) \in Y^n$. Given $\mathcal{D}$, we define the linear bounded operator $S : \widehat{\mathcal{H}} \to Y^n$ such that for all $v \in \mathcal{H}$, $Sv \coloneqq (v(\bar{q}_1),\dots,v(\bar{q}_n)) \in Y^n$, where $\bar{q}_i \in \mathcal{Q}$, and the adjoint $S^* : Y^n \to \widehat{\mathcal{H}}$ such that

$$\forall b \in Y^n, \qquad S^*b = \frac{1}{n}\sum_{i=1}^n F_{\bar{q}_i,b_i}. \tag{23}$$

With this in mind, the optimization problem (15) equivalently reads as $\min_{\mu\in\mathcal{H}} J_n(\mu)$, where

$$J_n(\mu) \coloneqq |S\mu^\star - \bar{z}|_{Y^n}^2 + \rho_n|\mu|_{\mathcal{H}}^2, \tag{24}$$

and

$$\bar{z} \coloneqq S\mu^\star + \bar{\tau} + \bar{e}, \tag{25}$$

where $\bar{\tau} \coloneqq (\bar{\tau}_1,\dots,\bar{\tau}_n) \in Y^n$ and $\bar{e} \coloneqq (\bar{e}_1,\dots,\bar{e}_n) \in Y^n$. Moreover, we define the linear bounded operator $\Gamma \coloneqq \iota^*\iota - S^*S$, with operator norm δ_n (that depends on n through S). As proven in [81, Lem 3.2], δ_n is bounded for all $n \in \mathbb{N}$ by[4]

$$\delta_n \le n^{-1}\zeta^2. \tag{26}$$

1) Instrumental results: We now prove preliminary results in a series of lemmata which will be used later in the proofs of the main theorems.

Lemma 2. *Suppose that Assumption 1 holds. Then,*

$$\left|W-\overline{W}_n\right|_{\mathcal{L}_2(\Omega,\mathbb{R}^{d_{\mathrm{y}}})}^2 = |\iota(\mu^\star-\hat{\mu}_n)|_{\mathcal{L}_2(\mathcal{Q},\mathbb{R}^{d_{\mathrm{y}}})}^2.$$

[4]Since $\delta_n \le n^{-1}\zeta^2$, a value of ρ_n that satisfies the condition $\rho_n > \delta_n$ always exist. In addition, for every value of ρ_n, it is always possible to increase the number of data n so that the condition is satisfied.

Proof. From the definition of ι, $\iota(\mu^\star(s)-\hat{\mu}_n(s)) = \mu^\star(s)-\hat{\mu}_n(s)$, for all $s \in \mathcal{Q}$.

Thus, using the definition of $|\cdot|_{\mathcal{L}_2}$

$$\begin{aligned}\left|W-\overline{W}_n\right|_{\mathcal{L}_2(\Omega,\mathbb{R}^{d_{\mathrm{y}}})}^2 &= \int_\Omega \left|W(\omega)-\overline{W}_n(\omega)\right|^2 d\omega \\ &= \int_{\mathcal{Q}} |\mu^\star(s)-\hat{\mu}_n(s)|^2 ds \\ &= |\iota(\mu^\star-\hat{\mu}_n)|_{\mathcal{L}_2(\mathcal{Q},\mathbb{R}^{d_{\mathrm{y}}})}^2.\end{aligned}$$

□

Lemma 3. *Let Assumptions 1 and 2 hold, and suppose that* $\mathbb{E}[e(t)] = 0$ *for all* $t \in \mathbb{T}$*. Then,* $\lim_{t\to\infty} \mathbb{E}[z(t)] - W(\omega(t)) = 0$.

Proof. For all (x_0,ω_0) in some neighborhood of $(0,0)$, Assumptions 1 and 2 imply that the system (4) satisfies

$$\forall t \in \mathbb{T}, \qquad \left|x(t)-\pi(\omega(t))\right| \le \beta\alpha^t\left|x(0)-\pi(\omega(0))\right|, \tag{27}$$

for some $\alpha \in (0,1)$ and $\beta \in \mathbb{R}_{\ge 0}$. Recall that the noisy measurement (11) can be equivalently rewritten, for all $t \in \mathbb{T}$, as $z(t) = W(\omega(t)) + \tau(t) + e(t)$, with $\tau : \mathbb{T} \to \mathbb{R}^{d_{\mathrm{y}}}$ defined in (13). Recall also that h is assumed sufficiently smooth, which further implies that h is locally Lipschitz with constant $\gamma \in \mathbb{R}_{\ge 0}$, namely $\left|h(x,\ell(\omega))-h(\pi(\omega),\ell(\omega))\right| \le \gamma\left|x-\pi(\omega)\right|$. Therefore, using (27), for all (x_0,ω_0) in some neighborhood of $(0,0)$, and for all $t \in \mathbb{T}$, we have that

$$\begin{aligned}|\tau(t)| &= \left|h(x(t),\ell(\omega(t))) - W(\omega(t))\right| \\ &\le \gamma\left|x(t)-\pi(\omega(t))\right| \\ &\le \gamma\beta\alpha^t\left|x(0)-\pi(\omega(0))\right|.\end{aligned} \tag{28}$$

Thus, $\lim_{t\to\infty}\alpha^t = 0 \implies \lim_{t\to\infty}\tau(t) = 0$. Finally, by the fact that $\mathbb{E}[e(t)] = 0$ for all $t \in \mathbb{T}$, we conclude that

$$\begin{aligned}&\lim_{t\to\infty} \mathbb{E}\Big[z(t)\Big] - W(\omega(t)) \\ &\quad= \lim_{t\to\infty} \mathbb{E}\Big[W(\omega(t)) + \tau(t) + e(t) - W(\omega(t))\Big] \\ &\quad= \lim_{t\to\infty} \tau(t) + \mathbb{E}\Big[e(t)\Big] = \lim_{t\to\infty} \mathbb{E}\Big[e(t)\Big] = 0.\end{aligned}$$

□

Lemma 4. *Consider* $\mathcal{D}$ *and suppose that Assumptions 1 and 2 hold. Then, the sequence* (14) *converges, that is* $\limsup_{n\to\infty} T_n < \infty$.

Proof. Using the inequality (28) and the fact that $\bar{\tau}_i = \tau(t_i)$ for all $i \in \mathbb{N}$, we have that

$$\forall i \in \mathbb{N}, \qquad |\bar{\tau}_i| \le \gamma\beta\alpha^{\bar{t}_i}\left|x(0)-\pi(\omega(0))\right|. \tag{29}$$

Then, by applying Cauchy-Schwartz inequality on the sequence (14) and by using (29), we have

$$T_n \le \sum_{i=1}^n\sum_{j=1}^n |\bar{\tau}_i||\bar{\tau}_j| \le \sum_{i=1}^n\sum_{j=1}^n \gamma^2\beta^2\alpha^{\bar{t}_i+\bar{t}_j}\left|x(0)-\pi(\omega(0))\right|^2.$$

Finally, since $\bar{t}_i$ is a divergent sequence and $|\alpha| < 1$, we conclude that

$$\limsup_{n\to\infty} T_n \le \gamma^2\beta^2\left|x(0)-\pi(\omega(0))\right|^2 \lim_{n\to\infty}\sum_{i=1}^n\sum_{j=1}^n \alpha^{\bar{t}_i+\bar{t}_j} < \infty.$$

□

2) Proof of Theorem 2: To begin with, since $\hat{\mu}_n$ is the minimizer of J_n, for all $\mu \in \mathcal{H}$, it is such that $J_n(\hat{\mu}_n) \leq J_n(\mu) \implies J_n(\hat{\mu}_n) \leq J_n(\mu^\star)$. Then, using (24) one obtains the inequality

$$|S\hat{\mu}_n - \bar{z}|^2_{Y^n} - |S\mu^\star - \bar{z}|^2_{Y^n} \leq \rho_n|\mu^\star|^2_{\mathcal{H}} - \rho_n|\hat{\mu}_n|^2_{\mathcal{H}}. \quad (30)$$

The left-hand side of (30) can be manipulated as

$$\begin{aligned}
&|S\hat{\mu}_n - \bar{z}|^2_{Y^n} - |S\mu^\star - \bar{z}|^2_{Y^n} \\
&= \langle S\hat{\mu}_n - \bar{z}, S\hat{\mu}_n - S\mu^\star\rangle_{Y^n} + \langle S\hat{\mu}_n - S\mu^\star, S\mu^\star - \bar{z}\rangle_{Y^n} \\
&= \langle S(\hat{\mu}_n - \mu^\star), S(\hat{\mu}_n - \mu^\star)\rangle_{Y^n} + 2\langle S(\hat{\mu}_n - \mu^\star), S\mu^\star - \bar{z}\rangle_{Y^n} \\
&= \langle \hat{\mu}_n - \mu^\star, S^*S(\hat{\mu}_n - \mu^\star)\rangle_{\mathcal{H}} - 2\langle \hat{\mu}_n - \mu^\star, S^*(\bar{z} - S\mu^\star)\rangle_{\mathcal{H}},
\end{aligned}$$

and thus

$$\begin{aligned}
\langle \hat{\mu}_n - \mu^\star, S^*S(\hat{\mu}_n - \mu^\star)\rangle_{\mathcal{H}} - 2\langle \hat{\mu}_n - \mu^\star, S^*(\bar{z} - S\mu^\star)\rangle_{\mathcal{H}} \\
\leq \rho_n|\mu^\star|^2_{\mathcal{H}} - \rho_n|\hat{\mu}_n|^2_{\mathcal{H}}. \quad (31)
\end{aligned}$$

By adding $|\iota(\mu^\star - \hat{\mu}_n)|^2_{\mathcal{L}_2}$ on both sides of (31), we have that

$$\begin{aligned}
&\rho_n|\hat{\mu}_n|^2_{\mathcal{H}} - \rho_n|\mu^\star|^2_{\mathcal{H}} + |\iota(\mu^\star - \hat{\mu}_n)|^2_{\mathcal{L}_2} \\
&\leq 2\langle \hat{\mu}_n - \mu^\star, S^*(\bar{z} - S\mu^\star)\rangle_{\mathcal{H}} + \langle \mu^\star - \hat{\mu}_n, \Gamma(\mu^\star - \hat{\mu}_n)\rangle_{\mathcal{H}} \\
&\leq 2|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}} + \langle \mu^\star - \hat{\mu}_n, \Gamma(\mu^\star - \hat{\mu}_n)\rangle_{\mathcal{H}} \\
&\leq 2|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}} + \delta_n|\mu^\star - \hat{\mu}_n|^2_{\mathcal{H}},
\end{aligned}$$

where in the second step we have used the Cauchy-Schwarz inequality, and in the third step we have used the definition of operator norm δ_n for the bounded linear operator Γ. Therefore, since $\rho_n|\hat{\mu}_n|^2_{\mathcal{H}} \geq 0$, using (26) we have that

$$\begin{aligned}
&|\iota(\mu^\star - \hat{\mu}_n)|^2_{\mathcal{L}_2} - \rho_n|\mu^\star|^2_{\mathcal{H}} \\
&\qquad \leq 2|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}} + \frac{\zeta^2}{n}|\mu^\star - \hat{\mu}_n|^2_{\mathcal{H}},
\end{aligned}$$

and, by Lemma 2, we obtain the inequality

$$\begin{aligned}
&\left|W - \overline{W}_n\right|^2_{\mathcal{L}_2} - \rho_n|\mu^\star|^2_{\mathcal{H}} \\
&\qquad \leq 2|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}} + \frac{\zeta^2}{n}|\mu^\star - \hat{\mu}_n|^2_{\mathcal{H}}. \quad (32)
\end{aligned}$$

Now, by taking the expectation on both sides of (32) and recalling that $|\mu^\star|^2_{\mathcal{H}}$ is deterministic, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left|W - \overline{W}_n\right|^2_{\mathcal{L}_2}\right] - \rho_n|\mu^\star|^2_{\mathcal{H}} \\
&\leq 2\mathbb{E}\left[|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}\right] + \frac{\zeta^2}{n}\mathbb{E}\left[|\mu^\star - \hat{\mu}_n|^2_{\mathcal{H}}\right],
\end{aligned}$$

and then, using Hölder inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\left|W - \overline{W}_n\right|^2_{\mathcal{L}_2}\right] &\leq \rho_n|\mu^\star|^2_{\mathcal{H}} + \frac{\zeta^2}{n}\mathbb{E}\left[|\mu^\star - \hat{\mu}_n|^2_{\mathcal{H}}\right] \\
&\quad + 2\sqrt{\mathbb{E}\left[|\mu^\star - \hat{\mu}_n|^2_{\mathcal{H}}\right]\mathbb{E}\left[|S^*(\bar{z} - S\mu^\star)|^2_{\mathcal{H}}\right]}. \quad (33)
\end{aligned}$$

Before investigating (33) in the limit, we first analyze each term on the right-hand side of (33).

I. By using equation (25), property (1), and Assumption 3, we obtain

$$\mathbb{E}\left[|S^*(\bar{z} - S\mu^\star)|^2_{\mathcal{H}}\right] \leq \frac{\zeta^2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[(\bar{e}_i + \bar{\tau}_i)^\top(\bar{e}_j + \bar{\tau}_j)\right].$$

Moreover, by the assumption on the noise and definition (14), we find a deterministic bound, *i.e.*

$$\mathbb{E}\left[|S^*(\bar{z} - S\mu^\star)|^2_{\mathcal{H}}\right] \leq \frac{\zeta^2}{n}\Sigma^2 + \frac{\zeta^2}{n^2}T_n. \quad (34)$$

II. By using first the Cauchy-Schwarz inequality and then the triangular inequality, the inequalities

$$\begin{aligned}
&\langle \hat{\mu}_n - \mu^\star, S^*(\bar{z} - S\mu^\star)\rangle_{\mathcal{H}} \\
&\qquad \leq |\hat{\mu}_n - \mu^\star|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}} \\
&\qquad \leq |\hat{\mu}_n|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}} + |\mu^\star|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}
\end{aligned}$$

hold. Since $\langle \hat{\mu}_n - \mu^\star, S^*S(\hat{\mu}_n - \mu^\star)\rangle_{\mathcal{H}}$ is positive, from (31) we have $\rho_n|\hat{\mu}_n|^2_{\mathcal{H}} - \rho_n|\mu^\star|^2_{\mathcal{H}} \leq 2\langle \hat{\mu}_n - \mu^\star, S^*(\bar{z} - S\mu^\star)\rangle_{\mathcal{H}}$. Therefore,

$$\begin{aligned}
|\hat{\mu}_n|^2_{\mathcal{H}} - |\mu^\star|^2_{\mathcal{H}} &\leq \frac{2}{\rho_n}|\hat{\mu}_n|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}} \\
&\quad + \frac{2}{\rho_n}|\mu^\star|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}. \quad (35)
\end{aligned}$$

Now, by adding $\frac{1}{\rho_n^2}|S^*(S\mu^\star - \bar{z})|^2_{\mathcal{H}}$ on both sides of (35) to complete the squares and then by applying the square root, one obtains

$$|\hat{\mu}_n|_{\mathcal{H}} \leq |\mu^\star|_{\mathcal{H}} + \frac{2}{\rho_n}|S^*(S\mu^\star - \bar{z})|_{\mathcal{H}}. \quad (36)$$

Moreover, since $|\hat{\mu}_n - \mu^\star|^2_{\mathcal{H}} \leq \left(|\hat{\mu}_n|_{\mathcal{H}} + |\mu^\star|_{\mathcal{H}}\right)^2$, using (36) and Young's inequality we have

$$|\hat{\mu}_n - \mu^\star|^2_{\mathcal{H}} \leq 8|\mu^\star|^2_{\mathcal{H}} + \frac{8}{\rho_n^2}|S^*(S\mu^\star - \bar{z})|^2_{\mathcal{H}}. \quad (37)$$

By performing the expectation on both sides of (37), using (34) and (14) we obtain

$$\mathbb{E}\left[|\hat{\mu}_n - \mu^\star|^2_{\mathcal{H}}\right] \leq 8|\mu^\star|^2_{\mathcal{H}} + \frac{8\zeta^2}{n\rho_n^2}\Sigma^2 + \frac{8\zeta^2}{n^2\rho_n^2}T_n. \quad (38)$$

By using (38) and performing the supremum limit yields

$$\begin{aligned}
\limsup_{n\to\infty}\frac{\zeta^2}{n}\mathbb{E}\left[|\hat{\mu}_n - \mu^\star|^2_{\mathcal{H}}\right] &\leq \limsup_{n\to\infty}\frac{8\zeta^4\Sigma^2}{(n\rho_n)^2} \\
&\quad + \limsup_{n\to\infty}\frac{8\zeta^4 T_n}{(n\rho_n)^2 n}. \quad (39)
\end{aligned}$$

Regarding the claim **R21**, from Lemma 4 together with (19), the right-hand side of (39) is zero, and thus

$$\limsup_{n\to\infty}\frac{\zeta^2}{n}\mathbb{E}\left[|\hat{\mu}_n - \mu^\star|^2_{\mathcal{H}}\right] = 0. \quad (40)$$

Similarly, regarding the claim **R22**, Lemma 4 together with (20) again imply (40).

III. Using (38) and (34) we have that

$$\begin{aligned}
&\mathbb{E}\left[|\mu^\star - \hat{\mu}_n|^2_{\mathcal{H}}\right]\mathbb{E}\left[|S^*(\bar{z} - S\mu^\star)|^2_{\mathcal{H}}\right] \\
&\qquad \leq \frac{8\zeta^2\Sigma^2}{n}|\mu^\star|^2_{\mathcal{H}} + \frac{8\zeta^2}{n^2}|\mu^\star|^2_{\mathcal{H}}T_n + \\
&\qquad\quad \frac{8\zeta^4\Sigma^4}{n^2\rho_n^2} + \frac{16\zeta^4\Sigma^2}{n^3\rho_n^2}T_n + \frac{8\zeta^4}{n^4\rho_n^2}T_n^2. \quad (41)
\end{aligned}$$

Regarding the claim **R21**, from Lemma 4 and (19), the supremum limit of the right-hand side of (41) converges to 0 as $n \to \infty$, i.e.,

$$\limsup_{n\to\infty} \mathbb{E}\Big[|\mu^\star - \hat{\mu}_n|_{\mathcal{H}}^2\Big]\mathbb{E}\Big[|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}^2\Big] = 0. \tag{42}$$

Similarly, regarding the claim **R22**, using Lemma 4 together with (20) we obtain again (42).

By collecting all the pieces together, we are finally ready to analyze the asymptotic behavior of (33). In particular, using (40) and (42), the bound

$$\limsup_{n\to\infty} \mathbb{E}\Big[\big|W - \overline{W}_n\big|_{\mathcal{L}_2}^2\Big] \leq \limsup_{n\to\infty} \rho_n |\mu^\star|_{\mathcal{H}}^2$$

holds for both claims **R21** and **R22**. Then, using (19), we derive the claim **R21**. Similarly, using (20), and the fact that $\mathbb{E}\big[|W - \overline{W}_n|_{\mathcal{L}_2}^2\big] \geq 0$ implies that the supremum limit to 0 yields convergence to 0, we derive the claim **R22**.

3) Proof of Theorem 3: To begin with, using (23), the triangular inequality, and (1), we note that

$$\begin{aligned}|S^*(S\mu^\star - \bar{z})|_{\mathcal{H}} &= |S^*(\bar{\tau} + \bar{e})|_{\mathcal{H}} = \left|\frac{1}{n}\sum_{i=1}^n F_{\bar{q}_i,\bar{\tau}_i+\bar{e}_i}\right|_{\mathcal{H}} \\ &\leq \sqrt{\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \bar{\tau}_i^\top k(\bar{q}_i,\bar{q}_j)\bar{\tau}_j} + \left|\frac{1}{n}\sum_{i=1}^n F_{\bar{q}_i,\bar{e}_i}\right|_{\mathcal{H}}.\end{aligned}$$

From Assumption 3 and definition (14) we also have that $\sum_{i=1}^n\sum_{j=1}^n \bar{\tau}_i^\top k(\bar{q}_i,\bar{q}_j)\bar{\tau}_j \leq \zeta^2 T_n$, which implies that

$$|S^*(S\mu^\star - \bar{z})|_{\mathcal{H}} \leq \zeta\frac{\sqrt{T_n}}{n} + \left|\frac{1}{n}\sum_{i=1}^n F_{\bar{q}_i,\bar{e}_i}\right|_{\mathcal{H}}. \tag{43}$$

Note that, by assumption on the noise $\bar{e}$ we have that $\mathbb{E}[F_{\bar{q}_i,\bar{e}}] = \mathbb{E}[k_{\bar{q}_i}\bar{e}] = k_{\bar{q}_i}\mathbb{E}[\bar{e}] = 0$, and, for all $m \in \mathbb{N}\setminus\{1\}$,

$$\begin{aligned}\mathbb{E}\big[|F_{\bar{q}_i,\bar{e}}|_{\mathcal{H}}^m\big] &= \mathbb{E}\Big[\big(\bar{e}^\top k(\bar{q}_i,\bar{q}_i)\bar{e}\big)^{\frac{m}{2}}\Big] \\ &\leq \zeta^m\mathbb{E}[|\bar{e}|^m] \leq \frac{m!}{2}(\zeta\Sigma)^2(\zeta\Psi)^{m-2}.\end{aligned}$$

Therefore, $F_{\bar{q}_i,\bar{e}}$ satisfies the assumptions of [82, Prop A.1]. Thus, the claim of [82, Prop A.1] holds in our case and yields that the inequality

$$\left|\frac{1}{n}\sum_{i=1}^n F_{\bar{q}_i,\bar{e}_i}\right|_{\mathcal{H}} \leq 2\log\left(\frac{2}{\varepsilon}\right)\left(\frac{\zeta\Psi}{n} + \frac{\zeta\Sigma}{\sqrt{n}}\right) \tag{44}$$

holds for all $\varepsilon \in (0,1]$ with probability at least $1-\varepsilon$. Applying (44) to (43) and fixing $\varepsilon = n^{-2} \in (0,1]$, we deduce that the inequality

$$\begin{aligned}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}^2 \leq \zeta\frac{\sqrt{T_n}}{n} \\ + 4\zeta\Psi\frac{\log(2n)}{n} + 4\zeta\Sigma\frac{\log(2n)}{\sqrt{n}}\end{aligned} \tag{45}$$

holds with probability at least $1-n^{-2}$. Now, using (36) and the fact that $|\hat{\mu}_n - \mu^\star|_{\mathcal{H}} \leq |\hat{\mu}_n|_{\mathcal{H}} + |\mu^\star|_{\mathcal{H}}$, we obtain

$$|\hat{\mu}_n - \mu^\star|_{\mathcal{H}} \leq 2|\mu^\star|_{\mathcal{H}} + \frac{2}{\rho_n}|S^*(S\mu^\star - \bar{z})|_{\mathcal{H}}. \tag{46}$$

Then, plugging (46) into (45) and using the complement rule from probability theory [83, Sec 2.1], the inequality

$$\begin{aligned}|\hat{\mu}_n - \mu^\star|_{\mathcal{H}} > 2|\mu^\star|_{\mathcal{H}} + 2\zeta\frac{\sqrt{T_n}}{n\rho_n} \\ + 8\zeta\Psi\frac{\log(2n)}{n\rho_n} + 8\zeta\Sigma\frac{\log(2n)}{\sqrt{n}\rho_n}\end{aligned} \tag{47}$$

holds with probability at most n^{-2}. Let P_n^μ be the probability of (47) to hold (that is $P_n^\mu \leq n^{-2}$). Then, we note that $\sum_{n=0}^\infty P_n^\mu \leq \sum_{n=0}^\infty \frac{1}{n^2} < \infty$. Therefore, when (19) holds, the inequality

$$\begin{aligned}&\limsup_{n\to\infty}|\hat{\mu}_n - \mu^\star|_{\mathcal{H}} \\ &\leq \limsup_{n\to\infty}\left(2|\mu^\star|_{\mathcal{H}} + 2\zeta\frac{\sqrt{T_n}}{n\rho_n} + 8\zeta\Psi\frac{\log(2n)}{n\rho_n} + 8\zeta\Sigma\frac{\log(2n)}{\sqrt{n}\rho_n}\right) \\ &= 2|\mu^\star|_{\mathcal{H}}\end{aligned} \tag{48}$$

holds. Since condition (21) holds by assumption, the inequality (48) follows directly from Lemma 4. Hence, by the Borel-Cantelli Lemma [83, Th 3.4.2], the condition $\sum_{n=0}^\infty P_n^\mu < \infty$ implies that (48) holds with probability one. Consequently, $|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}$ is asymptotically bounded almost surely. Now, using (32) and the probabilistic bound for (45) the inequality

$$\begin{aligned}\big|W - \overline{W}_n\big|_{\mathcal{L}_2}^2 &> \frac{\zeta^2}{n}|\mu^\star - \hat{\mu}_n|_{\mathcal{H}}^2 + \rho_n|\mu^\star|_{\mathcal{H}}^2 \\ &+ 2|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}\left(\zeta\frac{\sqrt{T_n}}{n} + 4\zeta\Psi\frac{\log(2n)}{n} + 4\zeta\Sigma\frac{\log(2n)}{\sqrt{n}}\right)\end{aligned} \tag{49}$$

holds with probability at most n^{-2}. Following similar reasoning as before, let P_n^W be the probability of (49) to hold (that is $P_n^W \leq n^{-2}$), and we note that $\sum_{n=0}^\infty P_n^W \leq \sum_{n=0}^\infty \frac{1}{n^2} < \infty$. Finally, using (48) and Lemma 4, we have that the condition

$$\begin{aligned}\limsup_{n\to\infty}\ |\hat{\mu}_n - \mu^\star|_{\mathcal{H}}\times \\ \left(\zeta\frac{\sqrt{T_n}}{n} + 4\zeta\Psi\frac{\log(2n)}{n} + 4\zeta\Sigma\frac{\log(2n)}{\sqrt{n}}\right) = 0\end{aligned}$$

holds almost surely as $|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}$ is asymptotically bounded almost surely. Therefore, we also have that the condition

$$\begin{aligned}&\limsup_{n\to\infty}\ \big|W - \overline{W}_n\big|_{\mathcal{L}_2}^2 \\ &\leq \limsup_{n\to\infty}\left(\frac{\zeta^2}{n}|\mu^\star - \hat{\mu}_n|_{\mathcal{H}}^2 + \rho_n|\mu^\star|_{\mathcal{H}}^2\right) \\ &\leq \limsup_{n\to\infty}\rho_n|\mu^\star|_{\mathcal{H}}^2\end{aligned} \tag{50}$$

holds almost surely as $|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}$ is asymptotically bounded almost surely. When (19) holds the inequality (50) yields

$$\limsup_{n\to\infty}\big|W - \overline{W}_n\big|_{\mathcal{L}_2}^2 \leq \bar{\rho}|\mu^\star|_{\mathcal{H}}^2,$$

and, from the Borel-Cantelli Lemma, we deduce that the claim **R31** holds. When (21) holds the inequality (50) yields

$$\limsup_{n\to\infty}\big|W - \overline{W}_n\big|_{\mathcal{L}_2}^2 \leq 0,$$

and, again, from the Borel-Cantelli Lemma and from the fact that $\big|W - \overline{W}_n\big|_{\mathcal{L}_2}^2 \geq 0$, we deduce that the claim **R32** holds because the supremum limit to 0 yields convergence to 0.

4) Proof of Theorem 4: To begin with, using the fact that

$$|\hat{\mu}_n|_{\mathcal{H}}^2 - |\mu^\star|_{\mathcal{H}}^2 = |\hat{\mu}_n - \mu^\star|_{\mathcal{H}}^2 - 2\langle \mu^\star, \mu^\star - \hat{\mu}_n\rangle_{\mathcal{H}},$$

from (31) we have that the inequality

$$\begin{aligned}\rho_n|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}^2 &\le 2\langle \hat{\mu}_n - \mu^\star, S^*(\bar{z} - S\mu^\star)\rangle_{\mathcal{H}} \\ &+ 2\rho_n\langle \mu^\star, \mu^\star - \hat{\mu}_n\rangle_{\mathcal{H}} - \langle \hat{\mu}_n - \mu^\star, S^*S(\hat{\mu}_n - \mu^\star)\rangle_{\mathcal{H}} \quad (51)\end{aligned}$$

holds. Adding $|\iota(\mu^\star - \hat{\mu}_n)|_{\mathcal{L}_2}^2$ to both sides of (51) we obtain

$$\begin{aligned}&|\iota(\mu^\star - \hat{\mu}_n)|_{\mathcal{L}_2}^2 + \rho_n|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}^2 \\ &\quad\le 2\langle \hat{\mu}_n - \mu^\star, S^*(\bar{z} - S\mu^\star)\rangle_{\mathcal{H}} + 2\rho_n\langle \mu^\star, \mu^\star - \hat{\mu}_n\rangle_{\mathcal{H}} \\ &\qquad + \langle \hat{\mu}_n - \mu^\star, \Gamma(\hat{\mu}_n - \mu^\star)\rangle_{\mathcal{H}}.\end{aligned}$$

By using Cauchy-Schwarz inequality, the definition of operator norm δ_n for Γ, and the fact that $\mu^\star = \iota^*\tilde{\mu}^\star$, we compute

$$\begin{aligned}&|\iota(\mu^\star - \hat{\mu}_n)|_{\mathcal{L}_2}^2 + \rho_n|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}^2 \\ &\qquad\le 2|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}} \\ &\qquad\quad + 2\rho_n\langle \tilde{\mu}^\star, \iota(\mu^\star - \hat{\mu}_n)\rangle_{\mathcal{L}_2} + \delta_n|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}^2\end{aligned}$$

and using Cauchy-Schwarz inequality once again yields

$$\begin{aligned}&|\iota(\mu^\star - \hat{\mu}_n)|_{\mathcal{L}_2}^2 - 2\rho_n|\tilde{\mu}^\star|_{\mathcal{L}_2}|\iota(\mu^\star - \hat{\mu}_n)|_{\mathcal{L}_2} \\ &\le (\delta_n - \rho_n)|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}^2 + 2|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}. \quad (52)\end{aligned}$$

Rearranging the left-hand side of (52), we have

$$\begin{aligned}&|\iota(\mu^\star - \hat{\mu}_n)|_{\mathcal{L}_2}^2 - 2\rho_n|\tilde{\mu}^\star|_{\mathcal{L}_2}|\iota(\mu^\star - \hat{\mu}_n)|_{\mathcal{L}_2} \\ &\qquad = \left(|\iota(\mu^\star - \hat{\mu}_n)|_{\mathcal{L}_2} - \rho_n|\tilde{\mu}^\star|_{\mathcal{L}_2}\right)^2 - \rho_n^2|\tilde{\mu}^\star|_{\mathcal{L}_2}^2,\end{aligned}$$

while the right-hand side of (52) yields

$$\begin{aligned}&(\delta_n - \rho_n)|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}^2 + 2|\hat{\mu}_n - \mu^\star|_{\mathcal{H}}|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}} \\ &\quad = -\left(\sqrt{\rho_n - \delta_n}|\hat{\mu}_n - \mu^\star|_{\mathcal{H}} - \frac{|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}}{\sqrt{\rho_n - \delta_n}}\right)^2 \\ &\qquad + \frac{|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}^2}{\rho_n - \delta_n}.\end{aligned}$$

Then, using the above derivations, the inequality (52) yields

$$\begin{aligned}&\left(|\iota(\mu^\star - \hat{\mu}_n)|_{\mathcal{L}_2} - \rho_n|\tilde{\mu}^\star|_{\mathcal{L}_2}\right)^2 \\ &\quad\le \rho_n^2|\tilde{\mu}^\star|_{\mathcal{L}_2}^2 + \frac{|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}^2}{\rho_n - \delta_n} \\ &\qquad - \left(\sqrt{\rho_n - \delta_n}|\hat{\mu}_n - \mu^\star|_{\mathcal{H}} - \frac{|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}}{\sqrt{\rho_n - \delta_n}}\right)^2 \\ &\quad\le \rho_n^2|\tilde{\mu}^\star|_{\mathcal{L}_2}^2 + \frac{|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}^2}{\rho_n - \delta_n}.\end{aligned}$$

By taking the square root of both sides of the above inequality, we obtain

$$\begin{aligned}|\iota(\mu^\star - \hat{\mu}_n)|_{\mathcal{L}_2} - \rho_n|\tilde{\mu}^\star|_{\mathcal{L}_2} &\le \sqrt{\rho_n^2|\tilde{\mu}^\star|_{\mathcal{L}_2}^2 + \frac{|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}^2}{\rho_n - \delta_n}} \\ &\le \rho_n|\tilde{\mu}^\star|_{\mathcal{L}_2} + \frac{|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}}{\sqrt{\rho_n - \delta_n}}.\end{aligned}$$

Now, using Lemma 2, we conclude that

$$|\iota(\mu^\star - \hat{\mu}_n)|_{\mathcal{L}_2} \le 2\rho_n|\tilde{\mu}^\star|_{\mathcal{L}_2} + \frac{|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}}}{\sqrt{\rho_n - \delta_n}}.$$

On the other hand, combining (44) and (43), we obtain that the inequality

$$|S^*(\bar{z} - S\mu^\star)|_{\mathcal{H}} \le \zeta\frac{\sqrt{T_n}}{n} + 2\zeta\log\left(\frac{2}{\varepsilon}\right)\left(\frac{\Psi}{n} + \frac{\Sigma}{\sqrt{n}}\right)$$

holds with probability at least $1 - \varepsilon$. Therefore, using the definition (22) we finally conclude that, for all $\varepsilon \in (0, 1]$, the inequality $\left|W - \overline{W}_n\right|_{\mathcal{L}_2}^2 \le \vartheta$ holds with probability at least $1 - \varepsilon$. Finally, using Definition 11 the proof is complete.

VI. Numerical Experiments

To validate the proposed methodology, we consider the problem of estimating the moment of the nonlinear RC ladder benchmark [68, Model 1] or [84]. The model is described by the equations

$$\begin{aligned}\dot{x}_1 &= -g(x_1) - g(x_1 - x_2) + u, \\ \dot{x}_i &= g(x_{i-1} - x_i) - g(x_i - x_{i+1}),\ \forall i \in \{2, \dots, 24\}, \\ \dot{x}_{25} &= g(x_{24} - x_{25}),\end{aligned}$$

where, for all $i \in \{1, \dots, 25\}$, $x_i \in \mathbb{R}$ is the i-th unmeasured state (the voltage in certain nodes of the electrical circuit), $y(t) = x_1(t) \in \mathbb{R}$ is the output, and $u(t) \in \mathbb{R}$ is the input voltage. The function $g : \mathbb{R} \to \mathbb{R}$ is the characteristic curve of a Shockley diode in parallel with a resistor with resistance 1 Ohm, *i.e.* $g(x) \coloneqq x + \exp(40x) - 1$. In the simulations, the initial conditions of the states are mutually independent normal random variables with mean 0 and variance 0.004.

We consider the problem of estimating the moment with two different signal generators:

G1. a linear signal generator that generates the first 4 non-zeros harmonics of a square wave of frequency 6 Hz;

G2. a nonlinear signal generator of order 2.

The input generator **G1** is of the form (3) with $d_\omega = 8$, $\omega(0) = \frac{1}{\pi}\left[0, 4, 0, \frac{4}{3}, 0, \frac{4}{5}, 0, \frac{4}{7}\right]$, $s(\omega) = S\omega$, $\ell(\omega) = L\omega$, and matrices

$$S \coloneqq \pi\begin{bmatrix} 0 & 12 & 0 & 0 & 0 & 0 & 0 & 0 \\ -12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 36 & 0 & 0 & 0 & 0 \\ 0 & 0 & -36 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 60 & 0 & 0 \\ 0 & 0 & 0 & 0 & -60 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 84 \\ 0 & 0 & 0 & 0 & 0 & 0 & -84 & 0 \end{bmatrix},$$

$$L \coloneqq \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}.$$

The input generator **G2** is also of the form (3) with $d_\omega = 2$, $\omega(0) = [0.8, 0.7]^\top$, and functions

$$\begin{aligned}s(\omega_1, \omega_2) &\coloneqq 50\begin{bmatrix} \omega_2 + \omega_2\omega_1\cos(\omega_2)\sin(\omega_1^2) \\ -\omega_1 - \omega_1^2\cos(\omega_2)\sin(\omega_1^2) \end{bmatrix}, \\ \ell(\omega_1, \omega_2) &\coloneqq 5\omega_2\,,\end{aligned}$$

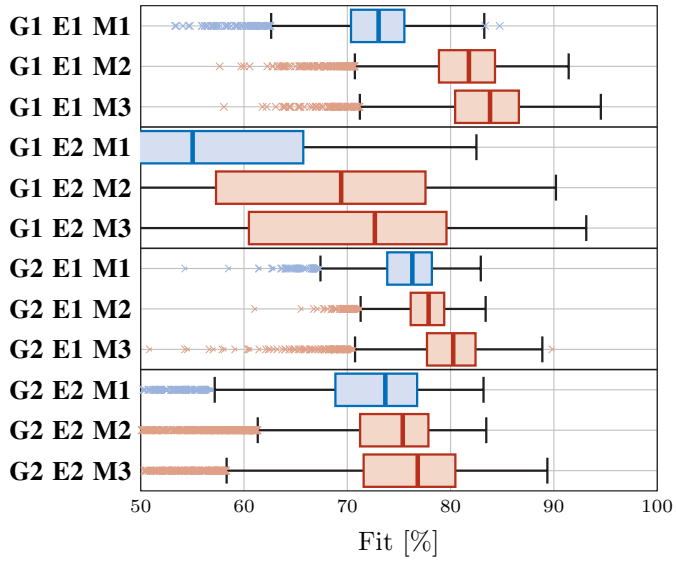

Fig. 2. Box plots of the performance indices for all considered cases. The blue boxes are for the cases without regularization (**M1**) while the red ones are for the cases with regularization (**M2** and **M3**).

describing a nonlinear oscillator. The noise e is zero-mean, normally distributed, and stationary, with variance $2.75 \cdot 10^{-5}$ for the simulations with **G1**, and variance $5 \cdot 10^{-4}$ for the simulations with **G2**, yielding a signal-to-noise ratio[5] in steady-state of approximately 5 in both cases. As the transient significantly impacts estimator performance (Section V), we conduct two experiments in each case:

E1. data are collected in steady-state;
E2. data are collected starting during the transient.

In all experiments we have collected $n = 100$ uniformly sampled data points in a time window Δ. In particular, $\Delta = [1, 2]$ for **G1**-**E1**, $\Delta = [0, 1]$ for **G1**-**E2**, $\Delta = [2, 3]$ for **G2**-**E1**, and $\Delta = [0, 1]$ for **G2**-**E2** (all the time units are in seconds). We tested the proposed algorithm in three different settings:

M1. with a polynomial kernel up to degree 5 [85, Prop 2.17] with $\rho = 0$, *i.e.* without regularization;
M2. with a polynomial kernel up to degree 5 with regularization;
M3. with the kernel (18) with $p = 1$ and $\tau_u = \gamma_u = 0$ as the two systems are strictly proper. Since $\tau_u = 0$, we set $\tau_w = 1$ without loss of generality. We also set $\alpha = 0$ because there is only one output.

It is worth noting that **M1** is equivalent to a not-recursive version of the method proposed in [31, Sec 3]. In all three methods, the hyperparameters of the kernel and ρ, unless specified otherwise, are set using Monte Carlo cross-validation [75] with 5 folds and 10 repetitions, following the procedure described in [86, Sec IV.B]. After the hyperparameters have been selected, the identified model is optimized using the method described in [86, Sec. IV.C], which enhances both the sparsity and performance of the estimated model. Then, as explained in Section IV-A, the estimated moment is used to define a model (10) that achieves moment matching with $d_\xi = 0$, output map (17), and

$$\bar{f}(\xi, u) := S\xi + \mathbf{1}_{25\times 1}(u - L\xi), \qquad \text{for } \mathbf{G1},$$
$$\bar{f}(\xi_1, \xi_2, u) := s(\xi_1, \xi_2) + 10\begin{bmatrix} \xi_1^2 u - \xi_1^2 \ell(\xi_1, \xi_2) \\ u - \ell(\xi_1, \xi_2) \end{bmatrix}, \qquad \text{for } \mathbf{G2}.$$

To statistically validate the proposed methodology, we have performed 5001 Monte Carlo experiments in which all the aforementioned random variables have different realizations. The results are reported in Figure 2 where we used the index

$$\text{Fit} = 1 - \frac{\sum_{i=1}^{n_v} |y(t_i) - \hat{y}(t_i)|}{\sum_{i=1}^{n_v} \left| y(t_i) - \frac{1}{n_v} \sum_{j=1}^{n_v} y(t_j) \right|}$$

with y the steady-state output of the model and $\hat{y}$ the steady-state output of the estimated reduced model, $n_v = 50000$ and $t_i = 10 + i\frac{5}{n_v}$. Here, we note that the regularization always provides a beneficial effect on the quality of the estimation. Furthermore, the proposed kernel (18), used in **M3**, provides better performance than the polynomial kernel, used in **M1** and **M2**, in all the considered cases.

Figures 3 and 4 show the estimated outputs of **G1**-**M3** and **G2**-**M3**, respectively. In more detail, Figure 3 compares the noiseless output y of **G1**-**M3** with the output of the estimated model with median performance, alongside the worst and best cases to visualize the uncertainty of the estimator. We note that, after the transient vanishes, the estimated model is able to match the output of the true model up to the uncertainty quantified in Theorem 4. Furthermore, the effect of the transient is readily visible in the second plot because it has a larger uncertainty despite using the same amount of data affected by the same noise. Similar conclusions also apply to **G2**-**M3** as shown in Figure 4. Finally, Figure 5 displays the graph of the estimated moment for **G2**-**E1**-**M3** with median performance. We can see that the choice of kernel ensures the derivation of a smooth estimator and a valid moment, as formally demonstrated in Proposition 1.

VII. Conclusions and Future Perspectives

We have presented a data-driven Moment Matching theory on RKHS. We have shown that data-driven moment matching in the RKHS provides a promising approach for estimating moments in several practical scenarios, such as noisy data and transient in the dataset, and has the advantage of identifying smoother moments from given finite datasets. We have been able to overcome the well-known limitations of ordinary least squares approaches, which typically suffer from the occurrence of ill-conditioned optimization problems. We have also introduced a new parameterized model and constructed an appropriate reproducing kernel for the moment estimation problem. Furthermore, we have investigated the conditions under which moment matching can be guaranteed both asymptotically and with a finite amount of data. Specifically, with different assumptions, we have shown that the estimate ensures: **i)** weak moment matching asymptotically in expectation; **ii)** moment matching almost everywhere asymptotically in expectation; **iii)** weak moment matching asymptotically, and; **iv)** moment matching almost everywhere asymptotically. Moreover, we have provided a probabilistic

[5]Estimated by computing the ratio between the mean square of the noiseless samples of y in steady-state and the variance of e.

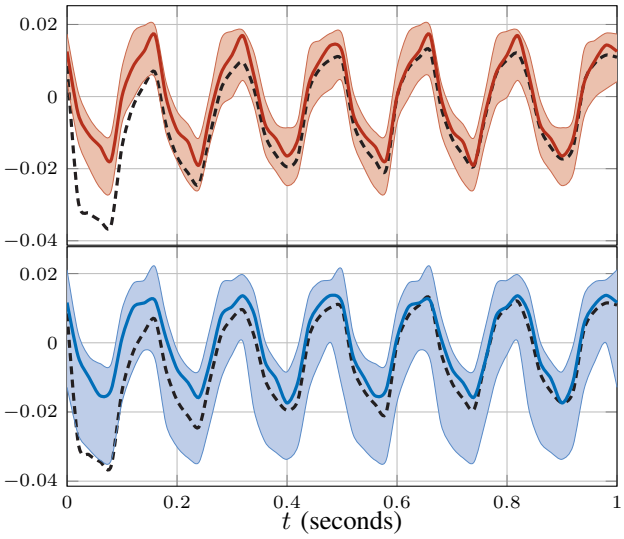

Fig. 3. Time history of the output of **G1**-**M3** (dashed line), the output of the estimated parametric model with the median performance of the Monte Carlo experiment (solid line), and the range of all the Monte Carlo experiments (shaded area). The top graph (in red) shows the result for **E1**, while the bottom graph (in blue) shows the result for **E2**.

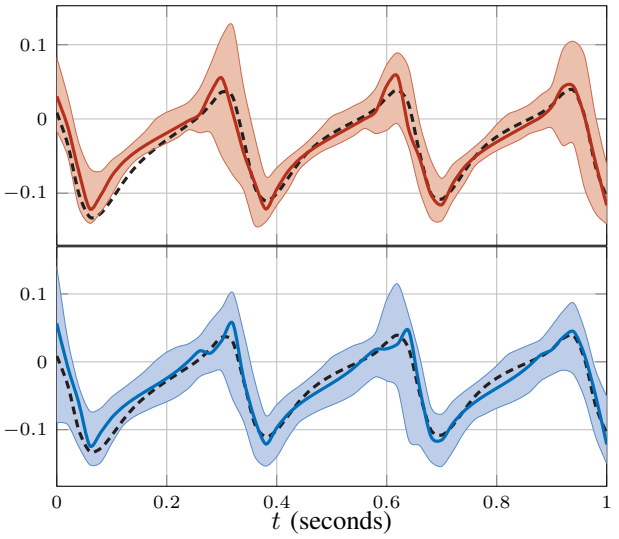

Fig. 4. Time history of the output of **G2**-**M3** (dashed line), the output of the estimated parametric model with the median performance of the Monte Carlo experiment (solid line), and the range of all the Monte Carlo experiments (shaded area). The top graph (in red) shows the result for **E1**, while the bottom graph (in blue) shows the result for **E2**.

bound that yields confidence in the quality of the estimated moments and ensures a priori weak moment matching. These results make a significant contribution to moment estimation from data by providing a priori guarantees on both the quality of the moment matching achieved for a given finite dataset and the accuracy of the parameterized model. Monte Carlo numerical experiments have been conducted on a benchmark nonlinear model to illustrate the statistical properties of the proposed method and its superior performance compared to related data-driven model reduction approaches.

Future work will focus on investigating the role of the transient effect more closely, particularly in embedding the transient term in the optimization problem itself using RKHS theory, which could lead to improved estimators by explicitly

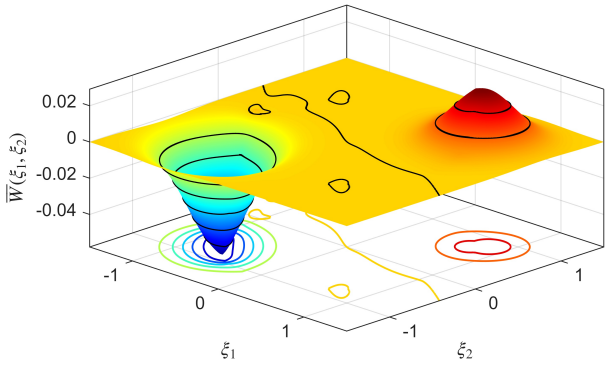

Fig. 5. Estimated moment for **G2**-**E1**-**M3** with median performance.

accounting for this source of error. This approach may enhance both theoretical understanding and practical performance in a finite sample regime. In addition, we will focus on developing an extension of the proposed algorithm to provide a computationally efficient algorithm that is able to adapt the estimated moments to newly available data in a manner similar to the recursive least squares method. Unlike the traditional recursive least squares approach, this method will accommodate an increasing number of features as more data becomes available, thereby presenting challenges to be addressed.

REFERENCES

[1] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *Natl. Sci. Rev.*, vol. 1, no. 2, pp. 293–314, 2014.
[2] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Comput. Sci.*, vol. 2, no. 3, 2021.
[3] J. C. Willems, "The behavioral approach to open and interconnected systems," *IEEE Control Syst. Mag.*, vol. 27, no. 6, pp. 46–99, 2007.
[4] A. Moreschini, M. Bin, A. Astolfi, and T. Parisini, "A generalized passivity theory over abstract time domains," *IEEE Trans. Autom. Control*, vol. 70, no. 1, pp. 2–17, 2025.
[5] M. E. J. Newman, "Communities, modules and large-scale structure in networks," *Nat. Phys.*, vol. 8, no. 1, pp. 25–31, 2011.
[6] J. P. Cunningham and B. M. Yu, "Dimensionality reduction for large-scale neural recordings," *Nat. Neurosci.*, vol. 17, no. 11, pp. 1500–1509, 2014.
[7] M. Newman, *Networks: An Introduction*. Oxford University Press, 2010.
[8] A. Moreschini, S. Monaco, and D. Normand-Cyrot, "Dirac structures for a class of port-Hamiltonian systems in discrete time," *IEEE Trans. Autom. Control*, vol. 69, no. 3, pp. 1999–2006, 2024.
[9] G. A. Baker, "Finite element methods for elliptic equations using nonconforming elements," *Math. Comput.*, vol. 31, no. 137, pp. 45–59, 1977.
[10] Y. Bar-Sinai, S. Hoyer, J. Hickey, and M. P. Brenner, "Learning data-driven discretizations for partial differential equations," *Proc. Natl. Acad. Sci.*, vol. 116, no. 31, pp. 15 344–15 349, 2019.
[11] A. C. Antoulas, *Approximation of large-scale dynamical systems*. SIAM, 2005.
[12] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Camb. Univ. Press, 2022.
[13] P. Benner, M. Ohlberger, A. Cohen, and K. Willcox, *Model Reduction and Approximation: Theory and Algorithms*. SIAM, 2017.
[14] A. Astolfi, C. Beck, D. Bhattacharjee, Y. Kawano, A. Moreschini, H. Sandberg, and J. M. A. Scherpen, "Forty plus years of model reduction and still learning," in *Proc. 63rd IEEE Conf. Decis. Control*, 2024, pp. 4480–4493.
[15] K. Glover, "All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds," *Int. J. Control*, vol. 39, no. 6, pp. 1115–1193, 1984.
[16] M. G. Safonov, R. Y. Chiang, and D. J. N. Limebeer, "Optimal Hankel model reduction for nonminimal systems," *IEEE Trans. Autom. Control*, vol. 35, no. 4, pp. 496–502, 1990.

[17] S. Kung and D. Lin, "Optimal Hankel-norm model reductions: Multivariable systems," *IEEE Trans. Autom. Control*, vol. 26, no. 4, pp. 832–852, 1981.

[18] B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Trans. Autom. Control*, vol. 26, no. 1, pp. 17–32, 1981.

[19] S. Lall and C. Beck, "Error-bounds for balanced model-reduction of linear time-varying systems," *IEEE Trans. Autom. Control*, vol. 48, no. 6, pp. 946–956, 2003.

[20] J. M. A. Scherpen, "Balancing for nonlinear systems," *Syst. Control Lett.*, vol. 21, no. 2, pp. 143–153, 1993.

[21] S. Gugercin and A. C. Antoulas, "A survey of model reduction by balanced truncation and some new results," *Int. J. Control*, vol. 77, no. 8, pp. 748–766, 2004.

[22] Y. Kawano and J. M. A. Scherpen, "Model reduction by differential balancing based on nonlinear Hankel operators," *IEEE Trans. Autom. Control*, vol. 62, no. 7, pp. 3293–3308, 2017.

[23] A. M. Burohman, B. Besselink, J. M. A. Scherpen, and M. K. Camlibel, "From data to reduced-order models via generalized balanced truncation," *IEEE Trans. Autom. Control*, vol. 68, no. 10, pp. 6160–6175, 2023.

[24] Z. Bai, "Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems," *Appl. Numer. Math.*, vol. 43, no. 1, pp. 9–44, 2002.

[25] P. Benner, S. Gugercin, and K. Willcox, "A survey of projection-based model reduction methods for parametric dynamical systems," *SIAM Rev.*, vol. 57, no. 4, pp. 483–531, 2015.

[26] A. Astolfi, "Model reduction by moment matching for linear and nonlinear systems," *IEEE Trans. Autom. Control*, vol. 55, no. 10, pp. 2321–2336, 2010.

[27] A. J. Mayo and A. C. Antoulas, "A framework for the solution of the generalized realization problem," *Linear Algebra Appl.*, vol. 425, no. 2, pp. 634–662, 2007.

[28] K. Gallivan, A. Vandendorpe, and P. Van Dooren, "Sylvester equations and projection-based model reduction," *J. Comput. Appl. Math.*, vol. 162, no. 1, pp. 213–229, 2004.

[29] E. J. Grimme, *Krylov Projection Methods for Model Reduction*. Univ. Ill. Urbana-Champaign, 1997.

[30] K. Gallivan, A. Vandendorpe, and P. Van Dooren, "Model reduction and the solution of Sylvester equations," in *Proc. 17th Int. Symp. Math. Theory Netw. Syst.*, 2006.

[31] G. Scarciotti and A. Astolfi, "Data-driven model reduction by moment matching for linear and nonlinear systems," *Automatica*, vol. 79, pp. 340–351, 2017.

[32] A. Moreschini and A. Astolfi, "Closed-loop interpolation by moment matching for linear and nonlinear systems," *IEEE Trans. Autom. Control*, vol. 70, no. 5, pp. 2918–2933, 2025.

[33] J. D. Simard and A. Moreschini, "Enforcing stability of linear interpolants in the Loewner framework," *IEEE Control Syst. Lett.*, vol. 7, pp. 3537–3542, 2023.

[34] A. Moreschini, J. D. Simard, and A. Astolfi, "Data-driven model reduction for port-Hamiltonian and network systems in the Loewner framework," *Automatica*, vol. 169, p. 111836, 2024.

[35] A. Isidori, *Nonlinear Control Systems*, 3rd ed. Springer-Verlag London, 1995.

[36] J. Carr, *Applications of Centre Manifold Theory*. Springer, 2012.

[37] G. Berkooz, P. Holmes, and J. L. Lumley, "The proper orthogonal decomposition in the analysis of turbulent flows," *Annu. Rev. Fluid Mech.*, vol. 25, no. Volume 25, pp. 539–575, 1993.

[38] K. Willcox and J. Peraire, "Balanced model reduction via the proper orthogonal decomposition," *AIAA J.*, vol. 40, no. 11, pp. 2323–2330, 2002.

[39] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *J. Fluid Mech.*, vol. 656, pp. 5–28, 2010.

[40] ——, "Dynamic mode decomposition and its variants," *Annu. Rev. Fluid Mech.*, vol. 54, no. Volume 54, pp. 225–254, 2022.

[41] C. W. Rowley and S. T. M. Dawson, "Model reduction for flow analysis and control," *Annu. Rev. Fluid Mech.*, vol. 49, no. 1, pp. 387–417, 2017.

[42] A. Ghadami and B. I. Epureanu, "Data-driven prediction in dynamical systems: recent developments," *Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci.*, vol. 380, no. 2229, 2022.

[43] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, "A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition," *J. Nonlinear Sci.*, vol. 25, no. 6, pp. 1307–1346, 2015.

[44] J. L. Proctor, S. L. Brunton, and J. N. Kutz, "Dynamic mode decomposition with control," *SIAM J. Appl. Dyn. Syst.*, vol. 15, no. 1, pp. 142–161, 2016.

[45] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor, *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. SIAM, 2016.

[46] D. Bhattacharjee, A. Moreschini, and A. Astolfi, "Signal generator agnostic moment matching," *IEEE Transactions on Automatic Control*, pp. 1–16, 2025, (Early Access).

[47] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

[48] N. Aronszajn, "Theory of reproducing kernels," *Trans. Am. Math. Soc.*, vol. 68, no. 3, pp. 337–404, 1950.

[49] S. Saitoh and Y. Sawano, *Theory of Reproducing Kernels and Applications*. Springer-Verlag, 2016.

[50] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2018.

[51] G. Wahba, *Spline Models for Observational Data*. SIAM, 1990.

[52] C. A. Micchelli and M. Pontil, "On learning vector-valued functions," *Neural Comput.*, vol. 17, no. 1, pp. 177–204, 2005.

[53] C. Carmeli, E. De Vito, and A. Toigo, "Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem," *Anal. Appl.*, vol. 04, no. 04, pp. 377–408, 2006.

[54] H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, and J. Audiffren, "Operator-valued kernels for learning from functional response data," *J. Mach. Learn. Res.*, vol. 17, no. 20, pp. 1–54, 2016.

[55] M. A. Álvarez, L. Rosasco, and N. D. Lawrence, "Kernels for vector-valued functions: A review," *Found. Trends Mach. Learn.*, vol. 4, no. 3, pp. 195–266, 2012.

[56] E. De Vito, V. Umanità, and S. Villa, "An extension of Mercer theorem to matrix-valued measurable kernels," *Appl. Comput. Harmon. Anal.*, vol. 34, no. 3, pp. 339–351, 2013.

[57] A. Moreschini, M. Scandella, and T. Parisini, "Non-convex learning with guaranteed convergence: Perspectives on stochastic optimal control," in *Proc. 63nd IEEE Conf. Decis. Control*, 2024, pp. 6002–6009.

[58] G. Pillonetto, T. Chen, A. Chiuso, G. De Nicolao, and L. Ljung, *Regularized System Identification: Learning Dynamic Models from Data*. Springer International Publishing, 2022.

[59] 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.

[60] G. Pillonetto, M. H. Quang, and A. Chiuso, "A new kernel-based approach for nonlinearsystem identification," *IEEE Trans. Autom. Control*, vol. 56, no. 12, pp. 2825–2840, 2011.

[61] 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.

[62] S. Formentin, M. Mazzoleni, M. Scandella, and F. Previdi, "Nonlinear system identification via data augmentation," *Syst. Control Lett.*, vol. 128, pp. 56–63, 2019.

[63] H. J. van Waarde and R. Sepulchre, "Kernel-based models for system analysis," *IEEE Trans. Autom. Control*, vol. 68, no. 9, pp. 5317–5332, 2023.

[64] M. Scandella, M. Bin, and T. Parisini, "Kernel-based identification of incrementally input-to-state stable nonlinear systems," in *Proc. 22nd IFAC World Congr.*, vol. 56, no. 2, 2023, pp. 5127–5132.

[65] J. Bouvrie and B. Hamzi, "Kernel methods for the approximation of nonlinear systems," *SIAM J. Control Optim.*, vol. 55, no. 4, pp. 2460–2492, 2017.

[66] Z. Hu, C. De Persis, and P. Tesi, "Learning controllers from data via kernel-based interpolation," in *Proc. 62nd IEEE Conf. Decis. Control*, 2023, pp. 8509–8514.

[67] A. Moreschini, M. Scandella, and T. Parisini, "Nonlinear data-driven moment matching in reproducing kernel Hilbert spaces," in *Proc. 22nd Eur. Control Conf.*, 2024, pp. 3440–3445.

[68] The MORwiki Community, "Nonlinear RC Ladder," MORwiki – Model Order Reduction Wiki, 2018, [Accessed: 2025-07-31]. [Online]. Available: http://modelreduction.org/index.php/Nonlinear_RC_Ladder

[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] R. Hermann and A. Krener, "Nonlinear controllability and observability," *IEEE Trans. Autom. Control*, vol. 22, no. 5, pp. 728–740, 1977.

[71] H. J. Sussmann and V. Jurdjevic, "Controllability of nonlinear systems," *J. Differ. Equ.*, vol. 12, no. 1, pp. 95–116, 1972.

[72] J. D. Simard, A. Moreschini, and A. Astolfi, "Parameterization of all differential-algebraic moment matching interpolants," *IEEE Trans. Autom. Control*, pp. 1–8, 2024.

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

[74] 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.
[75] Q.-S. Xu and Y.-Z. Liang, "Monte Carlo cross validation," *Chemom. Intell. Lab. Syst.*, vol. 56, no. 1, pp. 1–11, 2001.
[76] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. R. Stat. Soc.: B*, vol. 36, no. 2, pp. 111–133, 1974.
[77] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," *Technometrics*, vol. 16, no. 1, pp. 125–127, 1974.
[78] 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.
[79] D. J. C. MacKay, "Bayesian interpolation," *Neural Comput.*, vol. 4, no. 3, pp. 415–447, 1992.
[80] W. Rudin, *Real and Complex Analysis*. McGraw-Hill, 1987.
[81] E. De Vito, A. Caponnetto, and L. Rosasco, "Model selection for regularized least-squares algorithm in learning theory," *Found. Comput. Math.*, vol. 5, no. 1, pp. 59–85, 2004.
[82] G. Blanchard and N. Mücke, "Optimal rates for regularization of statistical inverse learning problems," *Found. Comput. Math.*, vol. 18, no. 4, pp. 971–1013, 2017.
[83] J. S. Rosenthal, *A First Look at Rigorous Probability Theory*, 2nd ed. World Scientific, 2006.
[84] M. Rewienski and J. White, "A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 22, no. 2, pp. 155–170, 2003.
[85] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2018.
[86] 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, 2020.

Alessio Moreschini (Member, IEEE) received the Ph.D. (Hons.) in Automatica from the University of Rome "La Sapienza," Italy and the Ph.D. in Systems and Control from Université Paris-Saclay, France, both in 2021. From 2021 to 2022, he was a Postdoctoral Fellow in the Dept. of Computer, Control, and Management Engineering at the University of Rome "La Sapienza." Since 2022, he has been a Research Associate in the Dept. of Electrical and Electronic Engineering at Imperial College London, U.K. He serves as an Associate Editor for Automatica and on the EUCA Conference Editorial Board. His research focuses on nonlinear systems and control theory, particularly passivity-based control, sampled-data systems, and model reduction.

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 Dept. of Management, Information and Production Engineering, University of Bergamo, Italy. From 2020 to 2024, he was with the Dept. of Electrical and Electronic Engineering, Imperial College London, UK. His research interests include system identification, health monitoring and Bayesian methods.

Alessandro Astolfi (Fellow, IEEE) graduated in Electronic engineering from the University of Rome in 1991. In 1992 he joined ETH-Zurich where he obtained a M.Sc. in Information Theory in 1995 and the Ph.D. degree with Medal of Honor in 1995. In 1996 he was awarded a Ph.D. from the University of Rome "La Sapienza". Since 1996 he has been with the Dept. of Electrical and Electronic Engineering of Imperial College London, where he is currently Professor of Nonlinear Control Theory. From 2010 to 2022 he served as Head of the Control and Power Group at Imperial College London and from 1998 to 2003 he was an Associate Professor at the Dept. of Electronics and Information of the Politecnico of Milano. Since 2005 he has also been a Professor at Dipartimento di Ingegneria Civile e Ingegneria Informatica, University of Rome Tor Vergata. His research focuses on mathematical control theory and applications, with particular emphasis on discontinuous stabilization, robust and adaptive control, observer design, and model reduction. He is the recipient of the IEEE CSS A. Ruberti Young Researcher Prize (2007), the IEEE RAS Googol Best New Application Paper Award (2009), the IEEE CSS George S. Axelby Outstanding Paper Award (2012), the Automatica Best Paper Award (2017), and the IEEE Transactions on Control Systems Technology Outstanding Paper Award (2023). He is a "Distinguished Member" of the IEEE CSS, IEEE Fellow, IFAC Fellow, IET Fellow, Member of the Academia Europaea, and of ITATEC. He served as Associate Editor for several journals; as Area Editor for the Int. J. of Adaptive Control and Signal Processing; as Senior Editor for the IEEE Trans. on Automatic Control; and as Editor-in-Chief for the European Journal of Control. He is currently Editor-in-Chief of the IEEE Trans. on Automatic Control (2018–). He served as Chair of the IEEE CSS Conference Editorial Board (2010-2017). He has served as Chair of the IEEE CSS Antonio Ruberti Young Researcher Prize (2015-2021); he is Vice Chair of the IFAC Technical Board (2020-2026).

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 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. 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.